Diss. ETH No. 24278

Perspectives on Hawkes Processes

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

MATTHIAS KIRCHNER

MSc Mathematics, ETH Zurich, Switzerland
born on 18.10.1978
citizen of Germany

accepted on the recommendation of

Prof. Dr. Paul Embrechts, examiner
Prof. Dr. Valérie Chavez-Demoulin, co-examiner
Prof. Dr. Alan Hawkes, co-examiner
Prof. Dr. Thomas Mikosch, co-examiner

2017

DOI: 10.3929/ethz-b-000161487
ISBN: 978-3-906327-81-5

Abstract

This thesis addresses Hawkes point processes in seven scientific papers. We build theoretical bridges between Hawkes processes and other mathematical concepts—such as time series, branching random walks, or graph theory. In **Paper A**, we represent monotype Hawkes processes as limits of time-series based point processes. We examine the corresponding time series, the integer-valued autoregressive (INAR) time series of infinite order, in some detail. Furthermore, we point out structural analogies between Hawkes processes and INAR time series. In **Paper B**, we represent multitype Hawkes processes as type/space projections of certain branching random walks. This representation allows to generalize the convergence result from Paper A to the multitype case. Furthermore, it opens the door to generalizations of Hawkes processes that might be interesting in applications. In **Paper C**, we introduce a nonparametric estimation procedure for multitype Hawkes processes: we discretize Hawkes-process data. From Paper A and Paper B, we know that the resulting bin-count sequences can be approximated by INAR time series. Thus, we estimate the INAR parameters by standard methods and¡ retranslate the results into the point process world. In **Paper D**, we represent multitype Hawkes processes as directed weighted graphs. These 'Hawkes graphs' summarize the branching structure of a Hawkes process in a compact, yet meaningful way. We point out how the graphical perspective is also fertile mathematically, implementation-wise, and pedagogically. Furthermore, we apply the estimation method from Paper C to infer the Hawkes graph from large datasets. We pay special attention to computational issues. In **Paper E**, we apply the methods and concepts from Paper C and Paper D to limit-order-book data. In particular, we extend our estimation procedure to the marked case. The various estimation results allow insights into market microstructure. In **Paper F**, we give the results of a simulation study, where we compare our estimation procedure with maximum-likelihood estimation. Finally, in **Paper G**, we consider a certain critical case of the monotype Hawkes process. We study the critical Hawkes process by applying results from critical cluster fields, renewal theory, and regular variation. We discuss a possible Poisson embedding and a Palm version of the critical Hawkes process. Our methods give possible directions for the open discussion of multitype critical Hawkes processes as well as of critical INAR times series.

Kurzfassung

Diese Doktorarbeit stellt sieben wissenschaftliche Abhandlungen zu Hawkesprozessen vor. Wir konstruieren theoretische Brücken zwischen Hawkesprozessen und anderen mathematischen Objekten wie Zeitreihen, Verzweigungsirrfahrten und Graphen. In **Paper A** stellen wir Hawkesprozesse als Grenzwerte von zeitreihenbasierten Punktprozessen dar. Wir untersuchen die entsprechende Zeitreihe – die ganzzahlige autoregressive (INAR) Zeitreihe von unendlicher Ordnung und zeigen strukturelle Analogien zwischen Hawkesprozessen und INAR Zeitreihen auf. In **Paper B** stellen wir Hawkesprozesse mit mehreren Punkttypen als Typ/Raum-Projektionen bestimmter Verzweigungsirrfahrten dar. Diese Darstellung erlaubt die Verallgemeinerung des Konvergenzresultats von Paper A zum Fall mit mehreren Punkttypen. Zudem führt diese Darstellung zu Verallgemeinerungen von Hawkesprozessen, die für Anwendungen interessant sein könnten. In **Paper C** stellen wir eine nichtparametrische Schätzmethode für Hawkesprozesse mit mehreren Punkttypen vor: Wir diskretisieren Daten eines Hawkesprozesses. Von Paper A und Paper B wissen wir, dass die resultierende Zahlenfolge durch eine INAR-Zeitreihe angenähert werden kann. Also schätzen wir die INAR-Parameter und übersetzen die Resultate wieder zurück in die Punktprozesswelt. In **Paper D** stellen wir Hawkesprozesse als gerichtete und gewichtete Graphen dar. Diese „Hawkesgraphen“ fassen die Verzweigungsstruktur eines Hawkesprozesses kompakt, aber doch aussagekräftig zusammen. Wir zeigen auf, inwiefern diese graphische Perspektive auch mathematisch, programmiertechnisch und pädagogisch fruchtbar ist. Ausserdem wenden wir die Schätzmethode von Paper C an, um den Hawkesgraph von grossen Datensätzen abzuleiten – mit besonderem Augenmerk auf Implementierungsfragen. In **Paper E** wenden wir die Methoden und Konzepte von Paper C und Paper D auf Orderbuchdaten an. Insbesondere erweitern wir unsere Schätzmethode auf den markierten Fall. Die verschiedenen Schätzresultate geben Einblick in die Mikrostruktur von Märkten. In **Paper F** präsentieren wir die Resultate einer Simulationsstudie, in der wir unsere Schätzmethode mit einer „Maximum Likelihood“-Schätzung vergleichen. Zuletzt betrachten wir in **Paper G** einen gewissen kritischen Fall des einfachen Hawkesprozesses. Wir untersuchen den kritischen Hawkesprozess, indem wir Resultate zu kritischen Clusterfeldern, Erneuerungstheorie und regulärer Variation anwenden. Wir erörtern zudem eine mögliche Poissoneinbettung sowie eine Palmversion des kritischen Hawkesprozesses. Unsere Methoden bieten sich für die noch offene Untersuchung von kritischen Hawkesprozessen mit mehreren Punkttypen sowie von kritischen INAR Zeitreihen an.

Acknowledgments

First and foremost, I would like to express my deep gratitude towards Paul Embrechts, my doctoral advisor, for giving me the great intellectual and social opportunity of a long-term research project at ETH RiskLab. Paul Embrechts introduced me to the beautiful topic of Hawkes processes. Under his guidance, I have developed my own language of scientific communication and judgement. Apart from my research, Paul Embrechts has always shown interest in my personal life, my family, and my music. His advice has made me a different person, and I consider Paul Embrechts to be my true academic father. Next, I want to thank my co-examinors: Valérie Chavez-Demoulin co-supervised my work from the very beginning. Her constructive and encouraging feedback helped me endure that first difficult period of academic dry spell. The contact with Alan Hawkes, the godfather of the examined stochastic process, is a particular honor. I thank him for his generosity. His openness towards my work and his personal kindness shall always be an example for me when dealing with younger students. Thomas Mikosch proofread my first papers in detail and thus facilitated that difficult first publication process. He was always there for me when I needed any special technical help. I learned a lot from him, his lecture notes, and his textbooks. Thomas Mikosch's way of communicating complex matters is exemplary.

I thank the institution ETH not only for providing a stimulating academic environment, but also for being a very reliable employer. Support processes were flawless: IT-support desk, Druckzentrum, ETH- and D-MATH-library, Polysnack, cleaning ladies, and personal administration. Not many scientists have the opportunity to work under such professional and well-organized conditions. I also express my gratitude towards the Department of Mathematics at the University of Bern for providing the perfect start to my mathematical curriculum—in particular, I thank Lutz Dümbgen who supported me from the very beginning. I would also like to mention my other employer of the last few years, the institute IVP NMS at the University of Education in Bern. In particular, I thank Martin Stadelmann, our dean, who always supported and appreciated my mathematical research.

I would like to acknowledge the tremendous work of R-programmers in the academic world. In particular, the authors of the package 'Matrix' and of the package 'igraph' made my life a lot easier. Marius Hofert was of invaluable assistance when I took my first footsteps in R. I also want to thank Rob Almgren and Isabel Marques da Silva who shared their expertise on high-frequency financial data, respectively, integer-valued time series so openly.

A special thanks goes to my long-standing office mate Philippe Deprez. Where would I have been without his good mood, his open ear, and his help in any possible way? I am also grateful towards our secretary, Galit Shoham, for being the kind angel of RiskLab. I thank my student collaborators Aritz Bercher and Silvan Vetter for their energy and time invested in our joint projects. I thank David Stefanovits, Thibault Vatter, and Phyllis Wan for their valuable feedback on parts of my work. I also like to thank my ETH friends Michel Baes, John Ery, Erwan Koch, as well as my former colleagues Laurent Huber, Edgars Jacobsons, Anne MacKay, Annina Saluz, and Mario Sikic who made my ETH days interesting beyond research. Due to my duties besides ETH, I was not the most social member of D-MATH Group 3. But I was forgiven and collaboration during exam sessions and otherwise were nothing but a pleasure.

I thank my mother, Rita Kirchner, who read every single word (but for the acknowledgments) of this thesis and gave it its current 'Durham-English flavor'. I also thank my father, Thomas Kirchner, who taught me to see the beauty in mathematics and music. Finally, I want to thank my family: my dearest wife, Martina, for her love, patience, and support, and my children, Kolja, Klara and Meret, who grew even faster than this thesis. As Barack Obama said at his farewell ceremony in spring this year: 'Of all that I have done ... I'm most proud to be your dad.' (Then Obama started crying.) In any case, it took me five years to realize this: the tree of life is the most beautiful Hawkes process!

M.K., June 2017 in Zurich.

Contents

Abstract **iii**

1 Introduction **1**

2 Hawkes processes **5**
2.1 Motivation 5
2.2 Autoregressive perspective 7
2.3 Branching perspective 8
2.4 Multitype Hawkes processes 10
2.5 Marked Hawkes processes 11
2.6 Literature 11

3 Main contributions **17**
3.1 Time series perspective (A, B) 17
3.2 Branching-random-walk perspective (B, G) 20
3.3 Perspectives on Hawkes-process estimation (C, D, E, F) 25
3.4 Graphical perspective (D, E) 29
3.5 Limit-order-book modeling with Hawkes processes (C, E) 31
3.6 Perspectives on the critical case (G) 33
3.7 Perspectives on future research 36

4 Accompanying papers **41**
Paper A. Hawkes and INAR(∞) processes 41
Paper B. Hawkes forests 81
Paper C. An estimation procedure for the Hawkes process 105
Paper D. Hawkes graphs 157
Paper E. Hawkes model specification for limit order books 191
Paper F. A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation 221
Paper G. A note on critical Hawkes processes 237

Bibliography **253**

Accompanying papers

A Matthias Kirchner.
Hawkes and INAR(∞) processes.
Stochastic Processes and their Applications, **162**(8):2494–2525, 2016.

B Matthias Kirchner.
Hawkes forests.
Submitted.

C Matthias Kirchner.
An estimation procedure for the Hawkes process.
Quantitative Finance, **17**(4):571–595, 2017.

D Paul Embrechts, Matthias Kirchner.
Hawkes graphs.
Theory of Probability and Its Applications, **62**(1):163–193, 2017.

E Matthias Kirchner, Silvan Vetter.
Hawkes model specification for limit order books.
Submitted.

F Aritz Bercher, Matthias Kirchner.
A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation.
Submitted.

G Matthias Kirchner.
A note on critical Hawkes processes.
Working paper.

1 Introduction

> "It has interested me for years. I have not made much progress, but my partial failures may stimulate quicker minds."
>
> Introduction of Greenwood (1946)

Hawkes processes are stochastic models for event streams. Their flexible, yet tractable structure makes Hawkes processes attractive for applications in areas as different as earthquake modeling, financial mathematics, or even crime prediction. Our thesis consists of seven scientific papers that contribute several relevant findings to the field—on the theoretical as well as on the applied side. The main object of our work is to build bridges between Hawkes processes and other mathematical concepts, such as time series, branching random walks, or graph theory. We hope that these new perspectives will deepen the understanding of Hawkes processes and will be inspiring for future research. We start with short summaries of the accompanying papers:

In **Paper A**, we represent monotype Hawkes processes as limits of time-series based point processes. We examine the involved time series model—that is, the integer-valued autoregressive time series of infinite autoregressive order (INAR(∞))—in some detail. We give standard autoregressive and moving-average representations. We calculate generating functions as well as second moments of the INAR(∞) model and point out their analogy with the corresponding entities of Hawkes processes. We conclude that Hawkes processes are continuous-time versions of INAR(∞) processes and, vice versa, INAR(∞) processes are discrete-time versions of Hawkes processes.

In **Paper B**, we represent multitype Hawkes processes as type-space projections of specific branching random walks of infinitely many particles. We call these branching random walks 'Hawkes forests'. The Hawkes-forest representation of a Hawkes process permits a generalization of the convergence result from Paper A to the multitype case: we define approximating (multivariate) INAR-based point processes with respect to (ω-wise) exactly the same underlying branching structure as the target Hawkes process. This leaves only the positions of the points for approximation—which is straightforward. General Hawkes forests yield much more

general point processes than Hawkes processes. We show, how these generalizations might be interesting for applications.

In **Paper C**, we introduce a nonparametric estimation procedure for multitype Hawkes processes: we discretize Hawkes-process data. From Paper A and Paper B, we know that the resulting bin-count sequences can be approximated by INAR time series. Thus, we calibrate INAR coefficients with respect to the bin counts and retranslate the results into the point process world. In the INAR context, we prove consistency and asymptotic normality for conditional-least-squares estimates. We also give corresponding covariance estimates. Simulation studies show that the inference results carry over to the Hawkes-process estimates; the coverage rates of confidence intervals are perfect. We propose selection methods for the estimation parameters (discretization and truncation). Finally, we consider a case study, where we apply the estimation procedure on order streams in a limit order book.

In **Paper D**, we represent multitype Hawkes processes as directed weighted graphs. These 'Hawkes graphs' summarize the branching structure of a Hawkes process in a compact, yet meaningful way. We point out how the graphical perspective is also fertile mathematically: for example, we give a graph-based criterion for subcriticality of the corresponding Hawkes process. Furthermore, the Hawkes graph allows an engineering-type approach when constructing or modifying event-streams structures. In a second step, we infer the Hawkes graph from event-stream data by applying the estimation method from Paper C. We pay special attention to computational issues. For example, we give most efficient algorithms to calculate covariance estimates—the computational bottleneck of our estimation procedure. The various estimation parameters make our method remarkably flexible. This flexibility allows applications of the procedure on event-streams of virtually any number of types. This makes the Hawkes graph a valid candidate for describing 'big data' in the event-stream context.

In **Paper E**, we apply the methods and concepts from Paper C and Paper D to limit-order-book data. In particular, we extend our estimation procedure to multitype *marked* Hawkes processes. With hardly any a priori assumptions, we derive a fully parametric Hawkes-based model for the event streams of market orders, limit orders, and cancelations. We find that the order-book imbalance describes the probability whether an event occurs on the bid side of the book in a perfect manner. This makes the imbalance an important summary of the state of a limit order book. Thus, we let the baseline intensities of our model depend on the imbalance. In other words, our model exhibits a mélange of past and state dependence. We indicate how the model may be used for order-type prediction.

In **Paper F**, we present the results of a simulation study, where we compare our estimation procedure with maximum-likelihood estimation. Computation-time wise, the advantages of our method are eminent; statistically, our method competes well. Furthermore, we find that MLE

estimates are considerably biased for nearly critical Hawkes processes and moderate sample sizes.

In **Paper G**, we consider a certain critical case of the Hawkes process. We observe that the law of a critical Hawkes process is 'cluster invariant'. This point process property is well-studied. Therefore, we obtain many properties of the critical Hawkes process from the relevant literature. Furthermore, we analyze the critical Hawkes process from a renewal-theory as well as from a branching-random-walk perspective. We present a possible Poisson embedding as well as a Palm version of the critical Hawkes process. Our considerations indicate that the existence of a critical Hawkes process is intimately related to the recurrence/transience dichotomy of an embedded random walk. Finally, we point out how the various approaches could be used for the open discussion of multitype critical Hawkes processes as well as of critical INAR(∞) time series.

The thesis is organized as follows: in Chapter 2, we present the objects of our field of study, namely Hawkes processes, introduce terminology, and give an overview of the relevant literature. In Chapter 3, we point out our main contributions to the field as well as directions for further research. The final chapter contains the accompanying papers.

2 Hawkes processes

> "The object of this study is to produce a class of theoretical models which may be applicable to a variety of problems."
>
> Hawkes (1971a)

First, we show why Hawkes processes are very natural mathematical objects when it comes to modeling event streams. Later, we give a more formal introduction, emphasizing the difference between the autoregressive and the branching perspective. Furthermore, we point out theoretical landmarks, present an overview of applications based on Hawkes processes, and give a short review of estimation methods.

2.1 Motivation

Hawkes processes are point processes. Point processes model the stochastic distribution of points over some set. We consider point processes on the real line. In this case, the real line is typically interpreted as 'time' and the points as 'events'. Thus, a Hawkes process is a model for event streams—like apples falling from a tree. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(M_p, \mathcal{M}_p)$ be the space of locally-finite counting measures on $\mathbb{R}$ endowed with the σ-algebra $\mathcal{M}_p$ generated by the sets $\{m \in M_p : m(A) = n\}$, $A \in \mathcal{B}_b(\mathbb{R})$, $n \in \mathbb{N}_0$. Then any random variable

$$N : (\Omega, \mathcal{F}) \to (M_p, \mathcal{M}_p) \tag{2.1}$$

is a *point process* (on $\mathbb{R}$). In particular, for $-\infty < a < b < \infty$, $N((a,b]) : (\Omega, \mathcal{F}) \to \mathbb{N}_0$ is a random variable. It counts the number of apples fallen in the time interval $(a,b]$. The law of N on $(M_p, \mathcal{M}_p)$ is in fact determined by the joint distribution of finite collections of random variables of this form. We often assume that the stochastic rules of the considered point process

N do not change over time, that is

$$\begin{aligned}\mathbb{P}\big[N(A_1) = k_1, N(A_2) = k_2, \ldots, N(A_n) = k_n\big] \\ = \mathbb{P}\big[N(A_1 + s) = k_1, N(A_2 + s) = k_2, \ldots, N(A_n + s) = k_n\big], \quad s \in \mathbb{R},\end{aligned} \tag{2.2}$$

holds for all $k_l \in \mathbb{N}_0$, $A_l \in \mathcal{B}_b(\mathbb{R})$, $l = 1, 2, \ldots, n \in \mathbb{N}$. In this case, we call N a *stationary* point process. Often, it is mathematically convenient (and reasonable from a modeling perspective) to assume that $\mathbb{P}\left[N(\{t\}) \in \{0, 1\},\ t \in \mathbb{R}\right] = 1$, i.e., we never observe more than one event in any instant of time. In this case, we say that N is a *simple* point process. For any point process N, we define its *intensity* λ_N by

$$\lambda_N(t) := \frac{\mathbb{E}\, N(\mathrm{d}t)}{\mathrm{d}t} := \lim_{\delta \downarrow 0} \frac{\mathbb{E}\, N((t, t+\delta])}{\delta} \ \big(\in [0, \infty] \big), \quad t \in \mathbb{R}, \tag{2.3}$$

whenever the (possibly infinite) limit exists . For stationary N, the existence of λ_N is given by Khintchin's Theorem and does not depend on t; see Daley and Vere-Jones (2009), Proposition 3.3.I. If N is in addition simple, then $\lambda_N \in [0, \infty]$ gives us the expected number of apples falling from the tree per time unit. If N is a simple point process and $N(A_i) \sim \mathrm{Pois}(\int_{A_i} \lambda(t)\mathrm{d}t)$, $i = 1, 2, \ldots, n \in \mathbb{N}$, independent for all mutually disjoint $A_i \in \mathcal{B}_b(\mathbb{R})$, where $\lambda_N : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, locally integrable, then N is a *Poisson random measure* or a *Poisson point process* with intensity λ_N. We call a stationary Poisson process a *homogeneous Poisson process*. This yields the 'most homogeneous distribution of points': given that a homogeneous Poisson point process produces n points in $A \in \mathcal{B}_b$, these n points are distributed independently and uniformly over A. However, many real-world events occur clumped, that is, inhomogeneously. Again, think of the apple tree: there might be times when more apples are falling; there might be windy periods, more or less gusty; there might even be hail emptying the whole tree! This clustering of events asks for generalizations of homogeneous Poisson processes. Roughly speaking, there are three possibilities:

1. The intensity depends on time; this leads to *inhomogeneous Poisson processes*. That is, the limits in (2.3) are not constant but yield a true function of time. With such a time-inhomogeneous model, we may catch seasonal effects (like the wind being stronger in the morning and in the evening). This approach however does not take into account that the wind strength is so variable that it is better modeled stochastically than deterministically. This yields the second possibility:

2. Conditional on some covariate process, the intensity is a deterministic function of this covariate; this leads to *doubly-stochastic point processes* or *Cox point processes*. Note that in this case, the (unconditional) intensity might again be constant—for example, if the covariate process is stationary. However, the counts in disjoint intervals will be dependent

in general. For the apple example, this might imply that we model the wind by some convenient stochastic process. Conditional on the state of this process, the *conditional intensity* adjusts accordingly. It is obvious that this kind of two-step model is the most realistic approach for the modeling of many event streams. But it may often be the case that data for such an explanatory process are not available. Maybe, we do not even have a clue where to look for such relevant covariate processes. This yields the third possibility:

3. We condition the intensity on earlier events. Why does this help? Say, we only have data of the fallen apples, but no information on the wind. Does this mean we have no information on the wind? No! Given that we have observed a lot of fallen apples in the near past, it is plausible to assume that the wind is still quite strong and that the intensity of the apple-fall process is consequently quite high. In other words, *we use the past of the process as a proxy for the unobserved covariate process* ('wind'). And the good thing about it: we do not even have to know anything about the covariate process at all. This is the concept of autoregression—well-known for time series.

Hawkes processes embody this third possibility; they can be interpreted as autoregressive point processes:

2.2 Autoregressive perspective

For autoregression in the point process context, we need more terminology: the *intrinsic history* of a point process N is the filtration $(\mathcal{H}_t^{(N)})_{t\in\mathbb{R}}$ defined by

$$\mathcal{H}_t^{(N)} := \sigma\Big(\big\{\omega\in\Omega : N(\omega)(A) = k\big\} : A\in\mathcal{B}_b\big((-\infty,t]\big),\, k\in\mathbb{N}_0\Big), \quad t\in\mathbb{R}\,. \tag{2.4}$$

Note that, by definition of a point process, the generating sets of the intrinsic history are elements of the basic σ-algebra $\mathcal{F}$ so that $\mathcal{H}_t^{(N)}\subset\mathcal{F}$, $t\in\mathbb{R}$. A *(monotype) Hawkes process* N is a simple and stationary point process with

$$\Lambda_N(t) := \lim_{\delta\downarrow 0}\frac{\mathbb{E}\left[N\big((t,t+\delta]\big)\middle|\mathcal{H}_t^{(N)}\right]}{\delta} = \eta + m\int_{(-\infty,t)} w(t-s)N(\mathrm{d}s), \quad t\in\mathbb{R}, \tag{2.5}$$

where $\eta>0$, $m\geqslant 0$, and $w:\mathbb{R}\to\mathbb{R}_{\geqslant 0}$, measurable, with $\int w(t)\mathrm{d}t = 1$ and $w(t)=0$, $t\leqslant 0$. Obviously, the *conditional intensity* in (2.5) is bounded from below by η; this is why η is often called *baseline intensity*. The function w is often called *decay kernel*, and the function $h:\mathbb{R}\to\mathbb{R}_{\geqslant 0}$, $t\mapsto mw(t)$, is often called *excitement function*. Assuming existence, taking expectations on both sides of (2.5), and applying stationarity, we immediately see that $\lambda_N = \mathbb{E}\,\Lambda_N(t) = \eta/(1-m)$. So any possible solution of (2.5) can only have finite intensity if $m\in[0,1)$ (for

$\eta > 0$). Before we discuss the solution to (2.5), we consider the 'autoregressive interpretation' of the Hawkes model in the sense of item 3 in Section 2.1:

Suppose we have $\hat{\eta} > 0$, $\hat{m} \in [0, 1)$, and a decay kernel $\hat{w}$ such that the model equation (2.5) fits some event-stream data N well. The first summand in (2.5), $\hat{\eta}$, stands for the unpredictable part of the data—it is independent of $\mathcal{H}_t^{(N)}$ for all $t \in \mathbb{R}$. The second summand, $\hat{m} \int_{(-\infty,t)} \hat{w}(t - s)N(\mathrm{d}s)$, stands for the predictable part. Thus, the coefficient $\hat{m}$ can be seen as a measure of how informative the past is for the prediction of the actual conditional intensity. If $\hat{m} = 0$, the model reduces to a 'completely random' homogeneous Poisson process (as, in this case, $\mathbb{E}[N((t, t + s])|\mathcal{H}_t] = s\eta = \mathbb{E}\,N((t, t + s])$, $s > 0$). When $\hat{m}$ increases, the weight of the autoregressive part increases. The shape of $\hat{w}$ tells us, which lags of past events have large explanatory power for the actual intensity. For example, if $\hat{w}$ is concentrated at 0, then only the most recent past matters for the present ('the wind changes quickly'), and if w is very heavy-tailed, then even events from very long ago carry relevant information for the state of the process ('the wind changes slowly'). The coefficient $\hat{m}$ being fixed, the baseline intensity is typically determined by $\hat{\eta} := (1 - \hat{m})\hat{\lambda}$ where $\hat{\lambda} > 0$ denotes the (unconditional) average intensity of the data. With this choice, the empirical and the theoretical unconditional intensity are equal. In this sense, the baseline intensity $\hat{\eta}$ compensates the part of the intensity that cannot be explained (in a linear way) from the past.

As we have pointed out in Section 2.1, in most cases (apples, crimes, credit defaults, earthquakes, fatalities, neurological activity, price jumps, traffic counts, ...) the 'true intensity' of events will depend on various complicated (often not observable or known) covariate processes. Often, the best guess we have about these covariates is the past of the observed event stream itself. So—as in time series autoregression—we use the past of the process as a proxy variable for these unobserved quantities. This makes the Hawkes process a flexible modeling tool. As a side remark, note that the analogy with autoregressive time series is not complete: in contrast to autoregressive times series, (linear) Hawkes processes only allow for positive excitement and no inhibition. In particular, we cannot obtain negative autocorrelation. In this sense, (linear, unmarked) Hawkes processes are a model class less rich than (linear) autoregressive time series.

2.3 Branching perspective

It is crucial to understand that the definition of the Hawkes process by equation (2.5) is *implicit*. The conditional intensity Λ_N on the left-hand side of the equation as well as the right-hand side of the equation depend on N. It is not a priori clear that (2.5) has a solution and that this solution is unique. So, despite the simple interpretation, the definition of the Hawkes process as a solution of a family of equations is quite involved. The proof of the existence of a solution

to (2.5) is constructive:

We start with a homogeneous Poisson process $\{T_k^{(0)}\}$ on $\mathbb{R}$ with intensity $\eta > 0$. These points form the *points of generation 0* or *immigration points*. From each generation-0 point $T_k^{(0)}$, we start an inhomogeneous Poisson process with intensity $mw(\cdot - T_k^{(0)})$. In other words, $T_k^{(0)}$ has Pois(m) *children* and each of these children is independently displaced from its *parent* $T_k^{(0)}$ with density w. All children points of all the generation-0 points form the process of *generation-1 points* $\{T_k^{(1)}\}$. For each of these generation-1 points we again find children that in turn form the *points of generation 2*, etc. All these *offspring and displacement operations* are performed independently. Consider the superposition of the points until generation $g \in \mathbb{N}_0$, that is, consider the point processes $(N^{(g)})_{g\in\mathbb{N}_0}$ defined by $N^{(g)}(A) := \sum_{g'=0}^{g} \#\{T_k^{(g')} \in A : k \in \mathbb{Z}\}$, $A \in \mathcal{B}_b(\mathbb{R})$. Obviously, $(N^{(g)}(A))_{g\in\mathbb{N}_0}$ is increasing for all $A \in \mathcal{B}_b(\mathbb{R})$, so we may consider the limit $N^{(\infty)}(A)$ ($\in \mathbb{N} \cup \{\infty\}$). One can show by induction that $\lambda_{N^{(g)}} = \eta(1 + m + m^2 + \cdots + m^g)$, $g \in \mathbb{N}$. Consequently, by monotone convergence, we find that

$$\frac{\mathbb{E}\, N^{(\infty)}(\mathrm{d}t)}{\mathrm{d}t} = \frac{\eta}{1-m}, \quad m < 1.$$

In other words, if $m < 1$, then $N^{(g)}(A)$ increases to an almost surely finite limit $N^{(\infty)}(A)$ for all $A \in \mathcal{B}_b(\mathbb{R})$. These limits define a point process $N^{(\infty)}$ with finite unconditional intensity $\eta/(1-m)$ if $m < 1$. In addition, one can show that the limit $N^{(\infty)}$ has conditional intensity (2.5).

This is the standard construction of a Hawkes process; see Hawkes and Oakes (1974) or Daley and Vere-Jones (2009), Example 6.3(c), for more formal presentations as well as for uniqueness arguments. In Paper D, Section 2.1, pp. 162, we give a similar construction and explain the connection to the prototypic Galton–Watson branching processes. Also note that the Hawkes process representation in Paper B, Proposition 19, p. 94, is the result of a most complete branching construction—keeping track of all genealogical lines. The construction explained above reveals the branching nature of the Hawkes process. Here, the parameters η, m, and w have different—maybe more concrete—meanings than in the autoregressive perspective. The new interpretation comes with new terminology: in the branching perspective, η is the *immigration intensity* that gives the expected number of immigrants per time unit, m is the *branching coefficient* that gives the expected number of offspring of each point, and w is the *displacement intensity* that gives the distribution of the distance between parent and child event. In the same spirit, we call $h : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, $t \mapsto mw(t)$, *reproduction intensity*.

We have seen that the Hawkes process has an autoregressive and a branching interpretation. It may often make sense to fit an autoregressive model to data. But the literal parent/offspring interpretation of the branching construction has to be handled with care. This needs similar diligence as causal interpretations of standard regression. Fitting Hawkes models to artificial

data from Cox processes (with no branching at all) can be very instructive in this respect; see Paper C, Section 4.3, page 137.

2.4 Multitype Hawkes processes

We also give the corresponding conditional-intensity equations for the multitype case. Now, each event is supplied with a type in $[d] := \{1, 2, \dots, d\}$, where $d \in \mathbb{N}$ denotes the number of types. A *d-type Hawkes process* is a simple d-type point process $\mathbf{N}$ on $\mathbb{R} \times [d]$ that solves the family of equations

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[\mathbf{N}\big((t, t+\delta] \times \{j\}\big) \middle| \mathcal{H}_t^{(\mathbf{N})}\right]}{\delta} = \eta_j + \sum_{i=1}^{d} \int_{(-\infty, t)} m_{i,j} w_{i,j}(t-s) \mathbf{N}(\mathrm{s} \times \{i\}), \quad j \in [d], t \in \mathbb{R}. \tag{2.6}$$

Here, for $(i, j) \in [d]^2$,

$$w_{i,j} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}, \quad \text{such that } \int w_{i,j}(t)\mathrm{d}t = 1 \text{ and } w_{i,j}(t) = 0, t < 0,$$

and $\eta_j, m_{i,j} \geqslant 0$. For the necessary generalizations of the underlying σ-algebra $\mathcal{F}$ and the history $(\mathcal{H}_t^{(\mathbf{N})}) \subset \mathcal{F}$, see Paper B, Definition 13, p. 91, and Section 3.1, pp. 93. Again, we may emphasize either the autoregressive perspective or the branching perspective. The terminology may depend on this point of view. From the autoregressive perspective, the term *multivariate* or *d-variate Hawkes process* is reasonable because we can interpret $\mathbf{N}$ as in (2.6) as d-tuple of univariate point processes $(N_1, \dots, N_d)$, where the components count the number of points of a specific type, that is, $N_i(\cdot) := \mathbf{N}(\cdot \times \{i\})$, $i \in [d]$. From a branching point of view, the term *d-type* or *multitype Hawkes process* is obviously more justified. If in doubt, we prefer the branching terminology because 'multivariate' may be misleading in that a 'multivariate Hawkes process' is not a 'multivariate point process' (on $\mathbb{R}^d$). As in the monotype/univariate case, we call $(w_{i,j})$ *displacement densities* or *decay kernels*, and $\eta = (\eta_1, \eta_2, \dots, \eta_d) \in \mathbb{R}^d_{\geqslant 0}$ (constant) *immigration* or *baseline intensities*—again depending on the perspective. In any case, for simulation and also for many calculations, the branching view is more convenient: independently for each type $i \in [d]$, we start with an homogeneous Poisson immigration processes with intensity η_i. Each type-i event (independently) triggers Pois($m_{i,j}$) type-j events. The random distance between these children and their parent has density $w_{i,j}$. These children have again children and so on. All these branching and displacement operations are applied independently. Similar to the monotype variate case, the condition for the superposition of all generations yielding

a multitype point process with finite unconditional intensity is that the spectral radius of the *branching matrix* $M := (m_{i,j})_{(i,j)\in[d]^2}$ is strictly less than one. For an insightful proof, see Paper D, Proposition 4, p. 164.

2.5 Marked Hawkes processes

One also considers *marked Hawkes processes*, where each event of a d-type Hawkes process is supplied with a mark (such as the strength of an earthquake, the height of a price jump, or the volume of a buy or sell order). To keep the model mathematically tractable, one often assumes that all marks are independent realizations from *mark distributions* F_i, $i \in [d]$, that depend on the type i of the corresponding event. The marks are included in the model (2.6) by multiplying the contribution $w_{i,j}(t - T_k^{(i)})$ of some past type-i event $T_k^{(i)}$ on the type-j intensity with a factor $g_{i,j}(Z_k^{(i)})$, where the functions $g_{i,j} : \text{range}(Z_k^{(i)}) \to \mathbb{R}_{\geqslant 0}$ are called *impact functions* or *boost functions*. The impact functions are normalized by the condition $\mathbb{E}\, g_{i,j}(Z_k^{(i)}) = 1$, $(i,j) \in [d]^2$. Without this condition the model would not be identifiable. Note that this normalizing condition includes an integrability condition for the mark distributions. Also note that, in this case with independent marks, the corresponding branching perspective involves a mixed Poisson distribution: indeed, each type-i point $T_k^{(i)}$ has a total number of $Y_k^{(i,j)}$ children, where $Y_k^{(i,j)}|Z_k^{(i)} \sim \text{Pois}(m_{i,j} g_{i,j}(Z_k^{(i)}))$ and $Z_k^{(i)} \sim F_i$ independently over $(i,k) \in [d] \times \mathbb{Z}$—whereas the displacement distributions stay unaffected by the marks.

2.6 Literature

Theory

Despite the huge and increasing number of applications of Hawkes processes, only few publications give truly new insight. This makes it relatively easy to list the main theoretical landmarks of Hawkes processes. We start with a short historical embedding.

The Hawkes process has been introduced in Hawkes (1971b,a). The accent of these publications lies on second order properties—in the spirit of the work in Bartlett (1963) on spectral analysis of stationary point processes. Note that already Section 2 of Bartlett (1963) considers a (monotype) point process that is a preform of the Hawkes process. For according choices of offspring and displacement distributions, one retrieves the process $N^{(1)}$ of the branching construction from Section 2.3 in our introduction. In a broader context, these works of Bartlett and Hawkes can be seen as part of the efforts of the (British) statistical community at that time to deal with alternatives to purely Poisson—that is, completely random—point processes. The goal was twosome: on the one hand, one aimed for statistical tests of departure from complete

randomness. On the other hand, one searched for corresponding alternative models. In this context, Cox (1955) may be the most important and influential single publication. It proposes various statistical tests and introduces doubly stochastic point processes; see Section 2.1. Bartlett's and Hawkes's works pursued a completely different and autonomous approach than Cox in that they did not account so much for exogenous influence but applied broadly autoregressive or branching ideas. A particularly interesting early line of problems in the event stream context can be found in the context of medical data analysis: Greenwood (1946) discussed if and how one could statistically decide, whether event-stream data stem from an infectious model or not. We believe that the problems and thoughts addressed by the latter work were motivation and inspiration for the work of Bartlett and Hawkes.

Hawkes and Oakes (1974) gives a cluster-process (or immigration–offspring) representation of the Hawkes process. In particular, this work shows how convenient this representation is for calculations. For example, an implicit equation for the probability generating functional of a Hawkes process is derived. In Adamopoulos (1975), these results are applied to derive some counting and interval properties. In the exponential-decay case, Oakes (1975) exploits the immigration–offspring representation further to derive some explicit formulas for distributional properties of the Hawkes process. Ozaki (1979) discusses the likelihood of Hawkes processes and its optimization. Also simulation is considered. The work contributes the important observation (proposed by H. Akaike) that in the case of exponential decay, the points of a Hawkes process can be constructed in a recursive manner. This fact can also be used to calculate conditional intensities, respectively, likelihoods in a recursive manner which is computationally beneficent. A full formalization of the algorithm is given in Section 1.4 of Liniger (2009). In Brémaud and Massoulié (1996), non-linear generalizations of multitype Hawkes processes are considered. More specifically, in the monotype case, the authors discuss point processes N on $\mathbb{R}$ with conditional intensities of the form

$$\Lambda_N(t) := \Phi\left(\int_{(-\infty,t)} h(t-s)N(\mathrm{d}s)\right), \quad t \in \mathbb{R}, \tag{2.7}$$

with $\Phi : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, $h : \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ measurable. Note that with the choice of $\Phi(x) = \eta + x$ together with $h \geqslant 0$ and $\eta > 0$, (2.7) corresponds to the standard linear Hawkes process as in (2.5) (with $m := \int h\mathrm{d}t$ and $w := h/m$). In the nonlinear case, a similar Poisson-embedding construction as described in Paper G, Section 2.2, pp. 244, is proposed. If Φ is α-Lipschitz, and $\int |h(t)|\mathrm{d}t =: m \leqslant \alpha^{-1}$, then the intensity of the approximating processes can be bounded by the intensity of a linear Hawkes process with branching coefficient $\alpha m\,(< 1)$. This upper bound allows to establish convergence results. One interesting aspect of these processes is that they allow for negative autocorrelation and, in particular, inhibition (as h may attain negative values)—in contrast to linear Hawkes processes. Note however that, in general, the branching

view becomes meaningless in this nonlinear setup. In Brémaud and Massoulié (2001), Hawkes branching processes without ancestors are discussed. The authors consider a family of monotype Hawkes processes $(N_\delta)_{\delta\in(0,1)}$, where N_δ has immigration intensity $\eta_\delta := \lambda\delta$, $\lambda > 0$, and branching coefficient $m_\delta := 1 - \delta$. They give conditions on the displacement density w under which N_δ converges to a non-trivial weak limit with intensity λ as $\delta \downarrow 0$. We will touch on this most interesting 'critical' case in Section 3.6, respectively, Paper F. In Brémaud et al. (2002), the rate of convergence to a stationary (possibly void) point process of a marked Hawkes process with respect to various starting conditions is discussed. The results are helpful in two ways: they allow to define stopping rules for burn-in times of simulations of Hawkes processes. Then, for the extinction case, one can apply the rates to the modeling of small epidemics. The Ph.D. thesis Liniger (2009) is the most complete and most self-contained theoretical introduction for (linear) marked multivariate Hawkes processes. It gives possible constructions, existence, and uniqueness a rigorous mathematical foundation. In particular, the cluster structure is formalized on a quite general level. The specialized terminology is used to derive various second moment measures of the Hawkes process. In Errais et al. (2010), a Markovian analysis of the Hawkes process with exponential decay is discussed. This paper considers Hawkes processes starting in 0. The authors note that for the special case of exponential decay kernels, the joint process of Hawkes process and conditional intensity forms a bivariate Markov process. They exploit the process from a Markovian point of view, derive the infinitesimal generator, give Dynkin formulas, calculate closed form expressions for moments of the conditional intensity etc. In Jaisson and Rosenbaum (2015), scaling limits of nearly critical Hawkes processes (with branching coefficient $m < 1$ near 1) are discussed. It is shown that Hawkes processes, accordingly scaled, converge to integrated Cox-Ingersoll-Ross processes. One can argue from this result that Hawkes-process based market-microstructure models are consistent with standard coarse-scale models (as the Heston model). Bacry and Muzy (2015) can be seen as a corollary and completion of Hawkes's first two papers. The authors note that Hawkes processes—like autoregressive time series—are in fact characterized by their second order properties.

Besides this list of publications, note that the monograph Daley and Vere-Jones (2009) contains dozens of examples using the Hawkes process. Although already Brillinger (1975) and the title of Ozaki (1979) speak of a *Hawkes' self-exciting point process*, it seems that the first edition of the mentioned monograph, Daley and Vere-Jones (1988), finally determined the naming of the (linear) autoregressive point process after its discoverer, Alan Hawkes.

Applications

In the seminal paper Hawkes (1971a), Alan Hawkes introduced 'mutually-exciting point processes' as a model class for epidemic-type data such as occurrences of shingles and chicken pox. Furthermore, he proposed neuron firing, radiation, and 'the computer' as possible appli-

cations for his model. However, the first statistical application of the Hawkes process occurred in earthquake modeling in Hawkes and Adamopoulos (1973). Chapter 7 of Lomnitz (1974) discusses the so-called 'Klondike process' as a possible earthquake model. This cluster process is quite similar to a Hawkes process. Vere-Jones (1975) is a philosophical–historical essay on earthquake modeling; it mentions the 'self-exciting process' in Section 5. In Ogata (1988), various models for earthquake data are compared—a specific marked Hawkes process yielding the best fit: the *epidemic-type aftershock sequence* (ETAS). It has been one of the most popular statistical earthquake models ever since. For an overview of the developments inside the ETAS model; see Ogata (1999).

The original idea of 'neuron firing' as an application of the Hawkes process has not been followed until an example in the mathematical paper Brémaud and Massoulié (1996). Today, Hawkes processes are a valid modeling option for these kinds of data; see, e.g., Reynaud-Bouret et al. (2014) and the references therein. Next to earthquake and neural spike train data, there are many more or less exotic fields of Hawkes-process applications; e.g., Reynaud-Bouret and Schbath (2010) proposes to model 'genomic events along DNA sequences' with Hawkes processes—one of the few examples, where the points are not interpreted as events in time but more as positions in space. In any case, there are surprisingly few applications to event streams that are obviously 'infectious'—like YouTube clicks in Crane and Sornette (2008) or virus spreading in Kim (2011). Most Hawkes process applications rather exploit the autoregressive structure like body counts in Lewis et al. (2012): the fact that many soldiers got killed in the near past does not kill other soldiers but it is a sign for (not observed) explanatory covariates being in a state that makes it more likely to get killed. It tells us that 'time is good for getting killed'. The same is true for applications in crime prediction as in Mohler et al. (2011).

A new line of Hawkes-process applications started with the papers Bowsher (2002) and Chavez-Demoulin et al. (2005), where Hawkes-process based models were used in the financial context for the first time. The textbook McNeil et al. (2005) contains an early Hawkes-process application on credit defaults. Countless applications of the Hawkes process in finance followed; see Bacry et al. (2015b) for a recent review of Hawkes processes in finance. Errais et al. (2010) and Aït-Sahalia et al. (2015) are examples for including Hawkes processes in the semi-martingale framework of stochastic finance: the applied model is a jump diffusion, where the jumps are modeled by a Hawkes process. In Embrechts et al. (2011), extremes of daily log returns are modeled by marked Hawkes processes. In the last years, E. Bacry, J.–F. Muzy, and coauthors published a most influential series of papers applying Hawkes processes to financial high-frequency data: see Bacry et al. (2012, 2013, 2014, 2015a); Bacry and Muzy (2015); Muzy and Bacry (2014). They show that their Hawkes models reproduce many stylized facts of this kind of data, e.g., the 'Epps effect', the 'lead lag' effect, and concave 'price impact' curves; see Bacry et al. (2013) and Muzy and Bacry (2014). Furthermore, they examine scaling limits and

thus show that their models are consistent with standard coarser-scale models.

Estimation

Maximum likelihood estimation

Maximum likelihood estimation (MLE) for Hawkes processes is discussed for the first time in Ogata (1978), Example 4, and Ozaki (1979). MLE for Hawkes processes is not straightforward; see Paper E, Section 2.2, p. 196: often, there will be edge effects at the beginning of the process and the likelihood will be incomplete; see the discussion in Chapter 1.4 of Liniger (2009). Furthermore, evaluation and optimization of Hawkes likelihoods is computationally demanding: it involves evaluation of conditional intensities at all observed time points. For each of the observed points, one has to evaluate the decay kernel(s) with respect to some potential parameters for all past points. That is, the complexity of the likelihood calculation is of order n^2, where n denotes the number of observed points. In the very special case when the decay kernel coincides with an exponential density of the form $w_\alpha : \mathbb{R} \to \mathbb{R}_{\geqslant 0},\ t \mapsto 1_{\{t>0\}}\alpha \exp(-\alpha t),\ \alpha > 0$, one can decrease this complexity to linear order. Indeed, denote by $T_n \in \mathbb{R}_{\geqslant 0},\ n \in \mathbb{N}$, the ordered event times of a (non-stationary) Hawkes process starting from zero. Clearly, $\Lambda(T_1)$ equals the baseline intensity η. Furthermore, a simple calculation shows that

$$\Lambda(T_n) = \Big(1 - \exp\big\{-\alpha(T_n - T_{n-1})\big\}\Big)\eta + \exp\big\{-\alpha(T_n - T_{n-1})\big\}\big(\Lambda(T_{n-1}) + m\alpha\big), \quad n > 1.$$

A similar recursion also holds for the multivariate case. This reduces the complexity of the likelihood to linear dependence on the number of observations and allows application of maximum likelihood estimation (MLE) also to larger event-stream data sets. This method is applied, e.g., in Lemonnier and Vayatis (2014) and Alfonsi and Blanc (2015). An obvious downside of this approach is the somewhat ad hoc choice of exponential decay. One can partly overcome this by considering convex combinations of exponential densities or Laguerre polynomials for the decay kernels; see, e.g., Akaike and Ogata (1982). Here, the values of the conditional intensities can also be calculated recursively. This approach, however, typically increases the number of parameters which comes with new questions. Another way of overcoming computational difficulties are expectation–maximization (E–M) algorithms. The application of E–M algorithms on MLE-estimation of point processes with hidden information (e.g., hidden Markov models) is discussed in Daley and Vere-Jones (2009), Section 10.3. Implementations of this concept for Hawkes processes can be found, e.g., in Lewis and Mohler (2011) or Wheatley (2016). Here, the genealogical structure of the process is treated as hidden data. This approach comes with the usual convergence issues of E–M algorithms. Also note that computational problems are still relevant. Furthermore, implementation is somewhat demanding. Obviously, for multitype

Hawkes processes all these issues become even more pronounced. Finally, note that there are no results for the distribution of MLE estimates. These drawbacks explain the increasing interest in alternative approaches to Hawkes estimation:

Alternative estimation methods

In Reynaud-Bouret and Schbath (2010) and Reynaud-Bouret et al. (2014), multitype Hawkes process parameters are fitted by minimizing a least-squares objective of the realized intensity subject to an L_1 penalty. The method is comparable to the LASSO method from linear regression and allows detection of zero-excitements. In Bacry et al. (2012, 2014), a moment estimator with respect to the bin-counts of Hawkes-process data is considered. Neither work provides theoretical results regarding the distribution of the proposed estimators. We discuss these methods in more detail in Paper C, Section 3.4, pp. 127. Finally note that during the final preparation of this thesis, we were informed of the publication Eichler et al. (2016). Here, the same approach as ours in Paper C is used: conditional-least-squares estimation of bin-count sequences. However, in contrast to our work, the mentioned publication lacks the connection with the corresponding INAR time series, and gives neither asymptotic normality results nor covariance-matrix estimates.

3 Main contributions

> "You know, Matthias, sometimes it is better to write a paper all over again rather than to start correcting."
>
> Paul Embrechts in summer 2013

In this chapter, we present the main contributions of our thesis to the field of Hawkes processes. Overall, the value of our work consists in pointing out and formalizing connections between Hawkes processes and other mathematical areas. The capital letters in the section titles refer to the accompanying papers. In the final section, we explain how our findings might be relevant for future research.

3.1 Time series perspective (A, B)

In Paper A and Paper B, we show that integer-valued autotegressive (INAR) times series are discrete-time versions of Hawkes point processes and, vice versa, Hawkes point processes are continuous-time versions of INAR time series. To the best of our knowledge, this connection has not been observed before. The INAR–Hawkes dictionary permits translation of results from one area to the other. We present the heuristics behind the approximation in the case of a univariate Hawkes process N with baseline intensity $\eta > 0$ and excitement function h with $\int h\mathrm{d}t < 1$: for some $\Delta > 0$, define the *bin-count sequence* $\tilde{X}_n^{(\Delta)} := N\big(((n-1)\Delta, n\Delta]\big)$, $n \in \mathbb{Z}$. For small $\Delta > 0$ and $p \in \mathbb{N}$, large, we obtain from (2.5) that

$$\mathbb{E}\left[\tilde{X}_n^{(\Delta)} \middle| \sigma\left(\tilde{X}_{n-1}^{(\Delta)}, \tilde{X}_{n-2}^{(\Delta)}, \dots\right)\right] \approx \Delta\eta + \sum_{k=1}^{p} \Delta h(\Delta k)\tilde{X}_{n-k}^{(\Delta)}, \quad n \in \mathbb{Z}. \tag{3.1}$$

More specifically, the bin counts approximately follow a Poisson autoregression of the form

$$\tilde{X}_n^{(\Delta)} | \tilde{X}_{n-1}^{(\Delta)}, \tilde{X}_{n-2}^{(\Delta)}, \dots, \tilde{X}_{n-p}^{(\Delta)} \overset{\text{approx.}}{\sim} \text{Pois}\left(\alpha_0 + \sum_{k=1}^{p} \alpha_k \tilde{X}_{n-k}^{(\Delta)}\right), \quad n \in \mathbb{Z}. \tag{3.2}$$

The well-known INAR(p) time series solves relation (3.2) exactly. To make the INAR–Hawkes correspondence even more complete, we introduce INAR processes of infinite autoregressive order:

Definition 1 (Paper A, Definition 2, pp. 45)**.** *For $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, let $\varepsilon_n \overset{\text{iid}}{\sim} \text{Pois}(\alpha_0)$, $n \in \mathbb{Z}$, and $\xi_l^{(n,k)} \sim \text{Pois}(\alpha_k)$, independently over $n \in \mathbb{Z}$, $k \in \mathbb{N}$, $l \in \mathbb{N}$, and also independent of (ε_n). An* integer-valued autoregressive time series of infinite order *(INAR(∞)) is a sequence of random variables $(X_n)_{n\in\mathbb{Z}}$ which is a solution to the system of stochastic difference equations*

$$X_n = \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \xi_l^{(n,k)} + \varepsilon_n, \quad n \in \mathbb{Z}. \tag{3.3}$$

We call α_0 immigration parameter*, (ε_n)* immigration sequence*, $\alpha_k \geqslant 0$, $k \in \mathbb{N}$,* reproduction coefficients*, and $K := \sum_{k=1}^{\infty} \alpha_k$* reproduction mean*.*

Note that in the first definitions of INAR(p) sequences in Al-Osh and Alzaid (1987) for $p = 1$, and Du and Li (1991) for $p \in \mathbb{N}$, the *counting sequences* $(\xi_l^{(n,k)})_l$ from Definition 1 have Bernoulli margins. This choice yields the binomial thinning operation from Steutel and van Harn (1979). We prefer the Poisson choice for the counting sequences so that (3.2) holds. This in turn leads to formulas that are simpler and that can be compared with their Hawkes counterparts more directly. Also note that the Poisson distribution is more convenient for our purpose: we want to interpret the INAR(p) model as an approximation of the bin-count sequence of a Hawkes process—and in the Hawkes model, an event can have potentially more than one direct offspring event in a future time-interval.

