Likelihood and bootstrap based inference for Hawkes processes

Carl-Emil Krarup Jensen*

Supervisor: Prof. Anders Rahbek†

June 18, 2025

Abstract

Hawkes processes have proven to be a versatile class of self-exciting point processes, with the exponential kernel variant among the most widely applied. Despite their popularity, statistical inferential tools for Hawkes models remain with room for exploration. In this paper, we introduce the Hawkes process as a Poisson cluster representation and give a series of results surrounding the large sample properties of the maximum likelihood estimator, score, and hessian of the Hawkes process with exponential kernel. Hereafter we introduce two bootstrap schemes for parameter estimation: The fixed intensity bootstrap and the recursive intensity bootstrap, as well as their nonparametric counterparts. We derive the asymptotic behavior of each method and evaluate their performance via Monte Carlo experiments across a variety of parameter configurations. Our findings show that the bootstrap schemes and their non parametric versions are more accurate in a broader range of scenarios compared to the recursive schemes. Furthermore we see how the bootstrap schemes still struggle with parameter estimation when the branching ratio is low. These results provide practical guidance for reliable inference in exponential kernel Hawkes modeling.

1 Introduction

Point processes provide a mathematical model for counting event occurrences over time. One subclass is the self-exciting point processes where past events determine the probability of events in the future. In 1971, Alan G. Hawkes introduced the self-exciting Hawkes process, in which the conditional intensity, is based on past events. In his paper, Hawkes suggested possible applications in epidemiology, neuroscience, and nuclear physics[11]. Since then, Hawkes processes have proven to be valuable modeling tools across an even broader range of applications, from forecasting crime in Los Angeles to volatility in financial markets ([19] and [9]). Hawkes processes admit many intensity-kernel choices and

*Department of Mathematics, University of Copenhagen, Denmark. gpl809@alumni.ku.dk

†Department of Economics, University of Copenhagen, Denmark anders.rahbek@econ.ku.dk

parameterizations [3]. In this paper we focus on the exponential-kernel Hawkes process, whose conditional intensity at time t,

$$\lambda(t) = \mu + \alpha \sum_{t_i<t} e^{-\beta(t-t_i)},$$

is parameterized by $\theta = (\mu, \alpha, \beta)$. We aim to introduce statistical methods for parameter inference and compare these.

Before doing any statistical analysis, we first build the probabilistic foundation of the process and introduce related point process constructions to make the model more intuitively accessible. Next we develop the asymptotic theory of likelihood based inference for the Hawkes process and extend it to several bootstrap schemes. Finally, we compare these schemes' ability to estimate the parameters of the exponential kernel Hawkes process via Monte Carlo simulations.

2 Probability theoretical background

In this section, we introduce point processes and their fundamental properties like atomic decomposition and the intensity measure. We then illustrate these concepts via the homogeneous and inhomogeneous Poisson processes. Finally, we present cluster processes and develop the notion of conditional intensity that leads to the Hawkes Process.

2.1 Point processes

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and let S be a local Borel space with corresponding σ-field $\mathcal{S}$. We define $N(B, \omega)$. Which plays the role as a local finite measure for $B \in \mathcal{S}$, and a fixed $\omega \in \Omega$, and conversely N plays the role as a random variable for for a fixed $B \in \mathcal{S}$ and $\omega \in \Omega$. We regard N as a random measure in the space of locally finite measures[1] on S denoted $\mathcal{M}_S$, equipped with with the sigma field generated by all evaluation maps $f_B : \mu \mapsto \mu(B)$, with μ locally finite and $B \in \mathcal{S}$.
A point process on S is defined as a random measure in $\mathcal{N}_\mathcal{S} \subset \mathcal{M}_S$, where $\mathcal{N}_\mathcal{S}$ is the space of locally finite integer valued measures on S. Using the atomic decomposition[2] we can write a general point process as

$$N(B) = \sum_{i=1}^{n} \delta_{X_i}(B), \quad n \in \mathbb{N}.$$

Where δ denotes the Dirac measure and $X_1, ..., X_n$ are random elements in S. We say the process is simple if $X_1, ..., X_n$ are distinct. The intensity of N is defined by $\Lambda = \mathbb{E}(N(B))$ on $(S, \mathcal{S})$ which we will sometimes refer to as the expectation measure[3]. We will put more weight on the intensity later, but for now we can regard it as the expected number of points in a set.

At first glance the above general definition may seem abstract, however, as we will be interested in point processes over time, it is helpful to think of B as a time interval $(0, t]$, $t > 0$, and the X_i being random variables determining when events occur in this time

[1] See A.1 in Appendix A.
[2] See A.2 in Appendix A.
[3] See[14], p. 321-322.

interval. When we consider time intervals, we will frequently use the notational shortcut $N((0,t]) = N_t$, and disregard ω since the randomness of the process is implicit. We now extend this to one of the most classical examples of point processes.

2.1.1 Poisson processes

To construct the homogeneous Poisson process on $\mathbb{R}_+$ we define $\mathscr{E}_i : \omega \mapsto \mathbb{R}_+$, for $i \in \mathbb{N}$, where $\mathscr{E}_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, $\lambda \in \mathbb{R}_+^*$. Setting $T_n = \sum_{i=1}^n \mathscr{E}_i$ allows us to define the Poisson "nuage"[4] $\Pi \subset \mathbb{R}_+$ with intensity parameter λ and the homogeneous Poisson counting process $N : \Omega \times \mathcal{S} \to (\mathbb{N} \cup \{\infty\})$ by

$$\Pi = \{T_n, n \in \mathbb{N}\}, \quad N(B, \Pi) = \#(\Pi \cap B).$$

For $B \in \mathcal{S}$. In this case $\mathcal{S}$ is the Borel sigma algebra on $\mathbb{R}_+$. As before we can write the atomic decomposition of the process as $N(B, \Pi) = \sum_{i=1}^{\infty} \delta_{T_i}(B)$. Note that since T_i is $\mathscr{A}$-measurable then $N(B, \Pi)$ is $\mathscr{A}$-measurable for every B [5]. The process can also be characterized by its probability mass function. If we let $B = (0, t]$, then we have

$$\mathbb{P}\big(N((0,t]) = n\big) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

i.e. N is a Poisson random variable. More generally, we write

$$\mathbb{P}\big(\{N(B_i) = n_i,\ i = 1, ..., k\}\big) = \prod_{i=1}^{k} \frac{\big(\lambda \nu(B_i)\big)^{n_i}}{n_i!} e^{-\lambda \nu(B_i)}, \tag{2.1}$$

where $B_i \in \mathcal{S}$, are pairwise disjoint and ν is the Lebesgue measure on $\mathbb{R}_+$.

To extend the class of Poisson processes on $\mathbb{R}_+$ we consider Poisson processes with time varying intensity $\lambda : \mathbb{R}_+ \to [0, \infty)$. Such are called inhomogeneous Poisson processes and are defined as in 2.1, where $\lambda\nu(B_i)$ is replaced with the integrated intensity

$$\Lambda(B_i) = \int_{B_i} \lambda(t) \nu(dt).$$ [6]

2.1.2 Cluster processes

To take one step towards the Hawkes process we introduce the general class of cluster processes and define the Poisson cluster process. The cluster process consists of, a point process N_c which produces points or "centers" x_i where each x_i births a "cluster" of children via new point process N. To formalize the cluster process we assume S is a Polish space e.g. $\mathbb{R}^d$. Let N_c be a point process on S with law Ξ_c and probability generating functional (p.g.fl) G_c [7]. We will refer to N_c as a cluster center.

Definition 2.1 *A finite point process N on S with cluster center process N_c on S and component processes $\{N(\cdot \mid x) : x \in S\}$ is a cluster process when for every bounded $B \in \mathcal{S}$*

$$N(B) = \int_S N(B \mid x) dN_c = \sum_i N(B \mid x_i).$$ [8]

[4] Nuage is a French word which translates to cloud. In french literature it's a common tool used to construct the Poisson process.

[5] See [7], p. 32.

[6] See [5], p. 18-21.

[7] See [5], p. 238. Probability generating functionals are defined in definition B.1 in Appendix B.

[8] The integral is a Lebsegue-Stieltjes integral wrt. the random counting measure N_c see A.4 in Appendix A.

If the component processes $N(B \mid x_i)$ are independent, we have an independent cluster process.
The intuition here is that the cluster center process generates a cloud of centers which each give birth to new points via the process N.

Remark 2.1 *If the following conditions are satisfied:*

$$\begin{aligned} &(i) \quad N_c \text{ is a Poisson process with intensity } \lambda \\ &(ii) \quad \text{Conditioned on } N_c,\ N(\cdot \mid x_i) \text{ are i.i.d} \\ &(iii) \quad N(B \mid x_i) < \infty, \quad \forall B \in \mathcal{S} \text{ and } i \geq 1 \end{aligned}$$

Then N is a Poisson cluster process.[9]

We now have the components ready to introduce the Hawkes process.

2.1.3 Hawkes processes

Now assume we have a stationary Poisson cluster process with parameter μ for the cluster centers. A parent generates offspring according to an inhomogeneous Poisson process with a rate $\phi : S \times \mathbb{R}_+ \to [0, \infty)$, $\phi(\cdot - x)$, and $||\phi(S)||_1 < 1$[10] to ensure that each branch goes extinct. Moreover we let $\phi(du)$ denote the intensity in a point $u \in S$. We denote this Poisson cluster process as N will and refer to it as a Hawkes process.
The clusters that arrive can be seen as independent realizations of a finite branching process[11]. These "branches" have the property that they evolve according to the same rule (p.g.fl.) as the contributing process. Let H_t denote the p.g.fl. for all individuals up to and including generation t starting from an initial ancestor $x \in S$. Furthermore let ξ satisfy the same conditions as in lemma B.1. We then consider the branching processes with the property

$$H_{t+1}(\xi \mid x) = \xi(x) G(H_t(\xi \mid \cdot) \mid x).$$

Note that the subscript denotes the generation and not the time interval. For more detail see Appendix B. Here G is the p.g.fl. of the offspring. Each branch of the process is then an inhomogeneous Poisson process. The evolution of the population described by a Hawkes process is therefore depending on the previous generations or its history $(\mathscr{F}_t)_{t \geq 0}$. To capture this property we extend the intensity notation.

Definition 2.2 *Let $\mathscr{F}_t$ denote the history of a point process N. The conditional intensity of N is defined as*

$$\lambda^*(t \mid \mathscr{F}_t) = \lim_{\Delta \to 0} \frac{\mathbb{E}(N([t, t+\Delta] \mid \mathscr{F}_t)}{\Delta}.$$[12]

To make notation more simple we will write $\lambda^*(t \mid \mathscr{F}_t) = \lambda^*(t)$ for the rest of this section and the next. Now suppose that ϕ is absolutely continuous wrt. the Lebesgue measure. We can then express the density of $\phi(du)$ as $\phi(du)du$. The conditional intensity is composed of the intensity of the cluster centers $\mu\Delta$ $\mu \geq 0$ and the intensity with

[9] See [5], p. 236-244.
[10] $||\cdot||_p$ where denotes the L^p norm with $p \geq 1$.
[11] See Appendix B
[12] See [12].

which each ancestor generates descendants $\phi(t-t_i)\Delta$. Parallel to [12] we can express the conditional intensity of the Hawkes process as

$$\begin{aligned}\lambda^*(t) &= \lim_{\Delta\to 0} \frac{\mu\Delta + \Delta\sum_{t_i\le t}\phi(t-t_i)}{\Delta} \\ &= \mu + \sum_{t_i<t}\phi(t-t_i) \\ &= \mu + \int_{-\infty}^{t}\phi(t-u)dN(u).\end{aligned}$$

We will refer to ϕ as the kernel function and, for further statistical analysis, be concerned with the exponential kernel given by

$$\phi(x) = \alpha e^{-\beta x},$$

where $\alpha, \beta \ge 0$. To develop a terminology surrounding this parameterization, it can be helpful to interpret the parameters.

- μ is the baseline intensity and can be seen as the rate with which immigrants arrive
- α is the expected number of descendants generated by each parent and can be seen as the jumping magnitude
- β is the decay rate determining how fast the process forgets it's history.

With the Hawkes process defined via its Poisson-cluster construction and the notation for conditional intensity established, we now explore its statistical properties.

3 Asymptotic results

We now present a series of results and assumptions to lay the foundation for our statistical analysis of a Hawkes process.
We assume that the conditional intensity from definition 2.2 is parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$ with $d = dim(\theta)$. We let $\theta_0 \in \Theta_0$ denote the true parameters of the process.

Proposition 3.1 *The likelihood function $L : \Theta \to \mathbb{R}^d$ of a regular point process over the interval $[0,t]$, $0 < t < \infty$ can be expressed as*

$$L = \prod_{i=1}^{N_t}\lambda^*(t_i,\theta)\exp\left(-\int_0^t \lambda^*(u,\theta)du\right)$$

and the log likelihood as

$$\ell_t(\theta) = \sum_{i=1}^{N_t}\log\lambda^*(t_i,\theta) - \int_0^t \lambda^*(t,\theta)\,dt.$$[13]

[13] See [5], p. 504.

Letting $\phi(\cdot, \theta)$ denote the parameterized kernel we can write down the log-likelihood of the Hawkes process as

$$\ell_t(\theta) = \sum_{i=1}^{N_t} \left(\log \left(\mu + \sum_{t_i < t} \phi(t - t_j, \theta) \right) - \int_{t_{i-1}}^{t_i} \mu + \sum_{t_i < t} \phi(t - t_j, \theta) dt \right).$$

A tool we will frequently use for analyzing the Hawkes process is the compensation formula. We here give a version of the result tailored for our further needs for this paper. A more thorough presentation of the compensation formula can be found in chapter 6 of [7].

Theorem 3.1 *Let N be a simple càdlàg point process on S adapted to the history $\mathscr{F}_t$ with conditional intensity λ^*. Let $\int_0^t \int_S \lambda^*(s)\mu(dx)ds < \infty$, with μ being a σ-finite measure. Then for all $t \in \mathbb{R}_+$ and all $B \in \mathcal{S}$*

$$t \mapsto M_t = N_t - \int_0^t \int_S \lambda^*(s)\mu(dx)ds$$

is a well defined $(\mathscr{F}_t)_{t \in \mathbb{R}_+}$-martingale càdlàg.[14]

We will refer to the second term of M_t as the compensator.
Using the compensation formula we can write

$$\begin{aligned} \mathbb{E}(dM_t \mid \mathscr{F}_{t-}) = \mathbb{E}(M_{t+} - M_{t-} \mid \mathscr{F}_{t-}) = M_{t-} - M_{t-} = 0 \\ \iff \mathbb{E}(N_{t+} - N_{t-} - \lambda^*(t)dt \mid \mathscr{F}_{t-}) = 0 \\ \iff \mathbb{E}(dN_t \mid \mathscr{F}_{t-}) = \lambda^*(t)dt. \end{aligned}$$

Where we use that M and N are càdlàg[15] and $\lambda^*(t)$ is independent of past history.
To ensure consistency of the MLE estimator $\hat{\theta} = \arg\min_\theta \ell(\theta)$ we employ a series of assumptions. We change the notation of the Hawkes process to $N(\cdot, \theta)$, to emphasize the dependence on the parameter θ. Before presenting the assumption we must develop the definition of statinarity, orderliness and ergodicity:

Definition 3.1 *A point process has stationary increments if it satisfies*

$$N((a + t, b + t]) \overset{\mathrm{d}}{=} N((a, b]).$$[16]

Definition 3.2 *A point process is said to be orderly if no more that one jump can happen at a time*

$$\mathbb{P}(N(t, t + \Delta) > 1) = o(\Delta).$$[17]

For point processes, different variations of ergodicity arise induced from various structures of the underlying process. However, in [24] Clinet and Yoshida conclude that the following definition is recurrent across most literature.

[14] See [7], p. 68.
[15] A locally finite point process on an interval closed from above is càdlàg, specifically stable Hawkes with exponential kernel, see A.6 in Appendix A.
[16] See assumption 2.2.I from [5].
[17] See [5], p. 28.

Definition 3.3 *A point process is ergodic if it satisfies, that for a mapping* $\pi : \Theta \to \mathbb{R}$

$$\frac{1}{T}\int_0^T \lambda(t,\theta)dt \xrightarrow{P} \pi(\theta) \quad when \quad T \to \infty.$$

For some stochastic intensity λ

We will be concerned with this definition for our assumption and later proof.

Assumption 3.1 [18]

(i) Θ *is compact, with true parameter* $\theta_0 \in \Theta_0 \subset \Theta$.

(ii) For each $\theta \in \Theta_0$, $N(\cdot,\theta)$ *is an orderly and ergodic point process with stationary increments. Moreover,*

$$\mathbb{E}\left(\sup_{m\geq 1} m\big(N((t,t+1/m],\theta) - N((0,t],\theta)\big)^2\right) < \infty.$$

(iii) The intensity process $\lambda^*(t,\theta)$ *is predictable (left continuous) in* t *, continuous in* θ*, and strictly positive.*

(iv) $\forall\, \vartheta \in \Theta\ \exists \varepsilon > 0$ *s.t.* $\forall\, \theta \in B[\vartheta,\varepsilon],$:
$|\lambda^*(t,\theta)| \leq \xi_1, \quad |\log \lambda^*(t,\theta)| \leq \xi_2, \quad \mathbb{E}(\xi_i^2) < \infty,\ i = 1,2.$

(v) $\lambda^*(t,\theta_1) = \lambda^*(t,\theta_2)\ \forall\, t \quad \Longleftrightarrow \quad \theta_1 = \theta_2.$

We now prove that these assumptions hold for the Hawkes process with exponential kernel.

(i) For $\theta = (\mu,\alpha,\beta)$ we have only assumed μ to be integrable and $||\phi||_1 < 1$ which are not sufficient conditions to make Θ compact. However we can construct a compact parameter space under which these conditions hold. For the stability condition we can analogous to what Alan G. Hawkes does in [11] write

$$\int_0^\infty \alpha e^{-\beta(u)}du = \frac{\alpha}{\beta} < 1.$$

Furthermore to ensure compactness we impose $0 < \beta_{\min} \leq \beta \leq \beta_{\max}$. Then for a $\delta > 0$, and $\mu_{\min}, \mu_{\max} \in \mathbb{R}$ define

$$\Theta = \{\mu,\alpha,\beta \in \mathbb{R}_+ \mid \mu_{\min} \leq \mu \leq \mu_{\max},\ \beta_{\min} \leq \beta \leq \beta_{\max},\ 0 \leq \alpha \leq (1-\delta)\beta\}.$$

For $[\mu_{\min}, \mu_{\max}]$ sufficiently wide and δ sufficiently small θ_0 will be in a compact subset of Θ.

(ii) We then look at the stationarity of the increments of a Hawkes process. It follows directly from [12] theorem 1 that the Hawkes process with exponential kernel and branching ratio $\alpha/\beta < 1$ is stationary.

Orderliness is one of the fundamental properties of the Hawkes process introduced in [11]. So by definition, the Hawkes process is orderly.

We build our proof of ergodicity on the results established in [4] and [12]. To do so we define the Hawkes process $\tilde{N}$ with conditional intensity $\tilde{\lambda}^*$. $\tilde{N}$ is defined such that it is identical to N on $(-,0]$ and by [12] the two inherit the same invariant law $\pi(\theta)$, i.e. i.e.

$$\mathbb{P}(N(B) \in A) = \mathbb{P}(\tilde{N}(B) \in A) \quad B \subset \mathbb{R}_+,\ A \subset \mathbb{N}.$$

[18] In [10] N is assumed to have "ergodic increments". The assumption to complete the proofs of theorem 3.2 and 3.3 seen in [10] and [20], can be modified to assuming N is ergodic.

.
Since exponential kernel satisfies that for some $\kappa > 0$, $\kappa \int_0^\infty \left|\alpha e^{-\beta t}\right| dt < 1$ we have from [4] Lemma 1 and page 15-16 that

$$\begin{aligned} \left|\lambda^*(t,\theta) - \tilde{\lambda}^*(t,\theta)\right| &\leq \int_s^t \alpha e^{-\beta(t-s)} d(N(s) - \tilde{N}(s)) \\ \implies \lim_{t\to\infty} \left|\lambda^*(t,\theta) - \tilde{\lambda}^*(t,\theta)\right| &= 0 \; a.s. \end{aligned}$$

We want to apply the strong law of large numbers for Markov chains on λ^*. In order to do so, we want to show that λ^* satisfies the Markov property. Denote the jump times $J_\lambda = \{t \in [a,b] : d\lambda(t) \neq 0\}$. We can decompose $\lambda^*(t)$ in the atom part and the continuous part[19] such that

$$\begin{aligned} d\lambda^* &= d\lambda_d + d\lambda_a \\ &= -\beta\left(\lambda^*(t,\theta) - \mu\right) dt + \sum_{t\in J_\lambda} \delta_t \alpha e^0 \\ &= -\beta\left(\lambda^*(t,\theta) - \mu\right) dt + \alpha dN. \end{aligned}$$

The intensity is constructed of a deterministic part and a pure jump type term[20]. Therefore λ^* is a piecewise deterministic Markov process[21]. From our construction of Hawkes process we know that it is adapted to its natural filtration $(\mathscr{F}_t)_{t\geq 0}$. Therefore observing the population at a time $s \geq 0$ tells you the state of the population and its history making the process independent of the time before s. We then have from [14] p. 278 that the intensity process satisfy the strong Markov property at every optional time i.e.

$$\mathbb{P}(\lambda^*(t,\theta) \in A \mid \mathscr{F}_s) = \mathbb{P}(\lambda^*(s,\theta) \in A \mid \lambda^*(s,\theta)) \quad s \leq t.$$[22]

Now define the sequence

$$X_i = \lambda^*(\Delta i, \theta), \quad i \in \mathbb{N},\ \Delta > 0,$$

where $\lambda^*(\Delta i, \theta)$ is started at $-\infty$ and then fixed at Δi. Then $(X_i)_{i\geq 0}$ is a Markov chain. Using assumption 3.2 from [23], we have that a stable Hawkes processes are Harris recurrent i.e. there exists a non zero σ-finite measure μ on $(S, \mathcal{S})$ such that

$$\mu(B) > 0 \implies \mathbb{P}(X_n \in B, \text{ for some } n \geq 1) = 1 \quad \text{c.f [15]}.$$

We can then apply the law of large numbers for positive Harris chains[23] so

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{n\to\infty} \int_0^\infty \pi(dx) \; a.s.$$

Where $\pi(dx)$ is the unique invariant initial distribution of X_n. By construction, $\lambda^*(\Delta i, \theta)$ has the same properties as $\tilde{\lambda}^*$ (transient law π). We will use this to invoke the convergence

[19] See [7] chapter 1.
[20] See [14] p. 277.
[21] This class of processes introduced in [6].
[22] Markovianity of Hawkes with exponential kernel is also given in [16].
[23] LLN for positive Harris chains is presented in [18], p. 421.

result from [4].
Set $T = n\Delta + \varepsilon$, with $\varepsilon \in [0, \Delta)$ and consider the integrated intensity

$$\int_0^T \lambda^*(t,\theta)dt = \sum_{i=1}^{n} \int_{(i-1)\Delta}^{i\Delta} \lambda^*(t,\theta)dt + \int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt.$$

We can bound the final term

$$\int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt \leq \sup_{t\geq 0} \lambda^*(t)(\Delta+\varepsilon)$$
$$\implies \frac{1}{T}\int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt = O\left(\frac{\Delta}{T}\right).$$

We can then write

$$\frac{1}{T}\int_0^T \lambda^*(t,\theta)dt - \frac{n}{T}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta} \left(\lambda^*(t,\theta) - X_i\right) dt + \int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt$$

Then for a fixed Δ we have

$$\int_{(i-1)\Delta}^{i\Delta} \left|\lambda^*(t,\theta) - \lambda^*(\Delta i,\theta)\right| dt \leq \Delta \sup_{t\in[(i-1)\Delta, i\Delta]} \left|\lambda^*(t,\theta) - \lambda^*(\Delta i,\theta)\right| \xrightarrow{i\to\infty} 0$$

Using the convergence result from [4]. The increments are vanishing so we can choose a $K, M > 0$ satisfying

$$\sup_{i\geq K}\left(\Delta \sup_{t\in[(i-1)\Delta, i\Delta]} \left|\lambda^*(t,\theta) - \lambda^*(\Delta i,\theta)\right|\right) \leq \Delta M.$$

Giving us

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta} \left|\lambda^*(t,\theta) - X_i\right| dt \\
&\leq \frac{1}{n}\sum_{i=1}^{K}\int_{(i-1)\Delta}^{i\Delta} \left|\lambda^*(t,\theta) - X_i\right| dt + \frac{(n-K)\Delta M}{n} \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Since the first term is finite. Putting together our convergence results we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta} \left(\lambda^*(t,\theta) - X_i\right) dt + \frac{1}{T}\int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt \\
&= \frac{n}{T}\frac{1}{n}\sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta} \left(\lambda^*(t,\theta) - X_i\right) dt + \frac{1}{T}\int_{n\Delta}^{n\Delta+\varepsilon} \lambda^*(t,\theta)dt \xrightarrow{T\to\infty} 0 \; a.s. \\
\implies &\frac{1}{T}\int_0^T \lambda^*(t,\theta)dt \xrightarrow{P} \pi(\theta).
\end{aligned}$$

Showing us that the Hawkes process with exponential kernel is ergodic in a Clinet and Yoshida sense.

For the next part of assumption *(ii)* we use that the point process takes the biggest amount of jumps on the largest interval so

$$\mathbb{E}\left(\sup_{m\geq 1} m\big(N((t,t+1/m]) - N((0,t])\big)^2\right) \leq \mathbb{E}(\sup_{m\geq 1} mN((t,t+1])^2)$$
$$= \sup_{m\geq 1}\big(mVar(N((0,1])^2 + m\mathbb{E}(N((0,1])^2\big) < \infty$$

c.f [5] p. 368.[24]
(iii) To see left continuity of $\lambda^*(t,\theta)$ in t we first note that between jumps the intensity is continuous and hence left continuous. Jumps happen when $t_i = t$, and $\lambda^*(t,\theta)$ only sums over jumps j when $t_i > t_j$. Therefore no jumps will occur when t approaches t_i from below. Moreover are all limit points are contained in the left limit, thus

$$\lambda^*(t_i-) = \lim_{s\to t_i-} \mu + \sum_{t_j<s} \alpha e^{-\beta(s-t_j)} = \mu + \sum_{t_j<t_i} \alpha e^{-\beta(t_i-t_j)} = \lambda^*(t_i).$$

Continuity of λ^* in θ and positivity is trivial.
(iv) Fix $\vartheta = (\mu_0, \alpha_0, \beta_0)$ and let $\varepsilon > 0$ such that $B[\vartheta, \varepsilon] \subset \Theta$. Compactness from assumption 1 ensures that $\theta = (\mu, \alpha, \beta) \in B(\vartheta, \varepsilon)$ satisfies $\mu \in [\mu_{\min}, \mu_{\max}]$, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\beta \in [\beta_{\min}, \beta_{\max}]$, so

$$\lambda^*(t,\theta) \leq \mu_{\max} + \alpha_{\max} \sum_{t_i<t} e^{-\beta_{\max}(t-t_i)} = \xi_1.$$

Where ξ_1 is positive with finite second moment c.f. [5] p. 368. We then obtain

$$|\lambda^*(t,\theta)| \leq \xi_1, \;\; \mathbb{E}(\xi_1^2) < \infty.$$

We can now bound $\log \lambda^*(t,\theta)$

$$\log \mu_{\min} \leq \log \lambda^*(t,\theta) \leq \log \xi_1$$
$$\implies |\log \lambda^*(t,\theta)| \leq \max\{|\log \mu_{\min}|, |\log \xi_1|\} = \xi_2.$$

Since ξ_1 has finite second moment and μ is constant ξ_2 has finite second moment.
(v) Let $\theta_1, \theta_2 \in \Theta$ with $\lambda^*(t,\theta_1) = \lambda^*(t,\theta_2)$. Consider a path of Hawkes process with exponential kernel. Now take the first two consecutive jump times of $0 < t_1 < t_2$. In the interval $t \in [0, t_1)$ we have

$$\mu_1 = \lambda^*(t,\theta_1) = \lambda^*(t,\theta_2) = \mu_2.$$

At the jump time $t = t_1$ we have

$$\alpha_1 = \lambda^*(t+,\theta_1) - \lambda^*(t-,\theta_1) = \lambda^*(t+,\theta_2) - \lambda^*(t-,\theta_2) = \alpha_2.$$

Now consider the interval $t \in (t_1, t_2)$, here we have

$$\mu_1 + \alpha_1 e^{-\beta_1(t-t_1)} = \mu_2 + \alpha_2 e^{-\beta_2(t-t_1)}$$
$$\implies \beta_1 = \beta_2.$$

[24] In this example the factorial moment measures are derived for a stable Hawkes process over a bounded set implying the finiteness of the 1st and second moment. Definition of factorial moments can be found in A.5.

This argumentation can be extended globally showing that

$$\lambda^*(t,\theta_1) = \lambda^*(t,\theta_2) \implies \theta_1 = \theta_2.$$

With the above assumptions verified for the Hawkes process with exponential kernel we give the next result which we will refer to as consistency of the MLE.