Theorem 2 (Paper A, Theorem 3, p. 46)**.** *Let $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, with reproduction mean $K := \sum_{k=1}^{\infty} \alpha_k < 1$. Then* (3.3) *has an almost surely unique stationary solution $(X_n)_{n\in\mathbb{Z}}$, where $X_n \in \mathbb{N}_0$, $n \in \mathbb{Z}$, and $\mathbb{E}\, X_n \equiv \alpha_0/(1-K)$, $n \in \mathbb{Z}$.*

We give autoregressive representations (Paper A, Proposition 8, p. 48) and moving-average representations (Paper A, Proposition 9, p. 49) as well as generating functions (Paper A, Proposition 6, p. 48), for the INAR(∞) model. These results are interesting in themselves. In the paper, we use these results to derive the following convergence theorem that formalizes the connection between INAR and Hawkes processes:

Theorem 3 (Paper A, Theorem 17, p. 54)**.** *Let N be a Hawkes process with immigration intensity η and piecewise-continuous reproduction intensity h. For $\Delta \in (0, \delta)$, let $\left(X_n^{(\Delta)}\right)$ be an INAR(∞) sequence with immigration parameter $\Delta\eta$ and reproduction coefficients $\Delta h(k\Delta)$, $k \in \mathbb{N}$. From the sequences $\left\{\left(X_n^{(\Delta)}\right)\right\}_{\Delta\in(0,\delta)}$, we define a family of point processes by*

$$N^{(\Delta)}(A) := \sum_{k:k\Delta\in A} X_k^{(\Delta)}, \quad A \in \mathcal{B}(\mathbb{R}),\ \Delta \in (0, \delta). \tag{3.4}$$

Then, we have that

$$N^{(\Delta)} \xrightarrow{\mathrm{w}} N \quad as \quad \Delta \downarrow 0.$$

In the lengthy and rather technical proof, we use the standard weak-convergence approach—as followed in the Hawkes context, e.g., by Brémaud and Massoulié (2001): first, tightness of the approximating family is established. By Prohorov's theorem, tightness yields weak subsequential limits for all subsequences. Then we show that all these potential weak subsequential limits have the same distribution as the Hawkes process. This establishes the result. In Paper B, we generalize the approximation result from Theorem 3 to the multivariate case. In addition, we state it as an almost-sure limit and in L_1. We also drop the technical piecewise-continuity (and in particular boundedness) assumptions on the reproduction intensities:

Theorem 4 (Paper B, Theorem 25, p. 100)**.** *Let* $\mathbf{N}$ *be a d-type Hawkes process as in* (2.6)*. Then there exist d-type point processes* $\mathbf{N}^{(\Delta)}$, $\Delta > 0$, *such that* $X^{(\Delta)}_{n,i} := \mathbf{N}^{(\Delta)}\big(\{\Delta n\} \times \{i\}\big)$, $n \in \mathbb{Z}$, $i \in [d]$, *defines a multivariate INAR sequence (see Paper B, Definition 20, p. 95) with immigration coefficients* $\alpha^{(\Delta)}_{0,i} = \Delta\eta_i$, $i \in [d]$, *and reproduction coefficients*

$$\alpha^{(\Delta)}_{i,j,k} = m_{i,j} \int_{(k-1)\Delta}^{k\Delta} w_{i,j}(t)\mathrm{d}t, \ (i,j,k) \in [d]^2 \times \mathbb{N},$$

and for all nonnegative continuous functions with compact support f *and for* $j \in [d]$*, we have that*

$$\int f(t)\mathbf{N}^{(\Delta)}(\mathrm{d}t \times \{j\}) \to \int f(t)\mathbf{N}(\mathrm{d}t \times \{j\}), \quad \Delta \downarrow 0, \tag{3.5}$$

almost surely, in L_1*, and (hence also) in distribution. In particular,*

$$\mathbf{N}^{(\Delta)} \xrightarrow{\mathrm{w}} \mathbf{N}, \quad \Delta \downarrow 0.$$

Next to these formal convergence theorems, we list structural analogies between INAR(∞) and Hawkes models in Paper A, Section 3.3, p. 55. We give possible consequences of the time series perspective on Hawkes processes in Section 3.7 of this introduction. At this point, the reader may already note that in view of the presented INAR–Hawkes analogy, analysts using the Hawkes model may consider to directly apply the INAR model in the first place—as most event data live on relatively discrete time grids. Also note that Theorem 4 is proven by completely different means than Theorem 3: for the proof of Theorem 4, we define the approximating point process sequences and the limit process on *exactly the same* underlying genealogical tree structure. This leaves only the positions of the points to be approximated—which is straightforward. This elegant proof is one of the reasons why in Paper B, we represent Hawkes processes in the formalism of branching random walks:

3.2 Branching-random-walk perspective (B, G)

In Paper B, we formalize the well-known branching construction of a Hawkes process described in Section 2.3 in branching-random walk (BRW) terminology. It turns out that this is the most adequate representation of a Hawkes process if we want to exploit its branching structure. The obvious connection between Hawkes processes and branching random walks has not been made in the relevant literature as yet. A possible reason for this ignorance may be that—as we will see—the space component from the BRW world becomes the time component in the Hawkes world, and the time component from the BRW world vanishes in the Hawkes world.

Hawkes trees and forests

We first sketch the terminology. For a more complete introduction to random trees, see Paper B, Section 2, pp. 85. Let $\mathcal{U}$ denote the set of all possible nodes; see Paper B, Definition 1, p. 85, for a possible construction. A rooted tree $\mathbf{t}$ is a set of nodes $\{u\} \subset \mathcal{U}$ such that $\varnothing \in \mathbf{t}$ and for all $u \in \mathbf{t}\backslash\{\varnothing\}$ there exists a unique parent node $u^- \in \mathbf{t}$. Next, each node u in a tree $\mathbf{t}$ is supplied with *two* labels, namely with one of $d \in \mathbb{N}$ types $\mathbf{l} : \mathbf{t} \to [d] := \{1, 2, \dots, d\}$ and, in addition, with a position $\mathbf{s} : \mathbf{t} \to \mathbb{R}$. We denote the space of $([d] \times \mathbb{R})$-labeled rooted trees by $\mathbb{T}^{([d]\times\mathbb{R})}$. We consider a special distribution on $\mathbb{T}^{([d]\times\mathbb{R})}$, where the positions $\mathbf{s}$ along the genealogical lines of the tree behave like a regime-switching random walk with the regimes depending on the actual node type: let $\mu := (\mu_i)_{i\in[d]}$ be a d-tuple of *offspring distributions* on $\mathbb{N}^d$ and $i_0 \in [d]$. In an (i_0, μ)-*random tree* $(\mathbf{T}, \mathbf{L})$, the root node is of type i_0 and every type-i node has k_1 type-1 children, k_2 type-2 children, ..., and k_d type-d children with probability $\mu_i(k_1, \dots, k_d)$. In Paper B, we always pick μ in such a way that the random tree is subcritical (and in particular almost surely finite); see Paper B, Definition 5, p. 87. We denote the space of finite $([d] \times \mathbb{R})$-labeled trees by $\mathbb{T}_f^{([d]\times\mathbb{R})}$. The positions are attached to the tree as follows:

Definition 5 (Paper B, Definition 11, p. 90)**.** *Let* $F = (F_{i,j})_{(i,j)\in[d]^2}$ *be a matrix of* displacement distributions *on* $\mathbb{R}$ *and let*

$$Y_u^{(i,j)} \sim F_{i,j}, \quad u \in \mathcal{U},\ (i,j) \in [d]^2,$$

be mutually independent random variables. Furthermore, let $(\mathbf{T}, \mathbf{L})$ *be a subcritical* (i_0, μ)*-tree (independent of the* $Y_u^{(i,j)}$*). Given* $(\mathbf{T}, \mathbf{L})$*, we define the* position $\mathbf{S}$ as

$$\mathbf{S} : \mathbf{T} \to \mathbb{R}, \quad u \mapsto \sum_{\varnothing < v \leqslant u} Y_v^{(\mathbf{L}(v^-),\mathbf{L}(v))}. \tag{3.6}$$

Any $\mathbb{T}_f^{([d]\times\mathbb{R})}$*-valued random variable with the same distribution as* $(\mathbf{T}, \mathbf{L}, \mathbf{S})$ *is an* (i_0, μ, F)-

subcritical d-type branching random walk*; in short,* (i_0, μ, F)-BRW. *We call* $(\mathbf{T}, \mathbf{L})$ *the* underlying tree *of the BRW* $(\mathbf{T}, \mathbf{L}, \mathbf{S})$. *If the displacement distributions are all chosen in such a way that* $F_{i,j}(0) = 0$, $(i, j) \in [d]^2$, *then we call* $(\mathbf{T}, \mathbf{L}, \mathbf{S})$ *an* (i_0, μ, F)-Hawkes tree.

We point out the connection with multitype Hawkes processes as in (2.6): consider immigrants $\{T_k\} \sim \mathrm{PRM}(\eta)$, $\eta := \eta_1 + \eta_2 + \ldots \eta_d$, independently supply them with a type $L_n \in [d]$, $\mathbb{P}[L_n = i] = \eta_i / (\sum_{j \in [d]} \eta_j)$, $i \in [d]$, and grow independent Hawkes trees $(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k)$, $k \in \mathbb{Z}$, from these immigrants. This yields a *Hawkes forest* $\mathbf{F}$:

Definition 6 (Paper B, Definition 14, p. 92)**.** *Let* $\{(T_k, L_k)\}_{k \in \mathcal{K} \subset \mathbb{Z}}$ *be a* d*-type point process with* immigration law Q*. For subcritical offspring distributions* $\mu = (\mu_i)_{i \in [d]}$ *and displacement distributions* $F = (F_{i,j})_{(i,j) \in [d]^2}$*, we define a* (Q, μ, F)-forest $\mathbf{F}$ *as the random set*

$$\mathbf{F} := \Big\{ \big(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k, (T_k, L_k) \big) : \ k \in \mathcal{K} \Big\}, \tag{3.7}$$

in such a way that, conditional on the immigration types $(L_k)_{k \in \mathcal{K}}$*, the* $(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k)$ *are independent* (L_k, μ, F)*-branching random walks as in Definition 5. If* $F_{i,j}(0) = 0$, $(i, j) \in [d]^2$*, we call the forest a* (Q, μ, F)-Hawkes forest*. Furthermore, given a forest* $\mathbf{F}$ *as in* (3.7)*, we consider its* projected d-type point measures

$$\mathbf{N}_{\mathbf{F}}\big(A \times \{j\}\big) := \sum_{k \in \mathbb{Z}} \#\big\{ u \in \mathbf{T}_k : \ \mathbf{L}_k(u) = j, \ T_k + \mathbf{S}_k(u) \in A \big\}, \quad A \in \mathcal{B}_b(\mathbb{R}), \ j \in [d]. \tag{3.8}$$

Note that these projected Hawkes forests $\mathbf{N}_{\mathbf{F}}$ are d-type point processes in the usual sense, that is, they satisfy a measurability condition as in (2.1) on the probability space underlying the original Hawkes forest; see Paper B, Proposition 15, p. 92. In Paper B, Proposition 19, p. 94, we note that if $\mu_j(k_1, \ldots, k_d) = \prod_{i \in [d]} \mu_{i,j}(k_i)$, $j \in [d]$, where $\mu_{i,j}$ denotes a Poisson distribution with mean $m_{i,j} \geqslant 0$, and if, in addition, for $(i, j) \in [d]^2$, $F_{i,j}(0) = 0$ and $F_{i,j}$ is absolutely continuous with density $f_{i,j}$, then $\mathbf{N}_{\mathbf{F}}$ as in (3.8) is indeed a d-type Hawkes process with immigration intensities η_i, $i \in [d]$, branching matrix $(m_{i,j})_{(i,j) \in [d]^2}$, and displacement densities $f_{i,j}$, $(i, j) \in [d]^2$. In a similar way, we construct multivariate INAR sequences from such Hawkes forests; see Paper B, Proposition 21, p. 95. The BRW representations of Hawkes and INAR processes are mathematically fertile as they allow us to treat branching structure and positions separately: the branching part can be analyzed like Galton Watson trees and the positions (conditional on the trees) with renewal theory. In particular, for a given Hawkes forest, we build a family of approximating forests *with respect to the same underlying trees*. That is, only the positions differ; types and trees are the same for the family of approximating forests and also for the original Hawkes forests. This yields the convergence Theorem 4 (already presented) as a corollary of the following, more general, approximation result for Hawkes forests.

Theorem 7 (Paper B, Theorem 23, p. 97)**.** *Let* $\mathbf{F} = \big\{(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k, (T_k, L_k))\big\}_{k\in\mathcal{K}}$ *be a* (Q, μ, F)*-Hawkes forest as in Definition 6 with immigration law* Q *such that, for* $i \in [d]$*, there exists* $\eta_i \geqslant 0$ *such that*

$$\mathbb{E}_Q \#\Big\{k :\ T_k \in \big((n-1), n\big], L_k = i\Big\} \leqslant \eta_i, \quad n \in \mathbb{Z}. \tag{3.9}$$

Given $\mathbf{F}$ *as above, define a family of* approximating Hawkes forests

$$\mathbf{F}^{(\Delta)} = \big\{(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k^{(\Delta)}, (T_k^{(\Delta)}, L_k))\big\}_{k\in\mathcal{K}}, \quad \Delta > 0,$$

where

$$T_k^{(\Delta)} := \left\lfloor \frac{T_k}{\Delta} \right\rfloor \Delta, \quad k \in \mathbb{Z},$$

are the approximative immigration points *(their types* L_k *do not depend on* Δ*) and*

$$\mathbf{S}_k^{(\Delta)}(u) := \sum_{\varnothing < v \leqslant u} Y_{v,k}^{(\mathbf{L}_k(v^-), \mathbf{L}_k(v))}, \quad u \in \mathbf{T}_k,\ k \in \mathbb{Z}, \Delta > 0,$$

are the approximative positions *with* approximative displacement variables

$$Y_{k,u}^{(i,j;\Delta)} := \left\lceil \frac{Y_{k,u}^{(i,j;\Delta)}}{\Delta} \right\rceil \Delta, \quad u \in \mathcal{U},\ k \in \mathbb{Z}, (i,j) \in [d]^2, \Delta > 0.$$

Let $\mathbf{N}_{\mathbf{F}^{(\Delta)}}$, $\Delta > 0$*, respectively,* $\mathbf{N}_{\mathbf{F}}$ *be the projected point measure of the approximating forests* $\mathbf{F}^{(\Delta)}$, $\Delta > 0$, *respectively of the forest* $\mathbf{F}$*. Then, for any nonnegative bounded continuous function with finite support,* $f : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$*, we obtain*

$$\int f(t)\mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t \times \{j\}) \quad \xrightarrow{\text{a.s.}} \quad \int f(t)\mathbf{N}_{\mathbf{F}}(\mathrm{d}t \times \{j\}), \quad j \in [d],$$

as $\Delta \downarrow 0$*, and the convergence also holds in* L_1*.*

Generalizations of Hawkes processes

Next to this approximation technique, Hawkes forests allow several promising generalizations of standard Hawkes processes (see also Section 3.7): the autoregressive view on the Hawkes process as in (2.5) suggests obvious generalizations such as non-linearities—as treated in Brémaud and Massoulié (1996). The representation as a projected point measure of a Hawkes forest in Paper B, Proposition 19, p. 94, suggests other generalizations and alternatives for the Hawkes process. To that aim, note that the Hawkes tree in Definition 5 is *much* more general than the Hawkes tree in a standard Hawkes process. In particular, the offspring distributions $\mu = (\mu_i)_{i\in[d]}$

in a Hawkes tree may be *much* more general as independent Poisson offspring. This yields possible alternatives to standard Hawkes models; see Paper B, Section 4.1, pp. 102. Analogous ideas are valid for INAR time series. Let us highlight three examples for such generalizations:

1. *Dependence of types*: in a Hawkes process, in general, any type-i node may have children of potentially all types. However, there might be applications where $\mu_i(k_1, k_2, \ldots, k_d) = 0$, if more than one k_l is greater than 0. That is, in this case the process does not allow simultaneous children of different types.
2. *Displacement distributions:* displacement distributions of a Hawkes process are absolutely continuous. However, in some applications, the displacements may be (partially) discrete or even deterministic.
3. *Filtration:* in a Hawkes forest, we have full information for each event, that is, node: we know the tree of the node, its parent as well as its children, its generation number, and, in particular, we know if it is an immigrant node or a child node. In the Hawkes model, we project all this information on $\mathbb{R} \times [d]$—all that remains are position and type of each node. This jump from full to nearly no information might be done more gradually. There could be applications, where one can distinguish immigrants from non-immigrants. In some case, we might even know the parents of each point. Or—a bit more restrictive—one might at least know the type of the parent of all points.

Critical Hawkes processes and Kesten trees

In Paper G, we study a critical version of the monotype Hawkes process, that is, a stationary point process N with intensity $\lambda > 0$, solving (2.5) for $\eta = 0$, $m = 1$, and some displacement density w. From a branching perspective, the critical Hawkes process can be seen as a process, where all points have a parent point; see Paper G, Definition 2, p. 243. That is, we have no immigrants and thus no obvious starting points for potential (rooted) trees. Furthermore, at least one of the involved underlying trees may not die out—otherwise we observe the trivial zero point process. This lack of root nodes together with the non-extinction condition makes application of the methods from Paper B not straightforward. We show that branching-random-walk constructions are still possible. To that aim, we work with so-called *Kesten trees* or *size-biased trees*, a generalization of Galton–Watson trees, where the nodes are either *normal* or *special*; see Lyons et al. (1995). The (independent) offspring operations of normal and special nodes have a different distribution. If $(p_k)_{k\in\mathbb{N}_0}$ is the offspring distribution of a normal node and $m \in (0, \infty)$ is the corresponding expectation, then $(kp_k/m)_{k\in\mathbb{N}_0}$ is the *size-biased offspring distribution* of a special node. In the standard monotype Hawkes case, where the normal offspring distribution is $\mathrm{Pois}(1)$, one can check that the size bias corresponds to adding $+1$ to a $\mathrm{Pois}(1)$ random variable. The root node $\varnothing$ of a Kesten tree is special. Normal nodes have only normal children whereas

special nodes have exactly one special child and otherwise normal children. Obviously, such a Kesten tree never dies out. In fact, it is well known that a Kesten tree is distributed like the corresponding Galton–Watson tree conditional on non-extinction. This deals with the non-extinction condition for the critical Hawkes process. Next, we have to overcome the lack of obvious root nodes. We propose two constructions. In the 'renewal-immigration construction', we let the process start with a single root node positioned at time 0. From this node, we grow a Kesten tree with position labels, where the displacement distributions are different for special and normal nodes. Finally, we consider the limit process. In the 'Palm-process construction', we also start with a root node at time 0; this time, we interpret this node as an arbitrarily chosen point of a critical Hawkes process. We reconstruct its parent, its grandparent, its great-grandparent, etc. together with all their offspring. This backward construction is related to the 'method of backward trees' in Kallenberg (1977). Denote such a Kesten tree—with respect to $\mathrm{Pois}(1)$ normal offspring and '$\mathrm{Pois}(1)+1$' special offspring—by $\hat{\mathbf{T}}$, its nodes by $\{\sigma\}_{\sigma\in\hat{\mathbf{T}}}$, and its root node by $\varnothing$. As before, we write σ^- for the unique parent node of $\sigma \in \hat{\mathbf{T}}\backslash\{\varnothing\}$. Every node is supplied with a position in $\mathbb{R}$ in a recursive (random) way—differently for the two constructions:

Renewal-immigration construction (Paper G, Section 2.3, p. 245)**.**

Define

$$S_\varnothing := 0, \quad S_\sigma := \begin{cases} S_{\sigma^-} + Y_\sigma^{(1)}, & \text{if } \sigma \in \hat{\mathbf{T}} \text{ is a normal node and} \\ S_{\sigma^-} + Y_\sigma^{(2)}, & \text{if } \sigma \in \hat{\mathbf{T}} \text{ is a special node and } \sigma \neq \varnothing, \end{cases} \tag{3.10}$$

where $Y_\sigma^{(i)} \sim F_i$, $i = 1, 2$, independent, with $F_1(0) = F_2(0) = 0$. The distribution F_1 coincides with F, the desired displacement distribution of the critical Hawkes process; the distribution F_2 controls the desired limiting average intensity $\lambda > 0$. The chain of nodes along the special nodes of $\hat{\mathbf{T}}$ form an *infinite spine*. We interpret the positions S_σ of these nodes along the infinite spine as a *renewal process of immigrants*. From each of these immigrants, we grow a standard Hawkes tree as in Definition 5. The construction in (3.10) yields a point process N defined by $N(A) := \#\{\sigma \in \hat{\mathbf{T}} : S_\sigma \in A\}$, $A \in \mathcal{B}_b(\mathbb{R})$. If the interarrival distribution F_2 has infinite mean, by the Renewal Theorem, the counting measure $N(\cdot + \tau)$ will observe no more renewal points in finite intervals as $\tau \to \infty$. However this does not imply that we cannot observe any points *at all* for large times. The displacement distribution F_1 can be chosen in such a way that it compensates the infinite dilution of immigrants: in Paper G, Example 6, p. 246, given some $\lambda > 0$, we propose specific distributions F_1 and F_2—both regularly varying and absolutely continuous—balanced in such a way, that $\mathbb{E}\,N([0,T])/T \to \lambda$ as $T \to \infty$.

Palm-process construction (Paper G, Section 2.4, pp. 247)**.**

Similarly, a Palm construction may be written in terms of Kesten trees. The underlying tree $\hat{\mathbf{T}}$ is exactly the same as for the case with renewal immigration in (3.10). However, the position labels change. Namely, we set

$$S_{\varnothing} := 0, \quad S_{\sigma} := \begin{cases} S_{\sigma^-} + Y_{\sigma}, & \text{if } \sigma \in \hat{\mathbf{T}} \text{ is a normal node and} \\ S_{\sigma^-} - Y_{\sigma}, & \text{if } \sigma \in \hat{\mathbf{T}} \text{ is a special node and } \sigma \neq \varnothing, \end{cases} \tag{3.11}$$

where $Y_{\sigma} \sim F$, iid, with F the (not necessarily absolutely continuous) displacement distribution. One can show that the first moment measure U_0 of the point process N_0 induced by the positions $\{S_{\sigma}\}_{\sigma \in \hat{\mathbf{T}}}$ from (3.11) is given by

$$U_0 = \tilde{U} * (U + U_- - \delta_0), \tag{3.12}$$

where U denotes the renewal measure induced by F, U_- the renewal measure of the corresponding backward renewal process, and $\tilde{U}$ the first moment measure of the occupation measure of the random walk induced by the *symmetrized displacement distribution* $\tilde{F}$, that is, the distribution of $Y_1 - Y_2$ with $Y_i \overset{\text{iid}}{\sim} F$, $i = 1, 2$. Note that U_0 is a symmetric measure and that $U_0(\{0\}) \geqslant 1$ with equality if F is absolutely continuous. We conjecture that U_0 is a locally finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ if and only if the random walk induced by $\tilde{F}$ is transient. We were not able to prove this in full generality. However, if F is regularly varying at infinity with index $\alpha \in (0, 0.5)$, we show that (3.11) yields positions $\{S_{\sigma}\}_{\sigma \in T}$ that define a locally finite point process with finite intensity; see Paper G, Example 8, p. 248. In any case, if (3.12) defines a locally finite measure U_0, then N_0 defines a (by definition locally finite) point process. The law of N_0 is the *Palm distribution* of a stationary process, the critical Hawkes process. Summarizing, we state the following

Conjecture 8 (Paper G, Conjecture 1, p. 242)**.** *A critical Hawkes process with displacement distribution F exists if and only if the random walk induced by the symmetrized displacement distribution is transient.*

We will present critical Hawkes processes from a more general perspective in Section 3.6 and establish the 'only if' part of Conjecture 8.

3.3 Perspectives on Hawkes-process estimation (C, D, E, F)

In Paper C, we introduce a nonparametric estimation method for multivariate Hawkes processes. The method is nonparametric in the sense that we do not have to choose a parametric

family of excitement functions: our estimation delivers estimates for the excitement functions on an equidistant grid. The method is easy to implement (and understand) and numerically flexible. Furthermore—maybe most importantly—we derive asymptotic normality and give formulas for the covariance matrix. To the best of our knowledge, neither asymptotic normality nor covariance estimates are available in the literature for any Hawkes estimation method (including MLE). Simulation studies confirm the effectiveness of our results; see Paper C, Section 3.3, pp. 123; Paper D, Section 4.2, pp. 182; Paper F, Section 3, pp. 229. Our estimator relies on discretization. We sketch the method in the following: let $\mathbf{N}$ be data from a multivariate Hawkes process. Fix some $\Delta > 0$ and consider the ($\mathbb{N}^d$-valued) bin-count sequences. From Theorems 3 and 4, we know that the distribution of these sequences is very similar to specific INAR sequences. In particular, the corresponding INAR coefficients are related to the excitement functions $(h_{i,j})_{(i,j)\in[d]^2}$ of the original Hawkes process as in a multivariate version of (3.1). Thus, we estimate the coefficients of the approximating INAR time series (by conditional least squares) and then retranslate the estimates into the Hawkes world accordingly. For the multivariate INAR model, we give the following result on conditional-least squares estimation.

Theorem 9 (Paper C, Theorem 10, p. 120)**.** *Let $(\mathbf{X}_n)$ be a d-dimensional INAR(p) sequence as in Paper C, Definition 3, p. 115, with immigration-parameter (column) vector $\mathbf{a}_0 \in \mathbb{R}^d_{\geqslant 0}\setminus\{0_d\}$, and reproduction-coefficient matrices $A_k \in \mathbb{R}^{d\times d}_{\geqslant 0}$, $k \in \{1,2,\ldots,p\}$, such that $\operatorname{spr}\left(\sum_{k=1}^p A_k\right) < 1$. Let*

$$\mathbf{B} := \left(A_1, A_2, \ldots, A_p, \mathbf{a}_0\right) \in \mathbb{R}^{d\times(dp+1)} \quad \text{and}$$
$$\hat{\mathbf{B}}^{(n)} := \hat{\theta}^{(p,n)}_{\mathrm{CLS}}\left((\mathbf{X}_k)_{k=1,\ldots,n}\right) \in \mathbb{R}^{d\times(dp+1)}$$

the CLS-estimator from Paper C, Definition 8, p. 119 with respect to the sample $(\mathbf{X}_k)_{k=1,\ldots,n}$. Then $\hat{\mathbf{B}}^{(n)}$ is a weakly consistent estimator for $\mathbf{B}$. Furthermore, let $\mathbf{Z}$ be the design matrix of the CLS-estimator with respect to $(\mathbf{X}_k)_{k=1,\ldots,n}$. Assume that the limit

$$\frac{1}{n-p}\mathbf{Z}\mathbf{Z}^\top \xrightarrow{\mathrm{p}} : \Gamma \in \mathbb{R}^{(dp+1)\times(dp+1)}, \quad n \longrightarrow \infty, \tag{3.13}$$

exists and is invertible. In addition, assume that the model is irreducible in the sense that $\mathbb{P}[\mathbf{X}_{0,i} = 0] < 1$, $i = 1,2,\ldots,d$. Then, for the asymptotic distribution of $\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) \in \mathbb{R}^{d^2p+d}$, one has, for $n \to \infty$,

$$\sqrt{n-p}\Big(\operatorname{vec}\big(\hat{\mathbf{B}}^{(n)}\big) - \operatorname{vec}\big(\mathbf{B}\big)\Big) \xrightarrow{\mathrm{d}} \mathcal{N}_{d^2p+d}\Big(0_{d^2p+d}, \left(\Gamma^{-1}\otimes 1_{d\times d}\right) W \left(\Gamma^{-1}\otimes 1_{d\times d}\right)\Big),$$

where

$$W := \mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_0\left(\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_0\right)^\top\right] \in \mathbb{R}^{(d^2p+d)\times(d^2p+d)} \tag{3.14}$$

with

$$\mathbf{u}_0 := \mathbf{X}_0 - \mathbf{a}_0 - \sum_{k=1}^{p} A_k \mathbf{X}_{-k} \quad \textit{and} \quad \mathbf{Z}_0 := \left(\mathbf{X}_{-1}^\top, \mathbf{X}_{-2}^\top, \ldots, \mathbf{X}_{-p}^\top, 1\right)^\top.$$

Here, the vec : $\mathbb{R}^{p\times q} \to \mathbb{R}^{pq\times 1}$ operator stacks the columns of its (matrix)-argument and $\otimes$ denotes the Kronecker product. In the first step of the proof, we use that the INAR sequence can be rewritten as a standard autoregressive with white-noise innovations. In Paper C, Proposition 6, p. 118, we show that

$$\mathbf{u}_n := \mathbf{X}_n - \sum_{k=1}^{p} A_k \mathbf{X}_{n-k}, \quad n \in \mathbb{Z}, \tag{3.15}$$

defines a white-noise sequence. Thus, we obtain the following

Corollary 10 (Paper C, Corollary 7, p. 119)**.** *Let* $(\mathbf{X}_n)$ *be the multivariate INAR(p) sequence and* $(\mathbf{u}_n)$ *the white-noise sequence from* (3.15)*. Then* $(\mathbf{X}_n)$ *solves the system of stochastic difference-equations*

$$\mathbf{X}_n = \mathbf{a}_0 + \sum_{k=1}^{p} A_k \mathbf{X}_{n-k} + \mathbf{u}_n, \quad n \in \mathbb{Z}\,.$$

That is, we can apply standard time series theory for multivariate INAR(p) sequences. Surprisingly, we did not find a proof for conditional-least-squares properties in this setup (without assuming iid innovations) in the relevant literature. This is why we supply a self-contained proof of Theorem 9; see Paper C, pp. 152. At this point, it may become even clearer, why we work with Poisson reproduction instead of binomial thinning for the INAR-sequences; see Definition 1 and Paper C, Definition 3, p. 115: in the multivariate case, the reproduction coefficients of the approximating sequences may well be larger than one—without any criticality issues. Furthermore, we want to approximate the number of children points of a past Hawkes event in a bin, and clearly a point can have more than one child in a future bin (for coarser Δ). We transform the INAR estimates appropriately and obtain the Hawkes estimator:

Definition 11 (Paper C, Definition 11, p. 121)**.** *Let* $\mathbf{N} = \left(N^{(1)}, N^{(2)}, \ldots, N^{(d)}\right)$ *be a d-variate Hawkes process with baseline-intensity vector* $\eta \in \mathbb{R}^d_{\geqslant 0}\setminus\{0_d\}$ *and excitement function* $H = (h_{i,j}) : \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{d\times d}_{\geqslant 0}$ *such that* (2.3) *in Paper C holds. Let* $T > 0$ *and consider a sample of the process on the time interval* $(0, T]$*. For some* $\Delta > 0$*, construct the* $\mathbb{N}_0^d$*-valued* bin-count

sequence *from this sample:*

$$\mathbf{X}_k^{(\Delta)} := \left(N^{(j)}\Big(\big((k-1)\Delta, k\Delta\big] \Big) \right)_{j=1,\dots,d}, \quad k = 1, 2, \dots, n := \lfloor T/\Delta \rfloor . \tag{3.16}$$

Define the multivariate Hawkes estimator *with respect to some support* s, $\Delta < s < T$, *by applying the CLS-operator from Paper C, Definition 8, p. 119, with maximal lag* $p := \lceil s/\Delta \rceil$ *on these bin-counts:*

$$\hat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta} \hat{\theta}_{CLS}^{(p,n)} \left(\left(\mathbf{X}_k^{(\Delta)} \right)_{k=1,\dots,n} \right). \tag{3.17}$$

The following additional notation clarifies what the entries of the $\hat{\mathbf{H}}^{(\Delta,s)}$ matrix actually estimate:

$$\left(\hat{H}_1^{(\Delta,s)}, \dots, \hat{H}_p^{(\Delta,s)}, \hat{\eta}^{(\Delta,s)} \right) := \hat{\mathbf{H}}^{(\Delta,s)} \quad \text{and} \quad \left(\hat{h}_{i,j}^{(\Delta,s)}(k\Delta) \right)_{(i,j)\in[d]^2} := \hat{H}_k^{(\Delta,s)}.$$

At this point note that in Paper C, the role of the indices $(i, j) \in [d]^2$ is switched in comparison to (2.6) and all our other work on multivariate/multitype Hawkes processes; see Paper D, Remark 6, p. 165. That is, in the context of Definition 11, $h_{i,j}$ models the excitement from component j to component i. Applying the results from Theorem 9 to the estimates in Definition 11, we obtain approximative distributional properties for all estimates; see Paper C, Remark 12, pp. 121. A bivariate simulation study confirms our findings: the estimates are normally distributed around the true values and the confidence intervals derived from the covariance estimates yield perfect coverage rates; see Paper C, Section 3.3, pp. 123.

In Paper F, we present another simulation study comparing MLE and our method. The computational and conceptual advantages of our method are obvious. In terms of mean squared error our method competes well. Note that our estimation procedure depends on the choice of the bin size Δ as well as on the chosen support parameter s. We propose several methods for a sensible choice of these parameters; see Paper C, Section 4, pp. 130. Furthermore, we discuss the sources of possible bias in some detail. In Paper E, we generalize the method to the marked case; see Section 2.5. That is, we also deliver nonparametric estimates of the impact functions on a grid. This is a simple conceptual step: we discretize not only time but also mark space. The resulting formulas are however quite involved notationwise. In Paper E, Section 2.5, pp. 199, we provide the necessary calculations and present the resulting formulas in all detail—ready for implementation.

In Paper D, we consider higher-dimensional (unmarked) Hawkes processes. We therefore pay special attention to computational issues. We give computationally efficient formulas for variance estimates (the numerical bottleneck of the procedure); see Paper D, Algorithm 14,

p. 174 and Algorithm 16, p. 177. Furthermore, we introduce a two-step procedure for the estimation tackling the computational difficulties: in a first step, we statistically check if there is significant excitement *at all* from one event type to another. Only in a second step, we aim to estimate the remaining nonzero excitement functions quantitatively. We explain, why Δ can be chosen quite coarse in this first step—thus simplifying computations. In addition, we introduce a significance parameter $\alpha \in (0, 1)$. It supplies the fraction of false positive excitement estimates. Thus, the smaller α, the fewer excitements we have to estimate in the second step (at the cost of ignoring some excitements). In other words, α determines the complexity of the second step. Summarizing, we use the bin size Δ and the significance parameter α in order to control the computational complexity of the method. This is discussed in Paper D, Section 3.2, pp. 172. The best way to analyze the excitement relations of a Hawkes process is a graphical representation:

3.4 Graphical perspective (D, E)

We have seen in Section 2.3 that one possible perspective on multitype Hawkes processes is the parent–children interpretation of a branching process. From this point of view, there are strong 'causal' effects between the different event streams—without going into details. Such causal interactions are often described as graphs; see Pearl (2009). In the context of many possible types, a desirable (and natural) assumption with respect to Hawkes models might be that the branching matrix is relatively sparse—say of order $O(d)$, where d denotes the number of types. This sparseness makes graphical descriptions even more interesting. In this spirit, Paper D introduces the Hawkes skeleton and the Hawkes graph. All graphical terminology used in the following definition is made precise in Paper D, Definition 7, p. 165. Also note that in the context of Paper D, we denote the branching matrix by $(a_{i,j})$ (before $(m_{i,j})$)—in order to emphasize its role as an **a**djacency matrix of a directed graph with weighted edges.

Definition 12. *Let* $\mathbf{N}$ *be a* d*-type Hawkes process with immigration intensities* $\eta_1, \eta_2, \ldots, \eta_d$ *and branching coefficients* $a_{i,j} (= \int h_{i,j}(t)\mathrm{d}t)$, $(i, j) \in [d]^2$*; see Paper C, Definition 2, p. 163 and Definition 5, p. 164. The* Hawkes graph skeleton $\mathcal{G}^*_{\mathbf{N}} = (\mathcal{V}^*_{\mathbf{N}}, \mathcal{E}^*_{\mathbf{N}})$ *of* $\mathbf{N}$ *consists of a set of vertices* $\mathcal{V}^*_{\mathbf{N}} = [d]$ *and a set of edges*

$$\mathcal{E}^*_{\mathbf{N}} := \big\{(i, j) \in \mathcal{V}^*_{\mathbf{N}} \times \mathcal{V}^*_{\mathbf{N}} : \ a_{i,j} > 0\big\}.$$

For $j \in [d]$*, we denote the* parent*, respectively,* ancestor set *of* j *with respect to the Hawkes skeleton* $\mathcal{G}^*_{\mathbf{N}}$ *by* $\mathrm{PA}_{\mathbf{N}}(j)$*, respectively,* $\mathrm{AN}_{\mathbf{N}}(j)$*. For the* Hawkes graph $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ *of* $\mathbf{N}$*, each vertex, respectively, edge of the corresponding Hawkes skeleton is supplied with a* vertex

weight*, respectively,* edge weight*:*

$$\begin{aligned}\mathcal{V}_{\mathbf{N}} &:= \Big\{(j;\eta_j) :\ j \in \mathcal{V}^*_{\mathbf{N}} \text{ and } \eta_j \text{ is the } j\text{-th immigration intensity of } \mathbf{N}\Big\},\\ \mathcal{E}_{\mathbf{N}} &:= \Big\{(i,j;a_{i,j}) :\ (i,j) \in \mathcal{E}^*_{\mathbf{N}} \text{ and } (a_{i,j})_{(i,j)\in[d]^2} \text{ is the branching matrix of } \mathbf{N}\Big\}.\end{aligned}$$

We call the branching matrix $A = (a_{i,j}) \in \mathbb{R}^{d\times d}_{\geqslant 0}$ *of* $\mathbf{N}$ *the* adjacency matrix *of* $\mathcal{G}_{\mathbf{N}}$.

i) A Hawkes graph $\mathcal{G}_{\mathbf{N}}$ *is* weakly, strongly, respectively, fully connected *if the corresponding skeleton* $\mathcal{G}^*_{\mathbf{N}}$ *is weakly, strongly, respectively, fully connected.*

ii) Vertex $(j;\eta_j)$ *of a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$ *is a* source*, respectively,* sink vertex*, if it is a source, respectively, sink vertex in the corresponding skeleton* $\mathcal{G}^*_{\mathbf{N}}$*. Furthermore,* $(j;\eta_j)$ *is a* redundant vertex *if* $\eta_j = 0$ *and, in addition,* $\eta_i = 0$ *for all* $i \in \mathrm{AN}_{\mathbf{N}}(j)$.

iii) For any walk $w \in \mathcal{W}_{\mathcal{G}_{\mathbf{N}}}$ $(:= \mathcal{W}_{\mathcal{G}^*_{\mathbf{N}}})$ *in a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$*, we define the* walk weight

$$|w| = |(i_0, i_1, \ldots, i_g)| := \begin{cases} 1, & g = 0, \text{ and}\\ \prod_{l=1}^{g} a_{i_{l-1},i_l}, & g > 0,\end{cases}$$

where a_{i_{l-1},i_l}, $l = 1,2,\ldots,g$, *are the edge weights from* $\mathcal{E}_{\mathbf{N}}$.

iv) A Hawkes graph is subcritical *if*

$$\sum_{w\in\mathcal{W}^{(i_0,i_0)}} |w| < \infty,\ i_0 \in [d], \text{ or, equivalently, } \sum_{\substack{w:\\ w \text{ closed walk in } \mathcal{G}_{\mathbf{N}}}} |w| < \infty. \tag{3.18}$$

The notion of a subcritical Hawkes graph in Definition 12 iv) might ask for further explanation. The following theorem clarifies things:

Theorem 13 (Paper D, Theorem 9, p. 167)**.** *Let* $\mathbf{N}$ *be a Hawkes process and let* $\mathcal{G}_{\mathbf{N}}$ *be the corresponding Hawkes graph. Then* $\mathbf{N}$ *is a subcritical Hawkes process if and only if* $\mathcal{G}_{\mathbf{N}}$ *is a subcritical Hawkes graph.*

Theorem 13 substitutes the eigenvalue-based criterion from Section 2.4 with a graph-based criterion. That is, we substitute an analytic problem with a more combinatorial problem. As long as the graph is sparse, the graph-based criterion is often tractable even without any computer-power. That is, when we build or estimate a (relatively sparse) Hawkes process, Theorem 13 provides a tool which allows to detect the walks, respectively, edges that are 'responsible' for global (near) criticality. If we have the possibility to influence the data-generating system (e.g., as regulators in a market) this tells us, where to adjust or even delete connections in order to stabilize the system. This engineering point of view is also present in the definition of specific coefficients that summarize the effect of the various event types on the whole

system (*cascade coefficients*) and on themselves (*feedback coefficients*); see Paper D, Definition 11, p. 170. Moreover, we found graph terminology very useful in terms of implementation: working with parent and ancestor sets of Hawkes graphs simplifies efficient implementation of Hawkes processes. E.g., in calculations of realized conditional intensities as in (2.6), it is enough to consider the parent set of j. For simulation, the Hawkes graph simplifies matters in a similar manner. The notion of ancestor sets becomes important if we are only interested in a model for the events of a specific type j. In this case, it suffices to consider the ancestor sets of j for our Hawkes model. The various notions of connectivity are also important with respect to modeling: e.g., if Hawkes graphs (estimates) are not weakly connected, one can model separately the event streams corresponding to the weakly connected subgraphs. Finally note that—when explaining a (calibrated) Hawkes model to students or to a board with little statistical background—Hawkes graphs are more meaningful (and entertaining) than formulas like (2.6). These applications make the Hawkes graph a compact, yet meaningful and useful summary of a Hawkes process.

3.5 Limit-order-book modeling with Hawkes processes (C, E)

We apply the estimation method from Paper C on limit-order-book (LOB) data in Paper E and also in Paper C itself. For an introduction to these kinds of data, see Gould et al. (2013) as well as Paper E, Section 3, pp. 200. In Paper E, we consider the six-type event stream of market bid orders (MB), market ask orders (MA), limit bid orders (LB), limit ask orders (LA), and cancelations on either side of the LOB (CB and CA). Note that, in our terminology, *a market bid order is a bid-side event, that is, a market sell order*. We apply the concepts and methods from Paper C and Paper D to specify a fully parametric and computationally tractable Hawkes model with very few a priori assumptions. In a nonparametric manner, we identify a set of non-zero edges in the Hawkes skeleton, power-law decay of the excitement functions (and thus for the decay kernels), linearity of impact functions, a general bid–ask symmetry of the data, as well as linear dependence of all conditional intensities on the *LOB imbalance* defined by

$$I(t) = \frac{Q^{(a)}(t) - Q^{(b)}(t)}{Q^{(b)}(t) + Q^{(a)}(t)} \quad (\in [-1, 1]),\ t \in \mathbb{R}, \tag{3.19}$$

where $Q^{(b)}(t)$, respectively, $Q^{(a)}(t)$, denotes the number of limit orders at best bid, respectively, best ask price; see Paper E, Section 4, p. 204. Finally, we summarize these findings in a parametric Hawkes model. For $O \in \mathcal{O} := \{\mathrm{MB}, \mathrm{MA}, \mathrm{LB}, \mathrm{LA}, \mathrm{CB}, \mathrm{CB}\}$ and $k \in \mathbb{Z}$, let $(T_k^{(O)}, V_k^{(O)})$ denote the k-th type-O order-book event with time stamp $T_k^{(O)}$ and volume $V_k^{(O)}$. We formulate

the final parametric model; see Paper E, Section 4.6, pp. 210:

Parametric model

For the conditional intensities at time $t \in \mathbb{R}$, we set

$$\begin{aligned}
\lambda_{\mathrm{MB}}^{(\theta_{MB})}(t) &= \eta_{MB}(1 + I(t-)),\\
\lambda_{\mathrm{LB}}^{(\theta_{LB})}(t) &= \eta_{LB}\big(1 - I(t-)\big) + \sum_{O\in\{\mathrm{MA,CA}\}} m_{O,\mathrm{LB}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{LB}})}(t - T_k^{(O)}),\\
\lambda_{\mathrm{CB}}^{(\theta_{\mathrm{CB}})}(t) &= \eta_{\mathrm{CB}}\big(1 + I(t-)\big) + \sum_{O\in\{\mathrm{LB,MA}\}} m_{O,\mathrm{CB}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{CB}})}(t - T_k^{(O)}),\\
\lambda_{\mathrm{MA}}^{(\theta_{MA})}(t) &= \eta_{MA}(1 - I(t-))\\
\lambda_{\mathrm{LA}}^{(\theta_{LA})}(t) &= \eta_{LA}\big(1 + I(t-)\big) + \sum_{O\in\{\mathrm{MB,CB}\}} m_{O,\mathrm{LA}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{LA}})}(t - T_k^{(O)}), \text{ and}\\
\lambda_{\mathrm{CA}}^{(\vartheta_{\mathrm{CA}})}(t) &= \eta_{\mathrm{CA}}\big(1 - I(t-)\big) + \sum_{O\in\{\mathrm{LA,MB}\}} m_{O,\mathrm{CA}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{CA}})}(t - T_k^{(O)}).
\end{aligned}$$

The decay kernels are parametrized by $\vartheta = (\alpha, x_m) \in (0,\infty)^2$, where

$$w^{(\vartheta)}(t) := \begin{cases} \alpha x_m^{\alpha}(x_m + t)^{-(1+\alpha)}, & t > 0,\\ 0, & \text{else.} \end{cases} \tag{3.20}$$

Furthermore, all parameters are bid–ask symmetric, that is,

$$\begin{aligned}
\theta_{\mathrm{MB}} &= (\eta_{\mathrm{MB}}) = (\eta_{\mathrm{MA}}) = \theta_{\mathrm{MA}},\\
\theta_{\mathrm{LB}} &= (\eta_{\mathrm{LB}}, m_{\mathrm{MA,LB}}, m_{\mathrm{CA,LB}}, \vartheta_{\mathrm{MA,LB}}, \vartheta_{\mathrm{CA,LB}})\\
&= (\eta_{\mathrm{LA}}, m_{\mathrm{MB,LA}}, m_{\mathrm{CB,LA}}, \vartheta_{\mathrm{MB,LA}}, \vartheta_{\mathrm{CB,LA}}) = \theta_{\mathrm{LA}},\\
\theta_{\mathrm{CB}} &= (\eta_{\mathrm{CB}}, m_{\mathrm{LB,CB}}, m_{\mathrm{MA,CB}}, \vartheta_{\mathrm{LB,CB}}, \vartheta_{\mathrm{MA,CB}}), \text{ and}\\
&= (\eta_{\mathrm{CA}}, m_{\mathrm{LA,CA}}, m_{\mathrm{MB,CA}}, \vartheta_{\mathrm{LA,CA}}, \vartheta_{\mathrm{MB,CA}}) = \theta_{\mathrm{CA}}.
\end{aligned}$$

In total, the model has 15 parameters. This is still quite a lot. However, they are relatively few, in comparison to $6^2(1 + 2 + 2) + 6 = 136$ potential parameters—working with the fully connected Hawkes skeleton with 6^2 edges—each supplied with one branching parameter, two decay parameters, and two impact parameters.

Roughly speaking, the literature on LOB modeling with point processes pursues two di-

rections: pure state dependence as, e.g., in Cont and de Larrard (2013), and pure past dependence as, e.g., in Bacry et al. (2014). By letting the Hawkes baseline intensities depend on the imbalance, our model in Paper E is a mélange of these approaches. Note that the inclusion of the imbalance in the model has its reason in descriptive analysis: we find that the 'imbalance/market-order side' relation exhibits a stretched inverted S-shape that can be linearly approximated by

$$\hat{\mathbb{P}}\left[\text{buy market order at time } t | \text{market order at time } t\right] \approx \frac{1 - I(t)}{2}$$

very well; see Paper E, Figure 3, p. 205. For the other order types, we find similar relations. We conclude that the imbalance is a relevant and compact summary for the state of the book. Our model gets rejected for large sample sizes by standard tests. However, it works well for prediction of the next order type; see Paper E, Section 6, pp. 218. Furthermore, the values of the calibrated parameters have reasonable interpretations such as 'reaction time' or 'proportional impacts'. In particular, we find that market orders 'drive the market': in the model, if we suppress market orders, the activity of the market would quasi stop. This makes sense from a naive economic point of view: if nobody is willing to actually buy or sell stocks, there is no need in sending offers, anymore. As regulator, one could exploit the various loops in the estimated Hawkes graph to detect the edges that contribute to potential criticality of the process. These could be regulated by taxes, delays, or other restrictions. Our results for the spectral radius and for the shape of the power-law decay are quite different than the results in the literature with values of $\alpha \in (0, 1)$ and the spectral radius being near to the critical case 1; see Hardiman, S.J. et al. (2013) or Bacry et al. (2014). We suppose that the inclusion of order size and—even more importantly—state dependence is the reason for these different results. Differently speaking, the often observed near criticality and long-range dependence property of Hawkes-model fits may presumably be explained by ignoring relevant covariates in the model fit. This lack of explanatory variables is compensated by large values for m and very heavy-tailed decay kernels of infinite expectations—similarly as one would expect in ordinary regression.

3.6 Perspectives on the critical case (G)

Data analysis applying Hawkes processes often report branching-coefficient estimates near 1. This partly explains the interest in possible critical cases of Hawkes processes, that is, in stationary point processes N with finite average intensity $\lambda > 0$ that solve a critical version of (2.5) of the form

$$\Lambda_N(t) = \int_{(-\infty,t)} f(t-s)N(\mathrm{d}s), \quad t \in \mathbb{R}, \tag{3.21}$$

for some displacement density f. Theorem 1 in Brémaud and Massoulié (2001) states that such a point process N exists if

$$\sup_{t\geqslant 0} f(t)t^{1+\alpha} \leqslant R \quad \text{and} \quad \lim_{t\to\infty} f(t)t^{1+\alpha} = r, \tag{3.22}$$

for some constants $r, R \in (0,\infty)$ and $\alpha \in (0, 0.5)$. Furthermore, Theorem 2 in the same reference states that (3.21) together with the target average intensity $\lambda > 0$ specifies at most one distribution. Critical Hawkes processes are also interesting from a purely theoretical point of view as their existence is somewhat paradoxical: for an arbitrary point of a Hawkes process, consider the number of its children, its grandchildren, its great-grandchildren etc. This sequence forms a Galton–Watson process with offspring distribution $\text{Pois}(m)$. It is well known that not only subcritical ($m < 1$) but also critical ($m = 1$) Galton–Watson processes die out almost surely; for example, see Theorem 6.1 in Harris (1963). Consequently, in a critical Hawkes process, any realized point will almost surely have only a finite number of descendants. For illustration, consider $F(0) = 1$ for the displacement distribution (for the sake of example, we disregard simplicity of the process). In this case, the position of each point coincides with the position of its descendants. Therefore, if the immigrants are thinned, then all descendants are bound to vanish—and consequently the whole process with them. In fact, Proposition 1 in Brémaud and Massoulié (2001) shows that $\int t\mathrm{d}F(t) < \infty$ already has the same effect: if the expectation of the displacements is finite, no nontrivial solution to (3.21) can exist. In contrast, the conditions in (3.22) guarantee that the displacements are balanced in such a way that the Hawkes families grow larger and larger to fill the larger and larger gaps between the immigrants. In Paper G, we rewrite (3.21) in terms of *critical cluster fields*. A critical cluster field $[F]$ operates on a point process N with distribution L and points $\{T_n\}_{n\in I}$, $I \subset \mathbb{Z}$, by

$$[F] \star \{T_n\}_{n\in I} := \bigcup_{n\in I} \bigcup_{k=1}^{K_n} \{T_n + X_{n,k}\}, \tag{3.23}$$

where $\{K_n, X_{n,k} : n \in \mathbb{Z}, k \in \mathbb{N}\}$ are independent random *cluster variables* (also independent of N) with $K_n \sim \text{Pois}(1)$ and $X_{n,k} \sim F$. We denote the resulting point process by $N_{[F]}$ and its distribution by $L_{[F]}$. We define critical Hawkes processes by applying this terminology:

Definition 14 (Paper G, Definition 2, p. 243)**.** *Let F be a distribution on $\mathbb{R}$ with $F(0) = 0$ and $[F]$ the induced critical cluster field. Assume that N is an ergodic solution to*

$$N = N_{[F]} \tag{3.24}$$

with finite and constant average intensity $\lambda > 0$. Then we call N a critical (F,λ)-Hawkes process.

If F is absolutely continuous with density f, then the critical cluster operation can be interpreted as starting an inhomogeneous Poisson process with intensitiy $f(\cdot - T_n)$ from each point T_n. Thus, one can show that the critical (F, λ)-Hawkes process indeed solves (3.21). Vice versa, (3.24) is more general than (3.21) in that the displacements are not necessarily absolutely continuous. Also note that (3.24) specifies a unique parent point for every point T_n of N: indeed, for each $n \in \mathbb{Z}$, there exist $n' \in \mathbb{Z}$ and $k \in \mathbb{N}$ such that $k \leqslant K_{n'}$ and $T_{n'} + X_{n',k} = T_n$, where the random variables $K_{n'}$ and $X_{n',k}$ stem from the clustering operation; see (3.23). That is, the critical Hawkes process is 'eating its own tail'. The trivial—yet crucial—observation is that (3.24) also holds in distribution. This distributional property of point processes is well examined; see Chapter 12 in Matthes et al. (1978). If $L = L_{[F]}$ holds for some point process law L, then L is called *cluster invariant with respect to* $[F]$. Furthermore, F is called *stable* if such a distribution L exists. From (3.24), we get that critical Hawkes processes are obviously cluster invariant with respect to the cluster induced by their displacement distribution. Thus, we obtain many necessary conditions for the existence of critical Hawkes processes as corollaries from standard results:

Theorem 15 (Paper G, Theorem 3, p. 244)**.** *Let F be a distribution on $[0, \infty)$ and $\lambda > 0$. Assume that a critical (F, λ)-Hawkes process N as in Definition 14 exists. Then the following holds:*

a) Definition 14 specifies a unique, infinitely divisible, and stationary distribution H on $(M_p, \mathcal{M}_p)$.

b) Let L be the distribution of a Poisson random measure with finite average intensity λ. For $g \in \mathbb{N}$, denote g independent clustering operations by $[F^{[g]}] \star \cdot$. Then, as $g \to \infty$, $L_{[F^{[g]}]}$ converges weakly to H.

c) The random walk induced by the symmetrized displacement distribution $\tilde{F}$ is transient.

We propose various constructions for critical (F, λ)-Hawkes processes such as a Poisson embedding (Paper G, Section 2.2, pp. 244), a renewal immigration construction (Paper G, Section 2.3, p. 245), and a backward construction of the corresponding Palm process (Paper G, Section 2.4, p. 247)—the latter two approaches were already presented in (3.10) and (3.11). All approaches indicate that transience of the random walk induced by the symmetrized displacement distribution is not only a necessary but also a sufficient condition for the existence of critical Hawkes processes; see Conjecture 8. In the end of Paper G, we point out how the presented methods could be fertile for the open discussion of critical multivariate Hawkes processes and of critical INAR time series. We will touch on this possibility in Section 3.7.

Though our contribution to the existence of critical Hawkes processes is incomplete, we have identified the distribution of any possible critical Hawkes process: it coincides with the

distribution of a cluster-invariant point process. From Theorem 15b), we get that this distribution can be constructed by starting with a Poisson random field of ancestors, and then applying iterated clustering—only considering children, then only grandchildren, then only great-grandchildren, etc. In other words, the points of a critical Hawkes process are related like 'cousins of a very high degree'.

3.7 Perspectives on future research

The ideas brought forward in this thesis open the door to further research projects on Hawkes processes:

1. For *any* $\mathbb{N}_0$-valued time series model, one can construct a point process model the way it is done for the approximating sequences in Theorem 17. So, studying integer-valued time series might be inspiring for developing and understanding point process models. For example, one might want to consider the point process corresponding to an integer-valued autoregressive moving-average (INARMA) time series. For INARMA time series, a moving-average part is added to the autoregressive part in the defining difference equations (2); see Fokianos and Kedem (2012). In fact, the corresponding point process is nothing else but the 'dynamic contagion process' as proposed in Zhao (2012).
2. Vice versa, it might also be inspiring for integer-valued time series theory to translate point process models into the discrete-time setup. For example, one might want to reuse the generalizing results on autoregressive point processes in Brémaud and Massoulié (1996) to the INAR context. Here, the autoregression of the point process is modeled not as an affine, but as a general Lipschitz function of the past of the process. The analogous generalization in time series theory is 'nonlinear Poisson autoregression' as studied in Fokianos and Tjøstheim (2012) for the case $p = 1$. In the latter paper, the authors also find a Lipschitz condition for the transfer function. Yet another idea might be to consider marked INAR sequences in analogy to marked Hawkes processes; see Liniger (2009).
3. A particularly interesting application of the Hawkes–INAR link would be to transfer the work from Paper G—respectively, Brémaud and Massoulié (2001)—on critical Hawkes processes to the time series case; see Paper G, Section 3.2, pp. 250. More explicitly, one could consider time series (X_n) solving

$$X_n = \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \xi_{n,k}^{(\alpha_k)}, \quad \mathbb{E}\, X_n \equiv \lambda > 0, \quad n \in \mathbb{Z}, \tag{3.25}$$

for some independent random variables $\{\xi_{n,k}^{(\alpha_k)}\}$ with $\xi_{n,k}^{(\alpha_k)} \sim \mathrm{Pois}(\alpha_k)$, $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, $n \in \mathbb{Z}$, where $\sum_{k\in\mathbb{N}} \alpha_k = 1$. Arguing in a similar manner as in Section 3.6 for the Hawkes

process, we can rewrite (3.25) in terms of a (critical) cluster operation with offspring distribution $\mathrm{Pois}(1)$ and displacement distribution (α_k). In analogy to Theorem 15, one can then argue that solutions to (3.25) necessarily specify a unique stationary time series distribution. Thus, our conjecture is that such a critical INAR(∞) process exists if and only if the symmetric random walk with step-size distribution $(\sum_{l=1}^{\infty} \alpha_l \alpha_{k+l})_{k\in\mathbb{Z}}$ ($\alpha_k := 0, k \leqslant 0$) is transient. One can show that for this transience condition we necessarily need $\sum_{k=1}^{\infty} \alpha_k k = \infty$. So in particular, we need $\alpha_k > 0$, infinitely often, and a critical INAR(p) process with $p < \infty$ cannot exist. This shows that the INAR(∞) processes introduced in Paper A are non-trivial extensions of INAR(p) processes with $p < \infty$. As a final point of interest, note that one can rewrite (3.25) as a standard autoregressive difference equation; see Paper A, Proposition 8, p. 48, for the subcritical case. In other words, critical INAR(∞) time series are nontrivial critical (='unit root') *and stationary* autoregressive processes of infinite variance.

4. Clearly, the estimation method from Paper C and Paper D awaits applications and possible extensions as in Paper E to the marked case.

5. In the simulation study presented in Paper F, we noted that—when the underlying Hawkes process is near criticality—also MLE is quite biased for moderate sample sizes. This asks for a more thorough analysis of MLE estimation of Hawkes parameters.

6. The concept of a Hawkes skeleton from Paper D is important for the estimation of high-dimensional Hawkes processes. No matter what estimation method we use: it will always be easier to estimate the skeleton first (simply because it carries less information). In a second step, we treat the estimated skeleton as the true skeleton and estimate the Hawkes model on the basis of this estimated skeleton. If the skeleton is sparse, then the Hawkes estimation will be much simpler. We believe that such a two-step estimation procedure is the method of choice when applying the Hawkes analysis to large event-stream datasets.

7. The Hawkes-graph concept from Paper D might be extended by including a second edge weight to the the branching coefficient, namely the (possibly infinite) expected displacement distance. This would include the time-domain to the graph representation.

8. It has been made clear that there are two quite different points of views on Hawkes processes. One is the autoregressive view, the other the branching view. Both approaches allow for different extensions/generalizations. The natural generalization of linear to non-linear autoregression has been treated in Brémaud and Massoulié (1996). In Paper B, Remark 26, p. 102, we propose several extensions that are natural directions when starting from the branching perspective of Hawkes processes. For example, the different offsprings of a type-i event might depend on each other (in the standard Hawkes case these are independently Poisson distributed). More specifically, a type-i event might in total

have either zero or one offspring. That is, no event could have offspring of several types. This kind of extension could be interesting for applications. E.g., in our context of limit-order-book modeling, one could argue that a limit order gets *either* executed *or* canceled; the standard Hawkes approach does not take these kinds of restrictions into account.

9. Our analysis in Paper E shows that both, past and state, are important explanatory entities for limit-order-book (LOB) event streams. Thus, we consider the imbalance in the book as state variable; see (3.19). As we work with 'large-tick assets'—that is, assets with relatively low prices—the bid–ask spread, the distance between best bid and best ask price, is constant to one tick. For more general assets, the spread may vary considerably. To include this in the model it would be natural to scale the LOB imbalance by the spread, that is, to introduce the *spread-scaled LOB imbalance* by

$$\tilde{I}(t) = \big(P^{(a)}(t) - P^{(b)}(t)\big)\frac{Q^{(a)}(t) - Q^{(b)}(t)}{Q^{(b)}(t) + Q^{(a)}(t)} \quad (\in \mathbb{R})$$

(for the notation, see Paper E, Section 3.2, pp. 201), and include a term $\eta + \tilde{\eta}\tilde{I}(t)$ for the baseline intensities.

10. Event streams from limit order books were very convenient data for our work in the sense that they provided many data points from a more or less stationary distribution to apply our Hawkes-based concepts on real-world data. So as a part of our research, we worked with large LOB datasets. If the goal is prediction of the near future, we would—in retrospective—approach the huge datasets of HFT data in a completely different manner: by bootstrapping! We propose to define some metric on the space of order book snapshots. For a given LOB snapshot, one selects a sample of all situations in the past that are in a reasonable distance from the observation. From this sample, one could then bootstrap the distribution of the near future of the event streams from this sample—possibly weighted by the corresponding distances to the actual situation. This approach seems extremely simple, but powerful—given the quasi infinite amount of data that is available. Furthermore, the method seems quite tractable from a computational point of view.

11. Some of our nonparametric LOB event-stream analysis exhibit periodicities of the excitement functions; see, e.g., Paper C, Figure 11, p. 145. A preference for 'absolute' round times can be ruled out because event times modulo suspicious lags are uniform. This leaves us with a preference for 'relative' round times. That is, on the level of the algorithms there seems to be something like 'round' reaction times. To the best of our knowledge this has not been noted before. This seems to be worth following up. Even if this statistical effect has a simple (e.g., technical or psychological) explanation, it might be possible that it can be used for statistical arbitrage strategies.

12. Branching-random-walk theory might be helpful for the discussion of critical or scaling limits of Hawkes processes. In the spirit of Paper B, one can treat the branching part and the displacements of the underlying trees separately. Then, for the branching part, there is a lot of theory for Galton-Watson trees. Given the tree, for the displacements one can apply standard renewal theory.

Paper

A

Matthias Kirchner.

Hawkes and INAR(∞) processes.

Stochastic Processes and their Applications,
162(8):2494–2525, 2016.

Hawkes and INAR(∞) processes

Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

In this paper, we discuss integer-valued autoregressive time series (INAR), Hawkes point processes, and their interrelationship. Besides presenting structural analogies, we derive a convergence theorem. More specifically, we generalize the well-known INAR(p), $p \in \mathbb{N}$, time series model to a corresponding model of infinite order: the INAR(∞) model. We establish existence, uniqueness, finiteness of moments, and give formulas for the autocovariance function as well as for the joint moment-generating function. Furthermore, we derive a branching-process as well as an AR(∞) and an MA(∞) representation for the model. We compare Hawkes process properties with their INAR(∞) counterparts. Given a Hawkes process N, in the main theorem of the paper we construct an INAR(∞)-based family of point processes and prove its convergence to N. This connection between INAR and Hawkes models will be relevant in applications.

Introduction

In this paper, we show that Hawkes point processes are continuous-time versions of integer-valued autoregressive time series and—vice versa—that integer-valued autoregressive time series are discrete-time versions of Hawkes point processes; see Theorem 17 for the main result of the paper. To start with, we outline the history of the concepts involved.

Standard time series theory for sequences of real-valued data points has been developed in seminal works like Whittle (1951) and Box and Jenkins (1970). This theory led to the natural question of time series models for count data. In the count-data context, the starting point is also a defining system of difference equations of the form "$X_n - \sum \alpha_k X_{n-k} = \varepsilon_n + \sum \beta_k \varepsilon_{n-k}$, $n \in \mathbb{Z}$". The main idea of the construction is to manipulate these equations in such a way that their solutions are integer-valued. This can be achieved by giving the error terms "(ε_n)" a distribution supported on $\mathbb{N}_0$ and by substituting all multiplications with thinning operations. In

the above spirit, autoregressive integer-valued (INAR) time series were defined and examined by McKenzie (1985) and Al-Osh and Alzaid (1987). The modern definition of the INAR model comes from Du and Li (1991). Latour (1997) generalizes the model to the multivariate case. For an exhaustive collection of properties of the INAR model; see Marques da Silva (2005). For a textbook reference; see Fokianos and Kedem (2012).

The Hawkes process was introduced in Hawkes (1971b,a) as a model for contagious processes such as measles infections or hijackings. As a point process in continuous time, the Hawkes process allows for the modeling of intensities which depend on the past of the process itself. Its alternative name, "self-exciting point process", stems from the fact that, given the occurrence of an event, intensity jumps upwards and then decays gradually. Theoretical cornerstones for the model are Hawkes and Oakes (1974) which establishes the representation as a cluster process, Ozaki (1979) which covers MLE estimation, Brémaud and Massoulié (1996) which extends the original model by generalizing the affine dependence on the past to Lipschitz dependence, Brémaud and Massoulié (2001) which proves the existence of a specific borderline case of the model, and Liniger (2009) which puts the subtleties of the definition and the construction on a solid mathematical foundation—especially for the marked multivariate case. For a textbook reference that covers many aspects of the Hawkes process; see Daley and Vere-Jones (2009).

To the best of our knowledge, the close connection between INAR and Hawkes processes has not been studied before. The correspondence between the model classes becomes even more direct if one applies infinite autoregression instead of finite autoregression for the time series model. This was our main motivation for generalizing the existing INAR(p) framework with $p < \infty$ to the case $p = \infty$. For the new INAR(∞) model, we give an explicit construction and show uniqueness; see Theorem 3. Then we derive three alternative descriptions of the process, namely a branching-process, an autoregressive, and a moving-average representation. Furthermore, we calculate basic quantities such as the joint moment-generating function and the autocovariance function. These are mainly presented for comparison with their Hawkes process counterparts. We observe that the equivalent branching-construction of INAR and Hawkes models form the core of the connection. This equivalence yields corresponding equations for generating functions, similar moment structures, and analogous stability criteria. Theorem 17 establishes a convergence result. In this theorem, for a given Hawkes process N, we construct a specific family of INAR(∞) sequences $\left\{\left(X_n^{(\Delta)}\right)_{n\in\mathbb{Z}}\right\}_{\Delta>0}$. From each member of this family, we derive a point process $N^{(\Delta)}$ by setting

$$N^{(\Delta)}\big((a,b]\big) := \sum_{n\colon \Delta n\in(a,b]} X_n^{(\Delta)}, \quad a < b.$$

The theorem states that $N^{(\Delta)}$ converges weakly to the Hawkes process N when Δ goes to zero.

This result is relevant for applications of INAR and Hawkes processes. In particular, the convergence theorem yields an estimation method for the Hawkes process by estimating the more tractable approximating INAR model instead. We work out this estimation method in Kirchner (2017a). Moreover, from a purely theoretical point of view, the presented line of thought is useful: the time series perspective on point processes as well as the point process perspective on (integer-valued) time series can be fertile for constructing and understanding event-data models; see Section 4.

The paper is organized as follows: Section 1 introduces the INAR(∞) model. Section 2 presents the Hawkes process. Section 3 establishes the convergence theorem. Furthermore, it collects structural analogies between the two model classes. In the final section, we discuss the broader interpretation of the INAR–Hawkes relation.

1 The INAR(∞) model

Throughout the paper, we consider a basic probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying all random variables involved.

1.1 Definition and existence

Definition 1. *For an $\mathbb{N}_0$-valued random variable Y and a constant $\alpha \geqslant 0$, the* reproduction operator $\circ$ *is defined by*

$$\alpha \circ Y := \sum_{n=1}^{Y} \xi_n^{(\alpha)},$$

where $\xi_n^{(\alpha)} \overset{\text{iid}}{\sim} \text{Pois}(\alpha)$, $n \in \mathbb{N}$, independently of Y. We refer to $\xi_n^{(\alpha)}$, $n \in \mathbb{N}$, as offspring variables *and to $\left(\xi_n^{(\alpha)}\right)$ as* offspring sequence.

Here and throughout the paper, we use the convention that $\sum_{n=p}^{q} a_n := 0$, $q < p$, for any sequence $(a_n)_{n\in\mathbb{Z}} \subset \mathbb{R}$.

Definition 2. *For $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, let $\varepsilon_n \overset{\text{iid}}{\sim} \text{Pois}(\alpha_0)$, $n \in \mathbb{Z}$, and $\xi_l^{(n,k)} \sim \text{Pois}(\alpha_k)$, independently over $n \in \mathbb{Z}, k \in \mathbb{N}$, $l \in \mathbb{N}$, and also independent of (ε_n). An* integer-valued autoregressive time series of infinite order *(INAR(∞)) is a sequence of random variables $(X_n)_{n\in\mathbb{Z}}$ which is a solution to the system of stochastic difference equations*

$$\varepsilon_n = X_n - \sum_{k=1}^{\infty} \alpha_k \circ X_{n-k} \tag{1}$$

$$:= X_n - \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \xi_l^{(n,k)}, \quad n \in \mathbb{Z}. \tag{2}$$

We call α_0 immigration parameter, (ε_n) immigration sequence, $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, reproduction coefficients, *and* $K := \sum_{k=1}^{\infty} \alpha_k$ reproduction mean.

In most situations, it is enough to use the reproduction notation from (1) in Definition 2 without explicitely writing out the offspring sequences as in (2)—keeping in mind that each "$\circ$" operates independently over $k \in \mathbb{N}$ and $n \in \mathbb{Z}$. Clearly, Definitions 1 and 2 depend on the choice of the distribution of the offspring variables. A more obvious option would have been sequences of Bernoulli variables. This would yield the binomial thinning operator from Steutel and van Harn (1979) which has in fact been the choice in the cited INAR(p) literature. The Poisson choice for the offspring sequences, however, again yields a Poisson distribution for "$X_n|X_{n-1}, X_{n-2}, \ldots$". This in turn leads to formulas that are simpler and that can be compared with their Hawkes counterparts more directly. We will address this issue in Sections 3.4 and 4. For the following existence theorem, any $\mathbb{N}_0$-valued distribution with finite first moments would do for the offspring variables. Throughout our paper, "stationary" is understood as "strictly stationary".

Theorem 3. *Let* $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, *with reproduction mean* $K := \sum_{k=1}^{\infty} \alpha_k < 1$. *Then* (2) *has an almost surely unique stationary solution* $(X_n)_{n\in\mathbb{Z}}$, *where* $X_n \in \mathbb{N}_0$, $n \in \mathbb{Z}$, *and* $\mathbb{E}\, X_n \equiv \alpha_0/(1-K)$.

Proof. See A.1. □

1.2 Branching structure

We highlight the branching nature of the solution to (2). As we will see, one can interpret this solution as a model for the size of a population, where each individual is alive exactly during one time-step. Each individual is either an immigrant or stems from a prior individual. This is similar to a Galton–Watson framework with immigration; see Section 5 in Seneta (1969). In contrast to the Galton–Watson setup, each INAR(∞) individual does not only have offspring at the next time-step but (potentially) at any future time. The proof of Theorem 3 formalizes this structure and then establishes that the construction indeed yields a process with the desired properties. The next proposition summarizes the branching representation of the INAR(∞) process. We emphasize the branching intuition of family processes consisting of generation processes by the unusual but suggestive notation for stochastic processes (F_n) and (G_n). The branching formulation will be useful for the derivation of the moment-generating function. It furthermore summarizes the most elegant and, at the same time, efficient way for simulating from the INAR(∞) model:

Proposition 4. *Let* (X_n) *be an INAR(*∞*) sequence with respect to an immigration parameter* $\alpha_0 > 0$ *and reproduction coefficients* $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, *so that* $\sum_{k=1}^{\infty} \alpha_k < 1$*; see Definition 2.*

Then

$$X_n \overset{\mathrm{d}}{=} \sum_{i\in\mathbb{Z}} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)}, \quad n \in \mathbb{Z}, \tag{3}$$

where $\varepsilon_i \overset{iid}{\sim} \mathrm{Pois}(\alpha_0), i \in \mathbb{Z}$, *and* $\left(F_n^{(i,j)}\right)$ *are independent (over* $i \in \mathbb{Z}$ *and* $j \in \mathbb{N}$*) copies of a branching process* (F_n) *defined by*

$$F_n := \sum_{g=0}^{\infty} G_n^{(g)}, \quad n \in \mathbb{Z}\,. \tag{4}$$

The generations (G_n) *in* (4) *are constructed recursively by*

$$G_n^{(0)} := 1_{\{n=0\}} \quad \textit{and} \quad G_n^{(g)} := \sum_{k=1}^{n} \alpha_k \circ G_{n-k}^{(g-1)} := \sum_{k=1}^{n} \sum_{m=1}^{G_{n-k}^{(g-1)}} \xi_m^{(n,k,g)}, \quad n \in \mathbb{Z}, \tag{5}$$

with $\xi_m^{(n,k,g)} \sim Pois(\alpha_k)$ *independently over* m, n, k, g*—and also independent of* $(\varepsilon_i)_{i\in\mathbb{Z}}$*. Furthermore, we have the following distributional equality for the generic family-process* (F_n) *from* (4)*:*

$$(F_n)_{n\in\mathbb{Z}} \overset{\mathrm{d}}{=} \left(1_{\{n=0\}} + \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} F_{n-i}^{(i,j)}\right)_{n\in\mathbb{Z}}. \tag{6}$$

Proof. See A.2. □

1.3 Moment generating function

From the representation of the INAR(∞) sequence as a superposition of shifted i.i.d. family processes in Proposition 4 above, we derive equations for the joint moment-generating function of the model. First, we fix some notation:

Definition 5. *For any sequence* $(t_n)_{n\in\mathbb{N}_0} \subset \mathbb{R}$*, let* $\mathrm{supp}\big((t_n)\big) := \sup\{n \in \mathbb{N}_0 : t_n > 0, k > n\}$ *be the* support of the sequence (t_n)*. Furthermore, for* $A \subset \mathbb{R}$*, let* $c_{00}(A) := \big\{(t_n)_{n\in\mathbb{N}_0} \subset A : \mathrm{supp}\big((t_n)\big) < \infty\big\}$*, the space of sequences in A with a finite number of nonzero values. For any time series* $(Y_n)_{n\in\mathbb{N}_0}$*, we define the* joint moment-generating function

$$M_{(Y_n)}\big((t_n)_{n\in\mathbb{N}_0}\big) := \mathbb{E}\exp\left\{\sum_{n=0}^{\infty} t_n Y_n\right\}, \quad (t_n) \in c_{00}\left(\mathbb{R}\right). \tag{7}$$

The somewhat unusual definitions above have been chosen for most direct comparison between the INAR(∞) joint moment-generating function and the Laplace functional of a Hawkes process; see Propositions 14 and 18.

Proposition 6. *Let (X_n) be an INAR(∞) sequence with respect to immigration parameter $\alpha_0 \geqslant 0$ and reproduction coefficients $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, such that $K = \sum_{k=1}^{\infty} \alpha_k < 1$; see Definition 2. Then there exists a constant $\delta > 0$ such that*

$$M_{(X_n)}\big((t_n)\big) \leqslant \exp\left\{ d\alpha_0 \frac{1+K}{2K} \right\} < \infty, \quad (t_n) \in c_{00}\big((-\infty, \delta]\big), \tag{8}$$

where $d := \operatorname{supp}\big((t_n)\big) + 1$ is the maximal number of nonzero values of (t_n). Furthermore, we have that

$$M_{(X_n)}\big((t_n)\big) = \exp\left\{ \alpha_0 \sum_{i\in\mathbb{Z}} \left(M_{(F_n)}\Big((t_{i+n})_{n\in\mathbb{N}_0}\Big) - 1 \right) \right\}, \tag{9}$$

where we set $t_m := 0$ for $m \leqslant 0$. Here, (F_n) denotes the generic family-process from Proposition 4. Its joint moment-generating function $M_{(F_n)}$ is the unique solution to

$$M_{(F_n)}\big((s_n)\big) = \mathrm{e}^{s_0} \exp\left\{ \sum_{k=1}^{\infty} \alpha_k \Big(M_{(F_n)}\big((s_{n+k})_{n\in\mathbb{N}}\big) - 1 \Big) \right\}, \quad (s_n) \in c_{00}((-\infty, \delta]). \tag{10}$$

Proof. See A.3. □

As a matter of fact, every moment-generating function that is finite in a neighborhood of zero has a Taylor series about zero. The coefficients of this series are the joint moments. Consequently, from Proposition 6, we obtain

Corollary 7. *Let (X_n) be an INAR(∞) sequence as in Theorem 3. For $m \in \mathbb{N}$ and $k_1, \ldots, k_m \in \mathbb{N}_0$, we have that*

$$\mathbb{E}\left[X_1^{k_1} X_2^{k_2} \cdots X_m^{k_m} \right] < \infty.$$

1.4 ARMA representations

We represent the INAR(∞) model as both, an autoregressive as well as a moving-average time series. These explicit representations allow the application of standard linear time-series theory on the INAR(∞) model.

Proposition 8. *Let $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, with $K = \sum_{k=1}^{\infty} \alpha_k < 1$, and let (X_n) be the corresponding INAR(∞) process; see Definition 2. Then*

$$u_n := X_n - \sum_{k=1}^{\infty} \alpha_k X_{n-k} - \alpha_0, \quad n \in \mathbb{Z}, \tag{11}$$

defines a stationary sequence (u_n) *with* $\mathbb{E}\, u_n \equiv 0$, $n \in \mathbb{Z}$, *and*

$$\mathbb{E}\left[u_n u_{n'}\right] = \begin{cases} 0, & n \neq n', \\ \frac{\alpha_0}{1-K}, & n = n'. \end{cases} \tag{12}$$

Furthermore, we have that

$$(X_n - \mu_X) - \sum_{k=1}^{\infty} \alpha_k \left(X_{n-k} - \mu_X\right) = u_n, \quad n \in \mathbb{Z}, \tag{13}$$

where $\mu_X := \mathbb{E}\, X_0 = \alpha_0/(1-K)$. *In other words,* (u_n) *is a (dependent) white-noise sequence and the time series* $(X_n - \mu_X)_{n\in\mathbb{Z}}$ *can be described in terms of a solution to an ordinary AR(∞) system of difference equations.*

Proof. See Appendix A.4 □

From Proposition 8, we find the Wold decomposition of the INAR(∞), that is, a representation as a standard MA(∞) time series. It will be most helpful for establishing the second-order properties of the process.

Proposition 9. *Let* $\alpha_k \geqslant 0$, $k \in \mathbb{N}_0$, *with* $\sum_{k=1}^{\infty} \alpha_k < 1$. *Then the corresponding INAR(∞) process from Definition 2 is a solution to the family of equations*

$$X_n - \mu_X = \sum_{k=0}^{\infty} \beta_k u_{n-k}, \quad n \in \mathbb{Z}, \tag{14}$$

where (u_n) *is the white-noise sequence from Proposition 8,* $\mu_X := \mathbb{E}\, X_0 = \alpha_0/(1-K)$, $\beta_0 := 1$, *and* $\beta_k := \sum_{i=1}^{k} \alpha_i \beta_{k-i}$, $k \in \mathbb{N}_0$. *Furthermore,* $\beta_k \geqslant 0$, $k \in \mathbb{N}_0$, *and* $\sum_{k=0}^{\infty} \beta_k = 1/(1-K) < \infty$.

Proof. See A.5. □

Note that the moving-average coefficients $(\beta_k)_{k\in\mathbb{N}_0}$ defined in Proposition 9 above correspond to $(\mathbb{E}\, F_k)_{k\in\mathbb{N}_0}$, the expected values of the family-process (4) after $k = 0, 1, 2, \ldots$ time steps. Also note that Proposition 9 is not a corollary of the standard result stating that ARMA(p,q) processes ($p, q < \infty$) are MA(∞) processes; for example, see Theorem 3.1.1 of Brockwell and Davis (1991). The argumentation for the case $p = \infty$ has to be more subtle: we have to prove that the (in general infinitely many) zeros of the involved power series can be bounded away from the unit circle.

1.5 Second order properties

The MA(∞) representation makes the second moments of the INAR(∞) sequence particularly tractable:

Proposition 10. *Let (X_n) be an INAR(∞) process with reproduction coefficients $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, such that $K := \sum_{k=1}^{\infty} \alpha_k < 1$, and immigration parameter $\alpha_0 \geqslant 0$. Furthermore, let $R(j) := \operatorname{Cov}(X_n, X_{n+j})$, $j \in \mathbb{Z}$, be the autocovariance function of the (stationary) sequence. Then*

$$R(j) = \frac{\alpha_0}{1-K} \sum_{k=0}^{\infty} \beta_k \beta_{k+|j|} \geqslant 0, \quad j \in \mathbb{Z}, \tag{15}$$

where $\beta_0 := 1$ and $\beta_k := \sum_{i=1}^{k} \alpha_i \beta_{k-i}$. In addition, we have that

$$\sum_{j=0}^{\infty} R(j) \leqslant \frac{\alpha_0}{(1-K)^3} < \infty. \tag{16}$$

Proof. See A.6. □

2 The Hawkes process

After the first section on the new INAR(∞) model, the following shorter section formally presents the well-known Hawkes point process. We treat point processes as random counting-measures and only consider point processes on $\mathbb{R}$. First, we fix some general notation and terminology. Then we recall the definition, the existence theorem, and selected properties of the (univariate) Hawkes process. For the general theory, we mainly follow Resnick (1987), Chapter 3. For the Hawkes part, our main references are the seminal papers Hawkes (1971b,a) and Hawkes and Oakes (1974).

2.1 Preliminaries

Let $\mathcal{B} := \mathcal{B}(\mathbb{R})$ be the Borel-sets in $\mathbb{R}$ and $\mathcal{B}_b := \{B \in \mathcal{B}(\mathbb{R}) : B \text{ bounded}\}$. A measure m on $\mathbb{R}$ is a *point measure* if $m(B) \in \mathbb{N}_0$, $B \in \mathcal{B}_b$. We denote the space of point measures on $\mathbb{R}$ by $M_p := M_p(\mathbb{R})$. Let $C_K^+ := C_K^+(\mathbb{R})$ be the space of nonnegative continuous functions on $\mathbb{R}$ with compact support. Point measures (m_n) *converge vaguely* to a point measure m if $\lim_{n\to\infty} \int f(t) m_n(\mathrm{d}t) \to \int f(t) m(\mathrm{d}t)$, $f \in C_K^+(\mathbb{R})$; we write $m_n \xrightarrow{\mathrm{v}} m$. Vague convergence yields the *vague topology* on M_p. The Borel σ-algebra generated by this topology, $\mathcal{M}_p := \mathcal{B}(M_p)$, coincides with the σ-algebra generated by the sets $\{m \in M_p : m(A) = k\}$, $A \in \mathcal{B}_b, k \in \mathbb{N}_0$; see Lemma 1.4. in Kallenberg (1983). Any measurable mapping $\Phi : (M_p, \mathcal{M}_p) \to (\mathbb{R}, \mathcal{B})$ such that $\lim_{n\to\infty} \Phi(m_n) = \Phi(m)$ whenever $m_n \xrightarrow{\mathrm{v}} m$, is *continuous with respect to the vague topology*. Our basic underlying probability space is $(\Omega, \mathcal{F}, \mathbb{P})$. A measurable mapping $N : (\Omega, \mathcal{F}) \to (M_p, \mathcal{M}_p)$, $\omega \mapsto N(\omega)$ is called *point process*. The *history* of a point process N is the filtration

$(\mathcal{H}_t^N)$, where, for $t \in \mathbb{R}$,

$$\mathcal{H}_t^N := \sigma\bigg(\Big\{\omega \in \Omega :\ N(\omega)\big((a,b]\big) \in B\Big\} :\ -\infty < a < b \leqslant t, B \subset \mathbb{N}_0\bigg). \tag{17}$$

We assume that $\mathcal{H}_t^N \subset \mathcal{F}$, $t \in \mathbb{R}$. Note that our definition of a point process allows multiple points, i.e., we may have that "$\mathbb{P}\left[N(\{t\}) > 1 | N(\{t\}) > 0\right] > 0$". Also note that, for $t \in \mathbb{R}$, the σ-algebra $\mathcal{H}_t^N$ includes all sets of the form

$$\Big\{\omega :\ N(\omega)\big(\{t_n\}\big) = k_n,\ N(\omega)\big((t_{n-1}, t_n)\big) = 0,\ n = 0, -1, -2, \dots\Big\}\Big\} \tag{18}$$

with $t =: t_0 \geqslant t_{-1} \geqslant \dots$ and $k_0, k_{-1}, \dots \in \mathbb{N}$.

2.2 Definition and existence

Definition 11. *For any point measure $m \in M_p$, define the* Hawkes intensity

$$\lambda(t|m) := \eta + \int_{\mathbb{R}} h(t-s)m(\mathrm{d}s), \quad t \in \mathbb{R},$$

where $\eta > 0$ is a constant and $h : \mathbb{R} \to \mathbb{R}_0^+$ is a nonnegative measurable function with $h(t) = 0$, $t \leqslant 0$. We refer to η as immigration intensity *and to h as* reproduction intensity.

The immigration intensity is often called *baseline intensity* and the reproduction intensity is often called *excitement function*. However, our objective is to highlight the similarity between Hawkes and INAR processes. Consequently, we make use of a joint branching-process terminology; see Definition 2.

Definition 12. *Let λ be a Hawkes intensity as in Definition 11. A* Hawkes process *is a point process N that is a solution to the family of equations*

$$\mathbb{E}\left[1_A N\Big(\big(a,b\big]\Big)\right] = \mathbb{E}\left[1_A \int_a^b \lambda(s|N)\mathrm{d}s\right], \quad a < b,\ A \in \mathcal{H}_a^N. \tag{19}$$

A priori, it is not clear whether this family of equations has a solution, whether any possible solution would be unique (in a distributional sense), and whether the distribution of a solution would be stationary. These questions are answered by the following proposition; see Hawkes and Oakes (1974):

Proposition 13. *Let λ be a Hawkes intensity with immigration intensity $\eta > 0$ and reproduction intensity h such that $\int_0^\infty h(t)\mathrm{d}t < 1$. Then there is precisely one stationary process that satisfies* (19).

The existence and uniqueness result above is established by the observation that the solution to (19) must be a cluster process or—more specifically—a branching process with immigration: the points are interpreted as individuals that are either immigrants or offspring. The immigrants (or cluster centers) stem from a homogeneous Poisson process with intensity η. These immigrants form generation zero of the following branching procedure: an immigrant at time $s \in \mathbb{R}$ triggers an inhomogeneous Poisson process with intensity $h(\cdot - s)$ where h is the reproduction intensity of the process as in Definition 11. These offspring individuals form generation one. Each of these first-generation individuals again triggers an inhomogeneous Poisson process in a similar way, etc., so that the families (or clusters) are generated by cascades of inhomogeneous Poisson processes.

2.3 The Laplace functional

The cluster and branching process point of view is also fertile beyond the results of Proposition 13. For example, it leads to equations for the *Laplace functional* of a Hawkes process. The Laplace functional Ψ_N of a point process N is a functional defined on the space of nonnegative measurable functions with compact support by

$$\Psi_N[f] := \mathbb{E} \exp\left\{ - \int_{-\infty}^{\infty} f(t) N(\mathrm{d}t) \right\}.$$

The next proposition is Theorem 2 in Hawkes and Oakes (1974)—with a slight modification as the original statement refers to the probability generating functional whereas we prefer the nowadays more common Laplace functional notion:

Proposition 14. *Let N be a Hawkes process with immigration intensity $\eta > 0$ and reproduction intensity h as in Proposition 13. Then the Laplace functional Ψ_N of N is*

$$\Psi_N[f] = \exp\left\{ \eta \int_{-\infty}^{\infty} \Big(\Psi_F\big[f(t + \cdot) \big] - 1 \Big) \, \mathrm{d}t \right\},$$

where Ψ_F is a functional that is the unique solution to

$$\Psi_F[f] = e^{-f(0)} \exp\left\{ \int_0^{\infty} \Big(\Psi_F\big[f(t + \cdot) \big] - 1 \Big) h(t) \, \mathrm{d}t \right\}.$$

In both equalities, f denotes an arbitrary measurable, nonnegative function with compact support.

3 Links between INAR(∞) and Hawkes processes

In the following section, we first explain how discrete-time INAR(∞) processes can approximate a continuous-time Hawkes process. After the convergence theorem, we establish a number of properties of the approximating sequence. Finally, we collect some structural analogies of the two models.

3.1 Preliminaries

Let Y_n, $n \in \mathbb{N}$, and Y be random variables with values in some topological space. The sequence (Y_n) *converges weakly* to Y if $\lim_{n\to\infty} \mathbb{E}\,\varphi(Y_n) = \mathbb{E}\,\varphi(Y)$ for all nonnegative continuous bounded functions φ. We define weak convergence of point processes with respect to the vague topology $\mathcal{M}_p$ on M_p; see Section 2.1. In this case, weak convergence of point processes is equivalent to convergence of their finite-dimensional distributions; see Daley and Vere-Jones (2009), Theorem 11.1.VII. General weak convergence theory, as developed in the monograph Billingsley (1968), considers sequences in metric spaces. Therefore it is important to note that *the vague topology is metrizable*; see Resnick (1987), Proposition 3.17. In other words, we may treat $(M_p, \mathcal{M}_p)$ as a metric space where necessary. A most helpful theorem in the weak-convergence context is the *continuous mapping theorem*; see Theorem 5.1 in Billingsley (1968). We apply it in the following form:

Proposition 15. *Let (N_n) and N be point processes such that $N_n \xrightarrow{\mathrm{w}} N$, $n \to \infty$. Furthermore, let $f : \mathbb{R} \to \mathbb{R}_0^+$ be a bounded, measurable function with compact support and with a set of discontinuities $D_f \in \mathcal{B}$ such that $\mathbb{P}[N(D_f) > 0] = 0$. Then, $\int f(t)N_n(\mathrm{d}t) \xrightarrow{\mathrm{w}} \int f(t)N(\mathrm{d}t)$ for $n \to \infty$.*

3.2 The convergence theorem

Next to the conditions on the reproduction intensity h from Definition 12 and Proposition 13, we introduce an additional assumption: we want h piecewise continuous. We say a function $f : \mathbb{R} \to \mathbb{R}$ is *piecewise continuous* if its set of discontinuities $D_f \subset \mathbb{R}$ is finite and for all $t_0 \in D_f$ the limits $\lim_{t\to t_0^-} f(t)$ and $\lim_{t\to t_0^+} f(t)$ exist and are finite. Combining all assumptions on h yields the following important technical

Lemma 16. *Let $h : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ be piecewise continuous function with $h(t) = 0$, $t \leqslant 0$, and $\int h(t)\,\mathrm{d}t < 1$. Then there exist constants $\delta > 0$ and $\tilde{K} < 1$ such that, for any $\Delta \in (0, \delta)$,*

$$K^{(\Delta)} := \Delta \sum_{k=1}^{\infty} h(k\Delta) \leqslant \tilde{K} < 1. \tag{20}$$

In the sequel of the section, let $\delta > 0$, $K^{(\Delta)} \leqslant \tilde{K} < 1$, and $\Delta \in (0, \delta)$, be as in the lemma above. We state the main mathematical result of our paper:

Theorem 17. *Let N be a Hawkes process with immigration intensity η and reproduction intensity h as in Lemma 16. For $\Delta \in (0, \delta)$, let $\left(X_n^{(\Delta)}\right)$ be an INAR(∞) sequence with immigration parameter $\Delta\eta$ and reproduction coefficients $\Delta h(k\Delta)$, $k \in \mathbb{N}$. From the sequences $\left\{\left(X_n^{(\Delta)}\right)\right\}_{\Delta\in(0,\delta)}$, we define a family of point processes by*

$$N^{(\Delta)}(A) := \sum_{k:k\Delta\in A} X_k^{(\Delta)}, \quad A \in \mathcal{B},\ \Delta \in (0, \delta). \tag{21}$$

Then, we have that

$$N^{(\Delta)} \xrightarrow{\text{w}} N \quad \textit{for} \quad \Delta \to 0.$$

Proof. See A.11. □

Our proof uses the standard weak-convergence approach—as followed in the Hawkes context, e.g., by Brémaud and Massoulié (2001). First, tightness of the approximating family is established. By Prohorov's theorem, tightness yields weak subsequential limits for all subsequences. Then we show that all those potential weak subsequential limits have the same distribution as the Hawkes process. This will establish the result. An alternative approach would be convergence of Laplace functionals that are given by

Proposition 18. *For some $\Delta \in (0, \delta)$, let $N^{(\Delta)}$ be as in Theorem 17. Let f be a nonnegative measurable function with compact support. Then the Laplace functional of $N^{(\Delta)}$ evaluated at f is*

$$\Psi_{N^{(\Delta)}}[f] = \exp\left\{\Delta\eta \sum_{i\in\mathbb{Z}} \left(\Psi_{F^{(\Delta)}}^{(\Delta)}\left(\left(f((i+n)\Delta)\right)_{n\in\mathbb{N}_0}\right) - 1\right)\right\}.$$

Here, the function $\Psi_{F^{(\Delta)}}^{(\Delta)}$ operating on sequences $(s_n)_{n\in\mathbb{N}_0} \in c_{00}([0,\infty))$ is a solution to

$$\Psi_{F^{(\Delta)}}^{(\Delta)}\left((s_n)_{n\in\mathbb{N}_0}\right) = e^{-s_0} \exp\left\{\sum_{k=1}^{\infty} \Delta h(k\Delta)\left(\Psi_{F^{(\Delta)}}^{(\Delta)}\left((s_{k+n})_{n\in\mathbb{N}_0}\right) - 1\right)\right\}.$$

Proof. See Appendix A.7. □

The similarities between the formulas in Proposition 18 above and the corresponding equations for the Hawkes process in Proposition 14 are striking. Still, rather than establishing the convergence result from Theorem 17 via the Laplace functionals, we choose a more direct reasoning on the process level that contains useful information on the approximating point process family as a by-product. The information necessary for the convergence proof is collected in the following lemmas:

Lemma 19. *For any $\Delta \in (0,\delta)$, let $N^{(\Delta)}$ be a point process as in Theorem 17. Then, for $A \in \mathcal{B}$, we have that*

$$A \cap \{k\Delta : k \in \mathbb{Z}\} = \varnothing \quad \Rightarrow \quad N^{(\Delta)}\big(A\big) = 0, \text{ almost surely.} \tag{22}$$

For the expectation, we find that

$$\mathbb{E}\, N^{(\Delta)}\big(\{k\Delta\}\big) = \Delta\frac{\eta}{1-K^{(\Delta)}} < \delta\frac{\eta}{1-\tilde{K}}, \quad k \in \mathbb{Z}, \tag{23}$$

and, for $a < b$,

$$\mathbb{E}\, N^{(\Delta)}\big([a,b]\big) < (b-a+2\delta)\frac{\eta}{1-\tilde{K}} < \infty. \tag{24}$$

Proof. See Appendix A.8. □

Lemma 20. *For $\big(N^{(\Delta)}\big)_{\Delta\in(0,\delta)}$, the approximating family of point processes from Theorem 17, we have that*

$$\sup_{\Delta\in(0,\delta)} \operatorname{Var}\big(N^{(\Delta)}(A)\big) < \infty, \quad A \in \mathcal{B}_b.$$

Proof. See Appendix A.9 □

A family of random variables $(Y_i)_{i\in I}$ is *uniformly integrable* if $\lim_{M\to\infty}\sup_{i\in I}\mathbb{E}\left[1_{|Y_i|>M}|Y_i|\right] = 0$. We obtain uniform integrability of the random variables in question as a corollary from Lemma 20 above:

Lemma 21. *Let $\big(N^{(\Delta)}\big)_{\Delta\in(0,\delta)}$ be the approximating family of point processes from Theorem 17 and $A \in \mathcal{B}_b$. Then we have that the family of random variables $\big(N^{(\Delta)}(A)\big)_{\Delta\in(0,\delta)}$ is uniformly integrable.*

A family of probability measures $\big(\mathbb{P}^{(i)}\big)_{i\in I}$ on $(M_p, \mathcal{M}_p)$ is *uniformly tight* if, for all $\varepsilon > 0$, there exists a compact set $K \subset M_p$ such that $\mathbb{P}^{(i)}[K^c] < \varepsilon$, $i \in I$.