Theorem 3.2 *Under assumption 3.1 the maximum likelihood estimator* $\hat{\theta}_T = \hat{\theta}(t_i; 0 \leq t_i \leq T)$ *converges to* θ_0 *in probability as* $T \to \infty$.[25]

To analyze the asymptotic behavior of the score and information we employ another series of assumptions.

Assumption 3.2

(i) $\lambda^*(t,\theta) \in \mathcal{C}^3(\Theta) \quad \forall\, t \geq 0,$
$\mathbb{E}\Big[\big(\partial_{\theta_i}\lambda^*(t,\theta)\big)^2\Big] < \infty, \quad \mathbb{E}\Big[\big(\partial^2_{\theta_i\theta_j}\lambda^*(t,\theta)\big)^2\Big] < \infty \quad \forall\, i,j.$

(ii) $I(\theta) > 0$, $I(\theta) = \mathbb{E}\big[h(t,\theta)\big]$ *and each entry of* $h(t,\theta_0)$ *has finite variance, where* $h(t,\theta) = \lambda^*(t,\theta)^{-1}\big(\partial_\theta\lambda^*(t,\theta)\big)\big(\partial_\theta\lambda^*(t,\theta)\big)^T$.

(iii) $\forall\,\vartheta \in \Theta\ \exists\,\varepsilon > 0 : \sup_{\theta\in B[\vartheta,\varepsilon]}\big|\partial^3_{\theta_i\theta_j\theta_k}\lambda^*(t,\theta)\big| \leq c_{ijk}(t), \ \sup_{\theta\in B[\vartheta,\varepsilon]}\big|\partial^3_{\theta_i\theta_j\theta_k}\log\lambda^*(t,\theta)\big| \leq d_{ijk}(t),$
where $c_{ijk}(t)$, $d_{ijk}(t)$ *are stationary ergodic processes with* $\mathbb{E}\big[c_{ijk}(t)\big] < \infty$,
$\mathbb{E}\big[\lambda^*(t,\theta)^2\, d_{ijk}(t)^2\big] < \infty$, $\forall i,j,k \geq 1$.

We now prove that these assumptions hold for the Hawkes process with exponential kernel. *(i)* Let $c_0, c_1, c_2, c_3 \geq 0$ and consider the intensity function in the parameter μ. We can define $\lambda^* : \mu \mapsto \mu + c_0$. It is clear that λ is $\mathcal{C}^3$ in μ. Now consider the intensity function in α $\lambda^* : \alpha \mapsto c_1 + \alpha c_2$. It is also clear that λ^* is $\mathcal{C}^3$ in α. Finally in β we can without loss of generality look at one term from the sum and define $\lambda^* : \beta \mapsto c_1 + c_3 e^{-\beta c_2}$ which is a $\mathcal{C}^3$ function in β. Thus, every term in the sum is $\mathcal{C}^3$ making λ^* $\mathcal{C}^3$ in β. So $\lambda^*(t,\theta) \in \mathcal{C}^3(\Theta)$. Now let $(\theta_1,\theta_2,\theta_3) = (\mu,\alpha,\beta)$ and take

$$\partial_\mu\lambda^*(t,\theta) = 1, \ \ \partial_\alpha\lambda^*(t,\theta) = \sum_{t_i<t} e^{-\beta(t-t_i)}, \ \ \partial_\beta\lambda^*(t,\theta) = -\alpha\sum_{t_i<t}(t-t_i)e^{-\beta(t-t_i)}.$$

It is clear that $\mathbb{E}\big((\partial_{\theta_1}\lambda(t,\theta))^2\big) < \infty$.

To see that the derivative wrt. β has finite second moment define $g : [0,\infty) \to \mathbb{R}$, $g(u) = ue^{-\beta u}$. Using the compensation formula we have

$$\begin{aligned}
-\alpha\sum_{t_i<t}(t-t_i)e^{-\beta(t-t_i)} &= -\alpha\int_0^t (t-s)e^{-\beta(t-s)}dN(s) \\
&= -\alpha\int_0^t (t-s)e^{-\beta(t-s)}d\lambda^*(s) - \alpha\int_0^t (t-s)e^{-\beta(t-s)}dM(s) \\
&\quad - \alpha\int_0^t g(t-s)d\lambda^*(s) - \alpha\int_0^t g(t-s)dM(s).
\end{aligned}$$

[25] See [20].

Since g is bounded on $[0,\infty)$ by some constant $C_1 > 0$ we can rewrite the compensator term

$$\begin{aligned}-\alpha \int_0^t g(t-s)d\lambda^*(s) &= -\alpha \int_0^t g(t-s)\lambda^*(s,\theta)ds \\ &\leq -\alpha C_1 \int_0^t \lambda^*(s,\theta)ds.\end{aligned}$$

But we have already seen that the stability condition $||\phi||_1 < 1$ ensures that the intensity has finite first moment. So

$$\mathbb{E}\left(\left(-\alpha \int_0^t (t-s)e^{-\beta(t-s)}d\lambda^*(s)\right)^2\right) \leq \left((\alpha C_1)^2 \left(\int_0^t \lambda^*(s,\theta)ds\right)^2\right) < \infty.$$

We now look at the martingale term. We can use Itô isometry since g is deterministic giving us

$$\mathbb{E}\left(\left(-\alpha \int_0^t g(t-s)dM(s)\right)^2\right) = \mathbb{E}\left(-\alpha \int_0^t g(t-s)^2 d\langle M\rangle_s\right).$$

Where $\langle M\rangle_t$ is the predictable quadratic variation. We already saw

$$\mathbb{E}(dM(t) \mid \mathscr{F}_{s-}) = \mathbb{E}(M_{t+} - M_{t-} \mid \mathscr{F}_{s-}) = \lambda^*(t,\theta).$$[26]

Thus $d\langle M\rangle_t = \lambda(t,\theta)dt$ giving us

$$\begin{aligned}\mathbb{E}\left(\left(-\alpha \int_0^t g(t-s)dM(s)\right)^2\right) &= \mathbb{E}\left(\alpha^2 \int_0^t g(t-s)^2\lambda^*(s,\theta)ds\right) \\ &\leq \mathbb{E}\left((\alpha C_1)^2 \int_0^t \lambda^*(s,\theta)ds\right) < \infty\end{aligned}$$

This shows us that

$$\mathbb{E}((\partial_{\theta_3}\lambda^*(t,\theta))^2) = \mathbb{E}\left(\left(-\alpha \sum_{t_i<t}(t-t_i)e^{-\beta(t-t_i)}\right)^2\right) < \infty.$$

To see that the derivative wrt α has finite second moment we can consider another function $f : [0,\infty) \to \mathbb{R}$, $f(u) = e^{-\beta u}$. This function can be bounded by $C_0 > 0$. We then have

$$\int_0^\infty f(t-s)^2 dN(s) \leq \mathbb{E}\left((C_0)^2 \left(\int_0^t \lambda^*(s,\theta)ds\right)^2\right) + \mathbb{E}\left((C_0)^2 \int_0^t \lambda^*(s,\theta)ds\right) < \infty.$$

This shows that $\mathbb{E}\big((\partial_{\theta_2}\lambda(t,\theta))^2\big) < \infty$, and consequently $\mathbb{E}((\partial_{\theta_i}\lambda^*(t,\theta))^2) < \infty$.

If we take $\partial^2_{\theta_i\theta_j}\lambda^*(t,\theta)$ then all derivatives will become constant except for

$$\partial^2_{\beta\alpha}\lambda^*(t,\theta) = \partial^2_{\alpha\beta}\lambda^*(t,\theta) = \sum_{t_i<t}(t-t_i)e^{-\beta(t-t_i)}, \;\; \partial^2_{\beta\beta}\lambda^*(t,\theta) = \alpha\sum_{t_i<t}(t-t_i)^2e^{-\beta(t-t_i)}.$$

[26] See [14] p. 500 for quadratic variation and predicable variation. Here computations are analogous to the ones seen below theorem 3.1.

By the previous calculations, it is clear that $\mathbb{E}(\partial^2_{\alpha\beta}\lambda^*((t,\theta))^2) < \infty$.
To see that $\partial^2_{\beta\beta}\lambda^*(t,\theta)$ has finite second moment define $f : [0,\infty) \to \mathbb{R}$, $f(u) = u^2e^{-\beta u}$. See that $f'(u) = e^{-\beta u}u(2-\beta u)$, so $f'(u) = 0 \iff u \in \{0, 2/\beta\}$ giving us that f can be bounded by some $C_2 > 0$. We then have from previous calculations that

$$\mathbb{E}\left(\left(\alpha\sum_{t_i<t}(t-t_i)^2e^{-\beta(t-t_i)}\right)^2\right) \leq \mathbb{E}\left((\alpha C_2)^2\left(\int_0^t \lambda^*(s,\theta)ds\right)^2\right) + \mathbb{E}\left((\alpha C_2)^2\int_0^t \lambda^*(s,\theta)ds\right) < \infty.$$

Giving us $\mathbb{E}\big((\partial^2_{ij}\lambda^*(t,\theta))^2\big) < \infty$.
(ii) We have already seen that $\mathbb{E}\big((\partial^2_{ij}\lambda^*(t,\theta))^2\big) < \infty$ so the variance of h_{ij} is well defined for all i, j. To see that $I(\theta) = \mathbb{E}(h(t,\theta)) > 0$ let $\mathbf{x} = (x_1,\ x_2,\ x_3) \in \mathbb{R}^3_+$ and look at

$$\begin{aligned} h(t,\theta) &= \mathbf{x}^T\frac{1}{\lambda^*(t,\theta)}(\partial_\theta\lambda^*(t,\theta))(\partial_\theta\lambda^*(t,\theta))^T\mathbf{x} \\ &= \lambda^*(t,\theta)(\mathbf{x}\partial_\theta(\log\lambda^*(t,\theta))^2 = 0 \iff \mathbf{x} = 0. \end{aligned}$$

Giving us that $I(\theta) > 0$.
(iii) Fix $\vartheta = (\mu_0,\ \alpha_0,\ \beta_0)$ and let $\varepsilon > 0$ such that $B[\vartheta,\varepsilon] \subset \Theta$ and $\theta \in B[\vartheta,\varepsilon]$. Looking at the third partial derivative of λ^* we see that the only non constant derivatives are

$$\partial_{\alpha\beta\beta}\lambda^*(t,\theta) = \sum_{t_i<t}(t-t_i)^2e^{-\beta(t-t_i)},\ \ \partial_{\beta\beta\beta}\lambda^*(t,\theta) = -\alpha\sum_{t_i<t}(t-t_i)^3e^{-\beta(t-t_i)}$$

We can bound $\alpha(t-t_i)^me^{-\beta(t-t_i)} \leq \alpha_{\max}(t-t_i)^me^{-\beta_{\min}(t-t_i)}$, $m \in \{1,2,3\}$. Since each third derivative is either constant or constructed by jumps we can define

$$c_{ijk}(t) = C\sum_{t_i<t}\alpha_{\max}(t-t_i)^3e^{-\beta_{min}(t-t_i)}$$

and let $C > 0$ be large enough such that

$$\left|\partial^3_{\theta_i\theta_j\theta_k}\lambda^*(t,\theta)\right| \leq c_{ijk}(t).$$

To see that $\mathbb{E}(c_{ijk}(t)) < \infty$, define $g : [0,\infty) \to \mathbb{R}$, $g(u) = u^3e^{-\beta u}$ and take $g'(u) = 3u^2e^{-\beta u} - u^3\beta e^{-\beta u}$. Then $g'(u) = 0 \iff u \in \{0, 3/\beta\}$ giving us that g is bounded. Following the same argumentation as we did when proving assumption 3.2 *(i)* we have $\mathbb{E}(c_{ijk}(t)) < \infty$. We know that $\{t_i\}_{i=0}^{N_t}$ arrive according to a stationary Hawkes process as seen in the verification of 3.1 *(ii)* thus, we have that $c_{ijk}(t)$ has a stationary distribution i.e.

$$C\sum_{a\leq t_i<t}(t-t_i)e^{-\beta(t-t_i)} \overset{\mathrm{d}}{=} C\sum_{a+\Delta\leq t_i<t+\Delta}(t-t_i)e^{-\beta(t-t_i)}.$$

By the mean ergodic theorem [14] p. 569, c_{ijk} is ergodic.
Now consider $\log\lambda^*(t,\theta)$. We have seen that all third order derivatives are bounded. In particular for $\theta \in \Theta$ we can define a new constant $C' > 0$ large enough such that

$$\left|\partial^3_{\theta_i\theta_j\theta_k}\log\lambda^*(t,\theta)\right| \leq \frac{C'\sum_{t_i<t}(t-t_i)^3e^{-\beta(t-t_i)}}{(\lambda^*(t,\theta))^3}.$$

Since $\lambda^*(t,\theta) \geq \mu_{\min}$ we have

$$\left|\partial^3_{\theta_i\theta_j\theta_k} \log \lambda^*(t,\theta)\right| \leq \frac{C' \sum_{t_i<t} e^{-\beta_{min}(t-t_i)}}{\mu^3_{\min}} = d_{ijk}(t).$$

Now we want to see that $\mathbb{E}(\lambda^*(t,\theta)^2 d_{ijk}(t)^2) < \infty$. We do this by using Cauchy Schwartz inequality. To do so, we check that d_{ijk} and λ^* has $4th$ moment.
Define $f : [0,\infty) \to \mathbb{R}$, $f(u) = u^3 e^{-\beta u}$, by using the properties of the Γ function we can write. Then

$$\begin{aligned}\int_0^\infty f(u)^4 du &= \int_0^\infty u^{12} e^{-4\beta u} du \\ &= \frac{\Gamma(13)}{(4\beta)^{13}} \\ &= \frac{12!}{(4\beta)^{13}} = C_3 < \infty.\end{aligned}$$

This allows us to bound d_{ijk} like we did in the proving of assumption 3.2 *(i)*. Looking at $\lambda^*(t,\theta)^2$ it follows from the computations seen in the proof of 3.2 *(i)* that

$$\begin{aligned}\mathbb{E}\left(\left(\alpha \sum_{t_i<t} e^{-\beta(t-t_i)}\right)^4\right) \leq \mathbb{E}\left((\alpha C_0)^4 \int_0^t \lambda^*(s,\theta) ds\right) \\ + \mathbb{E}\left((\alpha C_0)^4 \left(\int_0^t \lambda^*(s,\theta) ds\right)^4\right) < \infty.\end{aligned}$$

By Cauchy Schwartz we have

$$\mathbb{E}(\lambda^*(t,\theta)^2 d_{ijk}(t)^2) \leq \mathbb{E}(\lambda^*(t,\theta)^4)^{\frac{1}{2}} \mathbb{E}(d_{ijk}(t)^4)^{\frac{1}{2}} < \infty.$$

Following the same arguments as for c_{ijk} we have that d_{ijk} is stationary and ergodic.
We have now verified assumption 3.2 for the Hawkes process with exponential kernel leading us to the following results for the score and hessian.

Let $S_T = \partial_\theta \ell_T(\theta)$ and $H_T = \partial^2_\theta \ell_T(\theta)$ denote the score and the hessian. We can for the Hawkes process write

$$\begin{aligned}S_T(\theta) &= \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{N_T}\left(\log\left(\mu + \sum_{t_i<T}\phi(T-t_j,\theta)\right) - \int_{t_{i-1}}^{t_i} \mu + \sum_{t_i<t}\phi(T-t_j,\theta) dt\right)\right) \\ &= \int_0^T \frac{\partial_\theta \lambda^*(t,\theta)}{\lambda^*(t,\theta)} dN(t) - \int_0^T \partial_\theta \lambda^*(t,\theta) dt.\end{aligned}$$

Now notice that from the compensation formula we obtain:

$$\begin{aligned}\frac{\partial_\theta \lambda^*(t,\theta)}{\lambda^*(t,\theta)} dN(t) - \partial_\theta \lambda^*(t,\theta) dt &= \partial_\theta \log \lambda^*(t,\theta)(dN(t) - \lambda^*(t,\theta) dt) \\ &= \partial_\theta \log \lambda^*(t,\theta) dM(t) \\ \implies S_T(\theta) &= \int_0^T \partial_\theta \log \lambda^*(t,\theta) dM(t).\end{aligned}$$

Furthermore we have for the hessian

$$\begin{aligned} H_T(\theta) &= \frac{\partial}{\partial\theta}\int_0^T \partial_\theta \log\lambda^*(t,\theta)dM(t) \\ &= \int_0^T \partial_\theta^2 \log\lambda^*(t,\theta)dM(t) + \int_0^T \partial_\theta(\log\lambda^*(t,\theta))\partial_\theta(dM(t)) \\ &= \int_0^T \partial_\theta^2 \log\lambda^*(t,\theta)dM(t) - \int_0^T \lambda^*(t,\theta)^{-1}\big(\partial_\theta\lambda^*(t,\theta)\big)\big(\partial_\theta\lambda^*(t,\theta)\big)'dt \\ &= \int_0^T \partial_\theta^2 \log\lambda^*(t,\theta)dM(t) - \int_0^T h(t,\theta)dt. \end{aligned}$$

Where we use that $\partial_\theta(dM(t)) = -\partial_\theta\lambda^*(t,\theta)$, and write $h(t,\theta) = \lambda^*(t,\theta)^{-1}\big(\partial_\theta\lambda^*(t,\theta)\big)\big(\partial_\theta\lambda^*(t,\theta)\big)'$. Using Lemma 1 from [13] we can express the next result.

Theorem 3.3 *Under assumptions 3.1 and 3.2 (i), (ii).*

$$T^{-1/2}S_T(\theta_0) \xrightarrow{D} \mathcal{N}\big(0, I(\theta_0)\big) \text{ and } -T^{-1}H_T(\theta_0) \xrightarrow{P} I(\theta_0).$$

Moreover if 3.2 (iii) holds then when $T \to \infty$,

$$\sqrt{T}(\hat\theta - \theta_0) \xrightarrow{P} \mathcal{N}\big(0, I(\theta_0)^{-1}\big) \quad \text{and} \quad LR_T(\theta) := 2(\ell(\hat\theta) - \ell(\theta_0)) \xrightarrow{D} \chi^2_d.$$ [27]

With the above assumptions and asymptotic theory established we wish to examine the conditions necessary for bootstrap parameter estimators to reproduce the same results.

4 Bootstrap

In this section we introduce two of the main tools for parameter estimation touched in this paper: The fixed intensity bootstrap (FIB) and the recursive intensity bootstrap (RIB). While FIB builds on fixing the conditional intensity across bootstrap repetitions, the RIB considers the conditional intensity random estimating it at each bootstrap repetition. To make notation easier we will now denote the conditional intensity $\lambda(t,\theta)$ and the bootstrapped estimated conditional intensity $\lambda^*(t,\theta)$.

4.1 Fixed intensity bootstrap

To construct the FIB algorithm we perform a time change of the waiting times.

Proposition 4.1 *Let N be a simple point process adapted to the history $(\mathscr{F}_t)_{t\geq 0}$, with continuous $\mathscr{F}$ compensator A that is not a.s. bounded. Under the random time change $t \mapsto A(t)$ the transformed process*

$$\tilde N(t) = N(A^{-1}(t))$$

is a $\tilde{\mathscr{F}}$ adapted Poisson process with intensity parameter $\lambda = 1$. [28]

[27] See [10], p. 5-6.
[28] See [5], p. 550.

Now let $t_i \mapsto s_i$ where s_i is the integrated intensity evaluated from 0 to t_i. The sequence of waiting times $\{w_i\}_{i=1}^{N_T}$, $w_i = t_i - t_{i-1}$ can therefore be mapped to the sequence $\{v_i\}_{i=1}^{N_T}$, $v_i = s_i - s_{i-1}$, with

$$s_i - s_{i-1} = \int_{t_{i-1}}^{t_i} \lambda(t, \theta) dt = \Lambda(t_i, t_{i-1}, \theta).$$

By proposition 4.1 we have $v_i \overset{\text{iid}}{\sim} \text{Exp}(1)$. Particularly for the Hawkes process we can write the time transformed sequence

$$\begin{aligned} v_i &= \int_{t_{i-1}}^{t_i} \mu + \sum_{t_j < t} \phi(t - t_j) dt \\ &= w_i \mu + \int_0^{w_i} \sum_{j<i} \phi(w + t_{i-1} - t_j) dw. \end{aligned}$$

What classifies this bootstrap algorithm is that we compute θ_T^* using the sample and fix it. Hereafter we keep the intensity fixed conditionally on the original data across all bootstrap replications. We denote the intensity corresponding to the bootstrap true value parameter $\lambda(t, \theta_T^*)$ and $\Lambda(t, \theta_T^*) = \int_0^t \lambda(t, \theta_T^*) du$. Note that $\Lambda(t, \theta_T^*)$ is strictly increasing and invertible[29]. The intensity, thus depends on the original data sample $\{t_i\}$ and θ_T^*. Now let $n_T^* = \max\{k : s_k^* \leq \Lambda(T, \theta_T^*)\} = \max\{k : t_k^* \leq T\}$. For the FIB we consider the sequence of of event times $\{t_i\}_{i=1}^{n_T^*}$ in $[0, T]$. The FIB algorithm is defined in parallel to [10] as follows.

Algorithm 1 Fixed intensity bootstrap

1: Generate conditionally on the original data *i.i.d* sample $\{v_i^*\}$ of bootstrap transformed waiting times drawn from the Exp(1)-distribution, and compute event times $s_i^* = \sum_{j=1}^{i} v_i^*$.
2: Construct the bootstrap event time $t_i^* = \Lambda^{-1}(s_i^*, \theta_T^*)$ in the original scale for $i = 1, ..., n_T^*$, where the number of bootstrap events is n_T^*.
The associated bootstrap counting process is $N^*(t) = \sum_{i \geq 1} \mathbb{1}_{t_i^* \leq t}$, for $t \in [0, T]$
3: Define the bootstrap MLE $\hat{\theta}_T^* = \arg\max_{\theta \in \Theta} \ell_T^*(\theta)$ with bootstrap log likelihood $\ell_T^*(\theta) = \sum_{i=1}^{n_T^*} \log \lambda(t_i^*, \theta) - \int_0^T \lambda^*(t, \theta)\, dt$

When considering the algorithm it can be helpful to note that that $T^{1/2}(\hat{\theta}_T - \theta_0)$ is approximated by the empirical distribution $T^{1/2}(\hat{\theta}_T^* - \theta_T^*)$ which has been generated conditionally on the original data. In addition the FIB likelihood ratio statistic is given as $LR_T^*(\theta_T^*) = 2\big(\ell_T(\hat{\theta}_T^*) - \ell_T(\theta_T^*)\big)$. Furthermore the last term of the bootstrap log likelihood $\int_0^T \lambda(t; \theta)\, dt$ only depends on the original data, acting non random when we condition on the data. We also note that since the transformed waiting times are *i.i.d* exponentially distributed with parameter 1 conditionally on the original data, $N^*(t)$ is an inhomogeneous Poisson process with intensity $\lambda(t, \theta_T^*)$. Therefore FIB allows us to estimate the parameters of a simple point processes with well defined compensator, and the implementation is simple: After inverting $\Lambda(t, \theta_T^*)$, one can draw from the bootstrapped sample. In addition the bootstrap likelihood and estimator can easily be computed since $\lambda(t, \theta)$ is a function only depending on the original data.

[29] Since $\lambda(t, \theta_T^*) > 0$ is continuous and differentiable $\partial_t \Lambda(t, \theta_T^*) > 0$.

4.2 Recursive intensity bootstrap

The other bootstrap algorithm that we will consider is the recursive intensity bootstrap from in [10]. Contrary to FIB, the RIB uses the bootstrapped event times t_i^* and the analytic form of $\lambda(t,\theta)$ to compute the bootstrapped conditional intensity which we will denote $\lambda^*(t,\theta)$. So at each bootstrap repetition, the intensity is random for all $\theta \in \Theta$. And $\Lambda^*(t,\theta_T^*) = \int_0^t \lambda^*(u,\theta_T^*)du$. The same properties as the original intensity process $\lambda(t,\theta)$, such as differentiability wrt. θ are preserved in this bootstrap scheme.

Algorithm 2 Recursive intensity bootstrap

1: Same as in algorithm 1
2: For $i = 1, ..., n_T^*$ construct the bootstrap event time in original time scale recursively $t_i^* = \Lambda^{*-1}(s_i^*)$, where the number of bootstrap events is $n_T^* = \max\{k : s_k^* \leq \Lambda^*(T,\theta_T^*)\}$ the associated bootstrap counting process is $N^*(t) = \sum_{i\geq 1} \mathbb{1}_{t_i^* \leq t}$, for $t \in [0,T]$
3: Define the bootstrap MLE $\hat{\theta}_T^* = \arg\max_{\theta\in\Theta} \ell_T^*(\theta)$ with bootstrap log likelihood $\ell_T^*(\theta) = \sum_{i=1}^{n_T^*} \log \lambda^*(t_i^*,\theta) - \int_0^T \lambda^*(t,\theta)\, dt$

To clarify; in step 2 the sequence $\{t_i^*\}_{i=1}^{n_T^*}$ is generated recursively using $\{s_i^*\}_{i=1}^{n_T^*}$ from step 1 such that t_i^* is the solution to $s_i^* = s_{i-1} + \Lambda^*(t_1^*; t_{i-1}^*, \theta_T^*)$ for $i \geq 2$, with t_1^* being the solution to $s_1^* = \Lambda^*(t_1^*, \theta_T^*)$, given $t_1^*, ..., t_{i-1}^*$.

5 Validity of the bootstrap

We will now dive in to the asymptotic validity of the bootstrap algorithms and compare the asymptotic behavior of the MLE, score, hessian and likelihood ratio in parallel to what we did in section 2. To distinguish the bootstrap space from our original probability space we briefly establish some notation following the one from [8]. Let $X_1, ..., X_n$ be $i.i.d$ observed data defined on $(\Omega, \mathscr{A}, \mathbb{P})$ and let $x = (x_1, ..., x_n)$ be an observed sample. Denote the empirical measure $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. We then let $\mathbb{P}^*$ denote the product measure $\mathbb{P}^*(\cdot \mid x) = \hat{P}_n^{\otimes B}$ so that under $\mathbb{P}^*$ we can draw new samples $X_1^*, ..., X_B^* \overset{\text{iid}}{\sim} \hat{P}_n$. Throughout this section we assume consistency of the bootstrap parameter estimator in parallel to assumption 3.1, such that $\theta_T^* \xrightarrow{P} \theta_0$, and consequently $\mathbb{P}(\theta_T^* \in \Theta) \to 1$. We will further write $N^*(t) = N_T^*(t)$, since the distribution of $N^*(t)$ depends on T, and let $\mathscr{F}_{T,t}^* = \sigma\big(\{N^*(s) : 1 \leq s \leq t \leq T,\ 0 \leq T\}\big)$ denote the history of $N_T^*(t)$. Note that since we have $v_i^* \overset{\text{iid}}{\sim} \text{Exp}(1)$, and the compensator is assumed to be continuous and strictly increasing, the transformation back to the original time scale is a continuous mapping. The the conditional density of the bootstrapped waiting times are absolutely continuous wrt. the Lebesgue measure ν and the integrated intensity of the bootstrapped process $N_T^*(t)$ denoted $\Lambda^*(t,\theta_T^*)$ is well defined. Using the compensation formula we have; $M_T^*(t) = N_T^*(t) - \Lambda^*(t,\theta_T^*)$, where $M_T^*(t)$ is a $\mathscr{F}_{T,t}^*$ martingale conditionally on the original data. For further analysis it can be helpful to consider the transformed counting process $\tilde{N}_T^*(s) = \sum_{i\geq 1} \mathbb{1}_{s_{i^*} \leq s}$. Since $s_i^* = s_{i-1}^* - v_i^*$, the cdf of s_i^* conditioned on past events is given by,

$$\mathbb{P}\big(s_i^* \leq s | \mathscr{F}_{s_{i-1}^*}\big) = \mathbb{P}\big(v_i^* \leq s - s_{i-1}^* | \mathscr{F}_{s_{i-1}^*}\big) = 1 - e^{-(s - s_{i-1}^*)},$$

which is continuous for $s^*_{i-1} < s$ [30]. Since the inter arrival times of $\tilde{N}^*_T(s)$ are *i.i.d* exponentially distributed $\tilde{N}^*_T(t)$ is a homogeneous Poisson process with intensity parameter $\lambda = 1$,[31] and has the following relation to the untransformed bootstrap counting process for the FIB

$$N^*_T(t) = \sum_{i\geq 1} \mathbb{1}_{t_{i^*}\leq t} = \sum_{i\geq 1} \mathbb{1}_{s_{i^*}\leq \Lambda(t,\theta^*_T)} = \tilde{N}^*_T\big(\Lambda(t,\theta^*_T)\big) \equiv \tilde{N}^*_T(s) = N^*_T\big(\Lambda(s,\theta^*_T)^{-1}\big).$$

Let $\xi : \mathbb{R}_+ \to \mathbb{R}$ be a $\mathscr{F}_{T,t}$ measurable process. Using the compensation formula we can express the time continuous martingale

$$X^*_T(t) = \int_0^t \xi(u) dM^*_{T,N}(u) = \int_0^t \xi(u) dN^*_T(u) - \int_0^t \xi(u) d\big(\Lambda(u,\theta^*_T)\big).$$

Under the transformed event times, we can write the martingales as

$$\begin{aligned} X^*_T &= \int_0^t \xi(\Lambda(u,\theta^*_T)) d\big(M^*_{T,\tilde{N}}(u)\big) \\ &= \int_0^t \xi(\Lambda(u,\theta^*_T)) d\tilde{N}^*_T(u) - \int_0^t \xi(\Lambda(u,\theta^*_t)) d\big(\Lambda(u,\theta^*_T)\big). \end{aligned}$$

Which is now depending on the *i.i.d* Exp(1) draws, and therefore independent of the original data.
Above relations are the same for the RIB just where the intensity is computed at each repetition so we can replace $\Lambda(s,\theta^*_T)$ in the above formulas with $\Lambda^*(s,\theta^*_T)$. Then the martingale isn't independent of the original data.
Analogously to the derivations of the score and hessian in section 2, we have the FIB score and hessian evaluated in the bootstrapped true value θ^*_T given as

$$\begin{aligned} S^*_T(\theta^*_T) &= \int_0^T \partial_\theta \log \lambda^*(t,\theta^*_T) dM^*_{T,N}(t) \\ H^*_T(\theta^*_T) &= \int_0^T \partial^2_\theta \log \lambda^*(t,\theta^*_T) dM^*_{T,N}(t) - \int_0^T h(t,\theta^*_T) dt. \end{aligned}$$

Where $h(t,\theta^*_T) = \lambda^*(t,\theta^*_T)^{-1}(\partial_\theta \lambda^*(t,\theta^*_T))(\partial_\theta \lambda^*(t,\theta^*_T))^T$. Note that both depend on the bootstrap data only through N^*_T which is a inhomogeneous Poisson process with fixed conditional intensity when conditioning on the original data.