Lemma 22. *The family of the probability measures $\big(\mathbb{P}^{(\Delta)}\big)_{0<\Delta<\delta}$ on $(M_p, \sigma(\mathcal{M}_p))$ corresponding to the random point processes $\big(N^{(\Delta)}\big)_{0<\Delta<\delta}$ is uniformly tight.*

Proof. See Appendix A.10. □

3.3 Structural analogies

Besides the formal convergence result from Theorem 17, we point out a number of structural parallels between the Hawkes and the INAR(∞) model. The branching structure of the INAR(∞) model described after Theorem 3 is the same as the branching structure of the Hawkes

process described after Proposition 13. This similar underlying structure yields analogous equations for the moment-generating function of the INAR(∞) and the Laplace functional of the Hawkes process; see Proposition 6 and 14. Consequently, we can expect similar distributional properties. In the following, we compare the models more directly by considering a specific Hawkes process N together with its approximating family of INAR(∞) sequences, $\left(X^{(\Delta)}\right)$, $\Delta \in (0, \delta)$, obtained from Theorem 17.

i) The defining equations (2) and (19) have similar interpretations: taking expectations conditional on $\mathcal{H}_{n-1}^{X^{(\Delta)}} := \sigma\left(X_k^{(\Delta)} : k \leqslant n-1\right)$ on both sides of (2) yields

$$\frac{\mathbb{E}\left[X_n^{(\Delta)} | \mathcal{H}_{n-1}^{X^{(\Delta)}}\right]}{\Delta} = \eta + \sum_{k=-\infty}^{n-1} h\big((n-k)\Delta\big) X_k^{(\Delta)}, \quad n \in \mathbb{Z},$$

which is similar to the local version of (19)

$$\frac{\mathbb{E}\left[N(\mathrm{d}t) | \mathcal{H}_t^N\right]}{\mathrm{d}t} = \eta + \int_{-\infty}^{t} h(t-s) N(\mathrm{d}s), \quad t \in \mathbb{R}\,.$$

ii) The stability criteria from Theorems 3 and 13 correspond: for the time series case, $\sum_{k=1}^{\infty} \Delta h(\Delta k) < 1$ is a sufficient existence condition which in the Hawkes case becomes $\int_0^{\infty} h(t)\mathrm{d}t < 1$. Brémaud and Massoulié (2001) establishes the existence of a nontrivial Hawkes process, where the total weight of the reproduction intensity equals 1 and the immigration intensity is zero. The analogous statement for INAR(∞) sequences can be derived in a similar way.

iii) From the correspondence of the generating functions, we know that the moments of the INAR(∞) and the Hawkes process must be similar. The first moments are

$$\frac{\mathbb{E}\, X_n^{(\Delta)}}{\Delta} = \frac{\eta}{1 - K^{(\Delta)}}, \quad \text{respectively}, \quad \frac{\mathbb{E}\, N(\mathrm{d}t)}{\mathrm{d}t} = \frac{\eta}{1-K}.$$

For the first equality see Theorem 3; for the second equality see Hawkes (1971b).

iv) For the autocovariances of both models, i.e., for

$$R^{(\Delta)}(n) := \frac{\mathbb{E}\left[X_0^{(\Delta)} X_n^{(\Delta)}\right]}{\Delta^2} - \left(\frac{\mathbb{E}\, X_0^{(\Delta)}}{\Delta}\right)^2, \quad n \in \mathbb{Z},$$

respectively, for

$$r(t) := \frac{\mathbb{E}\left[\mathrm{d}N_0 \mathrm{d}N_t\right]}{(\mathrm{d}t)^2} - \left(\frac{\mathbb{E}\, N(\mathrm{d}t)}{\mathrm{d}t}\right)^2, \quad t \in \mathbb{R},$$

we find at the origin

$$R^{(\Delta)}(0) = \frac{1}{\Delta}\frac{\eta}{1-K^{(\Delta)}} + \sum_{k=1}^{\infty} \Delta h(k\Delta)R(k), \tag{25}$$

respectively,

$$r(0) = \frac{1}{\mathrm{d}t}\frac{\eta}{1-K} + \int_{0^+}^{\infty} h(s)r(s)\mathrm{d}s. \tag{26}$$

Implicit equations of Yule–Walker type are valid in both cases:

$$R^{(\Delta)}(n) = \sum_{k=1}^{\infty} h(\Delta k)R^{(\Delta)}(|n|-k)\Delta, \quad n \neq 0, \tag{27}$$

respectively,

$$r(t) = \int_{0}^{\infty} h(s)r(|t|-s)\mathrm{d}s, \quad t \neq 0. \tag{28}$$

Equations (25) and (27) are standard facts given the representation of the INAR(∞) sequence as a standard AR(∞) process in Proposition 8; (26) and (28) are derived in Hawkes (1971a).

3.4 The choice of the offspring distribution

As a last remark, we again refer to the choice of the offspring distribution in Definition 1. An obvious alternative to the Poisson distribution would have been the Bernoulli distribution; see the discussion after Definition 2. With the Bernoulli choice, each individual would have not more than one offspring at each future point in time instead of potentially unboundedly many. We want to indicate that in the limit (in the sense of Theorem 17 where all reproduction coefficients go to zero) this option would yield the same result: for $\Delta \in (0,1)$ and $(\alpha_k) \subset [0,1]$ such that $\sum \alpha_k < 1$, let $\xi_k^{(\Delta)} \overset{\text{iid}}{\sim} \text{Pois}(\Delta\alpha_k)$, $k \in \mathbb{N}$, and $\tilde{\xi}_k^{(\Delta)} \overset{\text{iid}}{\sim} \text{Bernoulli}(\Delta\alpha_k)$, $k \in \mathbb{N}$. Then one can easily show that

$$\lim_{\Delta\to 0^+} \frac{\mathbb{P}\left[\xi_k^{(\Delta)} = 0\right]}{\mathbb{P}\left[\tilde{\xi}_k^{(\Delta)} = 0\right]} = \lim_{\Delta\to 0^+} \frac{\mathbb{P}\left[\xi_k^{(\Delta)} = 1\right]}{\mathbb{P}\left[\tilde{\xi}_k^{(\Delta)} = 1\right]} = 1, \quad k \in \mathbb{N}.$$

So when Δ is very small, the offspring distribution candidates, Poisson and Bernoulli, become very similar. Roughly speaking, the limiting procedure in Theorem 17 is nothing else than a (complicated) superposition of limits of the form $\sum_{k=1}^{\infty} \xi_k^{(\Delta)}$. For these kinds of sums, we have that

$$\lim_{\Delta \to 0^+} \frac{\mathbb{P}\left[\sum_{k=1}^{\infty} \xi_k^{(\Delta)} = n\right]}{\mathbb{P}\left[\sum_{k=1}^{\infty} \tilde{\xi}_k^{(\Delta)} = n\right]} = 1, \quad n \in \mathbb{N}_0 .$$

If $\alpha_k > 0$ infinitely often, then—by the Poisson limit theorem—the two considered probabilities are even equal for all $\Delta \in [0, 1]$. In view of the above, one can expect that Theorem 17 with the Bernoulli offspring would yield the same limit as Poisson offspring, namely the Hawkes process.

4 Conclusion

The mathematical formulation of the correspondence between INAR and Hawkes processes in Theorem 17 has the following heuristic interpretation relevant for practical applications: let $\left(N^{(\Delta)}\right)$ be the approximating family of INAR(∞)-based point processes for a Hawkes process N as in Theorem 17. Then for $\Delta > 0$ (small), the finite-dimensional distributions of $N^{(\Delta)}$ are approximately equal to the finite-dimensional distributions of N. In particular, we find, for $n \in \mathbb{N}$,

$$\begin{aligned}
&\left(N\Big(\big(0, \Delta\big]\Big), N\Big(\big(\Delta, 2\Delta\big]\Big), \ldots, N\Big(\big((n-1)\Delta, n\Delta\big]\Big)\right) \\
&\overset{\mathrm{d}}{\approx} \left(N^{(\Delta)}\Big(\big(0, \Delta\big]\Big), N^{(\Delta)}\Big(\big(\Delta, 2\Delta\big]\Big), \ldots, N^{(\Delta)}\Big(\big((n-1)\Delta, n\Delta\big]\Big)\right) \\
&= \left(X_1^{(\Delta)}, X_2^{(\Delta)}, \ldots, X_n^{(\Delta)}\right), \qquad (29)
\end{aligned}$$

where $\left(X_n^{(\Delta)}\right)$ is the INAR(∞) sequence that the point process $N^{(\Delta)}$ is based on; see Theorem 17. So $\left(X_n^{(\Delta)}\right)$ is an approximative model for the bin-count sequences of the considered Hawkes process N. This point of view can be very fertile. For example, it leads to a nonparametric estimation procedure for the Hawkes process. Instead of fitting a Hawkes process directly, one fits the corresponding INAR(∞) model from Theorem 17 on the bin-counts for some small $\Delta > 0$. Kirchner (2017a) gives a detailed discussion of this estimation procedure including asymptotic properties, bias issues and techniques for an optimal bin-size choice Δ. Also note that the Hawkes bin-count sequence view in (29) on the INAR(∞) model is another argument for the choice of the Poisson instead of the Bernoulli distribution for the offspring sequences: clearly a Hawkes event can have potentially more than one direct offspring event in a future

time-interval.

For any $\mathbb{N}_0$-valued time series model, one can construct a point process model the way it is done in Theorem 17. So, studying integer-valued time series can be inspiring for developing and understanding point process models. For example, one might want to consider the corresponding point process of an integer-valued autoregressive moving-average (INARMA) time series. For INARMA time series, a moving-average part is added to the autoregressive part in the defining difference equations (2); see Fokianos and Kedem (2012). In fact, the corresponding point process is nothing else but the "dynamic contagion process" as proposed in Zhao (2012). Also for integer-valued time series theory, it is inspiring to translate point process models into the discrete-time setup. For example, one might want to translate the generalizing results on self-exciting point processes in Brémaud and Massoulié (1996) to the INAR context. Here, the self-excitement of the point process is modeled not as an affine but as a general Lipschitz function of the past of the process. The analogous generalization in time series theory is "nonlinear Poisson autoregression" as studied in Fokianos and Tjøstheim (2012) for the case $p = 1$. In the latter paper, the authors also find a Lipschitz condition for the transfer function. Yet another idea might be to consider marked INAR sequences in analogy to marked Hawkes processes; see Liniger (2009).

One can expect that the results of our paper also hold in a multivariate setup—with the obvious modifications. However, the notation would become even more tedious, so that we have decided to concentrate on the univariate case. In view of the many INAR/Hawkes-correspondences presented in this paper we conclude: INAR(∞) sequences are discrete-time versions of Hawkes processes and, vice versa, Hawkes processes are continuous-time versions of INAR(∞) sequences.

Acknowledgements

M.K. takes pleasure in thanking Thomas Mikosch for his most valuable comments on an earlier version of the paper and Paul Embrechts for guiding him to the topic of Hawkes processes and its applications to quantitative risk management. He also thanks an anonymous referee whose report improved some of the proofs as well as the final overall structure of the paper considerably. Furthermore, M.K. thanks Isabel Marquez da Silva for sharing her expertise on integer-valued time series models and Rita Kirchner as well as Anne MacKay for help with the editing. The author acknowledges financial support from RiskLab at the ETH Zurich and the Swiss Finance Institute.

A Proofs

A.1 Proof of Theorem 3.

Let $\varepsilon_i \overset{\text{iid}}{\sim} \text{Pois}(\alpha_0)$, $i \in \mathbb{Z}$, be the immigration terms from Definition 2. For the j-th potential immigrant at time $i \in \mathbb{Z}$, we define generation processes $\left(G_n^{(g,i,j)}\right)$, $g \in \mathbb{N}_0$, by the following recursive procedure:

$$G_n^{(0,i,j)} := 1_{\{n=0\}}, \quad n \in \mathbb{Z},\ i \in \mathbb{Z},\ j \in \mathbb{N} \tag{A.1}$$

$$G_n^{(g,i,j)} := \sum_{k=1}^{n} \alpha_k \circ G_{n-k}^{(g-1,i,j)}. \tag{A.2}$$

For all distributional properties of the construction, it will be enough to apply defining equation (A.2) for the generations. We will consider an explicit representation of the involved offspring sequences later. Note that $G_n^{(g,i,j)} = 0$ whenever $n < 0$. A family originating from the j-th individual immigrant at time i is the superposition of all the corresponding generation-processes:

$$F_n^{(i,j)} := \sum_{g=0}^{\infty} G_n^{(g,i,j)}, \quad n \in \mathbb{Z},\ i \in \mathbb{Z},\ j \in \mathbb{N}. \tag{A.3}$$

The candidate series $\left(\tilde{X}_n\right)$ for a solution of (2) is the superposition of all these families—modulo an appropriate shift in the time index:

$$\tilde{X}_n := \sum_{i=-\infty}^{n} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)}, \quad n \in \mathbb{Z}. \tag{A.4}$$

Note that only family and generation processes indexed by (i,j) with $j \in \{1,\ldots,\varepsilon_i\}$ come into play. Also note that all infinite series involved in the construction above are well-defined because their partial sums are nondecreasing. As a first step, we remind ourselves that $K = \sum_{k=1}^{\infty} \alpha_k$ and establish

$$\mathbb{E} \sum_{n=0}^{\infty} G_n^{(g,i,j)} = K^g, \quad g \in \mathbb{N}_0,\ i \in \mathbb{Z},\ j \in \mathbb{N}, \tag{A.5}$$

by induction: for $g = 0$, (A.5) is correct because we have $\mathbb{E}\sum_{n=0}^{\infty} G_n^{(0,i,j)} = \sum_{n=0}^{\infty} 1_{\{n=0\}} = 1 = K^0$. For $g > 0$, one can show that $\mathbb{E}\sum_{n=0}^{\infty} G_n^{(g,i,j)} = K \cdot \mathbb{E}\sum_{n=0}^{\infty} G_n^{(g-1,i,j)}$ and (A.5) follows. By (A.5), $\sum_{n=0}^{\infty} F_n^{(i,j)}$ has expectation $1/(1-K)$. In particular, $\sum_{n=0}^{\infty} F_n^{(i,j)}$ is almost surely finite. Since $F_n^{(i,j)}$ is a sum of integers, we have that $F_n^{(i,j)} \in \mathbb{N}_0$, $n \in \mathbb{Z}$, almost surely. For $\tilde{X}_n$, $n \in \mathbb{Z}$,

we find that

$$\mathbb{E}\,\tilde{X}_n = \sum_{i=-\infty}^{n} \mathbb{E} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)} = \alpha_0 \sum_{i=-\infty}^{n} \mathbb{E}\, F_{n-i}^{(i,1)} = \alpha_0\, \mathbb{E} \sum_{i=-\infty}^{n} F_{n-i}^{(1,1)} = \frac{\alpha_0}{1-K}. \tag{A.6}$$

Note that in the present work, nearly all random variables involved are nonnegative. Consequently, in most cases the interchange of summations as in the first equality above is covered by Tonelli's Theorem. Also Wald's Identity as applied in the second equality above will be applied without further notice. From (A.6), we get that $\tilde{X}_n$ is almost surely finite and therefore almost surely $\mathbb{N}_0$-valued. Note that, by construction, the generations $\left(G_n^{(i,j)}\right)$ and therefore the families $\left(F_n^{(i,j)}\right)$ are independently and identically distributed time series over $i \in \mathbb{Z}$ and $j \in \mathbb{N}$. Stationarity of $\left(\tilde{X}_n\right)$ then follows from the i.i.d. property of the immigration sequence (ε_i).

To show that our candidate sequence $\left(\tilde{X}_n\right)$ indeed solves (2), we have to work with an explicit representation for the reproductions involved in the (A.2). To that aim, let

$$\xi_{g,i,j,m}^{(n,k)} \sim \mathrm{Pois}(\alpha_k), \quad \text{independently over} \quad i,n \in \mathbb{Z} \text{ and } k,j,g,m \in \mathbb{N}\,. \tag{A.7}$$

In our branching terminology, $\xi_{g,i,j,m}^{(n,k)}$ denotes the number of offspring individuals at time n whose parent lived at time $n-k$. This parent belongs to the $(g-1)$-th generation of family (i,j). Furthermore, this parent is the m-th such individual. We repeat the defining recursion for the generation processes from above—this time we represent the involved offspring sequences explicitly:

$$G_n^{(0,i,j)} := 1_{\{n=0\}}, \quad n \in \mathbb{Z},\ i \in \mathbb{Z},\ j \in \mathbb{N} \tag{A.8}$$

$$G_n^{(g,i,j)} := \sum_{k=1}^{n} \alpha_k \circ G_{n-k}^{(g-1,i,j)} \tag{A.9}$$

$$:= \sum_{k=1}^{n} \sum_{m=1}^{G_{n-k}^{(g-1,i,j)}} \xi_{g,i,j,m}^{(i+n,k)}, \quad n \in \mathbb{Z},\ i \in \mathbb{Z},\ j \in \mathbb{N},\ g \in \mathbb{N}\,. \tag{A.10}$$

It is obvious that (A.10) justifies the distributional assumptions on (A.9) (i.e., (A.2)), used in the first part of this proof. For any $n \in \mathbb{Z}$, we find

$$\tilde{X}_n = \sum_{i=-\infty}^{n} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)} = \sum_{i=-\infty}^{n-1} \sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)} + \sum_{j=1}^{\varepsilon_n} F_0^{(n,j)} = \sum_{i=-\infty}^{n-1} \sum_{j=1}^{\varepsilon_i} \sum_{g=0}^{\infty} G_{n-i}^{(g,i,j)} + \varepsilon_n. \tag{A.11}$$

Note that the third summation really starts in $g = 1$ because $G_{n-i}^{(0,i,j)} = 1_{\{n-i=0\}} = 0$ whenever

$i \leqslant n-1$. For the triple sum in (A.11), we obtain

$$\sum_{i=-\infty}^{n-1}\sum_{j=1}^{\varepsilon_i}\sum_{g=1}^{\infty} G_{n-i}^{(g,i,j)} = \sum_{i=-\infty}^{n-1}\sum_{j=1}^{\varepsilon_i}\sum_{g=1}^{\infty}\sum_{k=1}^{n-i}\sum_{m=1}^{G_{n-i-k}^{(g-1,i,j)}} \xi_{g,i,j,m}^{(i+n-i,k)}$$
$$= \sum_{k=1}^{\infty}\left(\sum_{i=-\infty}^{n-k}\sum_{j=1}^{\varepsilon_i}\sum_{g=1}^{\infty}\sum_{m=1}^{G_{n-k-i}^{(g-1,i,j)}} \xi_{g,i,j,m}^{(n,k)}\right). \tag{A.12}$$

For (A.12), we use the fact that it is irrelevant whether we let k run over $\{1,2,\ldots,n-i\}$ or over $\mathbb{N}$ because $G_{n-i-k}^{(g-1,i,j)} = 0$ whenever $k > n-i$; by the same argument, we may let i run up to $n-k$ only. For fixed $n \in \mathbb{Z}$ and $k \in \mathbb{N}$, the term in the bracket is a sum of i.i.d. Pois(α_k) random variables $\xi_{n,j,g,m}^{(n+i,k)}$ over the (stochastic) index set

$$I^{(n,k)} := \left\{(g,i,j,m) \in \mathbb{Z}^4 : \ 1 \leqslant g,\ i \leqslant n-k,\ 1 \leqslant j \leqslant \varepsilon_i,\ 1 \leqslant m \leqslant G_{n-k-i}^{(g-1,i,j)}\right\}.$$

For the size of $I^{(n,k)}$, we obtain

$$\left|I^{(n,k)}\right| = \sum_{i=-\infty}^{n-k}\sum_{j=1}^{\varepsilon_i}\sum_{g=1}^{\infty} G_{n-i-k}^{(g-1,i,j)} \overset{(A.3)}{=} \sum_{i=-\infty}^{n-k}\sum_{j=1}^{\varepsilon_i} F_{n-i-k}^{(i,j)} \overset{(A.4)}{=} \tilde{X}_{n-k} \quad (<\infty,\ a.s). \tag{A.13}$$

Let $(\xi_m^{(n,k)})$ be the offspring sequences from Definition 2. Note that, for $n \in \mathbb{Z}$ and $k \in \mathbb{N}$,

$$\left(\xi_l^{(n,k)} : \ l = 1,2,\ldots,\tilde{X}_{n-k}\right) \quad \text{and} \quad \left(\xi_{g,i,j,m}^{(n,k)} : \ (g,i,j,m) \in I^{(n,k)}\right) \tag{A.14}$$

are equally distributed—no matter which order we choose for the second set. Also note that, for $(n,k) \neq (n',k')$, we have that $I^{(n,k)} \cap I^{(n',k')} = \varnothing$. Consequently, the independence properties over n and k necessary for the reproductions are preserved. We did indeed make correspondence (A.14) explicit. The reordering of the offspring sequences, however, is cumbersome. It involves a function that is recursively defined on multiple levels; its presentation would double the length of the whole proof—yielding hardly additional insight at that. So we chose to leave it with (A.14). Continuing with (A.11), we obtain that, for $n \in \mathbb{Z}$,

$$\tilde{X}_n \overset{(A.11)}{=} \sum_{i=-\infty}^{n-1}\sum_{j=1}^{\varepsilon_i}\sum_{g=0}^{\infty} G_{n-i}^{(g,i,j)} + \varepsilon_n$$
$$\overset{(A.12)}{=} \sum_{k=1}^{\infty}\sum_{(g,i,j,m)\in I^{(n,k)}} \xi_{g,i,j,m}^{(n,k)} + \varepsilon_n = \sum_{k=1}^{\infty}\sum_{l=1}^{|I^{(n,k)}|} \xi_l^{(n,k)} + \varepsilon_n \overset{(A.14)}{=} \sum_{k=1}^{\infty}\sum_{l=1}^{\tilde{X}_{n-k}} \xi_l^{(n,k)} + \varepsilon_n.$$

We conclude that $\left(\tilde{X}_n\right)$ indeed solves (2).

For uniqueness, consider two stationary solutions $(X_n,)$, (Y_n) of (2)—defined on the same probability space and with respect to the same immigration sequence (ε_n) and the same offspring sequences $\left(\xi_l^{(n,k)}\right)$, $n \in \mathbb{Z}$, $k \in \mathbb{N}$. It follows from (2) that $\mathbb{E}\,X_n = \mathbb{E}\,Y_n = \alpha/(1-K) < \infty$. Then

$$\begin{aligned}\mathbb{E}\left|X_n - Y_n\right| &\overset{(1)}{=} \mathbb{E}\left|\sum_{k=1}^{\infty}\left(\alpha_k \circ X_{n-k} - \alpha_k \circ Y_{n-k}\right)\right| \\ &\leqslant \mathbb{E}\sum_{k=1}^{\infty}\left|\alpha_k \circ X_{n-k} - \alpha_k \circ Y_{n-k}\right| = \sum_{k=1}^{\infty}\mathbb{E}\left|\alpha_k \circ X_{n-k} - \alpha_k \circ Y_{n-k}\right|. \end{aligned} \tag{A.15}$$

As the offspring at time n of X_{n-k} and Y_{n-k} is given by same offspring sequence, we have that

$$\begin{aligned}\left|\alpha_k \circ X_{n-k} - \alpha_k \circ Y_{n-k}\right| &= \left|\sum_{l=1}^{X_{n-k}} \xi_l^{(n,k)} - \sum_{l=1}^{Y_{n-k}} \xi_l^{(n,k)}\right| \\ &= \left|\sum_{l=\min\{X_{n-k},Y_{n-k}\}}^{\max\{X_{n-k},Y_{n-k}\}} \xi_l^{(n,k)}\right| \overset{\mathrm{d}}{=} \left|\sum_{l=1}^{|X_{n-k}-Y_{n-k}|} \xi_l^{(n,k)}\right|, \quad k \in \mathbb{N}. \end{aligned} \tag{A.16}$$

Plugging (A.16) in (A.15), we obtain

$$\mathbb{E}\left|X_n - Y_n\right| \leqslant \sum_{k=1}^{\infty}\mathbb{E}\sum_{i=1}^{|X_{n-k}-Y_{n-k}|}\xi_i^{(n,k)} = \sum_{k=1}^{\infty}\alpha_k\,\mathbb{E}\left|X_{n-k} - Y_{n-k}\right| \overset{\text{stat.}}{=} K\,\mathbb{E}\left|X_n - Y_n\right|, \quad n \in \mathbb{Z}.$$

As $K < 1$ by assumption and $\mathbb{E}\left|X_n - Y_n\right| < \infty$, we get that $\mathbb{E}\left|X_n - Y_n\right| = 0$ and therefore $X_n = Y_n$, $n \in \mathbb{Z}$, almost surely.

□

A.2 Proof of Proposition 4.

Equation (3) together with (4) and (5) is exactly the construction of a solution to the defining difference-equations (2) in the proof of Theorem 3; see (A.3) and (A.10). To establish (6), consider the process on the right-hand side:

$$\left(\tilde{F}_n\right)_{n\in\mathbb{Z}} := \left(1_{\{n=0\}} + \sum_{i=1}^{n}\sum_{j=1}^{G_i^{(1)}} F_{n-i}^{(i,j)}\right)_{n\in\mathbb{Z}}$$

We show that the process $\left(\tilde{F}_n\right)$ is constructed by the same (stochastic) recursion as (F_n); see (4) and (5). Then the equality in distribution follows. For $n \in \mathbb{Z}$, we define

$$\tilde{G}_n^{(0)} := 1_{\{n=0\}} \quad \text{and} \quad \tilde{G}_n^{(g)} := \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} G_{n-i}^{(g-1,i,j)}, \quad g \in \mathbb{N}, \tag{A.17}$$

where $\left(G_n^{(g,i,j)}\right)$, $g \in \mathbb{N}_0$, are the generation processes that constitute the family processes $\left(F_n^{(i,j)}\right)$, $i \in \mathbb{Z}$, $j \in \mathbb{N}$, in (6). In particular, $\left(G_n^{(g,i,j)}\right)$ are independent copies of the generations $\left(G_n^{(g)}\right)$ defined in (5). Then, by construction, $\tilde{F}_n = \sum_{g\geqslant 0} \tilde{G}_n^{(g)}$, $n \in \mathbb{Z}$. This establishes a representation for $\left(\tilde{F}_n\right)$ of the same form as (4) for (F_n). Next, we show that the summands $\left(\tilde{G}_n^{(g)}\right)$ follow the same recursion (5) as the original generations $\left(G_n^{(g)}\right)$: for $g = 0$, we have that $\tilde{G}_n^{(0)} = 1_{\{n=0\}}$, $n \in \mathbb{Z}$. So the starting value of the recursion for $\left(\tilde{G}_n^{(g)}\right)$ is the same as the starting value of the recursion (5) for $\left(G_n^{(g)}\right)$. For $g = 1$, recursion (5) is also analogue:

$$\tilde{G}_n^{(1)} = \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} G_{n-i}^{(0,i,j)} = \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} 1_{\{n-i=0\}} = G_n^{(1)} = \sum_{k=1}^{n} \alpha_k \circ 1_{\{n-k=0\}} = \sum_{k=1}^{n} \alpha_k \circ \tilde{G}_{n-k}^{(0)}.$$

And, for any $g \geqslant 2$, we find that

$$\tilde{G}_n^{(g)} = \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} G_{n-i}^{(g-1,i,j)} = \sum_{i=1}^{n} \sum_{j=1}^{G_i^{(1)}} \sum_{k=1}^{n-i} \alpha_k \circ G_{n-i-k}^{(g-2,i,j)} = \sum_{k=1}^{n} \sum_{i=1}^{n-k} \sum_{j=1}^{G_i^{(1)}} \alpha_k \circ G_{n-i-k}^{(g-2,i,j)},$$

where in the third equality we use that $G_{n-i-k}^{(g-2,i,j)} = 0$, $g \geqslant 2$, $n - i - k \leqslant 0$. At this point we avoid the explicit representation of the offspring sequences. We just remind ourselves that all reproductions involved are independent and establish

$$\tilde{G}_n^{(g)} = \sum_{k=1}^{n} \alpha_k \circ \sum_{i=1}^{n-k} \sum_{j=1}^{G_i^{(1)}} G_{n-i-k}^{(g-2,i,j)} = \sum_{k=1}^{n} \alpha_k \circ \tilde{G}_{n-k}^{(g-1)}, \quad n \in \mathbb{Z},\ g \geqslant 2.$$

In other words, the processes (G_n) and $\left(\tilde{G}_n\right)$ and, consequently, the processes (F_n) and $\left(\tilde{F}_n\right)$ are constructed by the same stochastic recursion. We conclude that $\left(F_n\right) \stackrel{\mathrm{d}}{=} \left(\tilde{F}_n\right)$. This establishes (6).

□

A.3 Proof of Proposition 6.

For (9), we first apply representation (3):

$$M_{(X_n)}\left((t_n)\right) = \mathbb{E}\exp\left\{\sum_{n=0}^{\infty} t_n X_n\right\} = \mathbb{E}\exp\left\{\sum_{n=0}^{d} t_n X_n\right\} = \mathbb{E}\exp\left\{\sum_{n=0}^{d} t_n \sum_{i\in\mathbb{Z}}\sum_{j=1}^{\varepsilon_i} F_{n-i}^{(i,j)}\right\}$$
$$= \mathbb{E}\prod_{i\in\mathbb{Z}}\prod_{j=1}^{\varepsilon_i}\exp\left\{\sum_{n=0}^{d} t_n F_{n-i}^{(i,j)}\right\}.$$

In the following, we set $t_m := 0$ whenever $m < 0$. Conditioning on the immigration sequence (ε_i) and exploiting its independence from the family processes $\left(F_n^{(i,j)}\right)$ yields

$$\begin{aligned}
M_{(X_n)}(t_1,\ldots,t_d) &= \mathbb{E}\prod_{i\in\mathbb{Z}}\prod_{j=1}^{\varepsilon_i}\mathbb{E}\left[\exp\left\{\sum_{n=0}^{d} t_n F_{n-i}^{(i,j)}\right\}\right] \\
&= \prod_{i\in\mathbb{Z}}\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\sum_{n\in\mathbb{Z}} t_n F_{n-i}\right\}\right]^{\varepsilon_i}\right] \\
&= \prod_{i\in\mathbb{Z}}\mathbb{E}\left[M_{(F_n)}\left((t_{i+n})_{n\in\mathbb{N}_0}\right)^{\varepsilon_i}\right] \\
&= \prod_{i\in\mathbb{Z}}\exp\left\{\alpha_0\left(M_{(F_n)}\left((t_{i+n})_{n\in\mathbb{N}_0}\right)-1\right)\right\} \\
&= \exp\left\{\sum_{i\in\mathbb{Z}}\alpha_0\left(M_{(F_n)}\left((t_{i+n})_{n\in\mathbb{N}_0}\right)-1\right)\right\}, \qquad \text{(A.18)}
\end{aligned}$$

where in the last but one step we applied the formula for the probability-generating function of a Poisson random variable. Up to now, (A.18) is only a formal representation of $M_{(X_n)}$ in terms of $M_{(F_n)}$. It is not clear yet, whether and when $M_{(X_n)}$ is finite. For (10), we apply representation (6) of (F_n) from Proposition 4:

$$M_{(F_n)}\left((s_n)\right) = \mathbb{E}\exp\left\{\sum_{n=0}^{\infty} s_n F_n\right\} \overset{(6)}{=} \mathbb{E}\exp\left\{\sum_{n=0}^{\infty} s_n\left(1_{\{n=0\}} + \sum_{k=1}^{n}\sum_{j=1}^{G_k^{(1)}} F_{n-k}^{(k,j)}\right)\right\}$$

We note that in the last term, the index k may run to ∞ instead of n, because $F_{n-k}^{(k,j)} = 0,\ j\in\mathbb{N}$, whenever $k > n$. After straightforward calculations, we obtain

$$\begin{aligned}
&M_{(F_n)}\left((s_n)_{n\in\mathbb{N}_0}\right) \\
&= e^{s_0}\,\mathbb{E}\exp\left\{\sum_{n=0}^{\infty}\sum_{k=1}^{\infty}\sum_{j=1}^{G_k^{(1)}} s_n F_{n-k}^{(k,j)}\right\} = e^{s_0}\exp\left\{\sum_{k=1}^{\infty}\alpha_k\left(M_{(F_n)}\left((s_{n+k})_{n\in\mathbb{N}_0}\right)-1\right)\right\}. \qquad \text{(A.19)}
\end{aligned}$$

Next we derive finiteness of $M_{(F_n)}$. Let (s_n) be a sequence with finite support and $s := \max\{s_n\}$ so that $\sum_{n=0}^{\infty} s_n F_n$ is bounded from above by sS, where $S := \sum_{n=0}^{\infty} F_n$ denotes the total number of individuals in the generic family (F_n). We remind ourselves of the defining equation (4) for the family process (F_n) and find that

$$S = \sum_{n=0}^{\infty} F_n = \sum_{n=0}^{\infty}\sum_{g=0}^{\infty} G_n^{(g)} = \sum_{g=0}^{\infty}\sum_{n=0}^{\infty} G_n^{(g)}. \tag{A.20}$$

We denote the total number of individuals in the g-th generation by $Y_g := \sum_{n=0}^{\infty} G_n^{(g)}$, $g \in \mathbb{N}_0$. The sequence $(Y_g)_{g\in\mathbb{N}_0}$ is the embedded generation process. Applying (5), we find that $Y_0 = 1$ and, for $g \geqslant 2$,

$$\begin{aligned} Y_g = \sum_{n=0}^{\infty} G_n^{(g)} = \sum_{n=0}^{\infty}\sum_{k=1}^{\infty} \alpha_k \circ G_{n-k}^{(g-1)} = \sum_{k=1}^{\infty}\sum_{n=0}^{\infty} \alpha_k \circ G_{n-k}^{(g-1)} \overset{\mathrm{d}}{=} \sum_{k=1}^{\infty} \alpha_k \circ \sum_{n=0}^{\infty} G_{n-k}^{(g-1)} = \sum_{k=1}^{\infty} \alpha_k \circ Y_{g-1} \\ = \sum_{k=1}^{Y_{g-1}} \xi_k^{(g)}, \end{aligned}$$

where $\xi_k^{(g)} \overset{\text{iid}}{\sim} \mathrm{Pois}(K)$, $k, g \in \mathbb{N}$. In other words, the embedded generation process (Y_g) is a standard Galton–Watson branching process. From (A.20), we see that $S \overset{\mathrm{d}}{=} \sum_{g=0}^{\infty} Y_g$. In other words, S is distributed like the cumulative limit of a standard Galton–Watson process. The moment-generating functions of such limits have been considered in the literature: as $K < 1$, by Theorem 2.1 in Nakayama et al. (2004), there exists a $\delta > 0$ such that $\mathbb{E}\exp\{\delta S\} < \infty$ if and only if there exists a $\tilde{\delta} > 0$ such that $\mathbb{E}\exp\left\{\tilde{\delta}\xi_1^{(1)}\right\} < \infty$. The latter is indeed the case because the moment-generating function of a Poisson variable is finite on $\mathbb{R}$. Furthermore, by Jensen's inequality, we see that $1 \leqslant \lim_{n\to\infty} \mathbb{E}\exp\{\delta S/n\} \leqslant \lim_{n\to\infty} \left(\mathbb{E}\exp\{\delta S\}\right)^{1/n} = 1$. So, for any given $\epsilon > 0$, we can assume the existence of a $\delta > 0$ such that $\mathbb{E}\exp(\delta S) < 1 + \epsilon$. In particular,

$$\exists\, \delta > 0 \text{ such that } M_{(F_n)}\big((s_n)\big) \leqslant \mathbb{E}\exp(\delta S) < 1 + (1-K)/(2K) \text{ for } (s_n)_{n\in\mathbb{N}_0} \in c_{00}\left((-\infty, \delta]\right), \tag{A.21}$$

where, as before, $K = \sum_{k=1}^{\infty} \alpha_k < 1$. We now have established finiteness of $M_{(F_n)}$ in a neighborhood of zero. It remains to establish finiteness of $M_{(X_n)}$. Our goal is to bound the series representation (A.18) of $M_{(X_n)}$. To that aim, we need to refine the bound (A.21) for $M_{(F_n)}\big((s_n)\big)$. To that aim, with the constant $\delta > 0$ from (A.21), for $i \in \mathbb{Z}$, we introduce the sequences $\left(\delta_n^{(i)}\right)_{n\in\mathbb{Z}}$ defined by

$$\delta_n^{(i)} := \begin{cases} \delta, & n = i,\, i-1,\, \ldots,\, i-d, \\ 0, & \text{else}, \end{cases} \tag{A.22}$$

where $d = \operatorname{supp}((t_n)) + 1$, as before, with (t_n) the considered argument sequence. Note that, for $i < 0$, we have that $\delta_n^{(i)} = 0$, $n \in \mathbb{N}_0$. Consequently, by definition of $M_{(F_n)}$, for $i < 0$, we have that $M_{(F_n)}\left(\left(\delta_n^{(i)}\right)_{n\in\mathbb{N}_0}\right) = 1$. Furthermore, observe that $\left(\delta_{n+k}^{(i)}\right)_n = \left(\delta_n^{(i-k)}\right)_n$, $i,k \in \mathbb{Z}$. For $(t_n) \in c_{00}\big((-\infty,\delta]\big)$, we have by component-wise monotonicity of $M_{(X_n)}$ that

$$\begin{aligned} M_{(X_n)}\big((t_n)\big) \leqslant M_{(X_n)}\left((\delta_{n+i}^{(d)})_{n\in\mathbb{N}_0}\right) &\overset{(A.18)}{=} \exp\left\{\sum_{i\in\mathbb{Z}} \alpha_0\Big(M_{(F_n)}\big((\delta_{n+i}^{(d)})_{n\in\mathbb{N}_0}\big) - 1\Big)\right\} \\ &= \exp\left\{\sum_{i=-\infty}^{d} \alpha_0\left(M_{(F_n)}\left((\delta_n^{(d-i)})_{n\in\mathbb{N}_0}\right) - 1\right)\right\} = \exp\left\{\alpha_0 \sum_{i=0}^{\infty} m_i\right\}, \end{aligned} \tag{A.23}$$

where we set $m_i := M_{(F_n)}\big((\delta_n^{(i)})_{n\in\mathbb{N}_0}\big) - 1$, $i \in \mathbb{Z}$. Note that, by (A.21), we have that

$$0 \leqslant m_i \leqslant (1-K)/(2K), \quad i \in \mathbb{Z}. \tag{A.24}$$

and, in particular, $m_i = 0$ for $i < 0$. In the following, we only consider m_i with $i > d$. In this case, we get from (A.22) that $e^{\delta_0^{(i)}} = e^0 = 1$ and we obtain the recursion

$$\begin{aligned} m_i = M_{(F_n)}\big((\delta_n^{(i)})_{n\in\mathbb{N}_0}\big) - 1 &\overset{(A.19)}{=} e^{\delta_0^{(i)}} \exp\left\{\sum_{k=1}^{\infty} \alpha_k\Big(M_{(F_n)}\big((\delta_{n+k}^{(i)})_{n\in\mathbb{N}_0}\big) - 1\Big)\right\} - 1 \\ &\overset{(A.22)}{=} \exp\left\{\sum_{k=1}^{i} \alpha_k m_{i-k}\right\} - 1, \quad i > d. \end{aligned} \tag{A.25}$$

For the sum in the exponential, we find that

$$\sum_{k=1}^{i} \alpha_k m_{i-k} \overset{(A.24)}{\leqslant} \sum_{k=1}^{\infty} \alpha_k (1-K)/(2K) = K(1-K)/(2K) < 1, \quad i \in \mathbb{Z}.$$

Therefore, we may apply the exponential inequality $\exp(x) \leqslant (1-x)^{-1}$, $x < 1$, in (A.25):

$$\begin{aligned} m_i \leqslant \frac{1}{1-\sum_{k=1}^{i} \alpha_k m_{i-k}} - 1 = \frac{\sum_{k=1}^{i} \alpha_k m_{i-k}}{1-\sum_{k=1}^{i} \alpha_k m_{i-k}} &\overset{(A.24)}{\leqslant} \frac{\sum_{k=1}^{i} \alpha_k m_{i-k}}{1-\sum_{k=1}^{i} \alpha_k (1-K)/(2K)} \\ &\leqslant \frac{2\sum_{k=1}^{\infty} \alpha_k m_{i-k}}{1+K}, \quad i > d. \end{aligned} \tag{A.26}$$

Summing both sides of (A.26) over $i > d$, we obtain

$$\begin{aligned}\sum_{i=d+1}^{\infty} m_i \leqslant \frac{2}{1+K} \sum_{i=d+1}^{\infty} \sum_{k=1}^{\infty} \alpha_k m_{i-k} &= \frac{2}{1+K} \sum_{k=1}^{\infty} \alpha_k \sum_{i=d+1}^{\infty} m_{i-k} \\ &\leqslant \frac{2}{1+K} \sum_{k=1}^{\infty} \alpha_k \sum_{l=0}^{\infty} m_l \\ &= \frac{2K}{1+K} \left(\sum_{l=0}^{d} m_l + \sum_{l=d+1}^{\infty} m_l \right). \end{aligned} \tag{A.27}$$

Keeping in mind that $1 - 2K/(1+K) > 0$, we solve (A.27) for $\sum_{i=d+1}^{\infty} m_i$ and find that

$$\sum_{i=d+1}^{\infty} m_i \leqslant \left(1 - \frac{2K}{1+K}\right)^{-1} \frac{2K}{1+K} \sum_{l=0}^{d} m_l \overset{(A.24)}{\leqslant} \frac{2K}{1-K} d \frac{1-K}{2K} = (1+d). \tag{A.28}$$

For the summation of $(m_i)_{i \in \mathbb{N}_0}$ over $i \in \mathbb{N}_0$, we finally obtain

$$\sum_{i=0}^{\infty} m_i \overset{(A.28)}{\leqslant} \sum_{i=0}^{d} m_i + d + 1 \overset{(A.24)}{\leqslant} (d+1) \left(\frac{1-K}{2K} + 1 \right) = (d+1) \frac{1+K}{2K}. \tag{A.29}$$

We conclude that, for all $(t_n) \in c_{00}((-\infty, \delta])$ with $\operatorname{supp}((t_n)) = d$, we have that

$$M_{(X_n)}(t_1, \ldots, t_d) \overset{(A.23)}{\leqslant} \exp\left\{ \alpha_0 \sum_{i=0}^{\infty} m_i \right\} \overset{(A.29)}{\leqslant} \exp\left\{ \alpha_0 d \frac{1+K}{2K} \right\} < \infty. \tag{A.30}$$

Uniqueness of $M_{(F_n)}$ follows by induction over the (finite) support of the argument sequence. In that sense, the implicit equation (A.19) specifies $M_{(F_n)}$ recursively.

□

A.4 Proof of Proposition 8.

The sequence values u_n, $n \in \mathbb{Z}$, are well-defined because the partial sums of $\sum_{k=1}^{\infty} \alpha_k X_{n-k}$ are nondecreasing and their expectations have a finite limit. Stationarity of (u_n) follows from stationarity of (X_n). Denote $\mathcal{F}_n := \sigma\{X_k : k \leqslant n\}$. From $\mathbb{E}\left[\alpha_k \circ X_{n-k} | \mathcal{F}_n\right] = \alpha_k X_{n-k}, k \in \mathbb{N}$, we get that

$$\mathbb{E}\left[u_n | \mathcal{F}_n\right] = \mathbb{E}\left[X_n - \alpha_0 - \sum_{k=1}^{\infty} \alpha_k X_{n-k} \mathcal{F}_n \right] = 0, \tag{A.31}$$

and, consequently, $\mathbb{E}\,u_n = 0$, $n \in \mathbb{Z}$. For the autocovariances of the errors, note that , for $n' < n$ (and then, by symmetry, for $n' \neq n$),

$$\mathbb{E}\left[u_n u_{n'}\right] = \mathbb{E}\left[\,\mathbb{E}\left[u_n u_{n'}|\mathcal{F}_{n-1}\right]\right] = \mathbb{E}\left[u_{n'}\underbrace{\mathbb{E}\left[u_n|\mathcal{F}_{n-1}\right]}_{\overset{(A.31)}{=}0}\right] = 0.$$

Finally, we have that

$$\mathrm{Var}(u_n) = \mathbb{E}\Big[\underbrace{\mathrm{Var}\left(u_n|\mathcal{F}_{n-1}\right)}_{\overset{(11)}{=}\mathrm{Var}(X_n|\mathcal{F}_{n-1})}\Big] + \mathrm{Var}\Big(\underbrace{\mathbb{E}\left[u_n|\mathcal{F}_{n-1}\right]}_{\overset{(A.31)}{=}0}\Big) = \frac{\alpha_0}{1-K}. \tag{A.32}$$

□

A.5 Proof of Proposition 9.

Let B be the backward shift operator defined by $B^k x_n := x_{n-k}$, $k \in \mathbb{Z}$, for any sequence $(x_n)_{n\in\mathbb{Z}}$. Consider the power series $\phi(z) := 1 - \sum_{k=1}^{\infty} \alpha_k z^k$. With these notations, we may rewrite (13) as $\phi(B)(X_n - \mu_X) = u_n$, $n \in \mathbb{Z}$, where (u_n) is the white-noise sequence from Proposition 8. One can show that the power series $\phi(z)$ is (absolutely) convergent for all $|z| \leqslant 1$. So ϕ is analytic on the open unit disc. Furthermore, we have that

$$|\phi(z)| = \left|1 - \sum_{k=1}^{\infty} \alpha_k z^k\right| \geqslant 1 - \left|\sum_{k=1}^{\infty} \alpha_k z^k\right| \geqslant 1 - \sum_{k=1}^{\infty} \alpha_k |z|^k \geqslant 1 - K > 0, \quad |z| \leqslant 1.$$

As $\phi(z) \neq 0$ for $|z| \leqslant 1$, we may define the function $\psi(z) := 1/\phi(z)$, $|z| \leqslant 1$, which is also analytic on the open unit disc and which, consequently, has a power-series representation $\psi(z) = \sum_{k=0}^{\infty} \beta_k z^k$, $|z| < 1$. As $1 = \psi(z)\phi(z)$ by definition, if follows that, for $|z| < 1$,

$$1 = \sum_{k=0}^{\infty} \beta_k z^k \left(1 - \sum_{l=1}^{\infty} \alpha_l z^l\right) = \sum_{k=0}^{\infty} \left(\beta_k - \sum_{j=1}^{k} \alpha_j \beta_{k-j}\right) z^k. \tag{A.33}$$

Comparing coefficients in (A.33), one obtains the recursion

$$\beta_0 = 1 \quad \text{and} \quad \beta_k = \sum_{j=1}^{k} \alpha_j \beta_{k-j}, \quad k \in \mathbb{N}\,.$$

We note that $\beta_k \geqslant 0$ because $\alpha_k \geqslant 0$. Formally, we can write

$$X_n - \mu_X = \psi(B) u_n, \quad n \in \mathbb{Z}\,. \tag{A.34}$$

For the well-definedness of the right-hand side of this equation, it suffices to show that $\sum_{k=0}^{\infty} |\beta_k| < \infty$; see Proposition 3.1.2 in Brockwell and Davis (1991). To that aim, we apply Wiener's Lemma; see Lemma IIc. in Wiener (1932). Let $\tilde{\phi}(\theta) = 1 - \sum_{k=1}^{\infty} \alpha_k e^{ik\pi\theta}$, $\theta \in (-\pi, \pi]$. The lemma states that if $\sum_{k=1}^{\infty} |\alpha_k| < 1$, then the Fourier series of the function $1/\tilde{\phi}(\theta)$ is absolutely convergent. By the same calculation as in (A.33), we find that the Fourier-coefficients of $1/\tilde{\phi}(\theta)$ are exactly the β_k, $k \in \mathbb{N}_0$, from our ψ function. From this we get

$$\sum_{k=0}^{\infty} |\beta_k| \overset{\text{Lemma}}{=} \frac{1}{\tilde{\phi}(0)} = \frac{1}{\phi(1)} = \frac{1}{1-K} < \infty.$$

We conclude that (A.34) is a meaningful family of equations. In other words, $(X_n - \mu_X)$ can be represented as a moving-average process with respect to the white-noise sequence (u_n).

□

A.6 Proof of Proposition 10.

With the notation from the moving-average representation of the INAR(∞) sequence in Proposition 9, we find that, for $j, n \in \mathbb{Z}$,

$$R(j) = \operatorname{Cov}(X_n - \mu_X, X_{n+j} - \mu_X) = \operatorname{Cov}\left(\sum_{k=0}^{\infty} \beta_k u_{n-k}, \sum_{k=0}^{\infty} \beta_k u_{n+j-k}\right).$$

From Proposition 8, we know that $\operatorname{Cov}(u_n, u_{n+j}) = 1_{\{j=0\}}\alpha_0/(1-K)$, $j \in \mathbb{Z}$,. Furthermore, from Proposition 9, we have that the coefficients β_k are absolutely summable. So (15) follows from Proposition 3.1.2. in Brockwell and Davis (1991). For the sum of the autocovariance sequence, we observe

$$\sum_{k=0}^{\infty} R(k) = \frac{\alpha_0}{1-K} \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \beta_i \beta_{i+k} = \frac{\alpha_0}{1-K} \sum_{i=0}^{\infty} \beta_i \sum_{k=0}^{\infty} \beta_{i+k} \leqslant \frac{\alpha_0}{1-K} \sum_{i=0}^{\infty} \beta_i \sum_{k=-i}^{\infty} \beta_{i+k} = \frac{\alpha_0}{(1-K)^3}.$$

The last equality re-uses the result $\sum_{i=0}^{\infty} \beta_i = 1/(1-K)$ from Proposition 9.

□

A.7 Proof of Proposition 18.

Plugging in definitions, we obtain

$$\Psi_{N^{(\Delta)}}[f] = \mathbb{E} \exp\left\{-\int_{\mathbb{R}} f(t) N^{(\Delta)}(\mathrm{d}t)\right\} = \mathbb{E} \exp\left\{-\sum_{n \in \mathbb{Z}} X_n^{(\Delta)} f(n\Delta)\right\}.$$

By stationarity of $(X_n^{(\Delta)})$, we may assume without loss of generality that $f(t) = 0$, $t < 0$. We apply formulas (9) and (10) for the joint moment-generating function of the INAR(∞) process $\left(X_n^{(\Delta)}\right)$ and the corresponding generic family process $\left(F_n^{(\Delta)}\right)$:

$$\begin{aligned}\Psi_{N^{(\Delta)}}[f] &= M_{\left(X_n^{(\Delta)}\right)}\Big(\big(-f(0), -f(\Delta), -f(2\Delta), \dots\big)\Big) \\ &\overset{(9)}{=} \exp\left\{\alpha_0 \sum_{i\in\mathbb{Z}} \left(M_{\left(F_n^{(\Delta)}\right)}\Big(\big(-f\big((n+i)\Delta\big)\big)_{n\in\mathbb{N}_0}\Big) - 1\right)\right\}.\end{aligned}$$

We set $\Psi^{(\Delta)}_{F^{(\Delta)}}\big((s_n)\big) := M_{\left(F^{(\Delta)}\right)}\big((-s_n)\big)$, $(s_n) \in c_{00}\big([0,\infty)\big)$. This establishes the lemma.

□

A.8 Proof of Lemma 19.

Claims (22) and (23) directly follow from the definition of $N^{(\Delta)}$ in (21) together with Proposition 3 and Lemma 16. For (24), we find that the number of grid points in the interval $[a,b]$ is less or equal $\lceil (b-a)/\Delta \rceil + 1$. To get rid of the ceiling function, we observe that $\lceil (b-a)/\Delta \rceil + 1 < (b-a)/\Delta + 2$. Combining this with the facts that $K^{(\Delta)} \leqslant \tilde{K}$ and $\Delta < \delta$ together with (22) and (23) yields inequality (24).

□

A.9 Proof of Lemma 20.