Lemma 5.1 *Under the following assumptions*

(i) Assumption 3.1 and 3.2 (i), (ii)

(ii) Consistency of the bootstrap parameter estimator; $\theta^*_T \xrightarrow{P} \theta_0$

(iii) $\forall \theta \in \Theta$, $\mathbb{E}(|\partial_{\theta_i} \lambda^*(t,\theta)|^{3+\eta}) < \infty$, *and*
$\mathbb{E}(|\partial^2_{\theta_i} \lambda^*(t,\theta)|^{2+\eta}) < \infty$ *for some* $\eta > 0$, *for all* $1 \leq i \leq d$

(iv) Either $\lambda^*_T(t,\theta) \geq \lambda^*_L > 0$ *a.s., or* $\partial_{\theta_i} \lambda^*_T(t,\theta) \leq c < \infty$ *a.s.*

It holds that

$$T^{-1/2} S^*_T(\theta^*_T) \xrightarrow[P]{d^*} \mathcal{N}\big(0, I(\theta_0)\big)$$

and

$$T^{-1} H^*_T(\theta^*_T) = -\int_0^T h(t,\theta^*_T) dt + o_p(1) \xrightarrow[P^*]{P} -I(\theta_0).$$

[30] See [10], p. 8-9.
[31] See [14], p. 282.

Theorem 5.1 *Under the same assumptions as in lemma 5.1 we have*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}^*(\sqrt{T}\big(\hat{\theta}_T^*-\theta_T^*\big)\leq x)-\mathbb{P}(\sqrt{T}(\hat{\theta}_T-\theta_0\big)\leq x)\right|\xrightarrow[P]{}0\quad as\quad T\to\infty$$

$$LR_T^*(\theta_T^*)=2\big(\ell_T(\hat{\theta}_T^*)-\ell_T(\theta_T^*)\big)\xrightarrow[P]{d_*}\chi_d^2.$$

For proof of the above two statements see [10]. Here we wont verify the assumptions for the Hawkes process with exponential kernel. However we will later test these results in our simulation study.
We can express the RIB score and hessian evaluated in bootstrapped true value parameter θ_T^* as

$$S_T^*(\theta_T^*)=\int_0^T\partial_\theta\log\lambda^*(t,\theta_T^*)dM_{T,N}^*(t)$$
$$H_T^*(\theta_T^*)=\int_0^T\partial_\theta^2\log\lambda^*(t,\theta_T^*)dM_{T,N}^*(t)-\int_0^T h^*(t,\theta_T^*)dt.$$

Where $M_{T,N}^*(t)$ denotes the martingale corresponding to the bootstrapped counting process with bootstrap estimated intensity.
Under an additional assumption which were not needed for the FIB we have the same asymptotic results as for the score and hessian in the true parameter θ_0.

Lemma 5.2 [32] *Under assumption 3.1, 3.2 (i), (ii), consistency of the bootstrap parameter estimator $\theta_T^*\xrightarrow{p}\theta_0$ and for $i,j=1,...,d$,*

$$\sup_{\vartheta,\theta\in\Theta_0}\frac{\partial^2}{\partial\theta_i\partial\theta_j}|h_\vartheta(t,\theta)|\leq e_{i,j}(t)\quad where\quad\mathbb{E}(e_{i,j}(t))<\infty,$$

it holds that

$$T^{-1/2}S_T^*(\theta_T^*)\xrightarrow[P]{d^*}\mathcal{N}\big(0,I(\theta_0)\big)$$

and

$$-T^{-1}H_T^*(\theta_T^*)=\int_0^T h^*(t,\theta_T^*)dt+o_p^*(1)\xrightarrow[P]{P^*}I(\theta_0).$$

Note that from assumption 3.2 (iii) $c_{ijk}(t)=c_{ijk}(t,\theta_0)$ and $c_{ijk}(t)=d_{ijk}(t,\theta_0)$ can be replaced by $\sup_{\theta\in\Theta_0}c_{ijk}(t,\theta)$ and $\sup_{\theta\in\Theta_0}d_{ijk}(t,\theta)$ such that

Assumption 5.1 $\forall\vartheta\in\theta\ \exists\varepsilon>0$ *such that*

$$\sup_{\theta\in B[\vartheta,\varepsilon]}\left|\partial^3_{\theta_i\theta_j\theta_k}\lambda^*(t,\theta)\right|+\sup_{\theta\in B[\vartheta,\varepsilon]}\left|\partial^3_{\theta_i\theta_j\theta_k}\log\lambda^*(t,\theta)\right|\leq\sup_{\theta\in\Theta_0}c_{ijk}(t,\theta)+\sup_{\theta\in\Theta_0}d_{ijk}(t,\theta).$$

We will again leave verification of the assumptions here as we will test the results in our simulation.

Theorem 5.2 *Under the conditions of lemma 5.2 and assumption 5.1, theorem 5.1 holds for the RIB.*

[32] Where $\xrightarrow[P]{d^*}$ denotes convergence in distribution under $\mathbb{P}^*$, and that the distribution converges in $\mathbb{P}$. This can be thought of as $\mathbb{P}(|\mathbb{P}^*(T^{-1/2}S_T^*(\theta_T^*\leq x)-F(x)|>\varepsilon)\xrightarrow{T\to\infty}0$, where F is the CDF of a $\mathcal{N}(0,I(\theta_0))$ rv. and $\varepsilon>0$.

6 Nonparametric bootstrap algorithms

We now extend to two previous bootstrap algorithms under model misspecification. Here we can't assume the waiting times $\{v_i\}_{i=1}^{n_T^*}$ to be $i.i.d$ Exp(1). In this context the previously introduced bootstrap schemes can't be used for inference. To circumvent this issue we introduce nonparametric versions of the schemes like the ones seen in [10].
The main principle for the nonparametric schemes involves resampling from the transformed waiting times $\{\hat{v}_i\}_{i=1}^{n_T}$ after fitting the point process model to the original data. Hereafter the $i.i.d$ bootstrapped waiting time samples can be drawn from $\{\hat{v}_i\}_{i=1}^{n_T}$.
To implement the nonparametric bootstrap schemes the waiting times $\hat{v}_i$ are rescaled such that the bootstrapped waiting times match (as a minimum) the mean of Exp(1) distribution. Then a draw $\{v_i^*\}_{i=1}^{n_T}$ from $\{\hat{v}_i\}_{i=1}^{n_T}$ has $\mathbb{E}^*(v_i^*) = 1$, which can be achieved by using the empirical mean $\bar{v}_T = n_T^{-1}\sum_{j=1}^{n_T}\hat{v}_j$ and defining the rescaled sample

$$\hat{v}_i^c = \frac{\hat{v}_i}{\bar{v}_T}, \quad i = 1, ..., n_T. \tag{6.1}$$

We note that $\hat{v}_i^c > 0$ so when drawing $i.i.d$ bootstrap samples from $\{\hat{v}_i^c\}_{i=1}^{n_T}$ conditioned on the original data the draws will have expected value 1. With this in place we can define the nonparametric bootstrap algorithms.

Algorithm 3 Nonparametric bootstrap algorithms

1: Generate a sample $\{v_i^*\}$ resampling with replacement from $\{\hat{v}_i^c\}_{i=1}^{n_T}$ such that $v_i^* = \hat{v}_{u_i}^c$, $i = 1, 2, ...,$, where $\{u_i\}$ is an $i.i.d$ discrete uniformly distributed sequence on $\{1, ..., n_T\}$.
The bootstrapped transformed event times are $s_i^* = \sum_{j=1}^{i} v_j^*$
2: Construct the bootstrap event times t_i^* and bootstrap MLE like in Algorithm 1 using Λ^{-1} or 2 with Λ^{*-1} depending on which scheme is wished for implementation

We note that if the model is correctly specified such that the waiting times are $i.i.d$ exponentially distributed with parameter 1, the empirical variance of the bootstrapped waiting times converge to one in probability $\mathbb{P}$. Therefore when taking sufficiently large samples, the bootstrap score is centered around zero and its variance match the inverse of the information. Furthermore without rescaling $\mathbb{E}^*(v_i^*) \underset{P}{\to} 1$, rather than $\mathbb{E}^*(v_i^*) = 1$. Unless $\mathbb{E}^*(v_i^* - 1) = o_P(T^{-1/2}T)$ the bootstrap score will have a random non zero mean driven by the term $\mathbb{E}^*(v_i^* - 1)T^{-1/2}$ so we need more conditions to ensure centering around zero of the asymptotic score[33].

6.1 Validity of the nonparametric schemes

Like we did with the previous schemes we investigate the validity of the nonparametric bootstrap schemes. The validity of these schemes haven't been proven for the Hawkes process with exponential kernel in literature. So to restrict ourselves we choose to give an intuition to the validity of the schemes on a simpler process and hereafter test how the schemes perform in a simulation study.

[33] These remarks are given in [10].

Consider a Poisson process N with intensity $\theta \in \mathbb{R}_+$. With interest in the inference of theta we recall the log likelihood

$$\ell_T(\theta) = \int_0^T \log \theta dN(t) - \int_0^T \theta dt = n_T \log \theta - \theta T,$$

and the associated bootstrap score $S_T(\theta) = n_T\theta^{-1} - T$, giving us the unique maximum likelihood estimator $\hat{\theta}_T = n_T/T$. Recall the transformed waiting times from the start of section 4, given by $\hat{v}_i = \hat{\theta} w_i$, with $w_i = t_i - t_{i-1}$.
The nonparametric bootstrap generates a sample $\{v_i^*\}$ like described in algorithm 3 resampling from $\{\hat{v}_i^*\}_{i=1}^{n_T}$. By construction $\mathbb{E}^*(v_i^*) = 1$ and

$$\begin{aligned} Var^*(v_i^*) &= \frac{n_T^{-1}\sum_{i=1}^{n_T} w_i^2}{\left(n_T^{-1}\sum_{i=1}^{n_T} w_i\right)^2} - 1 \\ &= n_T^{-1}\sum_{i=1}^{n_T} w_i^2\theta_0^2 - 1 + o_p(1) = 1 + o_p(1). \end{aligned}$$

Where we use the law of large number for randomly selected sums[34], $n_T/T = \theta_0 + o_p(1)$, and $w_i = v_i/\theta_0$, where $v_i \overset{\text{iid}}{\sim} \text{Exp}(1)$.
Now the v_i^* are transformed back to the original time using the mapping $w_i^* = v_i^*/\hat{\theta}_T$ giving us the bootstrap event times $t_i^* = \sum_{j=1}^{i} w_j^*$ and the bootstrapped counting process $N^*(t)$ defined as in algorithm 1 and 2. The Bootstrap log likelihood and score are given by

$$\ell_T^*(\theta) = \int \log \theta dN^*(t) - \int \theta dt = n_T^* \log \theta - \theta T,$$

and $S_T^*(\theta) = n_T^*\theta^{-1} - T$, with the total number of bootstrap events defined as earlier $n_T^* = \max\{k : \sum_{i=1}^{k} w_i \leq T\}$. The bootstrap score at the bootstrap true value parameter $\theta_T^* = \hat{\theta}_T$ is thus given as

$$S_T^*(\hat{\theta}_T) = n_T^*\hat{\theta}_T^{-1} - T = -\hat{\theta}_T^{-1}\sum_{i=1}^{n_T^*}(\hat{\theta}_T w_i^* - 1) = -\hat{\theta}_T^{-1}\sum_{i=1}^{n_T^*}(v_i^* - 1).$$

Because we draw from the rescaled sample constructed as in 6.1 we have $\mathbb{E}^*(v_i^* - 1) = 0$. Without rescaling the mean of $(v_i^* - 1)$ would be of the order $O_p\big(n_T^{-1/2}\big) = O_p\big(T^{-1/2}\big)$[35]. Therefore the rescaling eliminates the asymptotic bias of the bootstrap score function. To analyze the asymptotics of the score function we give a bootstrap version of Donsker's theorem as presented in [10].

Theorem 6.1 *Let $u_1^*, u_2^*, \ldots$ be bootstrap random variables which, conditional on the original data, are i.i.d. with mean 1 and variance $\hat{\kappa}_T$ and strictly greater than 0 a.s. Where the variance is a function of the original data. For each $T > 0$ and $s \in [0, 1]$, define the càdlàg counting process*

$$n_T^*(s) \;:=\; \max\Big\{\, k \geq 0 : \sum_{i=1}^{k} u_i^* \;\leq\; \lfloor Ts \rfloor \Big\}.$$

[34] This result is presented in A.8 Appendix A.

[35] Where $\mathbb{P}(|v_i^* - 1| > Mn_T^{-1/2}) < \varepsilon$, for all $\varepsilon > 0$ and and for all $n_T > N$ for N sufficiently large and for some $M > 0$. See [10].

Assume that as $T \to \infty$ $\hat{\kappa}_T \xrightarrow{P} \kappa > 0$, *and that a bootstrap version of Donsker's theorem holds:*

$$\frac{1}{\sqrt{T}} \sum_{i=1}^{\lfloor Ts \rfloor} \frac{u_i^* - 1}{\sqrt{\hat{\kappa}_T}} \xrightarrow[P]{d_*} B_s, \quad \text{when} \quad T \to \infty$$

where B_s *is a standard Brownian motion under* $\mathbb{P}$. *Then, as* $T \to \infty$,

$$\frac{n_T^*(s) - \lfloor Ts \rfloor}{\sqrt{\hat{\kappa}_T\, T}} \xrightarrow[P]{d_*} B_s \quad \text{when} \quad T \to \infty.$$

A direct application of this theorem let us outline the asymptotic behavior of the bootstrap score and hessian and their application. Let $T \to \infty$, then

$$\begin{aligned} T^{-1/2} S_T^*(\hat{\theta}_T) &\xrightarrow[P]{d^*} \mathcal{N}\big(0, \theta_0^{-1}\big) \\ T^{-1} H_T^*(\hat{\theta}_T) = \frac{n_T^*}{T\hat{\theta}_T^2} &\xrightarrow[P]{P^*} \frac{1}{\theta_0} \\ T^{1/2}(\hat{\theta}_T - \theta_T^*)^* &\xrightarrow[P]{d^*} \mathcal{N}\big(0, \theta_0\big) \\ LR^*(\theta_T^*) &\xrightarrow[P]{d^*} \chi_1^2. \end{aligned}$$

These results follow from [10], and are proven in A.9 Appendix A.
We have now established the asymptotic results for the parametric schemes and given an intuition to the validity of the nonparametric schemes. The performance of the schemes is tested through a simulation study.

7 Monte Carlo Simulation

To test the effectiveness of the bootstrap algorithms we perform a Monte Carlo simulation to estimate the parameters of the Hawkes process with exponential kernel. We simulate the event times t_i using the thinning algorithm described in [21].

Algorithm 4 Thinning algorithm for Hawkes process simulation

1: Generate $U_0 \sim \text{Unif}([0,1])$ and set $\lambda_0 = \mu_0$
2: Generate $u_0 = -\frac{\log U_0}{\lambda_0}$
3: If $u_0 \leq T$ set $t_1 = u_0$ else stop
4: Set $i = j = k = 0$ and $n = 1$
5: Set k to $k+1$ and $\lambda_k = \lambda(t_n|t_1, ...,) = \mu + \alpha \sum_{t_i<t} e^{-\beta(t-t_i)}$
6: Set j to $j+1$ and compute U_j
7: Set i to $i+1$ and generate $u_i = -\frac{\log U_j}{\lambda_k}$
8: Set $s_i = s_{i-1} + u_i$ if $s_i > T$ stop
9: Set j to $j+1$ and generate U_j
10: If $U_j \leq \frac{\lambda(s_i|t_1,...,)}{\lambda_k}$ set n to $n+1$, and $t_n = s_i$ and go to step 5
11: Set k to $k+1$, and $\lambda_k = \lambda(s_i|t_1, ...,)$ and go to step 6

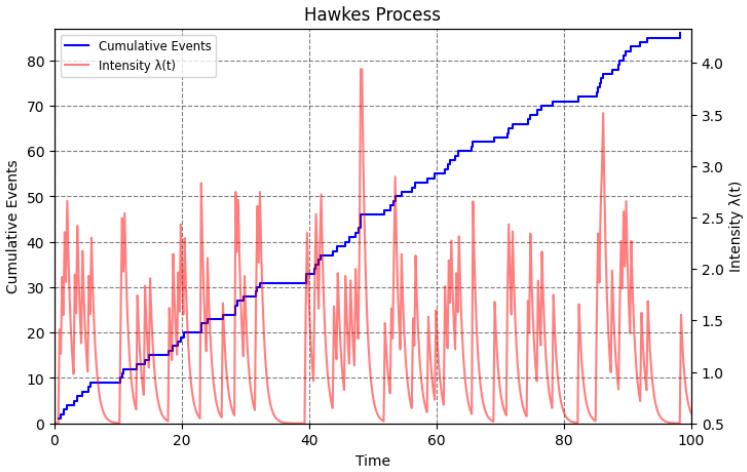
Like recommended in [10] we reparameterize the exponential kernel using the branching ratio $\eta = \alpha/\beta < 1$[36] to avoid numerical issues in estimation. Thus, the conditional

[36]The upper bound of the branching ratio ensures stability of the Hawkes process.

intensity of the Hawkes process is given by

$$\lambda(t, \alpha, \beta) = \mu + \sum_{t_i < t} \eta\beta e^{-\beta(t-t_i)}.$$

Furthermore to mimic the effect of "infinite memory" we simulate a series event times in the interval $[-M, 0)$, $M > 0$ of which we discard.



Hawkes process simulated via the thinning algorithm of Ogata for $(\mu, \alpha, \beta) = (0.5,\ 0.8\ , 1.0)$. The blue line plots the cumulated events on the left y axis and the red plots the conditional intensity on the right y axis.

The log likelihood for the reparameterized process is then

$$\ell_T(\theta) = \sum_{i=1}^{n_T} \log\left(\mu + \eta\beta \sum_{t_j < t_i} e^{-\beta(t_i - t_j)}\right) - \mu T - \eta\beta \int_0^T \sum_{t_i < t_j} e^{-\beta(t-t_j)} dt.$$

We fit the maximum likelihood estimator $\hat{\theta}_T$ using a limited memory Limited-memory Broyden-Fletcher-Goldfarb-Shanno from `scipy.optimize.minimize`. Other minimization methods could be explored to test the numerical framework. However numerical optimization methods are out of scope for this paper and L-BFGS suffices for our usage. We estimate the inverted lambda with a tolerance of $\varepsilon = 10^{-4}$ using Newtons method, where we set

$$t_{n+1} = t_n + \frac{s_i - \Lambda(t_n)}{\lambda(t_n)}, \quad t_0 = \frac{T s_i}{\Lambda(T)}.$$

If Newtons method doesn't converge i.e. $|t_n - s_i| \not< \varepsilon$ in 50 iterations (set to limit computational cost), the algorithm uses the bisection method as a backup. We estimate the hessian using finite differences to make computations light. This method is presented in chapter 10 of [22].

$$\partial^2_{\theta_i\theta_j} \ell(\hat{\theta}) \approx \frac{\ell(\hat{\theta} + h_i e_i + h_j e_j) - \ell(\hat{\theta} + h_i e_i - h_j e_j) - \ell(\hat{\theta} - h_i e_i + h_j e_j) + \ell(\hat{\theta} - h_i e_i - h_j e_j)}{4h_i h_j}.$$

$$i, j \in \{1, 2, 3\},\ h = \sqrt[3]{\varepsilon}(1 + \max\{|\hat{\theta}|, 1\}),$$

where we take $\varepsilon = 10^{-6}$. Other numerical methods could be implemented for the hessian, but we stick with finite differences for simplicity. The simulation is simplified in the

recursive computation of λ_k in step 7 and 10 of algorithm 4 where we fix $\lambda_k = \mu + \alpha^{1.2}\beta^{-0.1}$ for all $k \geq 0$. This form is chosen through iterative tests to reduce simulation time and sanity check failures (which we will return to). Fixing λ_k may contribute as a source of inaccuracy, however it will make computations light. We estimate the hessian and compute 95% confidence intervals for the parameter, respectively using bootstrap confidence intervals and asymptotic confidence intervals. We then measure the coverage by simulating these confidence intervals and checking if they cover the true parameter. We do this for a number of parameter configurations. To access the validity of the simulation we perform a "sanity check" where we check if $||\phi||_1 \geq 1$ and if the hessian evaluated in $\hat{\theta}_T$ is negative. We report back the sanity check failures for each parameter configuration and estimate the probability of sanity check failure $P_{CS}(T)$. To evaluate the sanity check we count the number of negative definite hessians over 100 simulations[37].

7.1 Results

μ_0	α_0	β_0	η_0	$p_{\text{SC}}(50)$	$p_{\text{SC}}(100)$
0.2	0.5	1.0	0.20	0.30	0.30
0.2	0.5	25.0	0.02	0.45	0.55
0.2	0.8	1.0	0.80	0.05	0.10
0.2	0.8	25.0	0.03	0.55	0.55
0.8	0.5	1.0	0.50	0.15	0.25
0.8	0.5	25.0	0.02	0.05	0.25
0.8	0.8	1.0	0.80	0.25	0.15
0.8	0.8	25.0	0.03	0.25	0.25

Parameter configurations and empirical probability of sanity check failure at $T = 50, 100$.

Our sanity checks roughly align with the ones seen in [10] for our configurations where $\beta = 1$. We see that there is a notable chance of the hessian being negative when $\beta = 25$. For the configurations where $\beta = 25$ we see that the sanity check has a higher failing rate for $T = 100$. This finding is not present in [10] and may be caused by the choice of λ_k which for larger β values will decay fast, making the thinning algorithm produce larger gaps between events. In addition the SC failures may also be caused by branching ratios being lower than the ones reported in [10]. Since there is a chance of the hessian being negative for a number of parameter configurations we compute 95% confidence if the hessian fails using `np.percentile` on θ_T^*.
In our simulation we set $K = 100$, do 199 bootstrap repetitions and check the coverage based on 20 simulations. Setting the number of simulations to 20 is sparse but is done to lower runtimes while simultaneously giving an indication of the coverage of the method for the given set of parameters.

[37] The whole simulation is available as an interactive notebook on Github, allowing for experimentation with different parameters and functions `https://github.com/CarlEmilKrarupJensen/Bootstrap-for-Hawkes-processes`.

Model 1A
($\mu = 0.2,\ \alpha = 0.5,\ \beta = 1.0$)

T	Method	μ	η	β	α
50	Asym	0.75	0.65	0.75	0.75
50	FIB	0.90	0.60	1.00	1.00
50	RIB	0.75	0.65	1.00	1.00
50	NP-FIB	0.80	0.65	1.00	0.85
50	NP-RIB	0.75	0.60	0.95	0.90
100	Asym	0.55	0.50	0.60	0.75
100	FIB	0.95	0.90	0.95	0.95
100	RIB	0.60	1.00	1.00	0.95
100	NP-FIB	0.95	1.00	0.95	0.85
100	NP-RIB	0.60	1.00	0.90	0.85

Model 1B
($\mu = 0.2,\ \alpha = 0.5,\ \beta = 25.0$)

T	Method	μ	η	β	α
50	Asym	0.10	0.10	0.10	0.10
50	FIB	0.55	0.50	0.15	0.55
50	RIB	0.35	0.60	0.20	0.55
50	NP-FIB	0.10	0.45	0.10	0.15
50	NP-RIB	0.15	0.45	0.15	0.20
100	Asym	0.25	0.25	0.15	0.25
100	FIB	0.60	0.75	0.20	0.75
100	RIB	0.50	0.80	0.20	0.75
100	NP-FIB	0.35	0.70	0.30	0.45
100	NP-RIB	0.30	0.75	0.30	0.40

Model 1C
($\mu = 0.2,\ \alpha = 0.8,\ \beta = 1.0$)

T	Method	μ	η	β	α
50	Asym	0.80	0.80	0.80	0.80
50	FIB	0.80	0.85	0.95	0.95
50	RIB	0.95	0.80	0.95	0.75
50	NP-FIB	0.80	0.75	0.95	0.85
50	NP-RIB	0.95	0.75	0.95	0.70
100	Asym	0.90	0.80	0.90	0.90
100	FIB	0.90	0.85	0.95	0.95
100	RIB	0.95	0.80	0.95	0.65
100	NP-FIB	0.90	0.85	0.95	0.80
100	NP-RIB	0.95	0.80	0.90	0.65

Model 1D
($\mu = 0.2,\ \alpha = 0.8,\ \beta = 25.0$)

T	Method	μ	η	β	α
50	Asym	0.40	0.35	0.20	0.40
50	FIB	0.70	0.75	0.45	0.70
50	RIB	0.55	0.80	0.35	0.70
50	NP-FIB	0.50	0.60	0.45	0.50
50	NP-RIB	0.50	0.70	0.40	0.55
100	Asym	0.30	0.30	0.20	0.30
100	FIB	0.70	0.70	0.30	0.65
100	RIB	0.65	0.80	0.30	0.70
100	NP-FIB	0.45	0.60	0.35	0.35
100	NP-RIB	0.40	0.55	0.35	0.45

Model 2A
($\mu = 0.8,\ \alpha = 0.5,\ \beta = 1.0$)

T	Method	μ	η	β	α
50	Asym	0.50	0.50	0.50	0.50
50	FIB	0.60	0.40	0.55	0.40
50	RIB	0.55	0.50	0.50	0.35
50	NP-FIB	0.60	0.50	0.55	0.45
50	NP-RIB	0.55	0.50	0.55	0.35
100	Asym	0.55	0.40	0.55	0.50
100	FIB	0.65	0.35	0.60	0.60
100	RIB	0.80	0.70	0.60	0.50
100	NP-FIB	0.65	0.65	0.50	0.35
100	NP-RIB	0.75	0.70	0.40	0.30

Model 2B
($\mu = 0.8,\ \alpha = 0.5,\ \beta = 25.0$)

T	Method	μ	η	β	α
50	Asym	0.30	0.30	0.05	0.35
50	FIB	0.45	0.40	0.35	0.40
50	RIB	0.45	0.50	0.30	0.40
50	NP-FIB	0.40	0.40	0.25	0.30
50	NP-RIB	0.40	0.45	0.25	0.30
100	Asym	0.15	0.15	0.00	0.10
100	FIB	0.55	0.55	0.25	0.40
100	RIB	0.35	0.40	0.25	0.25
100	NP-FIB	0.45	0.45	0.25	0.25
100	NP-RIB	0.35	0.30	0.20	0.30

Model 2C
($\mu = 0.8,\ \alpha = 0.8,\ \beta = 1.0$)

T	Method	μ	η	β	α
50	Asym	0.60	0.25	0.60	0.55
50	FIB	0.45	0.20	0.75	0.55
50	RIB	0.60	0.35	0.70	0.45
50	NP-FIB	0.50	0.20	0.60	0.45
50	NP-RIB	0.60	0.30	0.60	0.30
100	Asym	0.55	0.00	0.75	0.45
100	FIB	0.45	0.00	0.80	0.45
100	RIB	0.75	0.50	0.70	0.25
100	NP-FIB	0.35	0.50	0.65	0.25
100	NP-RIB	0.80	0.50	0.65	0.10

Model 2D
($\mu = 0.8,\ \alpha = 0.8,\ \beta = 25.0$)

T	Method	μ	η	β	α
50	Asym	0.15	0.15	0.00	0.00
50	FIB	0.40	0.45	0.00	0.30
50	RIB	0.25	0.45	0.05	0.25
50	NP-FIB	0.25	0.45	0.00	0.05
50	NP-RIB	0.25	0.40	0.00	0.10
100	Asym	0.15	0.10	0.15	0.10
100	FIB	0.65	0.65	0.30	0.25
100	RIB	0.30	0.40	0.30	0.25
100	NP-FIB	0.45	0.50	0.20	0.20
100	NP-RIB	0.35	0.50	0.15	0.20

Coverage of , FIB, RIB, NP-FIB and NP-RIB CIs across eight parameter configurations over different time intervals.