Let $A \in \mathcal{B}_b$ be a bounded Borel set. Note that

$$\operatorname{Var}\left(N^{(\Delta)}(A)\right) = \operatorname{Var}\left(\sum_{n:\, n\Delta\in A} X_n^{(\Delta)}\right) = \sum_{m:\, m\Delta\in A}\sum_{n:\, n\Delta\in A} \operatorname{Cov}\left(X_n^{(\Delta)}, X_m^{(\Delta)}\right) = \sum_{m:\, m\Delta\in A}\sum_{n:\, n\Delta\in A} R^{(\Delta)}(n-m),$$

where $R^{(\Delta)}$ is the autocovariance function of the INAR(∞) sequence from Proposition 10. From this proposition, we know that $R^{(\Delta)}(k) \geqslant 0$, $k \in \mathbb{Z}$ and $\sum_{k=0}^{\infty} R^{(\Delta)}(k) \leqslant \eta\Delta(1-K^{(\Delta)})^{-3}$. Applying these results yields

$$\begin{aligned}\operatorname{Var}\left(N^{(\Delta)}(A)\right) &\leqslant \sum_{m:\, m\Delta\in A}\sum_{n\in\mathbb{Z}} R^{(\Delta)}(n-m) \leqslant \sum_{m:\, m\Delta\in A} \frac{2\eta\Delta}{(1-K^{(\Delta)})^3} \\ &\leqslant \left(2 + \frac{\sup A - \inf A}{\Delta}\right)\frac{2\eta\Delta}{(1-\tilde{K})^3} \leqslant (2\delta + \sup A - \inf A)\,\frac{2\eta}{(1-\tilde{K})^3},\end{aligned}$$

where $\tilde{K} < 1$ does not depend on the choice of $\Delta \in (0,\delta)$; see Lemma 16.

□

A.10 Proof of Lemma 21.

The claim follows with Proposition 11.1.VI. from Daley and Vere-Jones (2009) if for all compact intervals $[a,b] \subset \mathbb{R}$ and for all $\epsilon > 0$ there exists an $M < \infty$ such that

$$\sup_{\Delta \in (0,\delta)} \mathbb{P}\Big[N^{(\Delta)}\big([a,b]\big) > M \Big] < \epsilon.$$

The uniform boundedness of these probabilities is a consequence of Lemma 19 and Markov inequality: for any $\epsilon > 0$ and $a < b$, let $M_\epsilon := (b-a+2\delta)\eta/(1-\tilde{K})$, where $\delta > 0$ and $\tilde{K} < 1$ as in Lemma 16. Then we have that

$$\mathbb{P}\Big[N^{(\Delta)}\big([a,b]\big) > M_\epsilon \Big] \leqslant \frac{\mathbb{E}\, N^{(\Delta)}\big([a,b]\big)}{M_\epsilon} < (b-a+2\delta)\frac{\eta}{M_\epsilon(1-\tilde{K})} = \epsilon, \quad \Delta \in (0,\delta).$$

□

A.11 Proof of Theorem 17.

As a consequence of Lemma 22, the family of probability measures $\big(\mathbb{P}^{(\Delta)}\big)_{\Delta\in(0,\delta)}$ that corresponds to the family of point processes $\big(N^{(\Delta)}\big)_{\Delta\in(0,\delta)}$ is relatively compact for weak convergence by Prohorov's theorem; see Daley and Vere-Jones (2009), Theorem A.2.4.I. So every sequence in $\big(\mathbb{P}^{\Delta}\big)_{\Delta\in(0,\delta)}$, respectively, $\big(N^{(\Delta)}\big)_{\Delta\in(0,\delta)}$, contains a weakly convergent subsequence. In particular, for any zero sequence in $(0,\delta)$, we can find a subsequence (Δ_n) such that $\big(N^{(\Delta_n)}\big)$ converges weakly to some point process N^*. If the distribution of N^* does not depend on the initial choice of the subsequence, it follows that the original sequence converges weakly to N^*; see Theorem 2.3. in Billingsley (1968). Reconsider the implicit defining-equation (19) from Definition 12. By Proposition 13, we know that this equation determines the distribution of the solving process. So, for the proof of Theorem 17, it suffices to show that any subsequential limit candidate N^* solves (19). Furthermore, one can show that it suffices to prove (19) for $A^* \in \mathfrak{B}_a^{N^*}$, where $\mathfrak{B}_a^{N^*}$ is a semiring of sets that generates the σ-algebra $\mathcal{H}_a^{N^*}$; see (17). A semiring is a class of sets $\mathcal{A}$ such that for any pair $A, B \in \mathcal{A}$ one has (i) $A \cap B \in \mathcal{A}$ and (ii) $(A \cup B)\backslash(A \cap B) = \cup_{i=1}^n A_i$ for some $n \in \mathbb{N}$, $(A_i) \subset \mathcal{A}$ and $A_i \cap A_j = \varnothing$, $i,j = 1,\ldots,n$. We consider, for any $a \in \mathbb{R}$ and any point process N,

$$\mathfrak{B}_a^N := \Big\{ \big\{\omega \in \Omega :\, N\big((s_1,t_1]\big)(\omega) \in D_1, \ldots, N\big((s_k,t_k]\big)(\omega) \in D_k\big\} : \\ -\infty < s_i < t_i \leqslant a,\, D_i \subset \mathbb{N}_0,\, k \in \mathbb{N} \Big\}. \tag{A.35}$$

One can check that the set system $\mathfrak{B}_a^N$ is indeed a semiring. Summarizing the above, for the proof of Theorem 17, it suffices to establish

$$\mathbb{E}\left[1_{A^*} N^*\big((a,b]\big)\right] = \mathbb{E}\left[1_{A^*} \int_a^b \lambda(s|N^*)\,\mathrm{d}s\right], \quad a < b,\ A^* \in \mathcal{H}_a^{N^*}. \tag{A.36}$$

First, let us establish a discrete version of (A.36) for the approximating sequence: set $N_n := N^{(\Delta_n)}$ for all Δ_n in the chosen subsequence. For $a < b$ and $A_n \in \mathfrak{B}_a^{N_n}$, we find that

$$\begin{aligned}
\mathbb{E}\left[1_{A_n} N_n\big((a,b]\big)\right] &= \mathbb{E}\left[1_{A_n} \sum_{k:\, k\Delta_n \in (a,b]} X_k^{(\Delta_n)}\right] \\
&= \mathbb{E}\left[1_{A_n} \sum_{k:\, k\Delta_n \in (a,b]} \left(\varepsilon_k^{(\Delta_n)} + \sum_{l=1}^{\infty} \big(\Delta_n h(l\Delta_n)\big) \circ X_{k-l}^{(\Delta_n)}\right)\right] \\
&= \mathbb{E}\left[1_{A_n} \sum_{k:\, k\Delta_n \in (a,b]} \left(\Delta_n \eta + \sum_{l=1}^{\infty} \Delta_n h(l\Delta_n) X_{k-l}^{(\Delta_n)}\right)\right], \quad n \in \mathbb{N}.
\end{aligned}$$

The last step follows by the observation that the immigrations $\varepsilon_k^{(\Delta_n)}$ as well as the reproductions that contribute to $X_k^{(\Delta_n)}$, $k\Delta_n > a$, are independent of $X_{k-1}^{(\Delta_n)}, X_{k-2}^{(\Delta_n)}, \ldots$ Rewriting the inner sum of the last term as an integral with respect to the random measure N_n, we obtain, for $a < b$,

$$\begin{aligned}
&\mathbb{E}\left[1_{A_n} N_n\big((a,b]\big)\right] \\
&= \mathbb{E}\left[1_{A_n} \sum_{k:\, k\Delta_n \in (a,b]} \Delta_n \left(\eta + \int_{-\infty}^{k\Delta_n} h(k\Delta_n - s) N_n(\mathrm{d}s)\right)\right], \quad A_n \in \mathfrak{B}_a^{N_n},\ n \in \mathbb{N}.
\end{aligned} \tag{A.37}$$

Note that here and throughout the proof the upper integration bounds in the Hawkes intensities do not require special attention due to the assumption $h(0) = 0$ for reproduction intensities h in Definition 11, respectively, Lemma 16. Now we show that (A.37) converges to (A.36) corresponding to the Hawkes process. For both sides of equation (A.37), this is achieved in three steps:

- First, we establish that the random variable in the expectation can be written as $\Phi(N_n)$, where $\Phi : (M_p, \mathcal{M}_p) \to (\mathbb{R}, \mathcal{B})$ denotes some measurable mapping with set of discontinuities $D_\Phi \subset M_p$.
- Next, we show that $\mathbb{P}\left[N^* \in D_\Phi\right] = 0$.
- Finally, we prove that the random variables in question are uniformly integrable.

By Proposition 15, the first two points together imply that $\Phi(N_n) \xrightarrow{\mathrm{w}} \Phi(N^*)$. The additional uniform-integrability property yields that the corresponding expectations also converge; see Theorem 5.4 in Billingsley (1968).

Left-hand side of (A.37)*:*
Consider the map

$$\Phi : (M_p, \mathcal{M}_p) \to (\mathbb{R}, \mathcal{B}),\ m \mapsto 1_{\{m((s_1,t_1])\in D_1,\ldots,m((s_k,t_k])\in D_k\}} m\big((a,b]\big); \tag{A.38}$$

see the definition of $\mathfrak{B}_a^N$ in (A.35) for the notation. We claim that Φ is vaguely continuous on $M_p \backslash \{m : m(D_\Phi) > 0\}$, where $D_\Phi := \Big(\{a,b\} \cup \bigcup_{i=1}^k \{s_i, t_i\}\Big)$. Indeed: the map $m \mapsto m((a,b])$ is vaguely continuous on $M_p \backslash \{m : m(\{a,b\}) > 0\}$ and, for $i = 1,\ldots,k$, the maps $m \mapsto m((s_i,t_i]\})$ are vaguely continuous on $M_p \backslash \{m : m(\{s_i,t_i\}) > 0\}$. The map $\mathbb{N}_0^k \ni (l_1,\ldots,l_k) \mapsto 1_{\{l_1\in D_1,\ldots,l_k\in D_k\}}$ is trivially continuous, so that $m \mapsto 1_{\{m((s_1,t_1])\in D_1,\ldots,m((s_k,t_k])\in D_k\}}$ is continuous on $M_p \backslash \bigcup_{i=1}^k \{s_i,t_i\}$. From Proposition 15, we have that $\Phi(N_n) \xrightarrow{\mathrm{w}} \Phi(N^*)$ if $\mathbb{P}\left[N^*(D_\Phi) > 0\right] = 0$. Because D_Φ is finite, it suffices to show that $\mathbb{P}\left[N^*\big(\{t\}\big) > 0\right] = 0$ for any $t \in \mathbb{R}$.

$$\mathbb{P}\big[N^*(\{t\}) > 0\big] = \mathbb{E}\, 1_{N^*(\{t\})>0} \leqslant \mathbb{E}\, N^*(\{t\}) \overset{\text{Lemma 21}}{=} \lim_{n\to\infty} \mathbb{E}\, N_n(\{t\}) \overset{(23)}{<} \Delta \frac{\eta}{(1-\tilde{K})}, \quad t \in \mathbb{R}, \Delta \in (0,\delta).$$

So $\mathbb{P}\left[N^*(\{t\}) > 0\right] = 0$ and, consequently, $\mathbb{P}\left[N^*(D_\Phi) > 0\right]$ must also be zero. This establishes $\Phi(N_n) \xrightarrow{\mathrm{w}} \Phi(N^*)$, respectively, $1_{A_n} N_n(a,b) \xrightarrow{\mathrm{w}} 1_{A^*} N^*(a,b)$. From Lemma 21, we know that $(N_n(a,b))$ is uniformly integrable, so $(1_{A_n} N_n(a,b))$ is also uniformly integrable. Combining weak convergence and uniform integrability yields convergence of expectations

$$\lim_{n\to\infty} \mathbb{E}\left[1_{A_n} N_n\big((a,b]\big)\right] = \mathbb{E}\left[1_{A^*} N^*\big((a,b]\big)\right].$$

We have established the convergence of the left-hand side of (A.37) to the left-hand side of (A.36).

Right-hand side of (A.37)*:*
Note that the right-hand side of (A.37) converges to the right-hand side of (A.36) if

$$\sum_{k\in(a,b]} \Delta_n\, \mathbb{E}\left[1_{A_n} \int_{-\infty}^{k\Delta_n} h(k\Delta_n - s) N_n(\mathrm{d}s)\right] \xrightarrow{n\to\infty} \int_a^b \mathbb{E}\left[1_{A^*} \int_{-\infty}^{t} h(t-s) N^*(\mathrm{d}s)\right] \mathrm{d}t. \tag{A.39}$$

As a first step for establishing (A.39), note that, for all choices of M with $-M < a$, and, for $t \in [a,b]$,

$$1_{A_n} \int_{-M}^{t} h(t-s) N_n(\mathrm{d}s) \xrightarrow{\mathrm{w}} 1_{A^*} \int_{-M}^{t} h(t-s) N^*(\mathrm{d}s), \quad n \to \infty. \tag{A.40}$$

This is due to a continuous-mapping argument similar to the one we have used for the left-hand side of (A.37). We establish that the variances of the random variables $\int_{-M}^{t} h(t-s)N_n(\mathrm{d}s)$, $n \in \mathbb{N}_0$, are uniformly bounded:

$$\begin{aligned}
&\operatorname{Var} \int_{-M}^{t} h(t-s) N_n(\mathrm{d}s) \\
&\leqslant \operatorname{Var} \sum_{l=1}^{\lceil M/\Delta_n \rceil} h(l\Delta_n) X_{-l}^{(\Delta_n)} = \sum_{l=1}^{\lceil M/\Delta_n \rceil} \sum_{m=1}^{\lceil M/\Delta_n \rceil} h(l\Delta_n) h(m\Delta_n) \operatorname{Cov}\left(X_{-l}^{(\Delta_n)}, X_{-m}^{(\Delta_n)}\right).
\end{aligned}$$

At this point, we write $R^{(\Delta_n)}$ for the autocovariance function of the INAR(∞) process $\left(X_l^{(\Delta_n)}\right)$. Applying Proposition 10 yields

$$\begin{aligned}
&\operatorname{Var} \int_{-M}^{t} h(t-s) N_n(\mathrm{d}s) \\
&\leqslant \sum_{l=1}^{\lceil M/\Delta_n \rceil} h(l\Delta_n) \sup h \sum_{m=1}^{\lceil M/\Delta_n \rceil} R^{(\Delta_n)}(|l-m|) \\
&\leqslant \sum_{l=1}^{\lceil M/\Delta_n \rceil} h(l\Delta_n) \sup h \sum_{i=-\infty}^{\infty} R^{(\Delta_n)}(i) \\
&\overset{(16)}{\leqslant} \sum_{l=1}^{\lceil M/\Delta_n \rceil} h(l\Delta_n) \sup h \frac{2\eta\Delta_n}{\left(1-K^{(\Delta_n)}\right)^3} \\
&\leqslant \left(\frac{M}{\Delta_n}+1\right) (\sup h)^2 \frac{2\eta\Delta_n}{\left(1-K^{(\Delta_n)}\right)^3} \\
&\overset{(20)}{\leqslant} (M+\Delta_n) (\sup h)^2 \frac{2\eta}{\left(1-\tilde{K}\right)^3} \\
&\overset{\Delta_n \leqslant \delta}{\leqslant} (M+\delta) (\sup h)^2 \frac{2\eta}{\left(1-\tilde{K}\right)^3} \leqslant c_M < \infty,
\end{aligned} \tag{A.41}$$

where c_M is a constant independent of n, respectively, Δ_n. We may conclude that the random variables $1_{A_n} \int_{-M}^{t} h(t-s)N_n(\mathrm{d}s)$, $n \in \mathbb{N}$, are uniformly integrable. Weak convergence together with uniform integrability yields convergence of expectations. We have established that, for M

with $-M < a$,

$$\lim_{n\to\infty} \mathbb{E}\left[1_{A_n}\int_{-M}^{t} h(t-s)N_n(\mathrm{d}s)\right] = \mathbb{E}\left[1_{A^*}\int_{-M}^{t} h(t-s)N^*(\mathrm{d}s)\right]. \tag{A.42}$$

For the proof of (A.39), we consider a truncated part of h and its remaining tail separately. For the truncated part, we use (A.42); for the tail part, we use the integrability condition $\int h\,\mathrm{d}t < 1$: for any $M > -a$, we have

$$\begin{aligned}
&\left|\sum_{k\in(a,b]} \Delta_n \mathbb{E}\left[1_{A_n}\int_{-\infty}^{k\Delta_n} h(k\Delta_n - s)N_n(\mathrm{d}s)\right] - \int_a^b \mathbb{E}\left[1_{A^*}\int_{-\infty}^{t} h(t-s)N^*(\mathrm{d}s)\right]\mathrm{d}t\right| \\
&\leqslant \left|\sum_{k\in(a,b]} \Delta_n \mathbb{E}\left[1_{A_n}\int_{-M}^{k\Delta_n} h(k\Delta_n - s)N_n(\mathrm{d}s)\right] - \int_a^b \mathbb{E}\left[1_{A^*}\int_{-M}^{t} h(t-s)N^*(\mathrm{d}s)\right]\mathrm{d}t\right| \\
&\quad + \sum_{k\in(a,b]} \Delta_n \mathbb{E}\left[1_{A_n}\int_{-\infty}^{-M} h(k\Delta_n - s)N_n(\mathrm{d}s)\right] \\
&\quad + \int_a^b \mathbb{E}\left[1_{A^*}\int_{-\infty}^{-M} h(t-s)N^*(\mathrm{d}s)\right]\mathrm{d}t.
\end{aligned} \tag{A.43}$$

Let $\varepsilon > 0$. We show that we can find M_ε and $N_\varepsilon \in \mathbb{N}$ such that each of the three summands in (A.43) is bounded by $\varepsilon/3$ for $n \geqslant N_\varepsilon$. First, consider the integrand of the last summand in (A.43). By arguing similarly to the m((a,b]) part in (A.38), we find that $\mathbb{E}\,N^*(\mathrm{d}t)/\mathrm{d}t \leqslant \eta/(1-\tilde{K})$. So we can choose $M_\varepsilon^{(1)} > 0$ so large that, for $t \in [a,b]$,

$$\begin{aligned}
&\mathbb{E}\left[1_{A^*}\int_{-\infty}^{-M_\varepsilon^{(1)}} h(t-s)N^*(\mathrm{d}s)\right] \\
&\leqslant \mathbb{E}\int_{-\infty}^{-M_\varepsilon^{(1)}} h(t-s)N^*(\mathrm{d}s) \leqslant \frac{\eta}{1-\tilde{K}}\int_{M_\varepsilon^{(1)}+a}^{\infty} h(s)\mathrm{d}s < \frac{\varepsilon}{3(b-a)}.
\end{aligned}$$

The summands of the second term in (A.43) can be bounded by $\varepsilon/(3\lceil b-a\rceil)$ in a similar and even more direct way—uniformly over n and possibly with respect to another $M_\varepsilon^{(2)} > 0$. We set $M_\varepsilon := \max\left\{M_\varepsilon^{(1)}, M_\varepsilon^{(2)}\right\}$. So taking the integral over the interval $[a,b]$ of the last summand, respectively, the Riemann sum of the second summand in (A.43), yields

$$\int_a^b \mathbb{E}\left[1_{A^*}\int_{-\infty}^{-M_\varepsilon} h_M(t-s)N^*(\mathrm{d}s)\right]\mathrm{d}t < \frac{\varepsilon}{3}, \tag{A.44}$$

respectively,

$$\sum_{k:\, k\Delta_n \in (a,b]} \Delta_n \, \mathbb{E}\left[1_{A^*} \int_{-\infty}^{-M_\varepsilon} h_M(k\Delta_n - s) N^*(\mathrm{d}s) \right] < \frac{\varepsilon}{3}. \tag{A.45}$$

For the first term in (A.43), denote

$$E_n(t) := \mathbb{E}\left[1_{A_n} \int_{-M_\varepsilon}^{t} h(t-s) N_n(\mathrm{d}s) \right] \quad \text{and} \quad E^*(t) := \mathbb{E}\left[1_{A^*} \int_{-M_\varepsilon}^{t} h(t-s) N^*(\mathrm{d}s) \right].$$

From (A.40), we already know that, for any choice of M_ε, $\lim_{n\to\infty} |E_n(t) - E^*(t)| = 0$, $t \in (a,b]$. However, the convergence of the Riemann-like sums $\sum_{k:\, k\Delta_n \in (a,b]} E_n(k\Delta_n)\Delta_n$ to the integral $\int_a^b E^*(s)\mathrm{d}s$ is nontrivial as the functions E_n are themselves part of the sequence. We write

$$\begin{aligned} &\left| \sum_{k:\, k\Delta \in (a,b]} \Delta_n E_n(k\Delta_n) - \int_a^b E^*(t)\mathrm{d}t \right| \\ &\leqslant \left| \sum_{k:\, k\Delta \in (a,b]} \Delta_n E_n(k\Delta_n) - \int_a^b E_n(t)\mathrm{d}t \right| + \left| \int_a^b E_n(t)\mathrm{d}t - \int_a^b E^*(t)\mathrm{d}t \right|. \end{aligned} \tag{A.46}$$

The second absolute difference in (A.46) converges to zero because of dominated convergence of (E_n). Indeed, for $t \in [a,b]$,

$$\begin{aligned} &\mathbb{E}\left[1_{A_n} \int_{-M_\varepsilon}^{t} h(t-s) N_n(\mathrm{d}s) \right] \\ &\leqslant \mathbb{E} \int_{-M_\varepsilon}^{b} h(t-s) N_n(\mathrm{d}s) \leqslant \sup h \sum_{k:\, k\Delta_n \in (-M_\varepsilon, b]} \mathbb{E}\, X_k^{(\Delta_n)} \leqslant \sup h \frac{\lceil M_\varepsilon + b \rceil \eta}{1 - \tilde{K}}. \end{aligned} \tag{A.47}$$

Note that $\sup h < \infty$ follows from the piecewise-continuity assumption. In view of the upper bound (A.47), we apply the dominated convergence theorem and choose $N_\varepsilon^{(1)} \in \mathbb{N}$ so large that, for the second absolute difference in (A.46), we have

$$\left| \int_a^b E_n(t)\mathrm{d}t - \int_a^b E^*(t)\mathrm{d}t \right| < \frac{\varepsilon}{6}, \quad n \geqslant N_\varepsilon^{(1)}. \tag{A.48}$$

For the first absolute difference in (A.46), we assume that, without loss of generality, the piecewise continuous function h is uniformly continuous on $(0,\infty)$. Otherwise, we note that any piecewise continuous function on $\mathbb{R}$ that is vanishing at infinity is uniformly continuous on each of its continuous pieces and do the following calculation once for every uniformly continuous

piece of h. Uniform continuity gives us a constant $\delta_h > 0$, so small, that, for any $t_0 > 0$,

$$|t - t_0| < \delta_h \ \wedge \ t > 0 \quad \Rightarrow \quad |h(t) - h(t_0)| < \frac{\varepsilon(1-\tilde{K})}{12\eta(b-a+\delta)(M_\varepsilon + b + \delta)}. \tag{A.49}$$

Now, choose $N_\varepsilon^{(2)}$ so large that

$$\Delta_n < \min\left\{\delta_h, \frac{\varepsilon(1-\tilde{K})}{12\eta(b-a+\delta)\sup h}\right\} \quad \text{for} \quad n \geqslant N_\varepsilon^{(2)}. \tag{A.50}$$

Here again, δ and $\tilde{K}$ are the constants from Lemma 16. Let $a \leqslant s < t \leqslant b$ with $t - s < \Delta_n (< \delta_h)$, then

$$\begin{aligned}
|E_n(t) - E_n(s)| &= \left| \sum_{k\Delta_n \in (-M_\varepsilon, t)} \mathbb{E}\left[1_{A_n} h(t - k\Delta_n) X_k^{(\Delta_n)}\right] - \sum_{k\Delta_n \in (-M_\varepsilon, s]} \mathbb{E}\left[1_{A_n} h(s - k\Delta_n) X_k^{(\Delta_n)}\right] \right| \\
&\leqslant \left(\sum_{k\Delta_n \in (-M_\varepsilon, s]} |h(t - k\Delta_n) - h(s - k\Delta_n)| + \sum_{k\Delta_n \in (s,t)} h(t - k\Delta_n) \right) \mathbb{E} X_0^{(\Delta_n)} \\
&\overset{(A.49)}{\leqslant} \left(\frac{M_\varepsilon + s + \Delta_n}{\Delta_n} \frac{\varepsilon(1-\tilde{K})}{12\eta(b-a+\delta)(M_\varepsilon + b + \delta)} + \sup h \right) \frac{\Delta_n \eta}{1 - K^{(\Delta)}} \\
&\leqslant \frac{\varepsilon}{12(b-a+\delta)} + \Delta_n \frac{\eta \sup h}{1 - \tilde{K}} \\
&\overset{(A.50)}{\leqslant} \frac{\varepsilon}{12(b-a+\delta)} + \frac{\varepsilon}{12(b-a+\delta)} = \frac{\epsilon}{6(b-a+\delta)}.
\end{aligned} \tag{A.51}$$

Summarizing the above calculation, we have established the existence of an $N_\varepsilon^{(2)} \in \mathbb{N}$ such that, for all $n \geqslant N_\varepsilon^{(2)}$, we have $|E_n(t) - E_n(s)| \leqslant \varepsilon/(6(b-a+\delta))$ whenever $|t-s| < \Delta_n$ and $s, t \in [a, b]$. For the first absolute difference in (A.46), we therefore get

$$\begin{aligned}
\left| \sum_{k:\, k\Delta_n \in (a,b]} \Delta_n E_n(k\Delta_n) - \int_a^b E_n(t)\mathrm{d}t \right| &\leqslant \sum_{k:\, k\Delta_n \in (a,b]} \int_{k\Delta_n}^{(k+1)\Delta_n} |E_n(k\Delta_n) - E_n(t)| \, \mathrm{d}t \\
&\leqslant \sum_{k:\, k\Delta_n \in (a,b]} \int_{k\Delta_n}^{(k+1)\Delta_n} \frac{\varepsilon}{6(b-a+\delta)} \mathrm{d}t \\
&\overset{(A.51)}{\leqslant} \frac{(b-a+\Delta_n)}{\Delta_n} \Delta_n \frac{\varepsilon}{6(b-a+\delta)} \\
&\leqslant \frac{\varepsilon}{6}, \quad n \geqslant N_\varepsilon^{(2)}.
\end{aligned} \tag{A.52}$$

Set $N_\varepsilon := \max\left\{N_\varepsilon^{(1)}, N_\varepsilon^{(2)}\right\}$. Combining (A.48) and (A.52), we get that

$$\left| \sum_{k:k\Delta\in(a,b]} \Delta_n E_n(t) - \int_a^b E^*(t)\mathrm{d}t \right| < \frac{\varepsilon}{3}, \quad n \geqslant N_\varepsilon. \tag{A.53}$$

Combining (A.44), (A.45) and (A.53), shows that (A.43) is smaller than the given ε, for $n \geqslant N_\varepsilon$ and $M := M_\varepsilon$, i.e.,

$$\lim_{n\to\infty} \sum_{k:k\Delta_n\in(a,b]} \Delta_n \,\mathbb{E}\left[1_{A_n} \int_{-\infty}^{\Delta_n} h(k\Delta_n - s)N_n(\mathrm{d}s)\right] = \int_a^b \mathbb{E}\left[1_{A^*} \int_{-\infty}^{t} h(t-s)N^*(\mathrm{d}s)\right] \mathrm{d}t.$$

We have established that the right-hand side of (A.37) also converges to the right-hand side of (A.36). With the result from Proposition 13 on the uniqueness property of (A.36), we find that every subsequential limit N^* has the same distribution as the Hawkes process N. We may then conclude that, for $\Delta \to 0$, the approximating sequence of point processes $\left(N^{(\Delta)}\right)$ converges weakly to the Hawkes process N.

□

Paper

B

Matthias Kirchner.

Hawkes forests.

Submitted.

Hawkes forests

Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

This paper starts with *multitype branching random walks*: each node of a random finite rooted tree is supplied with two labels, a type $\in \{1, 2, \ldots, d\}$ and a position $\in \mathbb{R}$. The number, the types and the positions of children nodes only depend on the type and the position of their parent node. Type and position of the root node are given. As a special case, we consider multitype branching random walks where children nodes are always positioned to the right of their parent node. We call this special case *Hawkes tree*. Growing Hawkes trees from a countable number of random immigrant points yields a multitype random forest, that we call *Hawkes forest*. By projecting such Hawkes forests on position–type space, we derive the multitype Hawkes point process as well as the multivariate integer-valued autoregressive time series (INAR). As an application, we generalize and sharpen a convergence theorem from Kirchner (2016): given a multitype Hawkes process, we construct an INAR-based sequence of point measures and show that it converges in a very strong sense to the Hawkes process. Finally, we point out how the Hawkes-forest formalism might be fertile for model building.

1 Introduction

In this paper, we observe that both processes, the multitype Hawkes point process and multivariate integer-valued autoregressive (INAR) time series, can be represented in terms of *multitype branching random walks* (BRW). These representations are conceptually fertile:

i) They allow to build a strong and elegant bridge between Hawkes and INAR processes.

ii) They suggest interesting modeling alternatives to Hawkes and INAR processes.

We give a sketch of the concepts involved: each node of a random Galton–Watson tree is supplied with two labels, namely with a type $\in [d] := \{1, 2, \ldots, d\}$ and with a position $\in \mathbb{R}$. Type and position of the root node are given. The number, the types, and the positions of children nodes only depend on the type and the position of the parent node. The *displacement distribution* controls the difference between a child-node position and the position of its parent. As a special case, we consider multitype branching random walks where the position of any (non-root) node lies right of its parent node's position, i.e., where the displacement distributions are concentrated on the positive half-line. We call this special case *Hawkes tree*. In a next step, we consider immigration point processes. Each of the (at most countably many) immigration points consists of a position $\in \mathbb{R}$ and a type $\in [d]$. The immigration points are used as root nodes of independent Hawkes trees. This collection of Hawkes trees forms a *multitype random forest*, that we call *Hawkes forest*. If the immigration process is a Poisson random measure, if the numbers of offspring follow Poisson laws, and if the displacement distributions are absolutely continuous, then the positions and types of the corresponding Hawkes forest (what we call *projected point measure*) actually form a Hawkes point process. In a similar manner, INAR time series can be derived from Hawkes forests. Furthermore, when we consider Hawkes forests with more general offspring and immigration laws, we can derive point processes or time series that might be interesting modeling alternatives to standard Hawkes and INAR models.

The branching structure of Hawkes and INAR processes is well known. In the point-process case, this branching structure has often been described since the seminal paper Hawkes and Oakes (1974). As a rule, the mathematical construction of both processes is mostly done generationwise, that is, applying an recursive branching mechanism. E.g., in Liniger (2009) and Embrechts and Kirchner (2017), the Hawkes process is formalized as a cascade of Poisson random measures. Similarly, the multivariate INAR time series in Latour (1997) is constructed generation by generation. The branching structure can also be convenient for calculations such as the derivation of generating functions in Hawkes and Oakes (1974) (for the point-process case) or Kirchner (2017a) (for the time-series case). However, to the best of our knowledge, the most general branching-random-walk representation of Hawkes and INAR processes has not been discussed anywhere as yet. One reason for this somewhat surprising gap in the literature may be the different interpretation of the objects: in the BRW world, the position label is typically interpreted as a 'space' attribute and the natural 'time' is usually the generation number. In contrast, in the Hawkes point process world, the position label itself is interpreted as 'time' and 'space' is more related to the type (or modeled by a further mark) whereas the generation number is ignored. These different interpretations may hide obvious similarities. In any case, the BRW terminology is mathematically attractive because the positions of the points and the tree structure can be treated separately. Here are the main contributions of our paper:

i) Definition of Hawkes trees and Hawkes forests as well as their projected point measures; see Definitions 7 and 14.

ii) Projected point measures of Hawkes forests are multitype point processes in the usual sense; see Proposition 15.

iii) Projected point measures of Hawkes forests can be approximated by discretized versions in an almost-sure sense as well as in L_1; see Theorem 23.

iv) Multitype Hawkes processes can be represented as projected point measures of certain Hawkes forests; see Proposition 19.

v) Multivariate INAR processes can be derived from the projected point measures of certain Hawkes forests; see Proposition 21.

vi) Multitype Hawkes processes can be approximated (almost surely and in L_1) by point processes that are based on multivariate INAR time series; see Theorem 25.

vii) Hawkes-forest terminology suggests numerous modeling alternatives to standard Hawkes or INAR models; see Remark 26.

The main body of the paper is organized as follows: in the longer second section, we give necessary terminology for trees, random trees, multitype random trees, and multitype branching random walks; we introduce Hawkes trees, Hawkes forests, and their projected point measures. In the third section, we represent multitype Hawkes processes as well as multivariate INAR processes as projected point measures of Hawkes forests; we give various convergence theorems. In the fourth section, we conclude with directions for natural alternatives to Hawkes and INAR models inside the Hawkes-forest framework. These alternatives might be relevant for model building.

2 Definitions

This section introduces the necessary notation for random trees and forests. We follow the formalism from Neveu (1986). For the multitype case, Stephenson (2016) has been helpful.

2.1 Multitype trees

Definition 1. *We identify* $\mathcal{U} := \bigcup_{g=0}^{\infty} \mathbb{N}^g$ *with the space of possible nodes, where we use the convention that* $\mathbb{N}^0 := \{\varnothing\}$.

i) For $u = (u_1, u_2, \ldots, u_g) \in \mathcal{U}\backslash\{\varnothing\}$, *the node* $u^- := (u_1, u_2, \ldots, u_{g-1}) \in \mathcal{U}$ *is the (unique)* parent *of u.*

ii) We set $|\varnothing| := 0$ *and, for* $u = (u_1, u_2, \ldots, u_g) \in \mathcal{U}\backslash\{\varnothing\}$*, we set* $|u| := g$.

iii) For $u = (u_1, u_2, \ldots, u_g)$, $v = (v_1, v_2, \ldots, v_{g'}) \in \mathcal{U}$*, we set* $uv := (u_1, u_2, \ldots, u_g, v_1, v_2, \ldots, v_{g'})$. *In particular,* $u\varnothing = \varnothing u = u$, $u \in \mathcal{U}$.

iv) For $u, v \in \mathcal{U}$*, we write* $u \leqslant v$ ($u < v$) *if there exists* $w \in \mathcal{U}$ ($w \in \mathcal{U}\backslash\varnothing$)*, such that* $uw = v$.

Trees are special subsets of $\mathcal{U}$. For the purpose of this paper, we only consider rooted finite trees:

Definition 2. *A* rooted, ordered, and finite tree $\mathbf{t}$ *is a finite subset of* $\mathcal{U}$ *such that*

i) $\varnothing \in \mathbf{t}$*, that is, the tree is rooted;*

ii) for all $u \in \mathbf{t}\backslash\{\varnothing\}$, $u^- \in \mathbf{t}$*, that is, the parent of any node in* $\mathbf{t}$ *(but the rootnode) can also be found in* $\mathbf{t}$*; and,*

iii) for all $u \in \mathbf{t}$*, there exists a* $n_u \in \mathbb{N}_0$ *in such a way that, for all* $n \in \mathbb{N}$*, we have that* $un \in \mathbf{t} \Leftrightarrow n \leqslant n_u$*, that is, the offspring of any node is finite and 'ordered without gaps'.*

We denote the set of rooted, ordered, and finite trees by $\mathbb{T}_f$.

In the remainder of the paper, we say 'tree' but we actually mean 'rooted, ordered, and finite tree'. Next, we assign each node of a tree to one of a finite number of types.

Definition 3. *For* $d \in \mathbb{N}$ *and* $[d] := \{1, 2, \ldots, d\}$*, a* d-type tree *is a tuple* $(\mathbf{t}, \mathbf{l})$ *with* $\mathbf{t} \in \mathbb{T}_f$ *and* $\mathbf{l} : \mathbf{t} \to [d]$. *We denote the space of (rooted, ordered, and finite)* d*-type trees by* $\mathbb{T}_f^{[d]}$. *For* $(\mathbf{t}, \mathbf{l}) \in \mathbb{T}_f^{[d]}$*, we define the* offspring-type functions

$$\mathbf{w}_{(\mathbf{t},\mathbf{l})}(u) : \mathbf{t} \to \mathcal{W}_d, \quad u \mapsto \big(\mathbf{l}(u1), \mathbf{l}(u2), \ldots, \mathbf{l}(un_u)\big),$$

where

$$\mathcal{W}_d := \bigcup_{n=0}^{\infty} [d]^n \tag{2.1}$$

denotes the set of all possible offspring-type vectors. Note that $\mathbf{w}_{(\mathbf{t},\mathbf{l})}(u)$ *specifies the types of the* $n_u \in \mathbb{N}_0$ *(ordered) offspring nodes* $u1, u2, \ldots, un_u$ *of* u *in* $\mathbf{t}$*. For any offspring-type vector* $w \in \mathcal{W}_d$*, we define the* type counts

$$n_j : \mathcal{W}_d \to \mathbb{N}_0^d, \quad \mathbf{w} = (w_1 w_2 \ldots w_n) \mapsto \#\Big\{k : k \in \{1, 2, \ldots, n\}, w_k = j\Big\}, \quad j \in [d]. \tag{2.2}$$

Finally, these type counts are collected in the type-count vector

$$\mathbf{n}(\mathbf{w}) := \big(n_1(\mathbf{w}), n_2(\mathbf{w}), \ldots, n_d(\mathbf{w})\big) \in \mathbb{N}_0^d\,.$$

Note that $n_j\big(\mathbf{w}_{(\mathbf{t},\mathbf{l})}(u)\big)$ gives the number of type-j offspring nodes of node u in the tree $(\mathbf{t},\mathbf{l})$. Also note that, in (2.1), we again apply the convention $[d]^0 = \{\varnothing\} \subset \mathcal{W}_d$. So if the the total number of offspring of some node $u \in \mathbf{t}$ is zero ($n_u = 0$), then $\mathbf{n}\big(\mathbf{w}_{(\mathbf{t},\mathbf{l})}(u)\big) = \mathbf{n}(\varnothing) = (0,\dots,0) \in \mathbb{N}_0^d$ is still well defined.

2.2 Random d-type trees

Definition 4. *Any probability distribution on $\mathcal{W}_d$ as in* (2.1) *is an* ordered offspring distribution. *Let $\nu = (\nu_i)_{i\in[d]}$ be a d-tuple of ordered offspring distributions. ν induces* offspring distributions *$\mu = (\mu_i)_{i\in[d]}$ on $\mathbb{N}_0^d$ by*

$$\mu_i(\mathbf{k}) := \sum_{\substack{\mathbf{w}\in\mathcal{W}_d:\\ n_j(\mathbf{w})=k_j,\ j\in[d]}} \nu_i(\mathbf{w}), \quad \mathbf{k} := (k_1,k_2,\dots,k_d) \in \mathbb{N}_0^d, \quad i \in [d]. \tag{2.3}$$

Furthermore, we define the branching matrix *of ν, respectively, μ as*

$$M := (m_{i,j})_{(i,j)\in[d]^2}, \quad \textit{with} \quad m_{i,j} := \sum_{k\in\mathbb{N}} k \sum_{\substack{\mathbf{k}\in\mathbb{N}^d:\\ k_j=k}} \mu_i(\mathbf{n}) \,(\leqslant \infty), \quad (i,j) \in [d]^2.$$

Before constructing random trees formally, we provide the reader with an interpretation of the quantities introduced:

i) The values of the ordered offspring distribution, $\nu_i(\mathbf{w}), \mathbf{w} = (w_1,w_2,\dots,w_n) \in \mathcal{W}_d$, will indicate the probability for a type-i node $(u,i) \in \mathcal{U}\times[d]$ to have n children, namely $(u1,w_1),(u2,w_2),\dots,(un,w_n)$.

ii) The values of the unordered offspring distribution, $\mu_i(\mathbf{k}), \mathbf{k} = (k_1,k_2,\dots,k_d) \in \mathbb{N}_0^d$, will indicate the probability for a type-i node to have k_1 type-1 children, k_2 type-2 children, …, and k_d type-d children.

iii) The branching-matrix entries, $m_{i,j}\,(\leqslant \infty)$, will indicate the expected number of type-j children of a type-i node.

iv) The row sums of the branching matrix, $\sum_{j=1}^d m_{i,j}\,(\leqslant \infty)$, will indicate the expected total number of children of a type-i node.

We only consider subcritical offspring distributions:

Definition 5. *A matrix $M \in \mathbb{R}^{d\times d}$ is* subcritical *if its spectral radius*

$$\rho(M) := \max\big\{|\lambda| :\ \lambda \textit{ eigenvalue of } M\big\}, \tag{2.4}$$

is strictly less than 1. (Ordered) offspring distributions are subcritical if the associated branching matrix is subcritical.

Lemma 6. *For any subcritical matrix $M \in \mathbb{R}_{\geqslant 0}^{d\times d}$, we have that $(1_{d\times d} - M)$ is invertible and that*

$$1_{d\times d} + M + M^2 + \cdots + M^g \overset{g\to\infty}{\longrightarrow} (1_{d\times d} - M)^{-1},$$

for any matrix norm.

Proof. For a proof, see for instance Watson (2015). □

Given subcritical ordered offspring distributions, we define a random subcritical multitype tree as follows:

Definition 7. *Let $\nu = (\nu_i)$ be subcritical ordered offspring distributions and $i_0 \in [d]$. A $\mathbb{T}_f^{([d])}$-valued random variable $(\mathbf{T}, \mathbf{L})$ with distribution*

$$\mathbb{P}_{i_0,\nu}\Big[(\mathbf{T},\mathbf{L}) = (\mathbf{t},\mathbf{l})\Big] := 1_{\{\mathbf{l}(\varnothing)=i_0\}} \prod_{u\in\mathbf{t}} \nu_{\mathbf{l}(u)}(\mathbf{w}_{(\mathbf{t},\mathbf{l})}(u)), \quad (\mathbf{t},\mathbf{l}) \in \mathbb{T}_f^{([d])}, \tag{2.5}$$

is a subcritical random d-type (i_0,ν)-tree*; in short, an (i_0,ν)-tree.*

An (i_0,ν)-tree is constructed recursively: the root node is given; it is labeled with some deterministic type $i_0 \in [d]$. The offspring vector $(u1, w_1), (u2, w_2), \ldots, (un_u, w_{n_u})$ of a given type-i node $u \in \mathcal{U}$ follows the ordered offspring distribution ν_i—and this offspring is generated independently of everything else. Formally:

Construction 8. *Let $\nu = (\nu_i)$ be subcritical ordered offspring distributions and $i_0 \in [d]$. Let*

$$\mathbf{W}^{(u,i)} = (W_1^{(u,i)}, \ldots, W_{N^{(u,i)}}^{(u,i)}) \sim \nu_i, \quad (u,i) \in \mathcal{U} \times [d], \tag{2.6}$$

be mutually independent $\mathcal{W}_d$-valued random variables. Define a random (i_0,ν)-tree $(\mathbf{T},\mathbf{L})$ generationwise as a function of (some of the) $\{\mathbf{W}^{(u,i)}\}$: the only node in generation 0 is the root node $\varnothing$; it is of type i_0:

$$\mathbf{T}^{(0)} := \{\varnothing\}, \quad \mathbf{L}^{(0)} : \mathbf{T}^{(0)} \to [d],\ \varnothing \mapsto i_0.$$

Each type-i node u in generation $(g-1)$ generates $N^{(u,i)}$ children in generation g:

$$\mathbf{T}^{(g)} := \bigcup_{u\in\mathbf{T}^{(g-1)}} \bigcup_{k=1,\ldots,N^{(u,\mathbf{L}^{(g-1)}(u))}} uk \tag{2.7}$$

The types of the offspring nodes uk in (2.7) *are determined by*

$$\begin{aligned}\mathbf{L}^{(g)} : \mathbf{T}^{(g)} &\to [d],\\ u &\mapsto W_k^{(u^-, \mathbf{L}_{g-1}(u^-))}, \quad \text{with } k \text{ s.t. } u = u^- k.\end{aligned}$$

Finally, define the random variable $(\mathbf{T}, \mathbf{L})$ *by*

$$\mathbf{T} := \bigcup_{g=0}^{\infty} \mathbf{T}^{(g)} \quad \text{and} \quad \mathbf{L} : \mathbf{T} \to [d], u \mapsto \mathbf{L}^{(|u|)}(u).$$

Proposition 9. *For* $(\mathbf{T}, \mathbf{L})$*, a (subcritical)* (i_0, ν)*-tree as in Definition 7, we have that*

$$\mathbb{E}_{i_0,\nu} \#\{u \in \mathbf{T}^{(g)} : \mathbf{L}(u) = j\} = m_{i_0,j}^{(g)} < \infty, \quad (i_0, j) \in [d]^2, \text{ and} \tag{2.8}$$

$$\mathbb{E}_{i_0,\nu} \#\{u \in \mathbf{T} : \mathbf{L}(u) = j\} = b_{i,j} < \infty, \quad (i_0, j) \in [d]^2, \tag{2.9}$$

where—as in Definition 5—$M = (m_{i,j})_{(i,j)\in[d]^2}$ *denotes the branching matrix of* ν*,* $(m_{i,j}^{(g)})_{(i,j)\in[d]^2} := M^g$*, and* $(b_{i,j})_{(i,j)\in[d]^2} := \sum_{g\geqslant 0} M^g = (1_{d\times d} - M)^{-1}$*. In particular, Construction 8 terminates with almost surely, the resulting tree* $(\mathbf{T}, \mathbf{L})$ *lives in* $\mathbb{T}_f^{[d]}$*, and has distribution* (2.5)*.*

Proof. Equation (2.8) follows by induction. Equation (2.9) follows from (2.8) together with Fubini's Theorem and Lemma 6. □

Note that the distribution of the number of children $N^{(u,i)}$ of a given node u in Construction 8 is also specified by the offspring distribution ν_i. Also note that, in general, the number of children $N^{(u,i)}$ is not necessarily independent of $W_k^{(u,i)}$, the type of the k-th child. However, in our main example of a random d-type tree, the Poisson tree, this independency holds:

Definition 10. *Let* $i_0 \in [d]$ *and let* $M = (m_{i,j}) \in \mathbb{R}_{\geqslant 0}^{d\times d}$ *be a (subcritical) matrix as in Definition 5. An* (i_0, M)-Poisson tree *is an* (i_0, ν)*-tree as in Definition 7 with ordered offspring distributions* $\nu = (\nu_i)$ *of the form*

$$\nu_i(\mathbf{w}) := \frac{n_1(\mathbf{w})!\, n_2(\mathbf{w})! \cdots n_d(\mathbf{w})!}{|\mathbf{w}|!} \prod_{j\in[d]} \nu_{i,j}\big(n_j(\mathbf{w})\big), \quad \mathbf{w} \in \mathcal{W}_d, i \in [d], \tag{2.10}$$

where $\nu_{i,j} = \mathrm{Pois}(m_{i,j})$*,* $(i, j) \in [d]^2$*, and the functions* n_j *denote the type counts from* (2.2)*.*

Note that by (2.10), in a Poisson tree, each type-i node u generates $\mathrm{Pois}(m_{i,j})$ type-j children nodes (independently over $j \in [d]$). Or, using the notation from Construction 8, each type-i node u generates a total of $N^{(u,i)} \sim \mathrm{Pois}(\sum_{j=1}^d m_{i,j})$ children nodes. Each of these children nodes uk, $k = 1, 2 \ldots, N^{(u,i)}$, is assigned to a type $j \in [d]$ with probability $m_{i,j}/\sum_{l=1}^d m_{i,l}$—independently of the other nodes and independently of $N^{(u,i)}$.

2.3 Branching random walks

Next, each node u in a tree $\mathbf{t} \in \mathbb{T}_f$ is supplied with *two* labels, namely, as before, with a type $\mathbf{l} : \mathbf{t} \to [d]$ and, in addition, with a position $\mathbf{s} : \mathbf{t} \to \mathbb{R}$. We denote the space of $([d] \times \mathbb{R})$-labeled rooted, ordered, and finite trees by $\mathbb{T}_f^{([d]\times\mathbb{R})}$. We consider a special distribution on $\mathbb{T}_f^{([d]\times\mathbb{R})}$, where the positions $\mathbf{s}$ along the genealogical lines of the tree behave like a regime-switching random walk with the regimes depending on the actual node type.

Definition 11. *Let $F = (F_{i,j})_{(i,j)\in[d]^2}$ be a matrix of* displacement distributions *on $\mathbb{R}$ and let*

$$Y_u^{(i,j)} \sim F_{i,j}, \quad u \in \mathcal{U},\ (i,j) \in [d]^2,$$

be mutually independent random variables. Furthermore, let $(\mathbf{T}, \mathbf{L})$ be a (ν, i_0)-tree (independent of the $Y_u^{(i,j)}$) as in Definition 7. Given $(\mathbf{T}, \mathbf{L})$, we define the position $\mathbf{S}$ as

$$\mathbf{S} : \mathbf{T} \to \mathbb{R}, \quad u \mapsto \sum_{\varnothing < v \leqslant u} Y_v^{(\mathbf{L}(v^-),\mathbf{L}(v))}. \tag{2.11}$$

*Any $\mathbb{T}_f^{([d]\times\mathbb{R})}$-valued random variable with the same distribution as $(\mathbf{T}, \mathbf{L}, \mathbf{S})$ is an (i_0, ν, F)-*subcritical d-type branching random walk*; in short, (i_0, ν, F)-*BRW. *We call $(\mathbf{T}, \mathbf{L})$ the* underlying tree *of the BRW $(\mathbf{T}, \mathbf{L}, \mathbf{S})$. If the displacement distributions are all chosen in such a way that $F_{i,j}(0) = 0$, $(i,j) \in [d]^2$, then we call $(\mathbf{T}, \mathbf{L}, \mathbf{S})$ an (i_0, ν, F)-*Hawkes tree.