All models except for 1A show poor coverage of the branching ratio η, especially for $T = 50$. Moreover we see that η is covered worse when η_0 is small. However it seems that η is covered better when μ is small.
We also see that μ is insufficiently covered for $\mu_0 = 0.8$ and that the baseline intensity is covered particularly bad in model 2D. We also note that the time frame has a significant impact on the coverage of μ where the coverage when $T = 50$ is lower than when $T = 100$, which is also seen in [10]. We further note that μ is worse covered when the branching ratio is small suggesting an interaction between the coverage of μ and η.
Looking at α we see that the coverage is insufficient for nearly all configurations and bootstrap schemes with the best coverage seen in model 1C and 1A where α is high and β is low. When the branching ratio shrinks, α is worse covered. This observation harmonizes with parts of the ones made in [10]. Here it is seen that a higher value of α_0 gives better coverage. However in [10] it isn't observed that the coverage of α decreases along with the branching ratio. Another observation to note is that α is worse covered when μ is larger.
We see that β is significantly worse covered for $\beta_0 = 25$ compared to $\beta_0 = 1$ with model 2D showing little to no coverage when $T = 50$. [10] also report poor coverage of β for small branching ratios, aligning with our observations. Though model 1B and 2D sees an increase in the coverage increasing for all schemes when $T = 100$ suggesting impact by finite sample distortion.
Overall the best performing model is 1C where μ_0 is low and η_0 is large. While the poorest performance is seen in model 1B, 2B and 2D where η_0 is low.
Comparing the inference methods we see that the bootstrap CIs cover the true parameter better than the asymptotic confidence intervals. Except for a few cases the schemes perform better when T is larger. This finite sampling distortion is also reported in [10] where it is most apparent for the recursive schemes. Furthermore the fixed intensity schemes seem to suffer from under coverage i the same parameter configurations. Likewise the the recursive schemes exhibit the same behavior. Overall the bootstrap based inference show clear improvement of coverage compared to asymptotic inference.

Besides the coverage lengths of of the confidence intervals were also computed to examine the accuracy of the schemes. There are a few cases where the branching ratio is low, and some schemes produce long confidence intervals. Although a large majority of the bootstrap produced confidence intervals are shorter than the asymptotic confidence intervals, showing that the bootstrap schemes in most cases, yield more accurate estimates

than the asymptotic estimator.

7.2 Discussion

Comparing our results with those of [10] it is important to note that not all the same parameter configurations were tested. While this simulation study focuses on the combination of parameters with high and low baseline intensity and small vs. big branching ratio [10] explores more parameterizations though not all combinations are reported. Furthermore, to avoid long run times we only looked at the cases of $T = 50$ and $T = 100$ while [10] simulates for $T = 1000$. Despite these dissimilarities, our results exhibit several of the same characteristics as the ones reported in [10]. Most notably, we showcase the superior precision of the bootstrap schemes in covering the different parameterizations compared to the asymptotic CIs. Specifically we saw that the nonparametric schemes performed well across the parameter configuration proposing validity of the asymptotic theory corresponding to the ones we saw for a simple Poisson process. Another central observation is the coverage of μ, α and β, being lower for small branching ratios. The undercoverage of α for small branching rations is unreported in [10] and might therefore be first be severe for really small branching ratios. Moreover it indicates that α is more consistently estimated across the configurations compared to the other parameters The under coverage for low branching ratios is likely explained by sparsity of events in these configuration, making parameter estimation difficult. However this low intensity setting is also a good way of uncovering the robustness of the schemes. While the simulation results presented here suffer from inaccuracy and high risk of sanity check failure, we are still able to effectively demonstrate how the bootstrap schemes improve coverage of the true parameter.
There are avenues to improve the simulation study established here. A higher computational budget would help test out performance with reduced finite sample distortion for a larger number of bootstrap replications and Monte Carlo simulations. Updating λ_k iteratively instead of fixing it will likely also improve accuracy. Moreover a parallel investigation of the numerical methods used (newtons method, L-BFGS and finite differences) could be done to further optimize the methodology. Our simulation study particularly emphasizes the challenge of inference under low branching ratios where the intensity decays fast. In this setting we have seen how inference inaccuracy dominates smaller samples, and therefore demanding longer time intervals to reach sufficient coverage. The difficulty of inference of Hawkes processes with exponential kernel and low branching ratio suggests more advanced schemes to improve precision.

8 Conclusion and future work

In this work, we have introduced Hawkes processes using a cluster process representation. We have established likelihood and bootstrap based results for parameter inference. Hereafter we have tested these in a simulation focusing on the impact of the branching ratio. The simulation demonstrated that the bootstrap schemes can significantly improve inference for Hawkes processes with exponential kernel. We were able to capture many of the overall characteristics presented in [10]. The schemes have shown to be effective for processes with high branching ratios, even when applied to smaller samples. However all four schemes exhibited reduced accuracy for smaller branching ratios, an issue amplified by finite sampling distortion.

There are a number of ways to extend the overall methodology and the bootstrap methods seen in this paper. First, it would be natural to look at more parameter configurations and consider longer time horizons T. For large T values, one could compare the performance of the schemes presented here to a neural model like the one seen in [17] and explore how the techniques compare wrt. T. In addition, the fixed intensity bootstrap could potentially be improved. One idea is to bootstrap the intensity itself before fixing it and then performing the FIB/NPFIB schemes. Creating bootstrap fixed intensity bootstrap schemes (BFIB/NPBFIB). This may improve the estimation of the intensity function and help reduce inaccuracy in small samples.
Another direction to consider is to broaden the class of Hawkes processes considered. E.g. by treating some of the models presented in [3], where large part of the realm of Hawkes processes is mapped out. It would be interesting to apply our framework to a variety of processes, like multivariate models, nonlinear models, or as suggested in [10]; processes with real valued marks - provided that they satisfy assumption 3.1, 3.2 and 5.1. Hereby paving the way for possible development of more general inference tools and potentially improve inference methods for more varieties of Hawkes processes.

9 Appendices

A

A.1 A measure μ on a localized Borel space S is said to be locally finite if $\mu(A) < \infty$, $\forall A \in \hat{\mathcal{S}}$, where $\hat{S}$ denotes the class of sets $B \in \mathcal{S}$ with compact closure [14].
A.2 For any measure $\mu \in \mathcal{M}_S$ μ can be decomposed as

$$\mu = \nu + \sum_{k=1}^{n} \beta_k \delta_{X_k} \quad n \in \mathbb{N}.$$

Where ν is a diffuse measure, and $\forall k$ $\beta_k > 0$, X_k distinct in S.
A.3 Let $F : [a, b] \to \mathbb{R}$ be an increasing right continuous function. There exists a unique positive measure μ defined on the Borel sets of $[a, b]$ such that

$$\mu([a, t]) = F(t) - F(a), \quad t \in [a, b].$$

We call μ the Stieltjes measure associated with F see [7], p. 8.
A.4 Let $M(\cdot \mid x)$ denote the expectation measure for the descendants of a contributing process $N(\cdot \mid x)$, with parent at x. And $M_{(t)}(\cdot)$ the expectation measure for the population at generation t. We can then in parallel to [5] write the expectation measure for the next generation

$$\begin{aligned} M_{(t+1)}(B \mid N_t) = \sum_{i=1}^{Z_t} M(B \mid x_{it}) &= \int_S M(B \mid x) N_t(dx) \\ \implies \mathbb{E}(M_{(t+1)}(B \mid N_t)) &= \mathbb{E}\left(\int_S M(B \mid x) N_t(dx) \right) \\ \implies M_{(t+1)}(B) &= \left(\int_S M(B \mid x) M_{(t)}(dx) \right). \end{aligned}$$

Where we take expectation over N_t. The above expression can be seen as the number of expected children in the next generation in some region $B \in \mathcal{S}$.

A.5 The first factorial moment measure is given by $M_{[1]}(\cdot) = M(\cdot) = \mathbb{E}(N(\cdot))$. Let $k_i \geq 1$ with $k_1 + ... + k_r = k$, $r \in \mathbb{N}$ and $A_1, ..., A_r$ disjoint sets in $\mathcal{S}$. Then for for $k > 1$ the kth factorial moment measure $M_{[k]} : \mathcal{S}^k \mapsto \mathbb{R}_+$ is defined as

$$M_{[k]}(A_1^{(k_1)} \times ... \times A_r^{(k_r)}) = \mathbb{E}(N(A_1)^{(k_1)} \cdots N(A_r)^{(k_r)})$$

Where $N(A)^{(k_i)} = N(A)(N(A) - 1) \cdots (N(A) - k_i + 1)$.
A.6 Let N be a locally finite point process (i.e. $\mathbb{P}(N((s,t]) < \infty) = 1$ for all, $s < t$). Then

- $\lim_{s \to t+} N((0,s]) = N((0,t])$ right continuous
- $\lim_{s \to t-} N((0,s]) = N((0,t))$ exists since N is locally finite.

Thus, N is right continuous with left limits making N càdlàg.
For the stable Hawkes process the baseline cluster center Poisson process is locally finite. Furthermore since $\|\phi\|_1 < 1$ this guarantees each branch produces only finitely many offspring in a bounded time interval. Since there are almost surely only finitely many nonzero branches on $[0,T]$ the overall Hawkes process N_t is locally finite and thus càdlàg on every bounded interval.
A.7 The gamma function $\Gamma(x)$ satisfies the following two properties

$$\Gamma(n) = (n-1)! \quad n \in \mathbb{N}$$
$$\Gamma(\alpha) = \beta^\alpha \int_0^\infty x^{\alpha-1} e^{-\beta x} dx \quad \alpha, \beta > 0.$$

A.8 Let N_n be an integer valued random variable and let $(X_i)_{i \geq 1}$ be an $i.i.d$ sequence of random variables independent of N_n with $\mathbb{E}(X_i) = \mu$ and finite second moment, such that $N_n/n \xrightarrow{P} 1$ when $n \to \infty$, then

$$\frac{1}{n} \sum_{i=1}^{N_n} X_i \xrightarrow{P} \mu \quad \text{when} \quad n \to \infty$$

and

$$\frac{1}{N_n} \sum_{i=1}^{N_n} X_i \xrightarrow{P} \mu \quad \text{when} \quad n \to \infty.$$

This result is introduced in [2]
A.9 To get the asymptotic result for the bootstrap score let $n_T^*(s)$ be defined as in theorem 6.1 and recall. $\frac{n_T^*(s) - \lfloor Ts \rfloor}{\sqrt{\kappa_T T}} \xrightarrow[P]{d^*} B_s$. Recall the bootstrap score

$$S_T^*(\hat{\theta}_T) = -\hat{\theta}_T^{-1} \sum_{i=1}^{n_T^*} (v_i^* - 1) = n_T^* \hat{\theta}_T^{-1} - T.$$

Dividing by $\sqrt{T}$ and applying theorem 6.1

$$\begin{aligned}\frac{1}{\sqrt{T}} S_T^*(\hat{\theta}_T) &= -\hat{\theta}_T^{-1} \frac{1}{\sqrt{\hat{\kappa}_T T}} \sum_{i=1}^{n_T^*} (v_i^* - 1) \\ &= \frac{1}{\hat{\theta}_T} \frac{n_T^* - \hat{\theta}_T T}{\sqrt{\hat{\kappa}_T T}} \\ &= \frac{1}{\hat{\theta}_T} \frac{n_T^* - \lfloor T\hat{\theta}_T \rfloor}{\sqrt{\hat{\kappa}_T T}} + o_p(1) \xrightarrow[P]{d^*} \frac{1}{\theta_0} B_{\theta_0} \sim \mathcal{N}(0, \theta_0^{-1}),\end{aligned}$$

using $\hat{\kappa}_T \to 1$ and consistency of the estimator.

For the Hessian we have

$$\begin{aligned}T^{-1} H_T^*(\hat{\theta}_T) &= \frac{n_T^*}{T\hat{\theta}_T^2} \\ &= \frac{n_T^* - T\hat{\theta}_T + T\hat{\theta}_T}{T\hat{\theta}_T^2} \\ &= \frac{1}{\hat{\theta}_T} + \frac{n_T^* - T\hat{\theta}_T}{T\hat{\theta}_T^2} \\ &= \frac{1}{\hat{\theta}_T} + O_p(T^{-1/2}) \xrightarrow[P]{P^*} \frac{1}{\theta_0}.\end{aligned}$$

To access the asymptotic distributional form of the estimator we can then do a Taylor expansion of the score around $\hat{\theta}_T$, for some $\tilde{\theta}_T \in \Theta$ between θ_T^* and $\hat{\theta}_T$.

$$\begin{aligned}S_T^*(\hat{\theta}_T) = S_T^*(\theta_T^*) + H_T^*(\tilde{\theta}_T)(\theta_T^* - \hat{\theta}_T) = 0 \\ \implies S_T^*(\theta_T^*) = -H_T^*(\tilde{\theta}_T)(\theta_T^* - \hat{\theta}_T) \\ \implies T^{-1/2} S_T^*(\theta_T^*) = -H_T^*(\tilde{\theta}_T)(\theta_T^* - \hat{\theta}_T) T^{-1/2} \\ \implies T^{-1/2} S_T^*(\theta_T^*) \left(T^{-1} H_T^*(\tilde{\theta}_T)\right)^{-1} = T^{1/2}(\hat{\theta}_T - \theta_T^*).\end{aligned}$$

With the previous convergence results we can write

$$\begin{aligned}T^{1/2}(\theta_T^* - \hat{\theta}_T) = -\frac{T^{-1/2} S_T^*(\theta_T^*)}{T^{-1} H_T^*(\tilde{\theta}_T)} \xrightarrow[P]{d^*} -\mathcal{N}(0, \theta_0^{-1})\theta_0 \\ \implies T^{1/2}(\theta_T^* - \hat{\theta}_T) \xrightarrow[P]{d^*} \mathcal{N}(0, \theta_0).\end{aligned}$$

Finally for the likelihood ratio asymptotics we can do a second order Taylor expansion of the log likelihood in $\hat{\theta}_T$

$$\ell_T(\theta_T^*) - \ell_T(\hat{\theta}_T) = -\frac{1}{2}(\theta_T^* - \hat{\theta}_T)^2 H_T^*(\theta_T^*) + o_p\big((\theta_T^* - \hat{\theta}_T)^2\big).$$

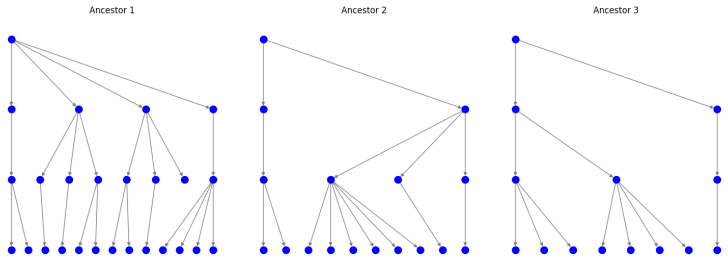
Giving us

$$\begin{aligned}LR^*(\theta_T^*) &= 2(\theta_T^* - \hat{\theta}_T)^2 H_T^*(\theta_T^*) + o_p(1) \\ &= T^{-1}(T^{1/2}(\theta_T^* - \hat{\theta}_T))^2 T T^{-1} H_T^*(\theta_T^*) + o_p(1) \xrightarrow[P]{d^*} B_{\theta_0}^2 \theta_0^{-1} \\ &\implies LR^*(\theta_T^*) \xrightarrow[P]{d^*} \left(\sqrt{\theta_0^{-1}} B_{\theta_0}\right)^2\end{aligned}$$

Since $B_{\theta_0}\theta_0^{-1/2}$ is standard normal we have $\left(B_{\theta_0}\theta_0^{-1/2}\right)^2 \sim \chi_1^2$, giving us the wished result.