Note that from (2.11), we get that the position of the root node $\varnothing$ is $\mathbf{S}(\varnothing) = 0$. Also note that a more technical and obvious choice-of-name for the Hawkes tree might be 'branching renewal process'—mimicking the naming 'branching random walk'. But as the main goal of our considerations is a BRW-discussion of the well-known Hawkes process, we have chosen the term Hawkes tree as more appropriate. Finally, note that, for the distribution $\mathbb{P}_{i_0,\nu,F}$ of an (i_0, ν, F)-BRW $(\mathbf{T}, \mathbf{L}, \mathbf{S})$, we obtain the factorization

$$\begin{aligned}&\mathbb{P}_{i_0,\nu,F}\Big[(\mathbf{T}, \mathbf{L}, \mathbf{S}) \in \{\mathbf{t}\} \times \{\mathbf{l}\} \times \{\mathbf{s} : \mathbf{t} \to \mathbb{R},\ \mathbf{s}(u) \in B_u,\ u \in \mathbf{t}\}\Big] \\ &= \mathbb{P}_{i_0,\nu}\Big[(\mathbf{T}, \mathbf{L}) = (\mathbf{t}, \mathbf{l})\Big]\, \mathbb{P}_F\Big[\mathbf{S}(u) \in B_u : u \in \mathbf{t}\Big], \quad (\mathbf{t}, \mathbf{l}) \in \mathbb{T}_f^{([d]\times\mathbb{R})},\ B_u \in \mathcal{B}(\mathbb{R}),\ u \in \mathbf{t},\end{aligned} \tag{2.12}$$

where the first factor can be calculated as in (2.5) and the second factor is a probability determined by the choice of the displacement distributions F. Equation (2.12) highlights the mathematical attractivity of the BRW approach: the position labels and the underlying trees can be treated separately. Next, we consider a special and particularly simple example of a branching random walk:

Definition 12. *Let $i_0 \in [d]$ and let $M \in \mathbb{R}_{\geqslant 0}^{d\times d}$ be a subcritical matrix. An (i_0, M, F)-*Poisson

branching random walk *is a BRW with displacement distributions F such that the underlying tree is an (i_0, M)-Poisson tree as in Definition 10.*

In the following, we consider countably many $[d]$-labeled immigrant points in $\mathbb{R}$. Each i_0-type immigrant serves as a root of a (i_0, ν)-BRW. The resulting object is a random forest. We pay special attention to the so-called projected point measure of a random forest—only considering the position of all points and disregarding the branching structure as well as the tree membership. We use the terminology of point processes:

Definition 13. *A* point process *N is a measurable mapping from some probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ into $(M_p, \mathcal{M}_p)$ where M_p denotes the space of locally finite counting measures on the bounded Borel sets $\mathcal{B}_b(\mathbb{R})$ in $\mathbb{R}$, and*

$$\mathcal{M}_p := \sigma\Big(\big\{m \in M_p : \ m(A) = n\big\} : \ A \in \mathcal{B}_b(\mathbb{R}), n \in \mathbb{N}_0\Big).$$

We write

$$\int_{-\infty}^{b} f(t)N(\mathrm{d}t) := \sum_{\substack{t \leqslant b: \\ N(\{t\})>0}} f(t)N(\{t\}) \quad (\leqslant \infty),$$

*for any measurable non-negative function f. A d-*type point process *$\mathbf{N}$ is a measurable mapping from some probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ into $(M_p^{([d])}, \mathcal{M}_p^{([d])})$ where*

$$M_p^{[d]} = \big\{m : \ \mathcal{B}_b(\mathbb{R}) \otimes 2^{[d]} \to \mathbb{N}_0, \ \text{s.t. } m(\cdot \times \{i\}) \in M_p, \ i \in [d]\big\}$$

and

$$\mathcal{M}_p^{[d]} := \sigma\Big(\big\{m \in M_p^{[d]} : \ m(A \times \{i\}) = n\big\} : \ A \in \mathcal{B}_b(\mathbb{R}), i \in [d], n \in \mathbb{N}_0\Big).$$

A d-type point process $\mathbf{N}$ is stationary *if $\mathbb{E}\,\mathbf{N}(A \times [d]) < \infty$, $A \in \mathcal{B}_b(\mathbb{R})$, and, for any $k \in \mathbb{N}$ and any $t \in \mathbb{R}$, $(A_l, I_l) \in \mathcal{B}_b(\mathbb{R}) \otimes 2^{[d]}$, $n_l \in \mathbb{N}$, $l = 1, 2, \ldots, k$,*

$$\mathbb{Q}\Big(\mathbf{N}(A_l \times I_l) = n_l, \ l = 1, 2, \ldots, k\Big) = \mathbb{Q}\Big(\mathbf{N}((A_l + t) \times I_l) = n_l, \ l = 1, 2, \ldots, k\Big).$$

We say that a d-type point process $\mathbf{N}$ is simple *if $\mathbb{Q}\big(\mathbf{N}(\{t\} \times [d]) \in \{0, 1\}, t \in \mathbb{R}\big) = 1$. We call $Q := \mathbb{Q} \circ \mathbf{N}^{-1}$ the* law *of the point process.*

Note that any d-type point process $\mathbf{N} : (\Omega, \mathcal{F}, \mathbb{Q}) \to (M_p^{([d])}, \mathcal{M}_p^{([d])})$ can be represented in the form

$$\big\{(T_k, L_k)\big\}_{k \in \mathcal{K}}, \quad L_k \in [d], T_k \in \mathbb{R},$$

for some indexing set $\mathcal{K} \subset \mathbb{Z}$. If we pick the indices such that $T_0 := \min\{t : \mathbf{N}(\{t\} \times [d]) > 0\}$ and such that $T_k \leqslant T_{k+1}$, for $k, k+1 \in \mathcal{K}$, then this representation is unique. Obviously, (T_k, L_k) are random variables on $(\Omega, \mathcal{F}, \mathbb{Q})$

Next, we add immigration to branching random walks:

Definition 14. *Let $\{(T_k, L_k)\}_{k\in\mathcal{K}\subset\mathbb{Z}}$ be a d-type point process with* immigration law Q. *For subcritical ordered offspring distributions $\nu = (\nu_i)_{i\in[d]}$ and displacement distributions $F = (F_{i,j})_{(i,j)\in[d]^2}$, we define a (Q,ν,F)-*forest $\mathbf{F}$ *as the random set*

$$\mathbf{F} := \Big\{ \big(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k, (T_k, L_k)\big) : k \in \mathcal{K} \Big\}, \tag{2.13}$$

*in such a way that, conditional on the immigration types $(L_k)_{k\in\mathcal{K}}$, the $(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k)$ are independent (L_k,ν,F)-branching random walks as in Definition 11. If $F_{i,j}(0) = 0$, $(i,j) \in [d]^2$, we call the forest a (Q,ν,F)-*Hawkes forest. *Furthermore, given a forest $\mathbf{F}$ as in* (2.13)*, we consider its* projected d-type point measures

$$\mathbf{N}_{\mathbf{F}}\big(A \times \{j\}\big) := \sum_{g\geqslant 0} \mathbf{N}_{\mathbf{F}}^{(g)}\big(A \times \{j\}\big), \quad A \in \mathcal{B}(\mathbb{R}),\ j \in [d], \tag{2.14}$$

with

$$\begin{aligned}&\mathbf{N}_{\mathbf{F}}^{(g)}(A \times \{j\})\\ &:= \sum_{k\in\mathbb{Z}} \#\Big\{ u \in \mathbf{T}_k^{(g)} : \big(T_k + \mathbf{S}_k(u)\big) \in A,\ L_k(u) = j \Big\}, \quad A \in \mathcal{B}(\mathbb{R}),\ g \in \mathbb{N}_0,\ j \in [d],\end{aligned}$$

where $\mathbf{T}_k^{(g)}$ collects the nodes u of the BRW with $|u| = g$; see Construction 8.

Note that $\mathbf{N}^{(0)}$ coincides with the immigration point process. Also note that the projected point measures are much less informative than the original forest: each point in a random forest carries information about its type, its position, the tree it belongs to, its line of ancestors inside this tree as well as its line of offspring. In contrast, the projected point measures only see the positions and types of the nodes; the branching structure gets lost—not even the immigration points (=root nodes) can be distinguished! One can show

Proposition 15. *The projected counting measures in Definition 14 are well-defined random variables. In other words, $\mathbf{N}_{\mathbf{F}}$ of the form as in* (2.14) *is a d-type point process in the sense of Definition 13.*

For local finiteness, one uses that the immigration point process is locally finite by Definition 13 and that the total number of nodes of any tree is almost surely finite by Proposition 9. For measurability, one uses repeatedly that the sum of point processes is again a point process. Note that the immigration law Q together with the offspring distributions ν and the displacement distributions F specify the finite-dimensional distributions and then the distribution of $\mathbf{N}_{\mathbf{F}}$.

Proposition 16. *Let* $\mathbf{F}$ *be a random forest with immigration point process* $\mathbf{N}_{imm}$ *such that, for all* $j \in [d]$, $\mathbb{E}\,\mathbf{N}_{imm}(A \times \{j\}) = \eta_j |A|$, $A \in \mathcal{B}(\mathbb{R})$, *for some* $\eta_j \in \mathbb{R}_{\geqslant 0}$. *Then we have that, for* $j \in [d]$,

$$\mathbb{E}\,\mathbf{N}_{\mathbf{F}}^{(g)}\big(A \times \{j\}\big) = \eta m_{\cdot,j}^{(g)} |A|, \quad A \in \mathcal{B}_b(\mathbb{R}), g \in \mathbb{N}_0, \tag{2.15}$$

and

$$\mathbb{E}\,\mathbf{N}_{\mathbf{F}}\big(A \times \{j\}\big) = \eta b_{\cdot,j} |A|, \quad A \in \mathcal{B}_b(\mathbb{R}), \tag{2.16}$$

where $\eta = (\eta_1, \ldots, \eta_d)$, $M = (m_{i,j})_{(i,j)\in[d]^2}$ *denotes the (subcritical) branching matrix of the underlying random trees,* $(m_{i,j}^{(g)})_{(i,j)\in[d]^2} := M^g$, *and* $(b_{i,j})_{(i,j)\in[d]^2} := \sum_{g\geqslant 0} M^g$ *as in Lemma 6.*

Proposition 16 follows from straightforward—if lengthy—calculations. Note that, in particular, the expectations in (2.15) and (2.16) do not depend on the displacement distributions F. An important example for Hawkes forests in this paper are Hawkes–Poisson forests:

Definition 17. *Let* M *be a subcritical branching matrix and let* $F = (F_{i,j})_{(i,j)\in[d]^2}$ *be displacement distributions on* $\mathbb{R}$. *A* (Q, M, F)-Poisson random forest *is a random forest such that the trees in Definition 14 are independent* (L_k, M, F)*-Poisson BRWs as in Definition 12. A Poisson forest with* $F_{i,j}(0) = 0$, $(i, j) \in [d]^2$, *is a* Hawkes–Poisson forest.

3 Applications

In this section, we derive d-type Hawkes point processes as well as multivariate INAR(∞) time series from Hawkes–Poisson forests. In particular, we give several approximation results.

3.1 Example: d-type Hawkes processes

For any d-type point process $\mathbf{N} : (\Omega, \mathcal{F}) \to (M_p^{[d]}, \mathcal{M}_p^{[d]})$ as in Definition 13, let $(\mathcal{H}_t^{(\mathbf{N})})_{t\in\mathbb{R}}$ be its intrinsic history, that is, let

$$\mathcal{H}_t^{(\mathbf{N})} := \sigma\Big(\big\{\omega \in \Omega : \mathbf{N}(A \times \{i\}) = k\big\} : A \in \mathcal{B}_b\big((-\infty, t]\big), k \in \mathbb{N}, i \in [d]\Big), \quad t \in \mathbb{R}.$$

Note that, by Definition 13 of a d-type point process, the generating sets of the intrinsic history are elements of $\mathcal{F}$ so that $\mathcal{H}_t^{(\mathbf{N})} \subset \mathcal{F}$, $t \in \mathbb{R}$.

Definition 18. *A d-type Hawkes process is a simple d-type point process* $\mathbf{N}$ *that solves*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[\mathbf{N}\big((t, t+\delta] \times \{j\}\big) \middle| \mathcal{H}_t^{(\mathbf{N})}\right]}{\delta} = \eta_j + \sum_{i=1}^{d} \int_{-\infty}^{t} m_{i,j} w_{i,j}(t-s) \mathbf{N}(\mathrm{d}s \times \{i\}), \quad j \in [d], t \in \mathbb{R}. \tag{3.1}$$

Here, for $(i, j) \in [d]^2$,

$$w_{i,j} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}, \quad \text{such that } \int w_{i,j}(t)\mathrm{d}t = 1 \text{ and } w_{i,j}(t) = 0, t < 0,$$

are displacement densities*,* $h_{i,j} := m_{i,j} w_{i,j}$ *are* reproduction intensities *with* $m_{i,j} \geqslant 0$ *such that the* branching matrix $M := (m_{i,j})$ *is subcritical. Furthermore,* $\eta = (\eta_1, \eta_2, \ldots, \eta_d) \in \mathbb{R}^d_{\geqslant 0}$ *are (constant)* immigration intensities*.*

Proposition 19. *Let w, M and* η *as in Definition 18. Let* Q_η *denote the law of a d-type immigration point process, such that the immigration points stem from a Poisson random measure with constant intensity* $\sum_{i=1}^d \eta_i$ *and each immigration point* T_k *is independently supplied with type* L_k *with probability* $\mathbb{P}[L_k = i_0] = \eta_{i_0} / \sum_{i=1}^d \eta_i$, $i_0 \in [d]$*. Define displacement distributions* $F_{i,j}(t) := \int_{-\infty}^t w_{i,j}(t), t \in \mathbb{R}$, $(i, j) \in [d]^2$*. Then the projected random point measure* $\mathbf{N}_\mathbf{F}$ *of the* (Q_η, M, F)*-Hawkes–Poisson forest* $\mathbf{F}$*,*

$$\mathbf{N}_\mathbf{F}\big(A \times \{j\}\big) := \sum_{k \in \mathbb{Z}} \#\big\{u \in \mathbf{T}_k : \mathbf{L}_k(u) = j,\, T_k + \mathbf{S}_k(u) \in A\big\}, \quad A \in \mathcal{B}_b(\mathbb{R}),\, j \in [d],$$

is a d-type Hawkes process with reproduction intensities $h_{i,j}$, $(i, j) \in [d]^2$*, and immigration intensities* η_i, $i \in [d]$*.*

Proof. By Proposition 15, $\mathbf{N}_\mathbf{F}$ is a d-type point process. Furthermore, because of absolute continuity of the immigration intensity and of the reproduction intensities, $\mathbf{N}_\mathbf{F}$ is a *simple* point process. Straightforward—if lengthy—calculations show that, for $a < b$ and $A \in \mathcal{H}_a^{\mathbf{N}_\mathbf{F}}$,

$$\begin{aligned}\mathbb{E}\left[1_A \mathbf{N}_\mathbf{F}\big((a, b] \times \{j\}\big)\right] &= \mathbb{E}\left[1_A \sum_{k \in \mathbb{Z}} \sum_{g \geqslant 0} \#\big\{u \in \mathbf{T}_k^{(g)} : \mathbf{L}_k(u) = j,\, T_k + \mathbf{S}_k(u) \in (a, b]\big\}\right] \\ &= \sum_{i \in [d]} \mathbb{E}\left[1_A \int_a^b \eta_j + \int_{-\infty}^t m_{i,j} w(t-s) \mathbf{N}_\mathbf{F}(\mathrm{d}s \times \{j\})\mathrm{d}t\right].\end{aligned}$$

From here, we can conclude that $\mathbf{N}_\mathbf{F}$ solves (3.1). So $\mathbf{N}_\mathbf{F}$ is a Hawkes process in the sense of Definition 18.

□

3.2 Multivariate INAR(∞) processes

Not only Hawkes processes but also their discrete-time analogue, integer-valued autoregressive time series (INAR), can be derived from Hawkes forests. We consider multivariate INAR(∞) time series as a generalization of both, of the multivariate INAR(p) ($p < \infty$) process from Latour (1997) and of the univariate INAR(∞) process from Kirchner (2016).

Definition 20. *Let $d \in \mathbb{N}$, $\alpha_{0,j} \geqslant 0$, $j \in [d]$, and $\alpha_{i,j,k} \geqslant 0$, $(i,j,k) \in [d]^2 \times \mathbb{N}$, such that the matrix $M = (m_{i,j})_{(i,j)\in[d]^2} := (\sum_{k=1}^{\infty} \alpha_{i,j,k})_{(i,j)\in[d]^2}$, is subcritical. Let*

$$(\mathbf{X}_n)_{n\in\mathbb{Z}} = \left((X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(d)})\right)_{n\in\mathbb{Z}}$$

be a sequence of $\mathbb{N}_0^d$-valued random variables, solving

$$X_n^{(j)} = \varepsilon_n^{(j)} + \sum_{i=1}^{d} \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}^{(i)}} \xi_{n,l}^{(i,j,k)}, \quad j \in [d], n \in \mathbb{Z}, \tag{3.2}$$

for independent random variables

$$\left\{\varepsilon_n^{(j)},\, \xi_{n,l}^{(i,j,k)} :\ (i,j) \in [d]^2, k \in \mathbb{N}, l \in \mathbb{N}, n \in \mathbb{Z}\right\}, \tag{3.3}$$

with

$$\xi_{n,l}^{(i,j,k)} \sim \mathrm{Pois}(\alpha_{i,j,k}) \quad \textit{and} \quad \varepsilon_n^{(j)} \sim \mathrm{Pois}(\alpha_{0,j}).$$

The random sequence $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ is a multivariate integer-valued autoregressive (INAR) time series. *We call $\alpha_{0,j} \geqslant 0, j \in [d]$,* immigration coefficients *and $\alpha_{i,j,k} \geqslant 0$, $(i,j,k) \in [d]^2 \times \mathbb{N}$,* reproduction coefficients.

One can show that (3.2) uniquely specifies a distribution on $\{(\mathbf{x}_n)_{n\in\mathbb{Z}} :\ \mathbf{x}_n \in \mathbb{N}_0\}$. We derive a multivariate INAR(∞) process $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ as in Definition 20 from a Hawkes forest:

Proposition 21. *Let $\alpha_{0,i} \geqslant 0$, $i \in [d]$, and $\alpha_{i,j,k} \geqslant 0$, $(i,j,k) \in [d]^2 \times \mathbb{N}$ such that $M = (m_{i,j})_{(i,j)\in[d]^2}$, with $m_{i,j} := \sum_{k\geqslant 1} \alpha_{i,j,k}$, is subcritical. Consider immigration points*

$$\left\{(T_k, L_k)\right\}_{k\in\mathbb{Z}},$$

from a d-type point process with law Q, such that

i) $Q\left[T_k \in \mathbb{Z} :\ k \in \mathbb{Z}\right] = 1$,

ii) $\#\left\{k \in \mathbb{Z} :\ T_k = n\right\} \overset{\text{iid}}{\sim} \mathrm{Pois}\left(\sum_{i=1}^{d} \alpha_{0,i}\right)$, $n \in \mathbb{Z}$, *and*

iii) $L_k, k \in \mathbb{Z}$, *i.i.d., with*

$$Q[L_k = i_0] = \frac{\alpha_{0,i_0}}{\sum_{i=1}^{d} \alpha_{0,i}}, \quad i_0 \in [d].$$

Furthermore, define (discrete) displacement distributions $F = (F_{i,j})$ *by*

$$F_{i,j}(t) := \sum_{k=1}^{\lfloor t \rfloor} \alpha_{i,j,k}/m_{i,j}, \quad (i,j) \in [d]^2,\ t \in \mathbb{R}\,.$$

Let $\mathbf{F}$ *be a* (Q, M, F)*-Hawkes–Poisson forest as in Definition 17 and let* $\mathbf{N}_{\mathbf{F}}$ *be the corresponding projected point measure. Then*

$$(\mathbf{X}_n)_{n\in\mathbb{Z}} := \big((X_{n,1}, X_{n,2}, \ldots, X_{n,d})\big)_{n\in\mathbb{Z}}$$

with

$$X_{n,i} = \mathbf{N}_{\mathbf{F}}(\{n\} \times \{i\}) \quad n \in \mathbb{Z},\ i \in [d],$$

solves (3.2) *(for some set of random variables as in* (3.3)*). In other words,* $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ *defines a multivariate INAR(*∞*) process with immigration coefficients* $\alpha_{0,j}$, $j \in [d]$, *and* reproduction coefficients $\alpha_{i,j,k}$, $(i,j,k) \in [d]^2 \times \mathbb{N}$.

The proof follows from straightforward calculations using the following well-known fact:

Lemma 22. *Let* $\alpha_k \geqslant 0$, $k \in \mathbb{N}$, *such that* $m := \sum_{k\geqslant 1} \alpha_k < \infty$, *and let* $\{K, Y_l : l \in \mathbb{N}\}$ *be mutually independent random variables with* $K \sim \mathrm{Pois}(m)$ *and* $\mathbb{P}[Y_l = k] = \alpha_k/m$, $l \in \mathbb{N}$. *Then*

$$\xi_k := \#\{Y_l = k : l = 1, \ldots, K\}, \quad k \in \mathbb{N},$$

are mutually independent Pois(α_k*) variables. (Note that* $K = 0 \Rightarrow \xi_k = 0$, $k \in \mathbb{N}$.*)*

Note that the branching construction of the INAR process formalized in Proposition 21 above is similar as in the existence proof of the univariate case in Kirchner (2016).

3.3 Convergence theorems

Using the concepts introduced in the previous sections, we give approximations of projected point measures of Hawkes forests. As a special case, this yields approximations of multitype Hawkes processes by multivariate INAR(∞)-based point processes. The basic idea is to define all processes involved as projections of forests that have *exactly the same underlying GWB-trees*. This leaves only the immigrant point positions and the displacement variables to be approximated—which is straightforward.

Theorem 23. *Let* $\mathbf{F} = \big\{(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k, (T_k, L_k))\big\}_{k\in\mathcal{K}}$ *be a* (Q, ν, F)*-Hawkes forest as in Definition 14 with immigration law* Q *such that, for* $i \in [d]$*, there exists* $\eta_i \geqslant 0$ *such that*

$$\mathbb{E}_Q \#\Big\{k :\ T_k \in \big((n-1), n\big], L_k = i\Big\} \leqslant \eta_i, \quad n \in \mathbb{Z}. \tag{3.4}$$

Given $\mathbf{F}$ *as above, define a family of* approximating Hawkes forests

$$\mathbf{F}^{(\Delta)} = \big\{(\mathbf{T}_k, \mathbf{L}_k, \mathbf{S}_k^{(\Delta)}, (T_k^{(\Delta)}, L_k))\big\}_{k\in\mathcal{K}}, \quad \Delta > 0,$$

where

$$T_k^{(\Delta)} := \left\lfloor \frac{T_k}{\Delta} \right\rfloor \Delta, \quad k \in \mathbb{Z},$$

are the approximative immigration points *(their types* L_k *do not depend on* Δ*) and*

$$\mathbf{S}_k^{(\Delta)}(u) := \sum_{\varnothing < v \leqslant u} Y_{v,k}^{(\mathbf{L}_k(v^-), \mathbf{L}_k(v))}, \quad u \in \mathbf{T}_k,\ k \in \mathbb{Z}, \Delta > 0,$$

are the approximative positions *with* approximative displacement variables

$$Y_{k,u}^{(i,j;\Delta)} := \left\lceil \frac{Y_{k,u}^{(i,j;\Delta)}}{\Delta} \right\rceil \Delta, \quad u \in \mathcal{U},\ k \in \mathbb{Z},\ (i,j) \in [d]^2, \Delta > 0.$$

Let $\mathbf{N}_{\mathbf{F}^{(\Delta)}}$, $\Delta > 0$*, respectively,* $\mathbf{N}_{\mathbf{F}}$ *be the projected point measure of the approximating forests* $\mathbf{F}^{(\Delta)}$, $\Delta > 0$, *respectively of the forest* $\mathbf{F}$*. Then, for any nonnegative bounded continuous function with finite support,* $f : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$*, we obtain*

$$\int f(t) \mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t \times \{j\}) \quad \xrightarrow{\text{a.s.}} \quad \int f(t) \mathbf{N}_{\mathbf{F}}(\mathrm{d}t \times \{j\}), \quad j \in [d],$$

as $\Delta \to 0$*, and the convergence also holds in* L_1*.*

Proof. For all following manipulations, we use that neither the underlying multitype trees $\{(\mathbf{T}_k, \mathbf{L}_k)\}_{k\in\mathcal{K}}$ nor the types of the immigration points $(L_k)_{k\in\mathcal{K}}$ of the approximating forests $\mathbf{F}^{(\Delta)}$, $\Delta > 0$, depend on Δ. That is, we only have to consider the approximations of the positions of the immigration points and the approximations of the displacements. Note that also the indexing set $\mathcal{K}$ does not depend on Δ. For the immigration points, we obtain

$$\big|T_k^{(\Delta)} - T_k\big| = \big|\lceil T_k/\Delta \rceil \Delta - T_k\big| = \Delta \big|\lceil T_k/\Delta \rceil - T_k/\Delta\big| \leqslant \Delta \to 0, \quad k \in \mathcal{K}.$$

Consequently, $T_k^{(\Delta)} \xrightarrow{\text{a.s.}} T_k$, as $\Delta \to 0$, $k \in \mathbb{Z}$. In analogy one can show for the displacement variables that

$$Y_{k,u}^{(i,j;\Delta)} \xrightarrow{\text{a.s.}} Y_{k,u}^{(i,j)}, \quad \Delta \to 0, \quad u \in \mathcal{U}, (i,j) \in [d]^2, k \in \mathbb{N}.$$

Furthermore, as all the trees $\mathbf{T}_k, k \in \mathbb{Z}$, are almost surely finite, we have that

$$\begin{aligned}\lim_{\Delta\to 0} \mathbf{S}_k^{(\Delta)}(u) = \lim_{\Delta\to 0} \sum_{0<v\leqslant u} Y_{k,v}^{(\mathbf{L}(v^-),\mathbf{L}(v);\Delta)} &= \sum_{0<v\leqslant u} \lim_{\Delta\to 0} Y_{k,v}^{(\mathbf{L}(v^-),\mathbf{L}(v);\Delta)} \\ &= \sum_{0<v\leqslant u} Y_{k,v}^{(\mathbf{L}(v^-),\mathbf{L}(v))} = \mathbf{S}_k(u),\end{aligned}$$

almost surely, $u \in \mathcal{U}, k \in \mathbb{Z}$. Consequently,

$$\lim_{\Delta\to 0} \big(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u)\big) = T_k + \mathbf{S}_k(u), \quad k \in \mathbb{Z}, u \in \mathbf{T}_k, \quad \text{almost surely.} \tag{3.5}$$

Note that we have the explicit bound

$$\big|T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u) - T_k + \mathbf{S}_k(u)\big| \leqslant \big(|u| + 1\big)\Delta, \quad u \in \mathbf{T}_k, k \in \mathbb{Z}\,. \tag{3.6}$$

In particular, using the notation from Construction 8, for any compact set $K \subset \mathbb{R}$, we have that

$$\begin{aligned}&\#\Big\{u \in \mathbf{T}_k^{(g)} :\ T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u) \in K\Big\} \\ &\quad\overset{(3.6)}{\leqslant} \#\Big\{u \in \mathbf{T}_k^{(g)} :\ T_k + \mathbf{S}_k(u) \in \big[\min K - (g+1)\Delta, \max K + (g+1)\Delta\big]\Big\} \\ &\quad\leqslant \#\Big\{u \in \mathbf{T}_k^{(g)} :\ T_k + \mathbf{S}_k(u) \in \big[\min K - (g+1), \max K + (g+1)\big]\Big\}\end{aligned} \tag{3.7}$$

for $\Delta \in (0,1]$ and $k \in \mathbb{Z}$. Arguing with Proposition 16, we can show that $\mathbb{E}\big[\mathbf{N}_{\mathbf{F}}(B \times [d])\big] < \infty$ for any $B \in \mathcal{B}_b(\mathbb{R})$. So, in particular, $\mathbf{N}_{\mathbf{F}}(B \times [d]) < \infty$ almost surely. Choose any non-negative bounded continuous function $f : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ with finite support $\operatorname{supp}(f)$ and set $f_{\max} := \max_{t\in\mathbb{R}} f(t)$. We observe that

$$\lim_{\Delta\to 0} \int_{\mathbb{R}} f(t)\mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t \times \{j\}) = \lim_{\Delta\to 0} \sum_{k\in\mathbb{Z}} \sum_{u\in\mathbf{T}_k:} 1_{\{j\}}\big(\mathbf{L}_k(u)\big)\, f\big(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u)\big), \quad j \in [d]. \tag{3.8}$$

We dominate the inner sum, $\sum_{u\in\mathbf{T}_k:} 1_{\{j\}}(\mathbf{L}_k(u))\, f(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u))$, $k \in \mathbb{Z}$, by

$$\begin{aligned}&\#\big\{u \in \mathbf{T}_k : T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u) \in \operatorname{supp}(f), \mathbf{L}_k(u) = j\big\}\, f_{\max} \\ &\qquad\overset{(3.7)}{\leqslant} \#\big\{u \in \mathbf{T}_k :\ T_k + \mathbf{S}_k(u) \in I_f(|u|)\big\}\, f_{\max}, \quad k \in \mathbb{Z},\ \Delta \in (0,1],\end{aligned} \tag{3.9}$$

where $I_f(g) := [\min\operatorname{supp}(f) - (g+1), \max\operatorname{supp}(f) + (g+1)]$, $g \in \mathbb{N}_0$. The latter term in (3.9)

is almost surely summable over $k \in \mathbb{Z}$ (for $0 < \Delta \leqslant 1$). Indeed,

$$\begin{aligned}
&\mathbb{E}\sum_{k\in\mathbb{Z}} \#\{u \in \mathbf{T}_k : \ T_k + \mathbf{S}_k(u) \in I_f(|u|), \mathbf{L}_k(u) = j\} \\
&= \mathbb{E}\sum_{k\in\mathbb{Z}}\sum_{g\geqslant 0} \#\{u \in \mathbf{T}_k^{(g)} : \ T_k + \mathbf{S}_k(u) \in I_f(g), \mathbf{L}_k(u) = j\} \\
&= \sum_{g\geqslant 0} \mathbb{E}\sum_{k\in\mathbb{Z}} \#\{u \in \mathbf{T}_k^{(g)} : \ T_k + \mathbf{S}_k(u) \in I_f(g), \mathbf{L}_k(u) = j\} \\
&= \sum_{g\geqslant 0} \mathbb{E}\, \mathbf{N}_{\mathbf{F}}^{(g)}\big(I_f(g) \times \{j\}\big) \\
&\overset{(2.15),(3.4)}{\leqslant} \sum_{g\geqslant 0} \eta m_{\cdot,j}^{(g)} |I_f(g)| \\
&\leqslant \sum_{i=1}^{d} \eta_i \sum_{g\geqslant 0} \big(\max \operatorname{supp}(f) - \min \operatorname{supp}(f) + 2(g+1)\big)\ ||M^g||_{\max},
\end{aligned}$$

where $||A||_{\max} := \max_{(i,j)\in[d]^2} |a_{i,j}|$, for any matrix $A = (a_{i,j})_{(i,j)\in[d]^2}$. This latter term is finite because $\sum_{g\geqslant 0} ||M^g|| g < \infty$ by the quotient criterion. Indeed, we have that

$$q_g := \frac{(g+1)\ ||M^{g+1}||_{\max}}{g\ ||M^g||_{\max}} = \rho(M)\frac{g+1}{g}\frac{\rho(M)^g}{||M^g||_{\max}}\frac{||M^{g+1}||_{\max}}{\rho(M)^{g+1}}.$$

The last two factors converge to 1 by Gelfand's formula, so that $\lim_{g\to\infty} q_g = \rho(M) < 1$ by subcriticality of M; see (2.4). Having thus established a dominating summable sequence, we may interchange the limit operation with the outer summation in (3.8). Fix any $k \in \mathbb{Z}$. For the inner sum in (3.8),

$$\sum_{u\in\mathbf{T}_k} 1_{\{j\}}(\mathbf{L}_k(u)) f(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u)),$$

note that, almost surely, for $\Delta > 0$ small enough,

$$\#\Big\{u \in \mathbf{T}_k : T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u) \in \operatorname{supp} f\Big\} = \#\Big\{u \in \mathbf{T}_k : T_k + \mathbf{S}_k(u) \in \operatorname{supp} f\Big\}\ (< \infty). \tag{3.10}$$

So, for the inner sum in (3.8), we find that, almost surely,

$$\begin{aligned}
&\lim_{\Delta\to 0} \sum_{u\in\mathbf{T}_k:} 1_{\{j\}}(\mathbf{L}_k(u)) f(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u)) \\
&\qquad \overset{(3.10)}{=} \sum_{\substack{u\in\mathbf{T}_k \\ T_k+\mathbf{S}_k(u)\in\operatorname{supp}(f)}} \lim_{\Delta\to 0} 1_{\{j\}}(\mathbf{L}_k(u)) f(T_k^{(\Delta)} + \mathbf{S}_k^{(\Delta)}(u)) \\
&\qquad \overset{(3.5)}{=} \sum_{\substack{u\in\mathbf{T}_k: \\ T_k+\mathbf{S}_k(u)\in\operatorname{supp}(f)}} 1_{\{j\}}(\mathbf{L}_k(u)) f(T_k + \mathbf{S}_k(u)).
\end{aligned}$$

We conclude that, almost surely,

$$\begin{aligned}\lim_{\Delta\to 0}\int_{\mathbb{R}} f(t)\mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t\times\{j\}) &= \sum_{k\in\mathbb{Z}}\sum_{u\in\mathbf{T}_k:} 1_{\{j\}}(\mathbf{L}_k(u))f(T_k+\mathbf{S}_k(u)),\\ &= \int_{\mathbb{R}} f(t)\mathbf{N}_{\mathbf{F}}(\mathrm{d}t\times\{j\}),\quad j\in[d].\end{aligned}$$

L_1-convergence follows with dominated convergence; the arguments are similar to the case above. Indeed, for $\Delta\in(0,1]$, we obtain the following bound:

$$\left|\int f(t)\mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t\times\{j\}) - \int f(t)\mathbf{N}_{\mathbf{F}}(\mathrm{d}t\times\{j\})\right| \leqslant f_{\max}\sum_{g\geqslant 0}\mathbf{N}_{\mathbf{F}}^{(g)}(I_f(g)\times\{j\}).$$

Integrability of the latter term follows as in the argumentation for the almost-sure case. □

With assumptions and notations from Theorem 23, we collect a couple of relevant corollaries:

Corollary 24. *For $m\in\mathbb{N}$ and $l=1,2,\ldots,m$, let $j_l\in[d]$ and $A_l\in\mathcal{B}_b(\mathbb{R})$ such that $\mathbb{P}\left[\mathbf{N}(\delta(A_l)\times\{j_l\})=0\right]=1$. Then we have that*

$$\begin{aligned}&\Big(\mathbf{N}_{\mathbf{F}^{(\Delta)}}\big(A_1\times\{j_1\}\big),\ldots,\mathbf{N}_{\mathbf{F}^{(\Delta)}}\big(A_m\times\{j_m\}\big)\Big)\\ &\qquad\xrightarrow{\text{a.s.}}\quad\Big(\mathbf{N}_{\mathbf{F}}\big(A_1\times\{j_1\}\big),\ldots,\mathbf{N}_{\mathbf{F}}\big(A_m\times\{j_m\}\big)\Big),\quad\Delta\to 0.\end{aligned}\tag{3.11}$$

Note that as $\mathbf{N}_{\mathbf{F}^{(\Delta)}}(A\times\{j\})\in\mathbb{N}_0, \Delta>0, A\in\mathcal{B}_b(\mathbb{R})$, there must almost surely be a $\Delta_0>0$ such that

$$\begin{aligned}&\Big(\mathbf{N}_{\mathbf{F}^{(\Delta)}}\big(A_1\times\{j_1\}\big),\ldots,\mathbf{N}_{\mathbf{F}^{(\Delta)}}\big(A_m\times\{j_m\}\big)\Big)\\ &\qquad=\quad\Big(\mathbf{N}_{\mathbf{F}}\big(A_1\times\{j_1\}\big),\ldots,\mathbf{N}_{\mathbf{F}}\big(A_m\times\{j_m\}\big)\Big),\quad\Delta<\Delta_0.\end{aligned}\tag{3.12}$$

Finally, we generalize Theorem 2 in Kirchner (2016) to the multitype case (with fewer assumptions at that). Using Poisson–Hawkes forests in Theorem 23, we obtain as a special case

Theorem 25. *Let $\mathbf{N}_{\mathbf{F}}$ be a Hawkes process—constructed as a projected point measure from a Hawkes forest $\mathbf{F}$ as in Proposition 19—with immigration intensities η_j, $j\in[d]$, and reproduction intensities $h_{i,j}=w_{i,j}m_{i,j}$, $(i,j)\in[d]^2$. For $\Delta>0$, let $\mathbf{N}_{\mathbf{F}^{(\Delta)}}$ be the projected point measure of the corresponding approximating forests $\mathbf{F}^{(\Delta)}$ as in Theorem 23. We have the following results:*

i) Let $X_{n,i}^{(\Delta)} := \mathbf{N}_{\mathbf{F}^{(\Delta)}}\big(\{\Delta n\} \times \{i\}\big)$, $n \in \mathbb{Z}$, $i \in [d]$. *Then*

$$\big(\mathbf{X}_n^{(\Delta)}\big)_{n\in\mathbb{Z}} := \big((X_{n,1}^{(\Delta)}, X_{n,2}^{(\Delta)}, \ldots, X_{n,d}^{(\Delta)})\big)_{n\in\mathbb{Z}}$$

defines a multivariate INAR process with immigration coefficients $\alpha_{0,i}^{(\Delta)} = \Delta\eta_i$, $i \in [d]$, *and reproduction coefficients*

$$\alpha_{i,j,k}^{(\Delta)} = m_{i,j} \int_{(k-1)\Delta}^{k\Delta} w_{i,j}(t)\mathrm{d}t, \quad (i,j,k) \in [d]^2 \times \mathbb{N};$$

see Definition 20.

ii) For all nonnegative continuous functions with compact support f *and for* $j \in [d]$, *we have that*

$$\int f(t)\mathbf{N}_{\mathbf{F}^{(\Delta)}}(\mathrm{d}t \times \{j\}) \to \int f(t)\mathbf{N}_{\mathbf{F}}(\mathrm{d}t \times \{j\}), \quad \Delta \to 0, \tag{3.13}$$

almost surely, in L_1, *and (hence also) in distribution. In particular,*

$$\mathbf{N}_{\mathbf{F}^{(\Delta)}} \to \mathbf{N}_{\mathbf{F}}, \quad \Delta \to 0, \quad \textit{fidi and weakly}.$$

Proof. For i), one argues as for Proposition 21. In ii), almost-sure and L_1 convergence follow from Theorem 23. L_1 convergence of (3.13) is corresponds to the definition of weak convergence of point processes (with respect to the topology $\mathcal{M}_p^{[d]}$). By Theorem 11.1.VII. in Daley and Vere-Jones (2009), weak convergence for point processes is equivalent to convergence of finite-dimensional (fidi) distributions, that is, the distributions of vectors as in (3.11) converge to the distribution of the limit vector. □

Note that, by definition of the $(\mathbf{X}_n^{(\Delta)})$ sequence in Theorem 25 and by statement *i*), we can obviously represent the approximating projected point measures in terms of INAR sequence:

$$\mathbf{N}_{\mathbf{F}^{(\Delta)}}(A \times \{j\}) = \sum_{n:n\Delta\in A} X_{n,j}^{(\Delta)}, \quad \Delta > 0, j \in [d].$$

So, it makes sense to say that the $\mathbf{N}_{\mathbf{F}^{(\Delta)}}$ are 'INAR-based' point processes. The approximation result in Theorem 25 establishes the theoretical basis for a nonparametric estimation method for multitype Hawkes processes that is worked out in Kirchner (2017a) and Embrechts and Kirchner (2017): given multitype event-stream data from a Hawkes model, we divide the dataset into bins of size Δ. By Theorem 25, these bin-counts are well approximated by a multivariate INAR(∞) time series. In the time series case, immigration and reproduction coefficients may be estimated

by standard methods like conditional least squares (CLS), Yule–Walker, or MLE. For CLS, consistency and asymptotic normality are established in Theorem 3.5 of Kirchner (2017a); we also provide variance estimates. After an appropriate rescaling these coefficient estimates yield estimates for Hawkes-process parameters. Note that, independently of our work, Eichler et al. (2016) pursued the same discretization approach for estimation of Hawkes processes—without observing the connection to multivariate INAR sequences at that.

4 Discussion

4.1 Generalizations of the Hawkes process

The autoregressive view on the Hawkes process as in (3.1) suggests obvious generalizations such as non-linearities—treated for the univariate case in Brémaud and Massoulié (1996). The Hawkes-process representation as a projected point measure of a Hawkes–Poisson forest in Proposition 19 suggests other generalizations and alternatives. To that aim, note that the Hawkes forest from Definition 14 is *much* more general than the Hawkes–Poisson forest from Definition 17. Namely, in a general Hawkes forest, the immigration law is left open and, more importantly, the unordered offspring distributions μ, respectively, the ordered offspring distributions ν in a Hawkes tree as in Definition 7 may be *much* more general than for the Hawkes–Poisson tree from Definition 10.

Remark 26. We propose possible alternatives to standard Hawkes models. Analogous ideas are valid for the INAR time series.

i) *Typewise number of offspring*: the typewise number of children of a node in a Hawkes–Poisson forest is Poisson distributed. In some applications, a different offspring distribution, e.g., Bernoulli, may be a better choice.

ii) *Total number of offspring*: in a Hawkes–Poisson forest, the total number of children of a node is also Poisson distributed. Again, there might be applications where a Bernoulli, a binomial, or a uniform distribution may be a better choice.

iii) *Dependency of types*: this is obviously related to the first two points. In a Hawkes–Poisson forest, in general, any type-i node may have children of potentially all types. However, there might be applications where $\mu_i(k_1, k_2, \ldots, k_d) = 0$ if more than one k_l is greater than 0. That is, μ_i does not allow children of different types.

iv) *Displacement distributions:* displacement distributions in a Hawkes–Poisson forest are absolutely continuous. However, in some applications, they may be (partially) discrete or even deterministic. Furthermore, if the goal is not so much actual prediction but more abstract modeling, then one may want to use displacement distributions that are supported

by $\mathbb{R}$ rather than only by the positive halfline. These ('non-causal') models might have larger explanatory power than standard ('causal') Hawkes processes.

v) *Immigration law*: for a standard Hawkes process, the immigrants form a Poisson process. However, the immigration law may be chosen as a more general (still stationary) point process.

vi) *Filtration:* in a Hawkes forest, we have all information for each node: we know the node's tree, its parent and its children, its generation number, in particular, we know if it is an immigrant node or a child node, and—if it is a child node—we know its place in the ordered offspring vector. In the Hawkes model, we project all this information on $\mathbb{R} \times [d]$. All that remains is the position and the type of each node. This jump from full to nearly none information seems somewhat extreme and might be done more gradually. There could be applications where one can distinguish immigrants from non-immigrants. In some case, we might even know the parents of each point. Or—a bit more restrictive—one might at least know the type of the parent of all points. Finally, there might be applications where the children of any point come in some natural order—which would make the concept of *ordered* offspring distributions fertile for modeling.

4.2 Conclusion

Using the formalism of branching random walks, we define random trees as well as random forests and consider their projected point measures. We pay special attention to the case when the underlying BRWs have displacement distributions concentrated on the positive halfline. We call these special cases Hawkes trees and Hawkes forests. This setup allows a number of approximation results of multitype Hawkes processes and INAR(∞)-based point processes. Furthermore, we show how the very general framework inspires alternatives to standard Hawkes and INAR processes in applications. Another interesting topic in this framework might be the study of critical and supercritical versions of multitype Hawkes and INAR(∞) processes. In any case, the disentanglement of the tree structure on the one hand and the position labels on the other hand, is mathematically attractive: for the branching structure, there is a lot of theory available, and, for the position-labels, standard renewal (or random-walk) theory applies.

Paper

C

Matthias Kirchner.

An estimation procedure for the Hawkes process.

Quantitative Finance, ***17****(4):571–595, 2017.*

An estimation procedure for the Hawkes process

Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

In this paper, we present a nonparametric estimation procedure for the multivariate Hawkes point process. The timeline is cut into bins and—for each component process—the number of points in each bin is counted. As a consequence of earlier results in Kirchner (2016), the distribution of the resulting 'bin-count sequences' can be approximated by an integer-valued autoregressive model known as the (multivariate) INAR(p) model. We represent the INAR(p) model as a standard vector-valued linear autoregressive time series with white-noise innovations (VAR(p)). We establish consistency and asymptotic normality for conditional least-squares estimation of the VAR(p), respectively, the INAR(p) model. After appropriate scaling, these time series estimates yield estimates for the underlying multivariate Hawkes process as well as corresponding variance estimates. The estimator depends on a bin-size Δ and a support s. We discuss the impact and the choice of these parameters. All results are presented in such a way that computer implementation, e.g., in R, is straightforward. Simulation studies confirm the effectiveness of our estimation procedure. In the second part of the paper, we present a data example where the method is applied to bivariate event-streams in financial limit-order-book data. We fit a bivariate Hawkes model on the joint process of limit and market order arrivals. The analysis exhibits a remarkably asymmetric relation between the two component processes: incoming market orders excite the limit-order flow heavily whereas the market-order flow is hardly affected by incoming limit orders. For the estimated excitement-functions, we observe power-law shapes, inhibitory effects for lags under 0.003 sec, second periodicities, and local maxima at 0.01 sec, 0.1 sec, and 0.5 sec.

1 Introduction

In this paper, we introduce a nonparametric estimation procedure for the multivariate Hawkes point process; see Definition 11 for the formal definition and Figure 1 for an illustrative summary of the main results. The Hawkes process is a model for event streams. Its alternative

name, 'self-exciting point process', stems from the fact that any event has the potential to generate new events in the future. Our estimator gives substantial information on this excitement: nonmonotonicities or regime switches in the excitement of the fitted Hawkes model can be detected; the estimates may also help with the choice of parametric excitement-functions. The asymptotic distribution of the estimator can be derived so that confidence bounds are at hand. Also note that the presented estimation method is numerically less problematic than the standard likelihood-approach. Last but not least, the figures generated from the estimation results are a graphical tool for representing large univariate and multivariate event data sets in a compact and at the same time informative way. In particular, the estimation results can be interpreted as measures for interaction and stability of empirical event-streams. This will be highlighted in the data example at the end of the paper where we apply the estimation procedure to the order arrival times in an electronic market.

The Hawkes process was introduced in Hawkes (1971b,a) as a model for event data from contagious processes. Theoretical cornerstones of the model are Hawkes and Oakes (1974), Brémaud and Massoulié (1996, 2001), Liniger (2009) and Errais et al. (2010). For a textbook reference that covers many aspects of the Hawkes process; see Daley and Vere-Jones (2009). The main theoretical reference for the following presentation is our own contribution Kirchner (2016), where we show that Hawkes processes can be approximated by certain discrete-time models.

By the omnipresence of 'event'-type data, the Hawkes process has become a popular model in many different contexts such as geology, e.g., earthquake modeling in Ogata (1988), internet traffic, e.g., youTube clicks in Crane and Sornette (2008), biology, e.g., genome analysis in Reynaud-Bouret and Schbath (2010), sociology, e.g., crime data in Mohler et al. (2011), or medicine, e.g., virus spreading in Kim (2011). A most active area of scientific activity today is financial econometrics with applications of Hawkes processes to the modeling of credit defaults in Errais et al. (2010), extreme daily returns in Liniger (2009), market contagion in Aït-Sahalia et al. (2015) and numerous applications to limit-order-book modeling such as high-frequency price jumps in Bacry et al. (2012) and Chavez-Demoulin and McGill (2012), order arrivals in Bacry et al. (2013), or joint models for orders and prices on a market microstructure level in Muzy and Bacry (2014). Early publications applying the Hawkes model in the financial context are Bowsher (2002), Chavez-Demoulin et al. (2005) and McNeil et al. (2005).

The paper is organized as follows: Section 2 explains how Hawkes processes can be approximated by specific integer-valued time series and how this approximation yields an estimation procedure. Section 3 defines the new Hawkes estimator formally, discusses its properties, and compares it to alternative estimation methods. Section 4 refines the procedure by giving methods for a reasonable choice of the estimation parameters. Section 5 presents the data example where the ideas of the paper are applied to the analysis of intraday financial data. The last

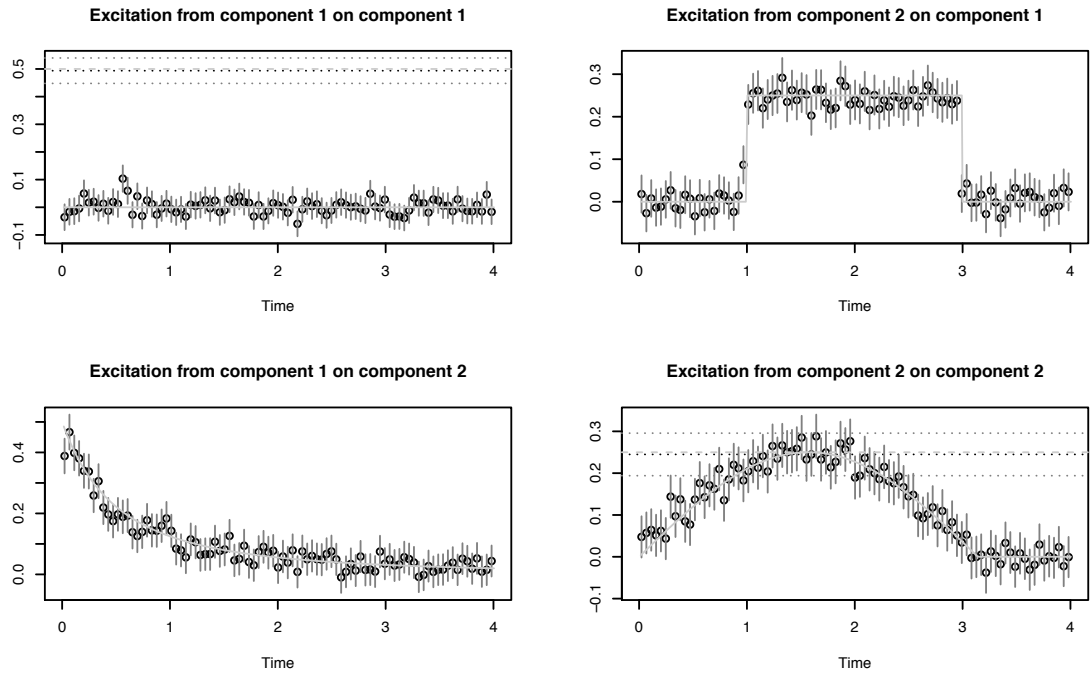


Figure 1: Summary of the main result of the paper. From the bivariate Hawkes model presented in Figure 2, around 100 000 events in each component are simulated. From this single large sample, we calculate the estimator from Definition 11. The black circles refer to estimated values of the excitement functions. The horizontal black dotted lines in the diagonal panels refer to the corresponding estimated baseline-intensity components. The vertical grey lines as well as the dotted horizontal grey lines refer to marginal 95%-confidence intervals; see Remark 12. All solid and dashed lightish-grey lines refer to the true underlying parameters; compare with Figure 2. Eyeball examination shows that the estimation method approximates the form of the true excitement functions well. Also the nonmonotonicities and the jumps are reproduced. The coverage rates of the confidence intervals seem just about right. There is no obvious bias. For a more quantitative analysis of the estimation method; see Section 3.3 and Figure 3.

section concludes with a discussion on the implications of the presented results. Appendix A contains proofs. Large parts of the paper are accompanied by examples with simulated data: in favor of a linear reading flow, we directly illustrate all new concepts with such examples—instead of devoting a separate section to simulations.