B

B.1 Finite branching processes The finite branching process provides a framework for modeling the evolution of populations. It is used in for example epidemiology [1], like what was intended for the Hawkes process. As we will see the two processes are also closely related mathematically.
Let a population be characterized by the state x_{it} on the space S and the count $Z_t \in \mathbb{N}$ over discrete time $t = 1, 2, ...$ The population making up generation t can be described by a finite point process N on S. Generation $t+1$ is built as the sum of the contributing processes $N(\cdot \mid x_{it})$. If we let generation t be characterized by the locations $\{x_{it} \mid i = 1, ..., Z_t\}$ and the count Z_t. The contributing processes to generation $t+1$ are mutually independent, independent of Z_t and independent of the history $\mathscr{F}_t = \sigma(\{N_s : 1 \leq s \leq t\})$. Whenever we refer to the "history" of a process throughout this paper the previous filtration is the one we mean.



Visualization of a finite branching process over 4 generations starting at 3 ancestors where each contributing process is Poisson.

An important property of this process is that the population will go extinct almost surely if the expectation measure[38] $M < 1$.[40] To better understand the mechanisms of the finite branching process, we give some of its properties.

Definition B.1 *Let $\xi : S \to \mathbb{R}$, be a bounded Borel measurable function and let $B \in \mathcal{S}$ then for a realization of a finite point process $\{x_i \mid i = 1, ..., N\}$, then the random product $\prod_{i=1}^{N(S)} \xi(x_i)$ is well defined, and if $|\xi(x)| \leq 1$ for all for all $x \in S$ then $\mathbb{E}(\xi(x))$ exists. The probability generating functional (p.g.fl.) of N is then given by*

$$G(\xi) = \mathbb{E}\left(\prod_{i=1}^{N(S)} \xi(x_i)\right).$$[41]

Now suppose that we are given a family of distributions $\mathcal{M}_{\mathcal{Y}}$, with $\mathcal{Y}$ being a Polish space[42], and the family is indexed by the random variables $x \in S$. We denote distributions in $\mathcal{M}_{\mathcal{Y}}$, $\mathcal{P}(\cdot \mid x)$. Suppose x has distribution $\Pi(\cdot)$ on $\mathcal{S}$. Then if $\mathcal{P}(A \mid x)$ is measurable for all $A \in \mathcal{B}(\mathcal{M}_{\mathcal{Y}})$ then there exists a process with distribution

$$\mathcal{P}(A) = \int_S \mathcal{P}(A \mid x)\Pi(dx)$$

on $\mathcal{M}_{\mathcal{Y}}$.

[38] The expectation measure and the expectation measure[39] for a generation is presented in A.4
[40] c.f. [14], p. 287.
[41] See [5] p. 141.
[42] Complete separable metric space.

Lemma B.1 *Suppose we are given a family of measurable point process on $\mathcal{Y}$, with p.g.f.ls. $G(\xi|x)$, where ξ is a real valued Borel function, with $1-\xi$ vanishing outside some bounded set*[43]*, and satisfying $0 \leq \xi(x) \leq 1$ for all $x \in S$. Further suppose we are given a measurable mapping $X : \Omega \to S$, where $\Pi(\cdot)$ is the probability distribution on $\mathcal{S}$ induced by x. The p.g.f.l. of a point process on $\mathcal{Y}$ is then given by*

$$G(\xi) = \int_S G(\xi|x)\Pi(dx).$$[44]

Analogous to the p.g.fl. above we can express the finite branching process over a set $B \in \mathcal{S}$ as:

$$N_{t+1}(B) = \sum_{i=1}^{Z_t} N(B \mid x_{it}) \quad B \in \mathcal{S}, t = 1, 2, ...$$

With Z_t is a finite point process and where $t+1$ is denotes the index of the generation. Here the offspring process $N(\cdot \mid x)$ may depend on the state x, where the number of children can be described via probability distributions $\{p_n(x) \mid n = 1, 2, ...\}$ while their joint spacial distribution conditional on having n children can be described by $\Xi_n(\cdot \mid x)$[45].

If we assume p_n and Π_n are measurable and let $\xi : S \to \mathbb{R}$, be measurable, then the p.g.fl. of the offspring G is measurable and:

$$G_{t+1}(\xi \mid N_t) = \prod_{i=1}^{Z_t} G(\xi \mid x_{it})$$

Recall $G(\xi \mid x) = \mathbb{E}\left(\prod_{y \in N(\cdot|x)} \xi(y)\right)$. Taking expectation over N_t gives us

$$\begin{aligned} G_{t+1}(\xi) &= \mathbb{E}\left(\prod_{i=1}^{Z_t} G(\xi \mid x_{it})\right) \\ &= G_t(G(\xi \mid x)). \end{aligned}$$

This property tells us that the populations evolves according to the same rule from generation to generation.

[43] Meaning that $1-\xi$ is 0 outside some bounded set.

[44] See [5], p. 235.

[45] $\Xi_n(\cdot)$ defined on $S^{(n)} = S \times ... \times S$ determines the joint distribution of points and is symmetrical, see [5], p. 121.

References

[1] Daniel Ahlberg. Epidemics and branching processes. Technical report, Stockholm University, May 2021.

[2] F. J. Anscombe. Large-sample theory of sequential estimation. *Biometrika*, 36(3/4):601–607, 1949.

[3] Emmanuel Bacry, Iacopo Mastromatteo, and JeanFrançois Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1:1550005, 2015.

[4] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996.

[5] D.J Daley and D. Vere Jones. An introduction to the theory of point processes. 1(1):18–28, 143–144, 151, 236–244, 367–368, 414, 504–505, 550, 1998.

[6] M. H. A. Davis. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):353–388, 1984.

[7] Thomas Duquesne. Document de travailpour le cours "modèles d'actifs avec sauts". master 2, finance. 1:32, 68, 2023.

[8] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[9] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6):933–949, 2018.

[10] Anders Rahbek Jacob Stærk-Østergaard Giuseppe Cavaliere, Ye Lu. Bootstrap inference for hawkes and general point processes. *Journal of Econometrics*, (1):1–33, 2022.

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

[12] Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, (1):493–494, 1974.

[13] SØREN TOLVER JENSEN and ANDERS RAHBEK. Asymptotic inference for nonstationary garch. (1):1003–1026, 2004.

[14] Olav Kallenberg. Foundations of modern probability. 1(3):15, 282, 277 287, 321–322, 500, 2021.

[15] Haya Kaspi and Avi Mandelbaum. On harris recurrence in continuous time. *Mathematics of Operations Research*, 19(1):211–222, February 1994.

[16] Patrick J. Laub, Young Lee, Philip K. Pollett, and Thomas Taimre. Hawkes models and their applications. *Annual Review of Statistics and Its Application*, 12(1):233–258, 2025.

[17] Kyungsub Lee. Recurrent neural network based parameter estimation of hawkes model on high-frequency financial data. *Finance Research Letters*, 55:103922, July 2023.

[18] Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[19] George E. Mohler, Martin B. Short, Spencer Malinowski, Michael Johnson, George E. Tita, Andrea L. Bertozzi, and Paul J. Brantingham. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[20] Yoshiko Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, (1):243–261, 1978.

[21] Yosihiko Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[22] Gilles Pagès. *Numerical Probability An Introduction with Applications to Finance*. Laboratoire de Probabilités, Statistique et Modélisation Sorbonne Université. Springer International, Cham, Switzerland, 1 edition, 2018.

[23] Mads Bonde Raad, Susanne Ditlevsen, and Eva Löcherbach. Stability and mean-field limits of age dependent hawkes processes. *Annales de l'institut Henri Poincaré (B) Probability and Statistics*, 56(3):1958–1990, 2020.

[24] Nakahiro Yoshida and Sylvain Clinet. Quasi likelihood analysis for point processes with a shift–volume intensity model. *Stochastic Processes and their Applications*, 127(4):1214–1255, 2017.

Deklaration for anvendelse af generative AI-værktøjer (studerende)

På Københavns Universitet udfører vi vores arbejde med **ansvarsfølelse** og **respekt for samfund, kulturarv, miljø og mennesker** omkring os.
Integritet, ærlighed og **transparens** er forudsætninger for akademisk arbejde. Vi forventer derfor, at eksamenspræstationer afspejler **den studerendes egen læring og selvstændige indsats**.
Akademisk arbejde baserer sig altid på andres indsigter, viden og bidrag, men altid med **grundig anerkendelse, respekt** og **kreditering** af dette arbejde.
Dette gælder også ved brug af generativ kunstig intelligens.

Vejledning

Brug af generativ AI ved eksamen

I henhold til KU's regler for brugen af værktøjer, der er baseret på generativ AI (GAI), skal du være transparent om din anvendelse af teknologien, fx i dit metodeafsnit og/eller ved at udfylde og vedlægge nedenstående deklarationsskabelon som bilag til skriftlige opgavebesvarelser.
Når du skriver din deklaration, er det vigtigt, at læseren får et tydeligt billede af, om og hvordan generativ AI har bidraget til det endelige produkt.

Hvis det er besluttet, at du skal bruge skabelonen i dit fag, skal du også benytte den, når du *ikke* har anvendt GAI-værktøjer som hjælpemiddel. I dette tilfælde skal du dog blot krydse af, at du ikke har brugt GAI, og behøver ikke at udfylde resten.
Ved at deklarere din brug af GAI-værktøjer sikrer du, at der ikke opstår udfordringer i forhold til reglerne om eksamenssnyd.

I kurser, hvor brugen af GAI er integreret i fagligheden, kan refleksion og kritisk vurdering af anvendelsen af GAI-værktøjer også indgå som en del af et metodeafsnit i din opgave. Spørg din underviser eller vejleder, hvis du er tvivl, om det er tilfældet i dit kursus.

Hvis generativ AI er objekt for din undersøgelse, vil det fremgå af dine forskningsspørgsmål, din metodebeskrivelse, din analyse og konklusion, hvilken rolle GAI spiller i din opgave. Hvis du

samtidig også bruger GAI som hjælpemiddel i processen, skal du deklarere denne anvendelse særskilt.

Opmærksomhedspunkter:

- Hvis GAI er et tilladt hjælpemiddel på dit kursus, må du anvende GAI til dialog og sparring under udarbejdelsen af din opgave, men du må **ikke** overlade udfærdigelsen af din opgavebesvarelse til GAI-værktøjer, selvom alle hjælpemidler er tilladt.
- Hvis materiale fra GAI inkluderes som kilde (direkte eller i redigeret form) i din besvarelse, gælder de samme krav om brug af citationstegn og kildehenvisning som ved alle andre kilder, da der ellers vil være tale om plagiat.
- Brug aldrig personhenførbare, ophavsretsbeskyttede eller fortrolige data i et AI-værktøj.
- Husk altid at undersøge gældende regler og retningslinjer for brug af generativ AI på KU.
- Læs kursusbeskrivelsen grundigt. Det er vigtigt at du ved, hvilke anvendelser der er tilladt i dit kursus. Der kan eventuelt være yderligere krav om dokumentation, fx at du skal beskrive dine centrale prompts og evt. kildemateriale (hvad har du givet af kontekst, hvad har du fodret værktøjet med, hvad har du bedt værktøjet om at gøre), beskrive outputtet (hvilke svar du fik af værktøjet), beskrive processen, f.eks. historik og iterationer (hvis du har skrevet frem og tilbage med værktøjet ad flere omgange for at komme frem til et brugbart svar).
- Tal med din underviser eller vejleder, hvis du er i tvivl.

Deklaration for anvendelse af generative AI-værktøjer (studerende)

☒ **Jeg/vi har benyttet generativ AI som hjælpemiddel/værktøj** *(sæt kryds)*

☐ **Jeg/vi har <u>IKKE</u> benyttet generativ AI som hjælpemiddel/værktøj** *(sæt kryds)*

Hvis brug af generativ AI er tilladt til eksamen, men du ikke har benyttet det i din opgave, skal du blot krydse af, at du ikke har brugt GAI, og behøver ikke at udfylde resten.

Oplist, hvilke GAI-værktøjer der er benyttet, inkl. link til platformen (hvis muligt):

ChatGPT4o

Beskriv hvordan generativ AI er anvendt i opgaven:

1) *AI er hovedsageligt blevet brugt til kildesøgning*
2) *Ydermere er AI blevet brugt til afklaring af spørgsmål vedr. eksempelvis notation i artikler samt udbedring af forståelse af nye begreber der er kommet frem i løbet af skriveprocessen.*
3) *ChatGPT4o er blevet brugt til at give forslag til at formatere koden af simulationen der præsenteres i kapitel 7. Der er linket til koden i opgaven.*
4) *Især i starten af projektet brugte jeg AI til at søge kilder der kunne bruges.*
5) *Alt output er læst med et kritisk blik. Især svarene på matematiske spørgsmål er blevet vurderet eftersom ChatGPT4o har en tendens til at videregive misinformation når det gælder disse.*

NB. GAI-genereret indhold brugt som kilde i opgaven kræver korrekt brug af citationstegn og kildehenvisning. Læs retningslinjer fra Københavns Universitetsbibliotek på KUnet.