2 Approximation of Hawkes processes

In this section, after defining the Hawkes process we introduce autoregressive integer-valued time series. We clarify how this model approximates the Hawkes model and how this approximation yields an estimation procedure.

2.1 The Hawkes process

From a geometric point of view, a (univariate) Hawkes process specifies a distribution of points on a line. Typically, the line is interpreted as 'time' and the points as 'events'. Self-exciting point process is the common alternative name for the Hawkes process. It highlights the basic idea of the model: given an event, the intensity—the expected number of events in one time unit—shoots up ('self-excites') and then decays ('forgets its past gradually'). The shape of this decay is specified by a function, namely the excitement function. The definition and the proof of existence of a Hawkes process are subtle matters. For rigorous theoretical foundation, we refer to Liniger (2009), Chapter 6. We assume a basic underlying probability space $(\Omega, \mathbb{P}, \mathcal{F})$, complete and rich enough to carry all random variables involved. On this probability space, we define stochastic point-sets $\mathcal{P} \subset \mathbb{R}$ of the form $\mathcal{P} = \{\dots, T_{-1}, T_0, T_1, \dots\}$ with $T_k \leqslant T_{k+1}$, $k \in \mathbb{Z}$, having almost surely no limit points. Furthermore, we assume that the σ-algebras

$$\mathcal{H}_t^{\mathcal{P}} := \sigma\left(\left\{\omega \in \Omega : \#\left(\mathcal{P}(\omega) \cap (a,b]\right) = n : \ n \in \mathbb{N}_0,\ a < b \leqslant t\right\}\right), \quad t \in \mathbb{R},$$

are subsets of $\mathcal{F}$. By setting

$$N_{\mathcal{P}}(A) := \#\left(\mathcal{P} \cap A\right), \quad A \in \mathcal{B}(\mathbb{R}),$$

any stochastic point-set $\mathcal{P}$ defines a random measure $N_{\mathcal{P}}$ on $\mathcal{B}(\mathbb{R})$, the Borel sets of $\mathbb{R}$. At this point, we drop the $\mathcal{P}$ index; the set $\mathcal{P}$ is completely specified by $N := N_{\mathcal{P}}$. In this paper, we call a random measure N of this kind *point process* and we call the filtration $(\mathcal{H}_t^N) := \left(\mathcal{H}_t^{\mathcal{P}}\right)$ *history* of the point process. The *conditional intensity* of a point process N is

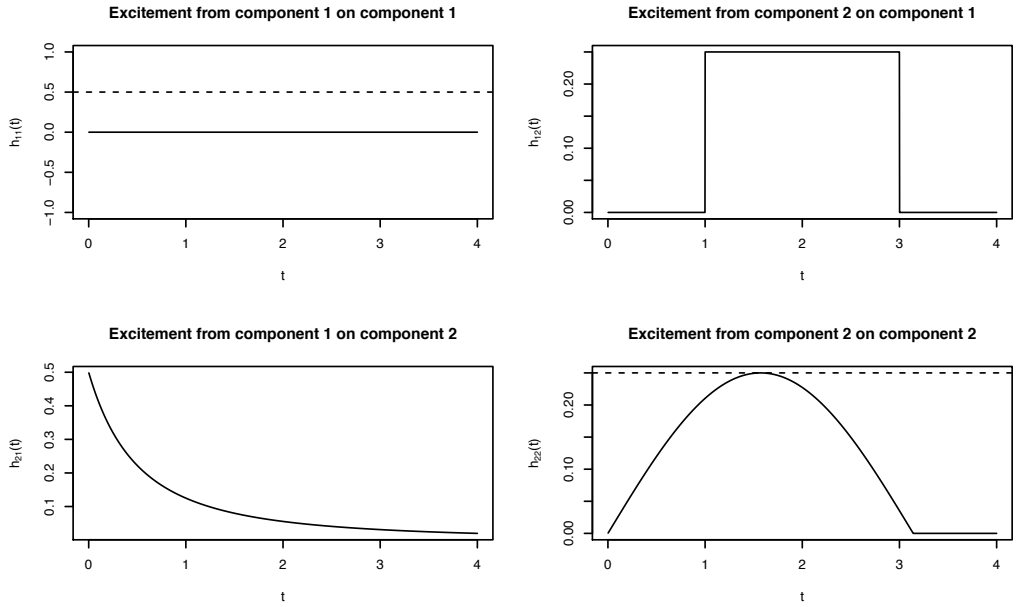
$$\Lambda_N(t) := \lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[N\left((t, t+\delta]\right) | \mathcal{H}_t^N\right]}{\delta}, \quad t \in \mathbb{R}. \tag{2.1}$$

A *Hawkes process* is a stationary point process N with conditional intensity

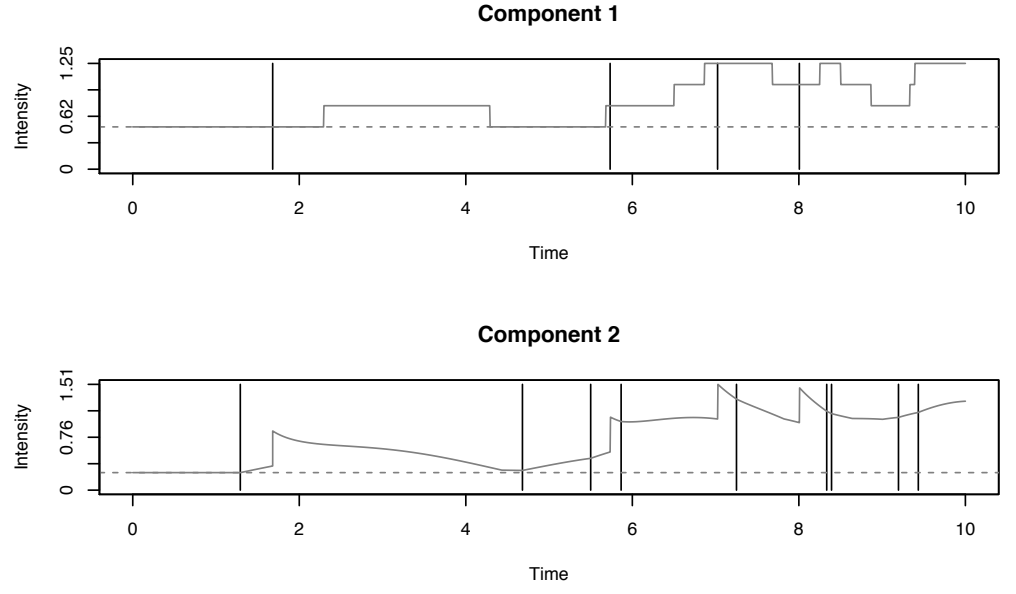
$$\Lambda_N(t) = \eta + \int_{-\infty}^{t} h(t-s) N(\mathrm{d}s), \quad t \in \mathbb{R}. \tag{2.2}$$

The constant $\eta \geqslant 0$ is called *baseline intensity*, and the function $h : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, measurable, is called *excitement function*. Existence-conditions are discussed below.

For $d \in \mathbb{N}$, a *d-variate Hawkes process* $\mathbf{N}$ is a process with d point processes on $\mathbb{R}$ as components, i.e., $\mathbf{N} = \left(N^{(1)}, \dots, N^{(d)}\right)^\top$. Each component process counts points from random point-sets $\mathcal{P}_1 \subset \mathbb{R}, \dots, \mathcal{P}_d \subset \mathbb{R}$. In this multivariate setup, the counting processes $N^{(k)}$, $k =$



(a) Model parameters of a bivariate Hawkes process. The solid lines refer to the excitement function $H = (h_{i,j})$. It consists of the two self-excitement functions $h_{1,1}(t) \equiv 0$ and $h_{2,2}(t) = 1_{t \leqslant \pi} 0.25 \sin(t)$ as well as the two cross-excitement functions $h_{1,2}(t) \equiv 1_{1<t\leqslant 3} 0.25$ and $h_{2,1}(t) = 0.5(1+t)^{-2}$. The dashed lines in the diagonal panels refer to the two components of the baseline intensity $\eta = (0.5, 0.25)$. The functions are chosen quite extreme for the sake of demonstration of the estimation method; see Figure 1.



(b) A realization of the two components of the process starting at time 0. The vertical lines refer to the events, the greyish solid lines refer to the realized conditional-intensity components, and the dashed lines refer to the baseline-intensity components. The cross-excitement from component 1 on component 2 and also the delayed rectangle impulse impact from component 2 on component 1 are particularly visible.

Figure 2: Illustration of a bivariate Hawkes process as described in Section 2.1. The upper panel shows the model parameters. The lower panel shows a realization.

$1, \ldots, d$, do not only self-excite but in general also interact with each other ('cross-excite'). The baseline intensity η is a d-variate vector in $\mathbb{R}^d_{\geqslant 0}$ and the excitement function is a measurable $d \times d$ matrix-valued function $H = (h_{i,j})_{1 \leqslant i,j \leqslant d} : \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{d \times d}_{\geqslant 0}$. The *conditional intensity of a d-variate Hawkes process* is $\mathbb{R}^d_{\geqslant 0}$-valued with

$$\mathbf{\Lambda}_{\mathbf{N}}(t) := \lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[\mathbf{N}((t, t+\delta])|\mathcal{H}_t^{\mathbf{N}}\right]}{\delta} = \eta + \int_{-\infty}^{t} H(t-s)\mathbf{N}\left(\mathrm{d}s\right), \quad t \in \mathbb{R}, \tag{2.3}$$

where, for $i = 1, \ldots, d$,

$$\left(\int_{-\infty}^{t} H(t-s)\mathbf{N}\left(\mathrm{d}s\right)\right)_i := \left(\sum_{j=1}^{d} \int_{-\infty}^{t} h_{i,j}(t-s)N^{(j)}\left(\mathrm{d}s\right)\right)_i \tag{2.4}$$

and $\mathcal{H}_t^{\mathbf{N}} := \sigma\left(\left\{\omega \in \Omega : \mathbf{N}\left((a,b]\right) = \mathbf{n}\right\}, \mathbf{n} \in \mathbb{N}_0^d,\ a < b \leqslant t\right)$. In other words, the entry $h_{i,j}(t)$ of the matrix $H(t)$ denotes the effect of any event $T_k^{(j)} \in \mathcal{P}_j$ in component j on the intensity of component i at time $T_k^{(j)} + t$. See Figure 2 for an example of a bivariate Hawkes process. In Hawkes (1971a), we find the following sufficient condition for existence: if

$$\mathrm{spr}(K) := \max\left\{|k| : k \text{ eigenvalue of matrix } K\right\} < 1, \tag{2.5}$$

where $K := \left(\int_0^\infty h_{i,j}(t)\mathrm{d}t\right)_{1 \leqslant i,j \leqslant d}$, then a process with conditional intensity as in (2.3) exists. The matrix K in (2.5) is sometimes referred to as *branching matrix* and the entries of K as *branching coefficients*. These terms reflect an alternative view on the process as a special cluster process (Hawkes and Oakes, 1974):

In each of the components of a d-variate Hawkes process, we observe cluster centers that stem from independent homogeneous Poisson processes with rates $\eta_1, \ldots, \eta_d$. These cluster centers are also called *immigrants* or *exogenous events*. Such an immigrant $I^{(j)} \in \mathbb{R}$ in component j triggers d inhomogeneous Poisson processes in components $i = 1, \ldots, d$ with intensities $h_{i,j}\left(\cdot - I^{(j)}\right)$, $i = 1, \ldots, d$. And each of these new points again produces d inhomogeneous Poisson processes in a similar way, so that the clusters are built up as a cascade of inhomogeneous Poisson processes. The non-immigrant events are called *offspring* or *endogenous events*. Disregarding the time component and only considering this immigrant–offspring structure, one actually has a branching process with immigration, where the number of direct offspring in component i from an event in component j is $\mathrm{Pois}\left(K_{ij}\right)$ distributed.

2.2 Parametrization and estimation of Hawkes processes

In most cases, the data analyst's choice of the excitement function H of a Hawkes process is a somewhat arbitrary parametric function—the main decision being between exponential functions or power-law functions. The function parameters are then estimated via standard likelihood maximization. Power-law decay of the excitement functions often turns out to be more 'realistic' in applications; exponential decay yields a likelihood that is numerically easier to handle by recursive representation; see Liniger (2009). In addition, exponential excitement functions are mathematically attractive because they yield a Markovian structure for the conditional intensity; see Errais et al. (2010). Even if the choice between exponential and power-law decay is handled carefully, these two functional families cannot catch regime switches or nonmonotonicities of excitement functions as in Figure 2. So it seems important to develop methods that can identify shapes of excitement in data with less stringent assumptions. Another motivation for our research on estimation of the Hawkes model stems from numerical issues—especially encountered in the multivariate case. A third gap that we aim to close with our paper is the derivation of the asymptotic distribution of the estimates.

Alternative estimation methods for the Hawkes process have been introduced in Lewis and Mohler (2011), Lemonnier and Vayatis (2014) , Reynaud-Bouret et al. (2014), Alfonsi and Blanc (2015), and Hansen et al. (2015). In particular, the method developed in Bacry et al. (2012, 2014), and Bacry and Muzy (2015) is similar to ours and can be interpreted in our approximation framework. We will discuss these alternative estimation approaches in Section 3.4.

2.3 Intuition of the approximation

The main idea is simple: given a (possibly multivariate) Hawkes process, we divide the time line into bins of size $\Delta > 0$ and count the number of events in each bin (for each component). These 'bin counts' form an $\mathbb{N}_0$-valued stochastic sequence ($\mathbb{N}_0^d$-valued in the d-variate case). The distribution of this sequence can be approximated by a well-known time series model. We present the heuristics behind the approximation in the case of a univariate Hawkes process N with baseline intensity $\eta > 0$ and excitement function h with $\int h\mathrm{d}t < 1$. For some $\Delta > 0$, we define the bin-counts $\tilde{X}_n^{(\Delta)} := N\big(((n-1)\Delta, n\Delta]\big)$, $n \in \mathbb{Z}$. We want to argue that for small $\Delta > 0$ and large $p \in \mathbb{N}$, we have that

$$\mathbb{E}\left[\tilde{X}_n^{(\Delta)} \middle| \sigma\left(\tilde{X}_{n-1}^{(\Delta)}, \tilde{X}_{n-2}^{(\Delta)}, \dots\right)\right] \approx \Delta\eta + \sum_{k=1}^{p} \Delta h(\Delta k)\tilde{X}_{n-k}^{(\Delta)}, \quad n \in \mathbb{Z}. \tag{2.6}$$

We divide the approximation above in three separate approximation-steps:

$$\begin{aligned}
\mathbb{E}\left[\tilde{X}_n^{(\Delta)} | \mathcal{H}^N_{(n-1)\Delta}\right] \left(\right. &\overset{(2.1)}{=} \int_{(n-1)\Delta}^{n\Delta} \mathbb{E}\left[\Lambda(t) | \mathcal{H}^N_{(n-1)\Delta}\right] \mathrm{d}t \left. \right) \\
&\overset{(2.2)}{\approx} \Delta\eta + \Delta \int_{-\infty}^{(n-1)\Delta} h(n\Delta - u) N(\mathrm{d}u) \qquad (2.7) \\
&\approx \Delta\eta + \Delta \int_{(n-p-1)\Delta}^{(n-1)\Delta} h(n\Delta - u) N(\mathrm{d}u) \qquad (2.8) \\
&\approx \Delta\eta + \sum_{k=1}^{p} \Delta h(\Delta k) \tilde{X}^{(\Delta)}_{n-k}, \quad n \in \mathbb{Z}. \qquad (2.9)
\end{aligned}$$

The estimator we are about to present ignores the three approximations above and treats them as equalities. In doing so, we make a distributional error (2.7), a cut-off error (2.8), and a discretization error (2.9). The term *distributional error* might demand further explanation: in (2.7), we treat the conditional intensity Λ as constant over $((n-1)\Delta, n\Delta]$. This is not true in general as h is typically not (piecewise) constant. In addition—and more importantly—(2.7) ignores the influence of possible events in the bin $((n-1)\Delta, n\Delta]$ on Λ. As an example, suppose we observe two events in a bin. In the original Hawkes model, the second of these events may very well be a result of the first event. But in the approximating model, we ignore this possibility and explain both of these events by events in earlier bins or by the constant term.

There is an integer-valued time series that solves the approximative bin-count equation (2.6) to the point: the integer-valued autoregressive model of order $p \in \mathbb{N}$, the INAR(p) model. The three different approximation errors (2.7), (2.8), and (2.9), contribute to the bias of our estimation method in different ways. We discuss these effects in Sections 4.1 and 4.2.

2.4 The INAR(p) model

The INAR(p) process was first proposed by Du and Li (1991) as a time series model for count data. For the history and an exhaustive collection of properties of the model; see Marques da Silva (2005). For a textbook reference; see Fokianos and Kedem (2012). The main idea of the construction is to manipulate the standard system of autoregressive difference-equations '$X_n - \sum \alpha_k X_{n-k} = \varepsilon_n$, $n \in \mathbb{Z}$' in such a way that its solution (X_n) is integer valued. This is achieved by giving the error terms a distribution supported on $\mathbb{N}_0$ and substituting all multiplications with independent $\mathbb{N}_0$-valued operations. The following notation borrowed from Steutel and van Harn (1979) makes the analogy particularly obvious.

Definition 1. *For an $\mathbb{N}_0$-valued random variable Y and a constant $\alpha \geqslant 0$ define the* reproduction operator $\circ$ *by*

$$\alpha \circ Y := \sum_{k=1}^{Y} \xi_k^{(\alpha)},$$

where $\xi_1^{(\alpha)}, \xi_2^{(\alpha)}, \ldots$ are i.i.d. and independent of Y with $\xi_1^{(\alpha)} \sim \text{Poisson}(\alpha)$. We use the convention that $\sum_{k=1}^{0} \xi_k^{(\alpha)} = 0$.

We immediately present the multivariate version of the reproduction operator and the multivariate version of the INAR(p):

Definition 2. *For a $d \times d$ matrix $A = (\alpha_{i,j})_{1 \leqslant i,j \leqslant d} \in \mathbb{R}_{\geqslant 0}^{d \times d}$ and an $\mathbb{N}^d$-valued random variable $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$, define the* multivariate reproduction operator $\circledast$ *by*

$$A \circledast \mathbf{X} := \begin{pmatrix} \sum_{j=1}^{d} \alpha_{1,j} \circ X_j \\ \ldots \\ \sum_{j=1}^{d} \alpha_{d,j} \circ X_j \end{pmatrix},$$

where the reproductions $(\alpha_{i,j} \circ \cdot\,)$ operate independently over $1 \leqslant i, j \leqslant d$.

Definition 3. *Let $d, p \in \mathbb{N}$, $A_k \in \mathbb{R}_{\geqslant 0}^{d \times d}$, $k = 1, \ldots, p$, $\mathbf{a}_0 \in \mathbb{R}_{\geqslant 0}^{d}$, and $(\varepsilon_n)_{n \in \mathbb{Z}}$ an i.i.d. sequence of vectors in $\mathbb{N}_0^d$ with mutually independent components $\varepsilon_{0,i} \sim \text{Pois}(\mathbf{a}_{0,i})$, $i = 1, \ldots, d$. A* d-variate INAR(p) sequence *is a stationary sequence $(\mathbf{X}_n)_{n \in \mathbb{Z}}$ of $\mathbb{N}_0^d$-valued random vectors; it is a solution to the system of stochastic difference-equations*

$$\mathbf{X}_n = \sum_{k=1}^{p} A_k \circledast \mathbf{X}_{n-k} + \varepsilon_n, \quad n \in \mathbb{Z},$$

where the '$\circledast$' operate independently over k and n and also independently of (ε_n). We refer to $\mathbf{a}_0$ as immigration-parameter vector *and to A_k, $k = 1, 2, \ldots, p$, as* reproduction-coefficient matrices.

This model has first been considered in Latour (1997). In the same paper we find that if all zeros of

$$z \mapsto \det\left(z 1_{d \times d} - \sum_{k=1}^{p} A_k\right), \quad z \in \mathbb{C}, \tag{2.10}$$

lie inside the unit circle, then a multivariate INAR(p) process as in Definition 3 exists.

Consider a univariate INAR(p) sequence (X_n) with immigration parameter α_0 and reproduction coefficients α_k, $k = 1, \ldots, p$. Note that the criterion from above now simply reads

$\sum_{k=1}^{p} \alpha_k < 1$. Under this condition, we have that $X_n | X_{n-1}, X_{n-2}, \dots \sim \text{Pois}\left(\alpha_0 + \sum_{k=1}^{p} \alpha_k X_{n-k}\right)$. In particular, $\mathbb{E}[X_n | \sigma(X_{n-1}, X_{n-2}, \dots)] = \alpha_0 + \sum_{k=1}^{p} \alpha_k X_{n-k}$—which is the exact version of (2.6). The INAR(p) sequence has a similar immigrant–offspring structure as the Hawkes process. In the time series case, the (possibly multiple) immigrants at each time step stem from i.i.d. Pois(α_0) variables. Each of these immigrants produces Pois(α_k) new offspring events at k time steps later. Each of these offspring events again serves as parent event for new offspring etc.

A more obvious choice for the distribution of the counting sequences in Definition 1 would be Bernoulli—yielding the original *thinning operation* from Steutel and van Harn (1979). Note, however, that for small reproduction coefficients, the Poisson and the Bernoulli approaches are very similar. Also note that the Poisson distribution is more convenient for our purpose: we want to interpret the INAR(p) model as an approximation of the bin-count sequence of a Hawkes process and in the Hawkes model, an event can have potentially more than one direct offspring event in a future time-interval. In addition, in the Poisson case, we do not have to exclude reproduction coefficients larger than one.

2.5 Approximation of the Hawkes process by the INAR(p) model

We examine the close relation between Hawkes point processes and INAR time series in Kirchner (2016). For a particularly obvious parallel, the reader may consider the analogy of the existence criteria (2.5) and (2.10). Our cited paper gives a precise convergence statement for the univariate case. After establishing existence and uniqueness of the INAR(∞) process as a generalization of Definition 3 with $d = 1$ and $p = \infty$, we prove

Theorem 4. *Let N be a univariate Hawkes process with baseline intensity $\eta > 0$ and piecewise-continuous excitement function $h : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ such that $\sum_{k=1}^{\infty} h(k\Delta)\,\Delta < 1$ for all $\Delta \in (0, 1)$. Furthermore, let $(X_n^{(\Delta)})$ be a univariate INAR(∞) sequence with immigration parameter $\alpha_0^{(\Delta)} := \Delta\eta$ and reproduction coefficients $\alpha_k^{(\Delta)} := \Delta h(k\Delta)$, $k \in \mathbb{N}$, and define a family of point processes by*

$$N^{(\Delta)}\big((a, b]\big) := \sum_{n:\, n\Delta \in (a,b]} X_n^{(\Delta)}, \quad a < b,\ \Delta \in (0, 1).$$

Then we have that, for $\Delta \downarrow 0$, the INAR(∞)-based family of point processes $\left(N^{(\Delta)}\right)$ converges weakly to the Hawkes process N.

Proof. This is Theorem 2 in Kirchner (2016). □

Note that weak convergence of point processes is equivalent to convergence of the corresponding finite-dimensional distributions; see Daley and Vere-Jones (2009), Theorem 11.1.VII. The other theoretical result that is important for our estimation purpose is the fact that INAR(∞) processes can be approximated by INAR(p) processes, $p < \infty$:

Proposition 5. *Let (X_n) be an INAR(∞) sequence with immigration parameter $\alpha_0 > 0$ and reproduction coefficients $\alpha_k \geqslant 0$, $k \in \mathbb{N}$. Furthermore, let $\big(X_n^{(p)}\big)$ be a corresponding INAR(p) sequence, where the reproduction coefficients are truncated after the p-th lag. That is, $\big(X_n^{(p)}\big)$ has immigration parameter $\alpha_0^{(p)} := \alpha_0$ and reproduction coefficients $\alpha_k^{(p)} := 1_{\{k \leqslant p\}}\alpha_k$, $k \in \mathbb{N}$. Then, for $p \to \infty$, the finite-dimensional distributions of $\big(X_n^{(p)}\big)$ converge to the finite-dimensional distributions of $\big(X_n\big)$.*

Proof. This can be derived by establishing the convergence of the corresponding moment-generating functions from Proposition 2 in Kirchner (2016). □

We have not worked out the multivariate versions of Theorem 4 and Proposition 5 above. However, the simulations presented further down in the paper support the assumption that both results also hold in the multivariate case. Under this assumption, we have the following approximation:

Basic approximation

Let $\mathbf{N}$ be a d-variate Hawkes with baseline-intensity vector η and excitement function H as in (2.3). Let $\big(\mathbf{X}_n^{(\Delta)}\big)$ be a d-variate INAR(∞) sequence with immigration-parameter vector $\mathbf{a}_0^{(\Delta)} := \Delta\,\eta$ and reproduction-coefficient matrices $A_k^{(\Delta)} := \Delta H(k\Delta)$, $k \in \mathbb{N}$. Furthermore, for $p \in \mathbb{N}$, let $\big(\mathbf{X}_n^{(\Delta,p)}\big)$ be a corresponding INAR(p) sequence with baseline intensity $\mathbf{a}_0^{(\Delta,p)} := \mathbf{a}_0^{(\Delta)}$ and p reproduction-coefficient matrices $A_k^{(\Delta,p)} := A_k^{(\Delta)}$, $k = 1, \ldots, p$. Then, for small $\Delta > 0$ and large $p\Delta > 0$, we have that

$$\begin{aligned}
&\Big(N\big((0,\Delta]\big), N\big((\Delta,2\Delta]\big), \ldots, N\big(((m-1)\Delta, m\Delta]\big)\Big) \\
&\overset{d}{\approx} \Big(X_1^{(\Delta)}, X_2^{(\Delta)}, \ldots, X_m^{(\Delta)}\Big) \\
&\overset{d}{\approx} \Big(X_1^{(\Delta,p)}, X_2^{(\Delta,p)}, \ldots, X_m^{(\Delta,p)}\Big), \quad m \in \mathbb{N}\,.
\end{aligned}$$

If $\operatorname{supp}(H) \subset [0, s]$ for some finite $s > 0$, then the second approximation becomes an equality for all $p \geqslant \lceil s/\Delta \rceil$.

The approximation summarized in the box above is the key observation for our *estimation procedure*:

(i) Choose a small bin-size $\Delta > 0$ and calculate the bin-count sequence of the events stemming from the Hawkes process.

(ii) Choose a large support $s := p\Delta$ and fit the approximating INAR(p) model to the bin-count sequence via conditional least-squares.

(iii) Interpret the scaled immigration-parameter estimate $\hat{\mathbf{a}}_0^{(\Delta,p)}/\Delta$ as the natural candidate for an estimate of η and, for $k \in \{1, 2, \ldots, p\}$, interpret the scaled reproduction-coefficient matrix estimates $\hat{A}_k^{(\Delta,p)}/\Delta$ as natural candidates for estimates of $H(k\Delta)$.

Before giving the formal definition of the estimator in the next section, we illustrate the power of the presented method in Figure 1.

3 The estimator

In this section, we first discuss estimation of the approximating INAR(p) process. Then we define our Hawkes estimator formally and collect some of its properties. We describe alternative estimation methods and compare them with our estimator. Furthermore, we present results of a multivariate simulation study that support our approach.

3.1 Estimation of the INAR(p) model

There are several possibilities to estimate the parameters of an INAR(p) process. As the margins are conditionally Poisson distributed, in principle, maximum-likelihood estimation (MLE) can be applied. In our context, however, numerical optimization of the likelihood is difficult, as the number of model parameters will typically be very large. A method-of-moments type estimator would be the Yule–Walker method (YW). A third method is the conditional least-squares estimation (CLS). We formulate the estimation in terms of CLS; see Section 3.4 for this choice. CLS-estimation in the univariate INAR(p) context has been discussed, e.g., in Du and Li (1991) and Zhang et al. (2010). In both papers, the reasoning is performed along the lines of Klimko and Nelson (1978), which was originally developed for CLS-estimation of time series with the very general structure '$\mathbb{E}[X_n|X_{n-1}, \ldots] = g_\theta(X_{n-1}, X_{n-2}, \ldots)$', where g_θ may be non-linear. However, as already noticed in Latour (1997), INAR(p) sequences can be represented as standard AR(p) models with white noise immigration terms. This yields ways for inference that are more direct.

Proposition 6. *Let $(\mathbf{X}_n)$ be a d-dimensional INAR(p) sequence as in Definition 3 with immigration-parameter vector $\mathbf{a}_0 \in \mathbb{R}^d_{\geqslant 0} \setminus \{0_d\}$ and reproduction-coefficient matrices $A_k \in \mathbb{R}^{d\times d}_{\geqslant 0}$, $k = 1, 2, \ldots, p$, such that* (2.10) *holds. Then*

$$\mathbf{u}_n := \mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^{p} A_k \mathbf{X}_{n-k}, \quad n \in \mathbb{Z},$$

defines a (dependent) white-noise sequence, i.e., $(\mathbf{u}_n)$ *is stationary,* $\mathbb{E}\,\mathbf{u}_n = 0_d$*,* $n \in \mathbb{Z}$*, and*

$$\mathbb{E}\left[\mathbf{u}_n\,\mathbf{u}_{n'}^\top\right] = \begin{cases} \operatorname{diag}\left(\left(1_{d\times d} - \sum_{k=1}^{p} A_k\right)^{-1}\right), & n = n', \\ 0_{d\times d}, & n \neq n'. \end{cases}$$

Proof. This can be shown by straightforward (if lengthy) calculations; see Appendix A.1. □

As a consequence of Proposition 6, a d-variate INAR(p) process can be represented as a standard d-variate autoregressive time series with (dependent) white-noise errors:

Corollary 7. *Let* $(\mathbf{X}_n)$ *be the multivariate INAR(p) sequence and* $(\mathbf{u}_n)$ *the white-noise sequence from Proposition 6. Then* $(\mathbf{X}_n)$ *solves the system of stochastic difference-equations*

$$\mathbf{X}_n = \mathbf{a}_0 + \sum_{k=1}^{p} A_k\,\mathbf{X}_{n-k} + \mathbf{u}_n, \quad n \in \mathbb{Z}\,.$$

Such vector-valued time series with linear autoregressive structure have early on been examined; see, e.g., Hannan (1970). However, estimation in a multivariate context requires cumbersome notation. In order to make our results comparable, we follow one reference throughout, namely the monograph Lütkepohl (2005). Adapting its notation is also the reason why we work with wide matrices—i.e., matrices having a number of columns in the order of the sample size—instead of the more common long matrices.

Definition 8. *Let* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *be an* $\mathbb{R}^d$*-valued sequence, where we interpret* $\mathbf{x}_k$ *as a column vector. Fix* p *and* $n \in \mathbb{N}$*,* $p < n$*, and define the* multivariate conditional least-squares estimator *as*

$$\begin{aligned} \hat{\theta}_{CLS}^{(p,n)} : \mathbb{R}^{d\times n} &\longrightarrow \mathbb{R}^{d\times(dp+1)} \\ (\mathbf{x}_1,\ldots,\mathbf{x}_n) &\longmapsto \hat{\theta}_{CLS}^{(p,n)}\,(\mathbf{x}_1,\ldots,\mathbf{x}_n) := \mathbf{Y}\,\mathbf{Z}^\top\left(\mathbf{Z}\,\mathbf{Z}^\top\right)^{-1}, \end{aligned}$$

where

$$\mathbf{Z}\,(\mathbf{x}_1,\ldots,\mathbf{x}_n) := \begin{pmatrix} \mathbf{x}_p & \mathbf{x}_{p+1} & \ldots & \mathbf{x}_{n-1} \\ \mathbf{x}_{p-1} & \mathbf{x}_p & \ldots & \mathbf{x}_{n-2} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_{n-p} \\ 1 & 1 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{(dp+1)\times(n-p)}$$

is the design matrix *and* $\mathbf{Y}\,(\mathbf{x}_1,\ldots,\mathbf{x}_n) := (\mathbf{x}_{p+1},\mathbf{x}_{p+2},\ldots,\mathbf{x}_n) \in \mathbb{R}^{d\times(n-p)}$.

Dealing with multivariate time series the following notations turn out to be useful:

Definition 9. *The vec$(\cdot)$-*operator *takes a matrix as its argument and stacks its columns. The binary $\otimes$-operator is the* Kronecker operator*: for an $m\times n$ matrix $A = (a_{i,j})$ and a $p\times q$ matrix B, $(A\otimes B)$ is the $mp \times nq$ matrix consisting of the block-matrices $a_{i,j}B$, $i = 1,\dots,m,\ j = 1,\dots,n$.*

The vec-notation arises because the estimator is matrix-valued and we have no notion of the covariance of a random matrix. As we will see the $\otimes$-notation is strongly related to the vec-operator. For a large collection of properties of these operators; see Appendix A of Lütkepohl (2005). The following theorem collects all relevant information for CLS-estimation of multivariate INAR(p) sequences. Together with the approximation results from Section 2.5, this theorem is the theoretical basis for our Hawkes estimation procedure.

Theorem 10. *Let $(\mathbf{X}_n)$ be a d-dimensional INAR(p) sequence as in Definition 3 with immigration-parameter (column) vector $\mathbf{a}_0 \in \mathbb{R}^d_{\geqslant 0}\setminus\{0_d\}$, and reproduction-coefficient matrices $A_k \in \mathbb{R}^{d\times d}_{\geqslant 0}$, $k \in \{1,2,\dots,p\}$, such that* $\operatorname{spr}\left(\sum_{k=1}^p A_k\right) < 1$. *Let*

$$\begin{aligned}\mathbf{B} &:= \left(A_1, A_2, \dots, A_p, \mathbf{a}_0\right) \in \mathbb{R}^{d\times(dp+1)} \quad \textit{and}\\ \hat{\mathbf{B}}^{(n)} &:= \hat{\theta}^{(p,n)}_{CLS}\left((\mathbf{X}_k)_{k=1,\dots,n}\right) \in \mathbb{R}^{d\times(dp+1)}\end{aligned}$$

the CLS-estimator with respect to the sample $(\mathbf{X}_k)_{k=1,\dots,n}$. Then $\hat{\mathbf{B}}^{(n)}$ is a weakly consistent estimator for $\mathbf{B}$. Furthermore, let $\mathbf{Z}$ be the design matrix from Definition 8 with respect to $(\mathbf{X}_k)_{k=1,\dots,n}$. Assume that the limit

$$\frac{1}{n-p}\mathbf{Z}\mathbf{Z}^\top \xrightarrow{\text{p}}: \Gamma \in \mathbb{R}^{(dp+1)\times(dp+1)}, \quad n\longrightarrow\infty, \tag{3.1}$$

exists and is invertible. In addition, assume that the model is irreducible in the sense that $\mathbb{P}[\mathbf{X}_{0,i} = 0] < 1$, $i = 1,2,\dots,d$. Then, for the asymptotic distribution of $\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) \in \mathbb{R}^{d^2p+d}$, *one has, for $n\to\infty$,*

$$\begin{aligned}&\sqrt{n-p}\Big(\operatorname{vec}\big(\hat{\mathbf{B}}^{(n)}\big) - \operatorname{vec}\big(\mathbf{B}\big)\Big)\\ &\qquad\qquad \xrightarrow{\text{d}}\ \mathcal{N}_{d^2p+d}\Big(0_{d^2p+d}, \left(\Gamma^{-1}\otimes 1_{d\times d}\right) W \left(\Gamma^{-1}\otimes 1_{d\times d}\right)\Big),\end{aligned}$$

where

$$W := \mathbb{E}\left[\left(\mathbf{Z}_0\otimes 1_{d\times d}\right)\mathbf{u}_0\left(\left(\mathbf{Z}_0\otimes 1_{d\times d}\right)\mathbf{u}_0\right)^\top\right] \in \mathbb{R}^{(d^2p+d)\times(d^2p+d)} \tag{3.2}$$

with

$$\mathbf{u}_0 := \mathbf{X}_0 - \mathbf{a}_0 - \sum_{k=1}^p A_k\mathbf{X}_{-k} \quad \textit{and} \quad \mathbf{Z}_0 := \left(\mathbf{X}_{-1}^\top, \mathbf{X}_{-2}^\top, \dots, \mathbf{X}_{-p}^\top, 1\right)^\top.$$

Proof. In view of Corollary 7, it suffices to prove Theorem 10 for the corresponding vector-valued autoregressive time series. So the distributional properties of the CLS-estimator can be derived similarly as in Lütkepohl (2005), pages 70–75, where independent errors are assumed. We provide a highly self-contained proof for the dependent white-noise case in Appendix A.2. □

Note that the condition $\mathbb{P}[\mathbf{X}_{0,i} = 0] < 1$, $i = 1, 2, \ldots, d$, in Theorem 10 above is purely technical: if we had $\mathbb{P}[\mathbf{X}_{0,i_0} = 0] = 1$ for some i_0, this would imply that in one component of our sample we cannot observe any events. We may exclude this case with a clear conscience.

3.2 The Hawkes estimator

Combining Theorem 10 with the basic approximation from Section 2.5 yields the following estimator for multivariate Hawkes processes:

Definition 11. *Let* $\mathbf{N} = \left(N^{(1)}, N^{(2)}, \ldots, N^{(d)}\right)$ *be a d-variate Hawkes process with baseline-intensity vector* $\eta \in \mathbb{R}^d_{\geqslant 0} \setminus \{0_d\}$ *and excitement function* $H = (h_{i,j}) : \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{d\times d}_{\geqslant 0}$ *such that* (2.5) *holds. Let* $T > 0$ *and consider a sample of the process on the time interval* $(0, T]$*. For some* $\Delta > 0$*, construct the* $\mathbb{N}^d_0$*-valued* bin-count sequence *from this sample:*

$$\mathbf{X}^{(\Delta)}_k := \left(N^{(j)}\Big(\big((k-1)\Delta, k\Delta\big]\Big)\right)_{j=1,\ldots,d}, \quad k = 1, 2, \ldots, n := \lfloor T/\Delta \rfloor . \tag{3.3}$$

Define the multivariate Hawkes estimator *with respect to some support* s, $\Delta < s < T$*, by applying the CLS-operator from Definition 8 with maximal lag* $p := \lceil s/\Delta \rceil$ *on these bin-counts:*

$$\hat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta}\hat{\theta}^{(p,n)}_{CLS}\left(\left(\mathbf{X}^{(\Delta)}_k\right)_{k=1,\ldots,n}\right). \tag{3.4}$$

We collect the main properties of the estimator in the following remark.

Remark 12. The following additional notation clarifies what the entries of the $\hat{\mathbf{H}}^{(\Delta,s)}$ matrix actually estimate:

$$\left(\hat{H}^{(\Delta,s)}_1, \ldots, \hat{H}^{(\Delta,s)}_p, \hat{\eta}^{(\Delta,s)}\right) := \hat{\mathbf{H}}^{(\Delta,s)} \tag{3.5}$$

From Theorem 10 on estimation of INAR(p) sequences together with the basic approximation in Section 2.5, we see that, for $0 < t < s$,

$$\left(\hat{H}^{(\Delta,s)}_{\lfloor t/\Delta \rfloor}\right)_{ij}, \quad i, j = 1, \ldots, d, \quad \text{respectively,} \quad \left(\hat{\eta}^{(\Delta,s)}\right)_i, \quad i = 1, \ldots, d,$$

are weakly consistent estimates (for $T \to \infty$, $\Delta \to 0$ and $s = \Delta p \to \infty$) for the excitement-function component value $h_{i,j}(t)$, respectively, for the baseline-intensity vector component η_i. Furthermore, we find from Theorem 10 that

$$\operatorname{vec}\left(\hat{\mathbf{H}}^{(\Delta,s)}\right) \overset{\text{approx.}}{\sim} \mathcal{N}_{d^2p+d}\left(\operatorname{vec}\left(\mathbf{H}\right), S^2\right), \tag{3.6}$$

with

$$S^2 := \frac{1}{\Delta^2(n-p)}\left(\Gamma^{-1} \otimes 1_{d\times d}\right) W \left(\Gamma^{-1} \otimes 1_{d\times d}\right),$$

where Γ and W are defined as in (3.1) and (3.2) with respect to the bin-count sequences. Substituting Γ and W with their empirical versions yields the covariance estimate

$$\hat{S}^2 := \frac{1}{\Delta^2}\left(\left(\mathbf{Z}\,\mathbf{Z}^\top\right)^{-1} \otimes 1_{d\times d}\right) \sum_{k=p+1}^{n} \mathbf{w}_k\,\mathbf{w}_k^\top \left(\left(\mathbf{Z}\,\mathbf{Z}^\top\right)^{-1} \otimes 1_{d\times d}\right), \tag{3.7}$$

where $\mathbf{Z}$ is the design matrix from Definition 8 with respect to the bin-count sequence and, for $k = p+1, p+2, \ldots, n$,

$$\mathbf{w}_k := \left(\left(\left(\mathbf{X}_{k-1}^{(\Delta)}\right)^\top, \left(\mathbf{X}_{k-2}^{(\Delta)}\right)^\top, \ldots, \left(\mathbf{X}_{k-p}^{(\Delta)}\right)^\top, 1\right)^\top \otimes 1_{d\times d}\right) \cdot \left(\mathbf{X}_k^{(\Delta)} - \Delta\hat{\eta}^{(\Delta,s)} - \sum_{l=1}^{p} \Delta\hat{H}_l^{(\Delta,s)}\,\mathbf{X}_{k-l}^{(\Delta)}\right).$$

Following formulas are useful for implementation of confidence intervals:

$$\operatorname{Cov}\left(\left(\hat{H}_{k_1}^{(\Delta,s)}\right)_{i_1,j_1}, \left(\hat{H}_{k_2}^{(\Delta,s)}\right)_{i_2,j_2}\right) = S^2_{(k_1-1)d^2+(j_1-1)d+i_1,\,(k_2-1)d^2+(j_2-1)d+i_2}, \tag{3.8}$$

for $i_1, i_2, j_1, j_2 \in \{1, \ldots, d\}$ and $k_1, k_2 \in \{1, \ldots, p\}$.

$$\operatorname{Cov}\left(\hat{\eta}_{i_1}^{(\Delta,s)}, \hat{\eta}_{i_2}^{(\Delta,s)}\right) = S^2_{pd^2+i_1,\,pd^2p+i_2}, \quad \text{for } i_1, i_2 \in \{1, ..., d\}.$$

Applying Remark 12 above together with Definitions 8 and 11, our Hawkes estimation procedure may be implemented in a straightforward manner. However, we emphasize that the resulting matrix $\hat{\mathbf{H}}^{(\Delta,s)}$ in (3.4) does not completely specify a fitted Hawkes model; it only yields pointwise estimates on a grid, whereas the true excitement-parameter is a function on $\mathbb{R}_{\geqslant 0}$; see Section 2.1. To complete the estimation, we have to apply some kind of smoothing method over the pointwise estimated values. We work with cubic splines, normal kernel smoothers, and local polynomial regression (`ksmooth()`, `smooth.spline()` and `loess()` in `R`). We find

that the results do not vary significantly. The choice of the estimation parameters bin-size Δ and support s has more impact. Therefore, we focus on the selection of these estimation parameters; see Section 4. The smoothing idea will be relevant in Section 4.2, where we discuss variance issues. In many applications, one can even avoid choosing and applying a smoothing method: practitioners might want to use our estimation procedure from Definition 11 for identifying or rejecting certain parametric models. For such purposes, the pointwise estimates suffice. The same is true if the estimation procedure is used as a mere tool for representing large event data sets; see Section 5.3. Finally, one is often only interested in the integral of the excitement; see the comments after (2.5). In this case, it makes more sense to directly add up the estimates rather than to take the detour over some smoothing method.

3.3 Simulation studies

We check the distributional properties of the Hawkes estimator collected in Remark 12 in a first simulation study. The results are summarized in Figures 3, 4, and 5. Note that at this point we omit the question of selection methods for the estimation parameters s and Δ . This issue will be discussed separately in Sections 4.1 and 4.2. Already in Figure 5, however, the impact of this choice on the estimation results is illustrated.

Bivariate estimation

We simulate 2 000 times from a bivariate Hawkes model with baseline intensity $\eta = (\eta_1, \eta_2)^\top = (0.5, 0.25)^\top$ and excitement function

$$H(t) = \begin{pmatrix} h_{1,1}(t) & h_{1,2}(t) \\ h_{2,1}(t) & h_{2,2}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1_{1<t\leqslant 3}0.25 \\ 0.5(1+t)^{-2} & 1_{t\leqslant\pi}0.2\sin(t) \end{pmatrix}; \tag{3.9}$$

see Figure 2 for this parametrization and Figure 1 for an estimation of a single realization. In each simulation, about 5 000 events in each component are generated and our Hawkes estimator (3.4) is calculated. We apply a bin size $\Delta = 0.2$ and a support parameter $s = 6$. These calculations yield 2 000 matrices of the form $\hat{\mathbf{H}}^{(\Delta,s)} \in \mathbb{R}^{2\times 121}$. We examine the estimations of $\eta_1 = 0.5$, i.e., the baseline-intensity for the first component, and the estimations of $h_{2,1}(1) = 0.125$, i.e., the cross-excitement on component 2 from component 1 after one time unit. These values correspond to the entries $\hat{\mathbf{H}}^{(\Delta,s)}_{1,121}$ and $\hat{\mathbf{H}}^{(\Delta,s)}_{2,9}$ in the estimator matrices. We find that the 2 000 estimates are distributed symmetrically around the true values. The means of the estimates correspond almost completely to the true values. QQ-plots support the asymptotic normality result. For both estimations, we also calculate the variance estimates from (3.7). Comparing the empirical variance of the 2 000 estimates with the 2 000 estimated variances confirms the analytic result. Furthermore, the empirical covering rates for the 95%-confidence intervals are

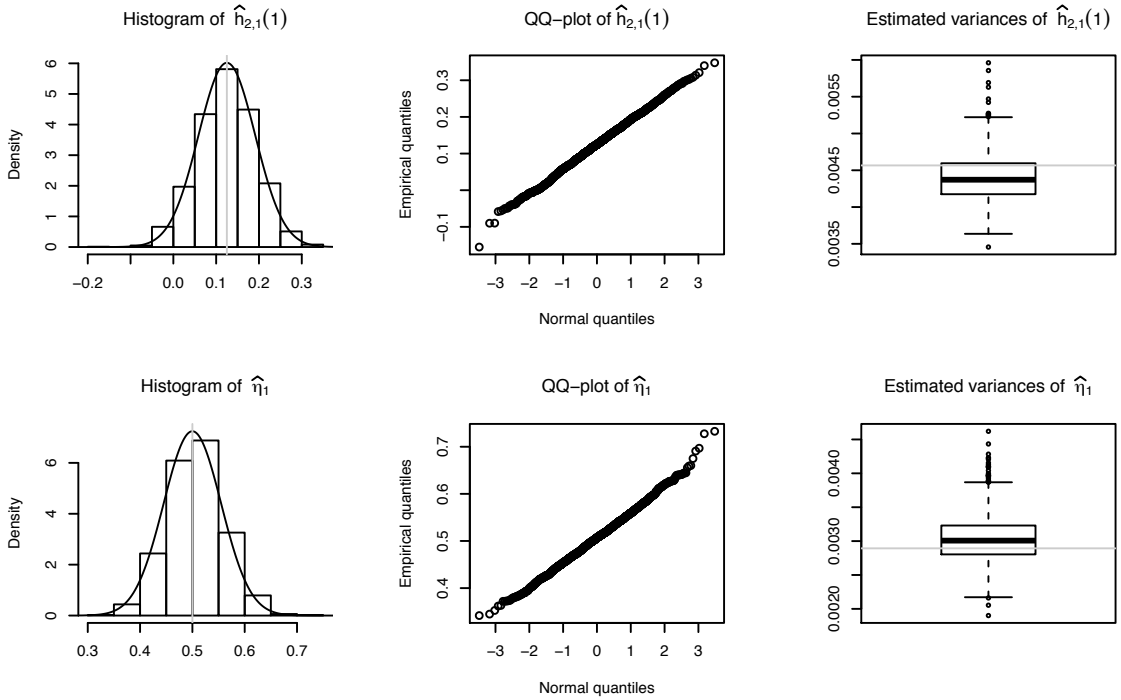


Figure 3: Illustration of the simulation study described in Section 3.3. The study confirms the distributional properties of our estimation procedure collected in Remark 12: We simulate 2 000 times from the bivariate Hawkes process introduced in Figure 2. In each simulation, we realize about 5 000 events in each component. For all of these samples, we calculate our estimator from Definition 11 as well as the covariance estimator from (3.7). These calculations depend on two parameters, the support s and the bin-size Δ. We apply $s = 6$ together with a relatively coarse bin-size $\Delta = 0.2$. The upper-row panels illustrate the estimation of $h_{2,1}(1) = 0.5(1+1)^{-2} = 0.125$; the lower-row panels illustrate the estimation of the baseline-intensity component $\eta_1 = 0.5$. *Left column panels:* the asymptotic normal densities around the true values (grey vertical lines) are added to the histograms. The grey vertical lines refer to the true values. The means of the estimates (not illustrated) would cover the true values. *Middle column panels:* the QQ-plots support the asymptotic normality result. *Right column panels:* the boxplots collect the 2 000 estimated variances; see (3.7). The horizontal grey lines refer to the empirical variance of the 2 000 estimates.

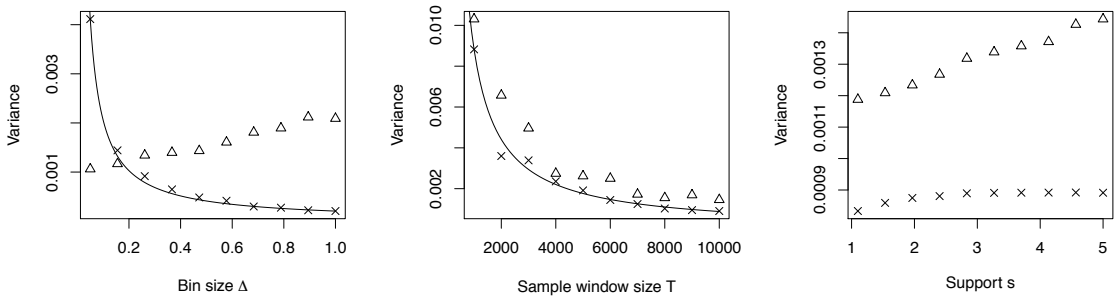


Figure 4: The Hawkes estimator from Definition 11 depends on the bin size Δ, on the size of the sample window T and on the support parameter s. We examine empirically how the variances of the estimates depend on these three parameters. We simulate a very large sample from a univariate Hawkes process with excitement function $h : t \mapsto 1_{t \leqslant 3}(1+t)^{-2}$ and baseline intensity $\eta = 1$. With respect to this single sample, we calculate the estimated variance for the estimates of $h(1) = 0.25$ (crosses) and $\eta = 1$ (triangles) using different Δ, T and s; see (3.7). The solid lines in the two left panels are $\Delta \mapsto c_1 \Delta^{-1}$, respectively, $T \mapsto c_2 T^{-1}$, for some constants $c_1, c_2 > 0$. The curves fit the variance estimates of the excitement-function estimate well. In contrast, the variance of the baseline estimate (triangles) is relatively constant with respect to Δ. In the right panel, we see that the larger the support parameter s, the larger the variances become—this seems natural, as we estimate more parameters with respect to the same sample-size.

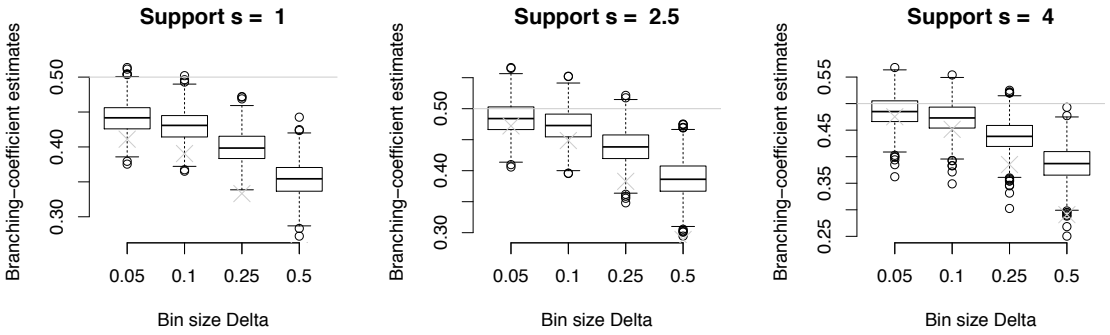
94.5% for the baseline-intensity estimate, respectively, 94.8% for the excitement-value estimate. Note that the applied estimation parameters $\Delta = 0.2$ and $s = 6$ are considerably 'wrong': the bin-size is quite large and the true support of H would be ∞. We may interpret the successful estimation as a sign for the robustness of the method with respect to the estimation parameters.

Variance of the estimates

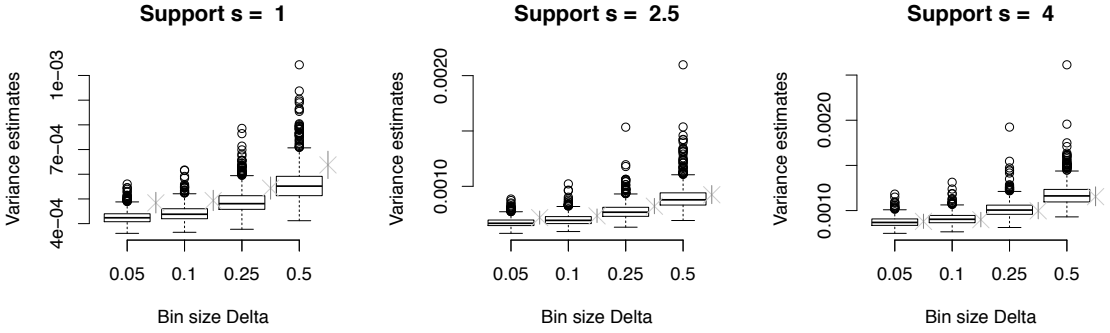
Separately, we examine the impact of the choice of the bin-size Δ, the support s and the size of the sample window $[0, T]$ on the variances of the estimates; see Figure 4. For various Δ, s and T, we calculate (3.7), the estimated covariance of the estimator matrix with respect to a single very large sample of a univariate Hawkes process. We find that the excitation and baseline estimation variances with respect to sample windows $[0, T]$ are proportional to T^{-1}. Variances slightly increase if we increase the support parameter s. The variance of the baseline intensity estimate with respect to Δ is roughly constant in Δ. However, the variance of the excitement estimate with respect to Δ is proportional to Δ^{-1}. Albeit this relation, we will see in Section 4.2 that the excitement estimates are still meaningful for very small values of Δ.

Estimation of the branching coefficient

A particularly meaningful statistic of a Hawkes process is its branching matrix; see the explanations below (2.5). In the univariate case, this matrix provides the (unconditional) probability



(a) The boxplots summarize the estimates. The grey horizontal lines refer to the true branching coefficient 0.5. The crosses refer to the Riemann-type sum $\sum_{k=1}^{\lceil s/\Delta \rceil} \Delta h(k\Delta)$. The plots illustrate the three different errors of our method presented in Section 2.3. *Cut-off error*; see (2.8): for the estimates in the left panel and middle panel, we applied a too small support. This results in an underestimation of the branching coefficient. *Discretization error*; see (2.9): the boxplot centers catch $\sum_{k=1}^{\infty} \Delta h(k\Delta)$ (crosses) rather than $\int h \mathrm{d}t$. *Distributional error* (2.7): the coarser the value of Δ, the greater the distance from the average estimate to the Riemann-type sums (crosses); see the explanations in Section 2.3.



(b) The boxplots collect the variance estimates. The crosses on the right of each boxplot refer to the corresponding empirical variance of the 1 000 estimates together with a bootstrapped 95%-confidence interval. We furthermore observe that the variance of the branching-coefficient estimates is relatively stable over smaller Δ—this is in contrast with the point-wise estimates, whose variance behaves as Δ^{-1}; see Figure 4. This means, if the goal of the estimation is a branching-coefficient estimate, there is no reason not to choose Δ as small as possible. The variance increases when the support is chosen larger. This is intuitive: we estimate more values with the same amount of data. Furthermore, we observe the center of the variance *estimates* (boxplots) gets nearer and nearer to the empirical variances (crosses) the larger the support becomes. For large s and (relatively) small Δ the average of the estimates catches the empirical variance perfectely.

Figure 5: Estimation of the branching coefficient; see Section 3.3: we simulate 1 000 realizations from a univariate Hawkes process with excitement function $h(t) := \exp(-2t)$. Each realization consists of about 1 000 events. For each realization we calculate the Hawkes estimator $\hat{H}^{(\Delta,s)}$ from Definition 11 with respect to $\Delta \in \{0.05, 0.1, 0.25, 0.5\}$ and $s \in \{1, 2.5, 4\}$. From each of these estimates, we derive the branching-coefficient estimate $\Delta \sum_{k=1}^{\lceil s/\Delta \rceil} H_k^{(\Delta,s)}$. We also calculate the corresponding variance estimates derived from (3.8).

that an arbitrarily chosen event has a parent event—that is, that the event is an endogenous event. This explains the alternative notion *rate of endogeneity*. With the notation from (3.5), we define the natural branching-coefficient estimate $\hat{K}^{(\Delta,s)} := \Delta \sum_{k=1}^{p} \hat{H}_k^{(\Delta,s)}$, where $p = \lceil s/\Delta \rceil$. The corresponding variance-estimate can be calculated from (3.7), respectively, (3.8). Note that efficient ways to calculate these particular variance-estimates are given in Embrechts and Kirchner (2017). From the same univariate Hawkes model as above, we simulate 1 000 times and calculate the branching-coefficient estimate as well as the corresponding variance-estimates. We do this with respect to different bin-sizes Δ and different support-parameters s. Figure 5 illustrates the results. In particular, it visualizes the cut-off error, the discretization error, and the distributional error as discussed in Section 2.3. It is no surprise that the empirical variance of the estimates also depends on the estimation parameters Δ and s. The variance estimates adapt accordingly. However, if the support s is chosen (much) too small, then the variance estimates are biased. On the other hand, the bin size Δ hardly influences the bias of the variance estimator: if the support is chosen large enough, then the variance estimates catch the empirical variance well—even for very coarse bin-sizes. We performed this relatively large simulation study on an ordinary laptop. This did not allow to apply really small bin-sizes. But in smaller simulation studies one finds that the estimation bias becomes essentially invisible for $\Delta \leqslant 0.01$ (not illustrated). Also note that integrating a smoothed version of the estimates typically captures the true branching-coefficient even better. However, the distribution of this kind of estimates seems less tractable and—again—as long as Δ is small, the difference between the methods vanishes.

3.4 Alternative estimation methods

There are other approaches to Hawkes estimation that propose alternatives to straightforward MLE. We first give an overview. Then we discuss a specific nonparametric method that is closely related to our estimation procedure.

Overview

In Lewis and Mohler (2011), the authors consider the estimation of univariate Hawkes processes with time-varying baseline intensities. Their simulation study shows that the convergence of the algorithm strongly depends on the underlying true model. It is not clear how feasible this approach would be in the multivariate set-up. In Lemonnier and Vayatis (2014), the authors approximate the excitement functions as well as the time-varying baseline intensities of a multivariate Hawkes process by sums of exponential functions. They show that the approximating model is a Markov process. This reduces the complexity of the likelihood to linear dependence on the number of observations and allows to apply MLE. In Alfonsi and Blanc (2015), the authors incorporate a bivariate marked Hawkes model into a larger model for the price jumps on

a market microstructure level. In this larger model, the Hawkes process models the times when a trade on either side of the market takes place. The authors also consider (i.i.d.) marks that influence the intensity. As in Lemonnier and Vayatis (2014), they approximate the excitement functions by sums of exponential functions. This allows to calibrate the model parameters via MLE. In Hansen et al. (2015), the authors concentrate on the identification of the nonzero excitement functions of a multivariate Hawkes process by minimizing a least-squares objective of the realized intensity subject to an l_1-penalty. The method is comparable to the LASSO method from time series. In Reynaud-Bouret et al. (2014), this method is revisited with respect to diagnostic tests. Note that our method can also deal with the problem of identification of nonzero excitement functions; see the explanations after (5.1). This idea is worked out in Embrechts and Kirchner (2017), where we present a simulation study with fairly high-dimensional Hawkes processes ($d = 10$). Our approach has the advantage that the tuning parameter of the estimation has an interpretation in terms of significance of the observed excitement whereas, as a rule, the LASSO method offers no help when it comes to choosing the penalization parameter.

Finally, we refer to the alternative nonparametric Hawkes estimation approach first introduced in Bacry et al. (2012). In Bacry et al. (2014) the method is further developed and extended for the marked case. Bacry and Muzy (2015) presents an application of the method in the context of high-frequency order flows and price jumps. As this method is closely related to our approach, we discuss it in more detail:

Bacry–Muzy method

The starting point of the Bacry–Muzy method (BM-method) is an integral equation of Wiener–Hopf type from Hawkes (1971b). This equation relates the autocovariance density of a multivariate Hawkes process with its excitement function. It is the analogue of the Yule–Walker equations for discrete-time processes. The BM-method basically consists of two steps:

1. The autocovariance density (respectively, something very related and similar) of the event-stream data is estimated on a grid by a discretization scheme that depends on some bin-size h. Smoothing the estimated values yields continuous-time functions. These functions are plugged into a Wiener–Hopf type integral equation as described above.
2. For a maximal support $A > 0$ and some $Q \in \mathbb{N}$, the Wiener–Hopf equation is discretized and solved for values of the excitement function by a quadrature with Q quadrature points.

We are not sure if this double discretizing and smoothing—once for the autocovariance density and then for the numerical solution of the Wiener–Hopf equation—is necessary: directly applying Yule–Walker estimation to the bin-count sequences would presumably yield similar results—and less calculation and consideration at that.

In Bacry and Muzy (2015), the authors show that their procedure can be extended for the estimation of marked processes as well: under the assumption that the *mark functions* (Bacry and Muzy, 2015), respectively, *impact functions* (Liniger, 2009) only take a finite number of values, one can express the marked process as a higher dimensional Hawkes process without marks. This latter process is calibrated by the method for the unmarked case, and the estimates are appropriately transformed to the original marked model. In Kirchner and Vetter (2017), we apply this method in combination with our estimation method.

As for results, in most cases the estimates of the BM-method and of our method are presumably quite similar if we set $\Delta := h$, $s := A$, and $p := Q := \lceil A/h \rceil$. Or alternatively, if we want to mimic the double smoothing from the BM-method, we set $\Delta := h$ and $s := A$ as before, and, in addition, $\tau := \lceil A/Q \rceil$. Here, $\tau > \Delta$ denotes a bandwidth over which our excitement estimates from Definition 11 are smoothed; see Section 4.2. Note that for the BM-method, the choice of A and the corresponding cut-off error (2.8) are not discussed whereas the choice of s will be discussed in Section 4.1. Despite the similarities to the BM-method, we believe that our method offers important new insights into Hawkes process estimation:

1. **Asymptotic distribution:** neither the BM-method nor the approaches mentioned in the overview present theoretical results regarding the distribution of the estimates. In contrast, we give (3.6) and (3.7) that open the door to confidence intervals and testing. For a particularly fertile application of our distributional results; see Embrechts and Kirchner (2017). Also note that in time series theory, the asymptotic distribution for YW-estimates is typically derived by noting—in a first step—that YW-estimates and CLS-estimates are asymptotically equivalent and then—in a second step—applying the asymptotic distribution of the CLS-estimates. Similarly, we can expect that the path towards the asymptotic distribution of the BM-method leads over our CLS-approach.
2. **Simplicity:** our method is the natural extension of the naive way to estimate general intensities of point processes as step-functions: namely normalizing the number of events in a bin by the length of the bin. Our method is also simpler in the sense that it can be explained without introducing concepts like autocovariance densities or Wiener–Hopf integral equations together with the theory of their solutions. Given the approximation insight from Kirchner (2016), we only need the elementary concept of linear least-squares. Furthermore, for the BM-method, there are more choices involved that will influence the estimation: in addition to the bin-size choice h and the support choice A one has to choose a smoothing method for the autocovariance function estimate as well as the quadrature for the solution of the integral equation (and its parameters). In contrast, our method depends on only two parameters—with quite tractable effects at that. As a minor remark, note that the output of our calculations (3.5) includes the baseline-intensity estimates whereas for the BM-method one has to do some further calculations.

3. **Efficiency and bias:** both methods include the cut-off error, the discretization error, and the distributional error as presented in Section 2.3. In the BM-method, by the two-step procedure, these errors seem less tractable and an additional (if small) error occurs from the Gaussian quadrature. Furthermore, in time series theory, for small sample-sizes, CLS-estimates have a smaller variance than YW-estimates because the YW-method depends on estimates of the autocovariance at large lags. These estimates typically have a very large variance as there is so little data for their estimation. The same can be expected for the point process world. Finally, it is well-known that YW- and CLS-estimates can become quite different for (multivariate) time series near unit roots. Citing from Section 4.4 in Reinsel (1997): 'When the vector AR process is ... near nonstationarity, it is known that the CLS-estimator ... still performs consistently, whereas the Yule–Walker estimator may behave much more poorly with considerable bias, ...' Given the interest in Hawkes models near criticality, e.g., in Hardiman, S.J. et al. (2013) or Jaisson and Rosenbaum (2015) this is another advantage of our CLS-based method. It would be interesting to perform a large simulation-study that compares the two methods more systematically and quantitively.

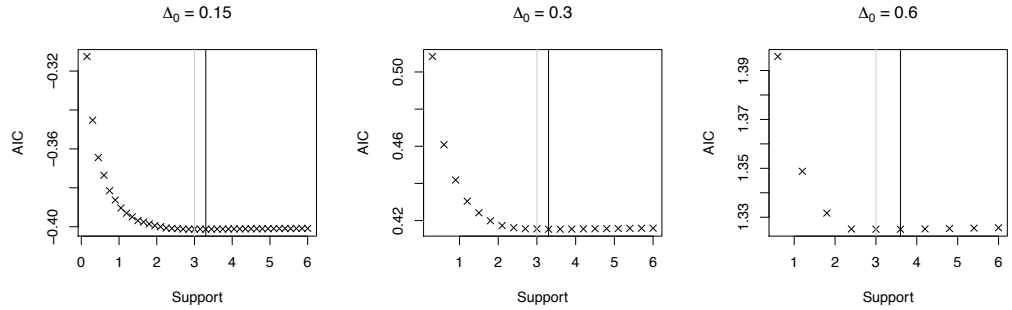
4 Refinements

Our Hawkes estimator $\hat{\mathbf{H}}^{(\Delta,s)}$ from Definition 11 depends on a bin size $\Delta > 0$ and on a support $s > 0$. In the following section, we present procedures for sensible choices of these parameters. Furthermore, we discuss numerical and diagnostic issues.

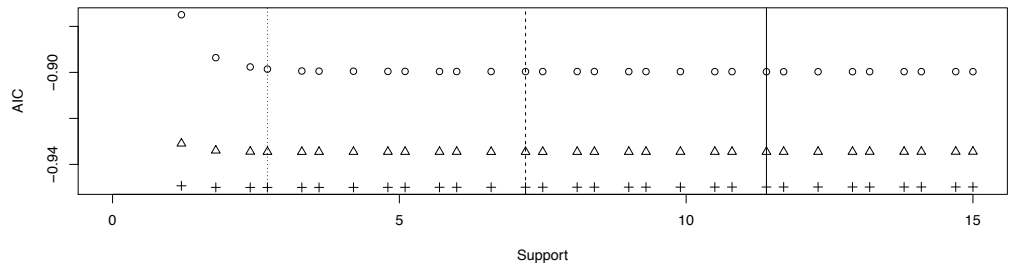
4.1 Choice of support

Estimating the support of the excitement function of a Hawkes process corresponds to estimating the largest lag of a nonzero reproduction-coefficient (matrix) of the approximating INAR sequence. In view of the VAR(p) representation of INAR(p) sequences from Corollary 7, we can use any model-selection procedure stemming from traditional time series analysis; see Chapter 4.3 of Lütkepohl (2005) for an overview of such procedures in the multivariate context. For comparison of different order-selection methods for univariate INAR(p) sequences; see Marques da Silva (2005). As a most common example, we apply Akaike's information criterion (AIC); see Akaike (1973).

We work in the setup of Definition 3.2. The starting point is a sample of a d-variate Hawkes process on $[0,T]$ together with a preliminary bin-size $\Delta_0 > 0$. In our experience, a preliminary bin-size of about one event on an average per bin and component is a good choice; see the argumentation on the distributional error in Section 4.2. With respect to this Δ_0, we calculate



(a) We consider a univariate Hawkes process with excitement function $h(t) := 1_{t\leqslant 3}\exp(-t)$. The true support of the excitement is 3 (grey vertical lines). About 40 000 events are simulated from this model. Following the ideas brought forward in Section 4.1, we apply automatic support-selection on this single large sample using the AIC-criterion with three different values for the preliminary bin-size $\Delta_0 > 0$. The value where the minimum AIC-value is attained (black vertical lines) hardly depends on Δ_0 and though all three bin-sizes are rather coarse, the true support is estimated correctly up to few 'Δ-ticks'.



(b) We examine the infinite support case. To that aim, we consider simulated data from three univariate Hawkes process with excitement functions $h_\alpha(t) := 0.5\alpha\exp(-\alpha t)$, where $\alpha = 1.1$ (circles for AIC-values and solid vertical line for AIC-minimizing support), $\alpha = 1.5$ (triangles and dashed line) and $\alpha = 2$ (crosses and dotted line). The larger α, the lighter the tail of the function and, as desired, the smaller our estimated support; see Section 4.1. Note that the cut-off error (2.8) is in all three cases so small ($< 10^-3$ and much less) that it will typically be negligible in comparison to the estimation standard errors.

Figure 6: Simulation study on the choice of the support parameter of our estimator from Definition 11; see Section 4.1. Figure 6(a) illustrates the case where the true underlying support is finite and Figure 6(b) illustrates the case where it is infinite.

the bin-count sequence(s) from the data as in (3.3). Now let $n_0 := \lfloor T/\Delta_0 \rfloor$, $p_0 \in \mathbb{N}$, and let $s_0 := p_0\Delta_0 > 0$ be some very large support, e.g., $s_0 = T/10$. Then, for $p \in \{1, 2, \dots, p_0\}$, we calculate Akaike's information criterion

$$\mathrm{AIC}^{(\Delta_0)}(p) := \log\left(\det \hat{\Sigma}^{(\Delta_0)}(p)\right) + \frac{2pd^2}{n_0 - p}, \tag{4.1}$$

where $\hat{\Sigma}^{(\Delta_0)}(p) := \sum_{k=p+1}^{n_0} \hat{\mathbf{u}}_k^{(\Delta_0,p)} \hat{\mathbf{u}}_k^{(\Delta_0,p)\top} / (n_0 - p)$. Here, with the notation from Remark 12,

$$\hat{\mathbf{u}}_k^{((\Delta_0),p)} := \mathbf{X}_k^{(\Delta_0)} - \Delta_0 \hat{\eta}^{(\Delta_0, p\Delta_0)} - \sum_{l=1}^{p} \Delta_0 \hat{H}_l^{(\Delta_0, p\Delta_0)} \mathbf{X}_{k-l}^{(\Delta_0)}, \quad k = p+1, p+2, \dots, n_0,$$

denote the estimated prediction-error vectors with estimated coefficient-matrices from the fit of the approximating INAR(p)-model with respect to p lags; see Lütkepohl (2005) for the multivariate AIC-formula (4.1). Finally, we choose $\hat{p}^{(\Delta_0)} := \mathrm{argmin}_{p \leqslant p_0} \mathrm{AIC}(p)$ as estimated maximal lag for the approximating INAR model, respectively, we choose $\hat{s}^{(\Delta_0)} := \hat{p}^{(\Delta_0)}\Delta_0$ as support parameter in the calculation of the Hawkes estimator (3.4). The estimate $\hat{s}^{(\Delta_0)}$ depends on Δ_0. However, given our approximation result in Theorem 4, we expect the estimates to be quite stable over the bin-size choices. Indeed, the examples below confirm this view: for large sample-sizes, we always get that $\hat{s}^{(\Delta_0)} \in [s \pm 2\Delta_0)$, where s denotes the true underlying (finite) support of the data generating excitement function.

In parametric Hawkes setups, the support of the Hawkes excitement function is typically chosen infinite. Our estimation procedure, however, assumes finite excitement. Note that from (2.5), we get that the excitement functions necessarily vanish for large times. In other words, the influence of the tail of the excitement on the model is negligible; see (2.8) and Proposition 5. The only question remaining is how we can choose a support $p \in \mathbb{N}$, respectively, $s > 0$, large enough so that the truncated model with the truncated excitement is a good approximation for the true model. We will see in the examples below that in this infinite-support case the AIC-approach from above also turns out to be helpful. As a side remark note that the standard parametric approach is not free from cut-off errors either, as we only observe data in finite time-windows.

Example with exponential decay

We simulate from a univariate Hawkes model with excitement function $h(t) := 1_{t \leqslant 3} \exp(-t)$ and calculate the AIC-minimizing support with respect to different preliminary bin-sizes Δ_0. All results catch the true support very well; the estimated results are remarkably stable with respect to the choice of Δ_0. Next, we consider the case of infinite support $t \mapsto 0.5\alpha \exp(-\alpha t)$ with respect to three different decay parameters $\alpha \in \{1.1, 1.5, 2\}$. Again, we simulate large sam-

ples for each of the three α-values and then calculate the three corresponding AIC-minimizing support estimates. The smaller α, the larger the tail of the true excitement function becomes and—as desired—the larger the support estimates $\hat{s}^{(\mathrm{AIC})}(\alpha)$ get. For all choices of α, the ignored excitement weights, $0.5\alpha\int_{\hat{s}^{(\mathrm{AIC})}(\alpha)}^{\infty}\exp(-\alpha t)\mathrm{d}t$, are very small (less than 10^{-3}). That is, in the exponential-decay case, our method controls the cut-off error (2.8) well.

Examples with power-law decay

Next, we consider the case where the excitement function is governed by an extremely slowly decaying power-law such as

$$h(t) := \beta(\alpha-1)(1+t)^{-\alpha} \tag{4.2}$$

with $\alpha - 1 > 0$, small, and $\beta \in (0,1)$. We apply the AIC-based support-choice method from above to a large realization of a Hawkes process with excitement of form (4.2). The realizations of $\hat{s}$ increase when α decreases (not illustrated)—as desired and just like in the (infinite) exponential-decay case. However, the cut-off error $\int_{\hat{s}}^{\infty} h(r)\mathrm{d}r$ is now typically large. E.g., for $\alpha = 1.15$ and $\beta = 0.5$, our method yields a support parameter $\hat{s} \approx 20$—quite independently of the choice of Δ_0. This estimated support leaves a very large cut-off error of $\int_{20}^{\infty} h(t)\mathrm{d}t \approx 0.32$. In other words, we miss nearly two thirds of the total excitement ($= 0.5$). How is it possible that our selection method opts for such a faulty model? The AIC-based support selection chooses $\hat{s}$ such that the increase of local explanatory power by looking even further back into the past is relatively small. In other words, despite the large cut-off error, the truncated model $(\eta^{(\hat{s})}, h^{(\hat{s})})$ is locally very similar to the Hawkes model (η, h), where $h^{(\hat{s})}(t) := 1_{t\leqslant s}h(s)$ and $\eta^{(\hat{s})} := \eta + \eta/(1-\beta)\int_s^{\infty} h(t)\mathrm{d}t = \eta(1+\beta/(1-\beta)(1+\hat{s})^{1-\alpha})$. It would be interesting to discuss this truncation-approximation more quantitively as the cut-off comes with enormous computational advantages (for all estimation methods). We will touch the specific issues arising in the context of Hawkes excitement-functions with power-law decay again in Section 5.2.

Bivariate example

Finally, we consider a bivariate Hawkes model with the excitement function H from (3.9). We realize a single large sample from this model. Then we simulate another sample from a truncated version of the model, with

$$H^{(\mathrm{tr})}(t) = \begin{pmatrix} h_{1,1}(t) & h_{1,2}(t) \\ h_{2,1}^{(\mathrm{tr})}(t) & h_{2,2}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1_{1<t\leqslant 3}0.25 \\ 1_{t\leqslant 4}0.5(1+t)^{-2} & 1_{t\leqslant\pi}0.2\sin(t) \end{pmatrix}.$$

The AIC-minimizing support estimate is 9.5 for the original model and 4.2 for the truncated model. So the AIC-approach is able to discriminate between these cases.

4.2 Choice of bin size

Below, we discuss the choice of the bin size $\Delta > 0$ for the Hawkes estimator $\hat{\mathbf{H}}^{(\Delta,s)}$ from Definition 11. At first sight, one can interpret the choice of the bin size Δ as a bias/variance trade-off: the smaller Δ, the smaller the potential bias stemming from the model approximation, i.e., the smaller the errors (2.7) and (2.9). At the same time, due to the $1/\Delta$ factor in the calculation of the estimator matrix $\hat{\mathbf{H}}^{(\Delta,s)}$ from (3.4), its (componentwise) variance increases when Δ decreases. In a simulation study, we simulate 100 times from a univariate Hawkes model with excitement function $h(t) = \exp(-1.1t)$. We suppose a reasonable support $s > 0$ has already been chosen by a procedure as described in Section 4.1. For each sample, we calculate the Hawkes estimator with respect to three different bin sizes $\Delta \in \{0.1, 0.5, 1\}$. Figure 7 collects the estimation results in boxplots. The bias/variance trade-off is obvious. Note, however, that we had to choose the bin-size quite large to make the bias visible at all. In the following, we discuss various issues related to bin-size choice. Overall, we want to argue that one should choose the bin size Δ as small as (computationally) possible.

Estimation of baseline intensities and branching coefficients

If we are only interested in estimates of aggregated values such as baseline-intensity components or branching coefficients, the bin-size $\Delta \downarrow 0$ leaves the variance of the estimates nearly uneffected. Indeed: let $\hat{K}^{(\Delta,s)}$ denote the branching-matrix estimate where each entry is estimated as the branching coefficient in the univariate example from Section 3.3. Then we have that

$$\hat{\Lambda} := \hat{\eta}^{(\Delta,s)}(1 - \hat{K}^{(\Delta,s)})^{-1} \tag{4.3}$$

is essentially equal to $\sum_{n=1}^{\lceil T/\Delta \rceil} \mathbf{X}_n^{(\Delta)} / T = \mathbf{N}((0,T])/T$ and therefore (4.3) does not depend on Δ. The same is true for $\hat{\eta}^{(\Delta,s)}$ and $\hat{K}^{(\Delta,s)}$: with respect to the same point process data, $\hat{\eta}^{(\Delta)}$ and $\hat{K}^{(\Delta)}$ hardly vary in Δ. Figures 4 and 5 confirm this argumentation: if anything, the variances of baseline-intensity and branching-coefficient estimate decrease for $\Delta \downarrow 0$. Consequently, if we only want to estimate baseline intensities and branching coefficients, there is no reason why we should not choose the bin size Δ as small as computationally possible.

Estimation of the excitement function

If we want to estimate specific values of the excitement functions, we face an increasing variance for decreasing Δ; see the Figures 4 and 7. However, we should keep in mind that the final goal of our analysis may be the estimation of the excitement-function components h_{ij}— and not only for a finite number of their values $h_{ij}(k\Delta)$, $k = 1, 2, \ldots, p$. When we apply some smoothing method on these values, a smaller Δ typically leads to an 'averaging' over more point

estimates. This averaging balances the increase in pointwise variance. At first sight, this seems an odd thing to do: the bandwidth of the smoothing method seems more or less equivalent to the bin size of the first bin-wise aggregation. In particular, it looks as if the bias that we avoided by applying a small bin-size Δ for the original estimation is reintroduced by choosing a coarser bandwidth τ for the smoothing of the estimates. This is only partly right. To understand this, we have to reconsider the errors in the approximative model equation from Section 2.3: a too large bin-size Δ effects two of the approximation errors, namely, the distributional error (2.7) and the discretization error (2.9). The crucial observation is that the distributional error is *not* reintroduced by the smoothing method. We repeat the source of the distributional error: for example, supporse that we observe three events in a bin. In the approximating bin-count model, we explain all of these three events by events in earlier bins. But the last of the three considered events is very likely a result of the two previous events in the bin in question itself! We do not account for this in our approximative model equation. Consequently, we overestimate the influence of the past on the bin (or overestimate the baseline intensity). This (important) part of the bias becomes smaller and smaller with decreasing Δ and nearly vanishes if Δ is chosen so small that no bin contains more than one event. This distributional error is *not* reintroduced by any aggregating smoothing method applied on the pointwise estimates. This effect can be easily observed empirically (not illustrated): for a single large simulated sample from a Hawkes process, we calculate the pointwise Hawkes estimator from Definition 11 with respect to three different bin-sizes Δ. Applying a cubic smoothing-spline procedure on the results we get some function estimates. The bias of these smoothed functions vanishes for $\Delta \downarrow 0$ and, at the same time, their variance does not increase. We conclude: if the goal of the estimation procedure is a completely specified Hawkes model, then the smallest Δ that is computationally convenient may be chosen.

Bias correction

After explaining the relation between support and bin-size choice on the one side with cut-off error (2.8) and distributional error (2.7) on the other side, we next propose a heuristic bias correction idea that mainly corrects the discretization error (2.9). The method can be applied in the special (but typical) case when the underlying true excitement function is decreasing and convex. We demonstrate the idea in the univariate case: up to now, we interpreted the k-th entry of $\hat{\mathbf{H}}^{(\Delta,s)}$, i.e., $\hat{h}_k^{(\Delta,s)}$, as an estimate for $h(k\Delta)$, the excitement from an event at some time $t \in \mathbb{R}$ on the conditional intensity at time $t + k\Delta$; see Remark 12. But what $\hat{h}_k^{(\Delta,s)}$ measures is in fact an *aggregation of the excitement from one bin to the k-th next bin*. As a consequence, rather than

estimating $h(k\Delta)$, the term $\hat{h}_k^{(\Delta,s)}$ estimates more an average of the form

$$a_k := \frac{1}{2\Delta} \int_{(k-1)\Delta}^{(k+1)\Delta} h(t)\mathrm{d}t, \quad k = 1, 2, \ldots, p.$$

For convex functions h, a_k is always larger than $h(k\Delta)$. If at the same time, h is decreasing, we have $a_k \approx h\left((k-0.5)\,\Delta\right)$. Following these heuristics, a large part of the discretization error can be corrected by shifting the estimation grid of the excitement function by 0.5Δ to the left. In other words, if we interpret $\hat{h}_k^{(\Delta,s)}$ as an estimator for $h\big((k-0.5)\Delta\big)$, this typically corrects a large part of the bias. This bias correction may be useful if we have to choose Δ quite coarse.

Computational issues

If we choose a very small bin-size Δ, computation time becomes an issue. The calculations in (3.4) require the construction of the design matrix $\mathbf{Z}$ from Definition 8 with about T/Δ rows and about $d \cdot s/\Delta$ columns. Here, T is the size of the time window, d is the dimension of the process, and s is the support parameter of the estimation. Then the matrix $\mathbf{Z}\mathbf{Z}^\top$ has to be inverted. This square matrix is approximately of size $\lceil d \cdot s/\Delta \rceil \times \lceil d \cdot s/\Delta \rceil$. In short, the smaller Δ, the larger the matrices involved. Note, however, that, for a very small bin-size Δ, the corresponding design-matrix is very sparse. Specialized software makes construction and manipulation of sparse matrices numerically efficient; see Bates and Maechler (2015).

Note that if d is large, we might not be able to choose Δ small enough which leaves significant bias in our estimates. In this case, we propose the following two-step procedure: first we choose a too large (but computationally feasible) bin size Δ_1. With respect to this preliminary Δ_1, we calculate the estimates of all branching coefficients including confidence bounds with respect to some significance level α; see Section 3.3. If the confidence intervals include zero, that is, if the corresponding excitement is not significant, we assume that there is no excitement at all. Given that the underlying true 'excitement-graph' is sparse, the 'excitement-graph estimate' will typically also be sparse. This reduces the dimensionality of the problem remarkably. The reduction allows to reestimate the remaining excitements in a second step *with respect to a much smaller bin-size* Δ_2. In this sense, the bin-size can be applied as a parameter controlling the computational complexity of the method. This two-step procedure is worked out and demonstrated in Embrechts and Kirchner (2017). Also note that the calculation of the covariance matrix estimate (3.7) can be computationally even more challenging than the calculation of the estimator itself. For that reason, we also provide particularly efficient ways for the calculation of the variance of the branching coefficient estimates in the cited paper.

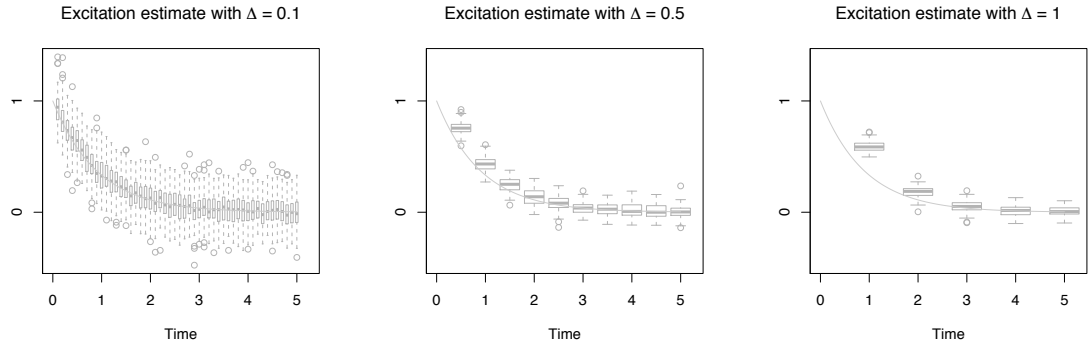


Figure 7: Illustration of the bias/variance trade-off in the choice of the bin size Δ; see Section 4.2. We simulate 100 realizations of a Hawkes process. For each of these 100 samples, we calculate the estimator from Definition 11 with respect to three different bin-sizes $\Delta \in \{0.1, 0.5, 1\}$. The estimates are collected in boxplots. The grey lines denote the true excitement function $h(t) = \exp(-1.1t)$. A larger Δ leads to a larger bias. This is particularly obvious in the first boxplot of the right panel. Note that the bin sizes had to be chosen quite coarse to make this bias visible. A smaller Δ leads to larger pointwise variance of the excitement function value estimates.

Choosing Δ small enough

We now understand that the trade-off related to the bin-size choice is not so much a bias/variance trade-off but rather a bias/computational-issues trade-off! To check if we have chosen Δ small enough, we propose to calculate the (biased) estimate of the baseline intensity vector $\eta := (\eta_i)_{1 \leqslant i \leqslant d}$ for a decreasing sequence of bin sizes $\Delta_0 > \Delta_1 > \Delta_2 > \dots$

The variance of the baseline-intensity estimates $\hat{\eta}_i^{(s,\Delta_n)}$, $n = 0, 1, \dots$ is approximately constant over the different bin-sizes; see Figure 4. This makes the estimates comparable. For $i = 1, 2, \dots, d$, we plot the values $\left(\hat{\eta}_i^{(s,\Delta_n)}\right)_{n=0,1,\dots}$ against $(\Delta_n)_{n=0,1,\dots}$. Typically, one observes a monotone convergence in n to some constant (or d constants for $d > 1$). Plotting confidence intervals around the point estimates indicates when the bias is negligible in comparison to the random noise of the estimate. We will apply this method in the concluding data-example.

4.3 Diagnostics

We see a certain danger in the application of our nonparametric Hawkes estimator from Definition 11. Reasonable graphical results as in Figure 1 might be used as an argument in favor of the Hawkes process as the true model. But this conclusion would be a misuse of the method. In fact, the proposed estimator depends only on second-order properties of the data. So, we have to expect that there is a whole family of point processes that generate the same excitement estimates, although only one of these processes is a genuine Hawkes process. As an example, consider a continuous-time, nonnegative, stationary Markov chain that has the same second-order properties as some given Hawkes process. We use this Markov chain as a stochastic

intensity for another point process; see Daley and Vere-Jones (2009), Example 10.3(e). The resulting doubly-stochastic point process is a point process with different distributional properties than the corresponding Hawkes process. But our estimator will still yield the same results in both cases. As another example, consider a time-reversed Hawkes process. Clearly, this is not a Hawkes process anymore. However, the time-reversed version has the same autocovariance density as the original process and therefore our estimator will again yield the same result.

This means, the application of our estimation approach always ought to be followed by a model test. A most common basis for such a test in our context is a multivariate version of the random time-change theorem for point processes; see Meyer (1971); Brown and Nair (1988): for points $\left(T_k^{(i)}\right)_{k\in\mathbb{Z}}$, $i = 1,\dots,d$, from a d-variate point process with conditional intensity $\Lambda = \left(\Lambda^{(i)}\right)_{i=1,\dots,d}$, one has that $\int_{T_k^{(i)}}^{T_{k+1}^{(i)}} \Lambda^{(i)}(t)\,\mathrm{d}t \sim \mathrm{Exp}(1)$ independently over $i = 1,\dots,d$ and $k \in \mathbb{Z}$. So, after having fit the Hawkes process to point process data, we calculate the corresponding conditional-intensity estimate and time-transform the interarrival times. These transformed interarrival times ought to be compared with theoretical Exp(1)-quantiles in a QQ-plot. Next to this graphical method one ought to apply a Kolmogorov–Smirnov test and an independence test to the transformed interarrival times.

5 Data application

There are two contexts of growing importance where large event-data sets are not the exception but the rule: internet traffic and high-frequency data in financial econometrics. The paper concludes with an exemplary application of the estimation procedure to the latter.

5.1 The data

The data we use stem from the *limit order book (LOB)* of an electronic market. LOBs match buyers and sellers of a specific asset. We will consider a certain future contract. Whoever wants to buy or sell one or several of these contracts has to send his or her orders to the LOB. An order basically consists of two pieces of information: it names (a) the maximal (respectively, minimal) price at which the sender is willing to buy (respectively, to sell) and (b) the desired quantity in terms of numbers of contracts. If the order is matched to another order, the trade is executed. Such orders that immediately find counterparts are called *market orders*. All other incoming orders are stacked in the LOB; these are called *limit orders*. Limit orders either wait for getting executed by a new incoming matching (market) order or—and this happens relatively often—they are withdrawn after some time. The empirical process of time points when orders arrive we call *order flow*. Such an order flow can be modeled by a point process. In particular, our estimation method from Definition 11 allows to analyze the order flow in a Hawkes setup.

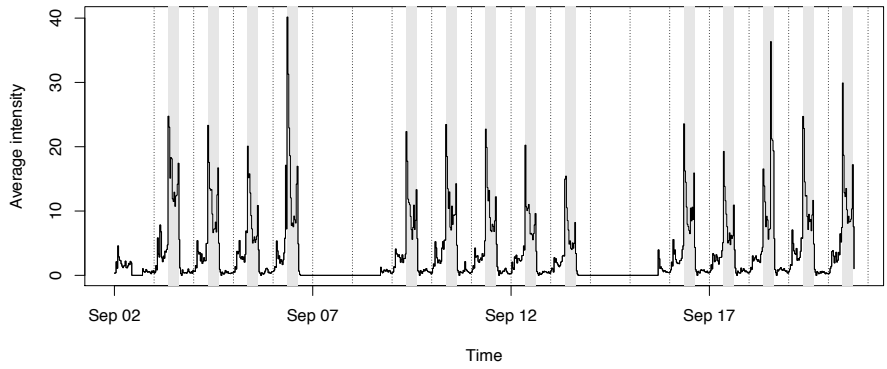


Figure 8: Illustration of our data set; see Section 5.2. The time is Chicago local time. The dotted vertical lines refer to midnights. The black solid line shows the average number of order-book events per second in 30 min windows, as explained in Section 5.1. For our estimation, we only consider the regular trading hours, that is, 8:30am–3:00pm, Chicago time (grey stripes). Inside these stripes, we observe the characteristic U-shape of the intensity. Also note the two preceding smaller U-shapes. These correspond to the regular trading hours of the exchanges in Europe and Asia. The vanishing activity on the first Monday of our data is due to a US-holiday (Labor Day) on September 2, 2013.

For a detailed survey of order-book quantitative analysis; see Gould et al. (2013). Financial intraday histories are attractive for econometric research as there is so much data available. However, by the very differing data qualities, results are sometimes hard to compare. To clarify our starting point, we explain in some detail the context and the preparation of the data.

We consider a sample of the LOB of E-mini S&P 500 futures with most current maturity. The enormous liquidity makes the data attractive for quantitative analysis. Samples of these particular data have also been analyzed in the Hawkes setting, e.g., by Filiminov and Sornette (2012) and Hardiman, S.J. et al. (2013). Our particular data sample was provided by TickData inc. It stems from September 2013. We have a separate data set for quotes and for trades. A new entry in the quotes data corresponds to one of the following three events:

(i) Arrival of some (not marketable) limit order

(ii) Arrival of some market order, i.e., a trade takes place

(iii) Cancellation of some limit orders

In the trade data set, we see the traded price and the number of contracts traded. In both data sets, we observe ties, i.e., multiple events with identical millisecond time-stamps. These ties require special consideration as our model, the Hawkes model, does not allow for simultaneous jumps. We had the opportunity to compare our data with a snapshot of the fully reconstructed LOB. This complete data provide ‘match tags’ for each order. This additional information shows that nearly all multiple events are in fact orders from one single market-participant. This

confirms our point of view. We therefore consider each time stamp in the data sets only once. After the reproduction procedure, we derive two one-dimensional event data sets from our data:

- the *trade data* $\mathcal{T}$ and
- the (pure) *limit-order data* $\mathcal{L}$ that collects all the times when a new non-marketable limit order has arrived or a limit order has been canceled.

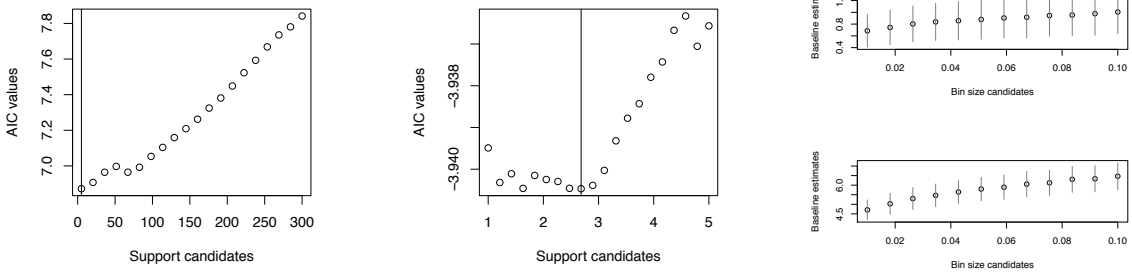
In regular trading hours, i.e., between 8:30am and 3:15pm (Chicago local time), we observe about 5 events per second in the trade data $\mathcal{T}$, and about 12 events per second in the limit-order data $\mathcal{L}$. At Chicago night time, all of these average intensities are up to twenty times smaller. All interarrival-times processes exhibit significant autocorrelation at large lags. This rules out simple standard homogenous Poisson point processes as models as well as other renewal processes. On the other hand, the autocorrelation may also stem from nonstationarities in the underlying true model; see Mikosch and Stărică (2000).

5.2 Bivariate estimation of the market/limit order process

With our nonparametric method from Definition 11, we fit a bivariate Hawkes process $(N^{(\mathcal{T})}, N^{(\mathcal{L})})$ on 30 min-samples of the data $(\mathcal{T}, \mathcal{L})$ in the first three weeks of September 2013. We first illustrate our approach in detail for a single 30 min-sample, namely on data from Friday, 2013/09/06, 10:00am–10:30am (Chicago time). In this specific sample, we observe about 20 000 trades and 40 000 limit orders. Our estimation procedure from Section 3.2 depends on a choice of support and on a choice of bin size. For a sensible choice of these parameters, we apply the methods from Sections 4.1 and 4.2:

Choice of support

As a first step, we calculate the Hawkes estimator with respect to a relatively large preliminary bin-size of $\Delta_0 = 0.5$ sec and for various support candidates between 1 and 300 seconds. As proposed in Section 4.1, we compare the corresponding AIC-values. This coarse analysis shows that the AIC-optimal support is surely less than 20 seconds; see Figure 9(a). Repeating the analysis with respect to a much finer bin-size $\tilde{\Delta}_0 = 0.01$ sec on the interval $(0\,\text{sec}, 20\,\text{sec})$, we find an AIC-minimizing support of about 2.7 sec; see Figure 9(b). Let us note that the obtained minimum is much more clear-cut than in the controlled simulation study from Section 4.1 illustrated in Figure 6. We set $s = 3$ sec. In other words, our support analysis indicates that the process forgets its past after three seconds. This preliminary result is already interesting: it can be interpreted such that—in this sample—the algorithms that drive the market take not more than the last three seconds of the LOB-history into account.

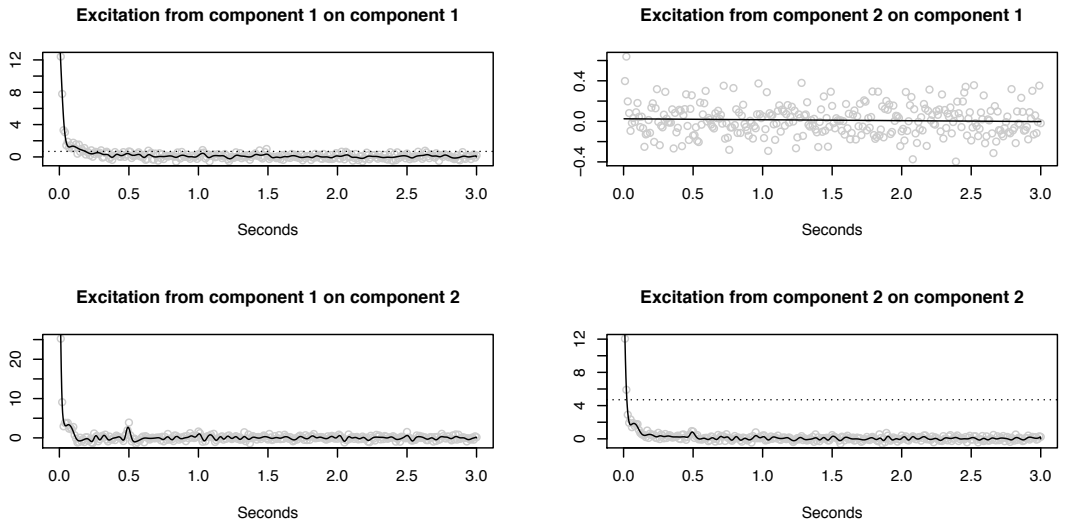


(a) Support analysis with respect to a very coarse preliminary bin-size $\Delta_0 = 1$ sec; see Section 4.1. The estimator from Definition 11 is calculated for different support candidates (in seconds). The corresponding AIC-values are calculated as in (4.1). We establish a quite short AIC-optimal support of the excitement function.

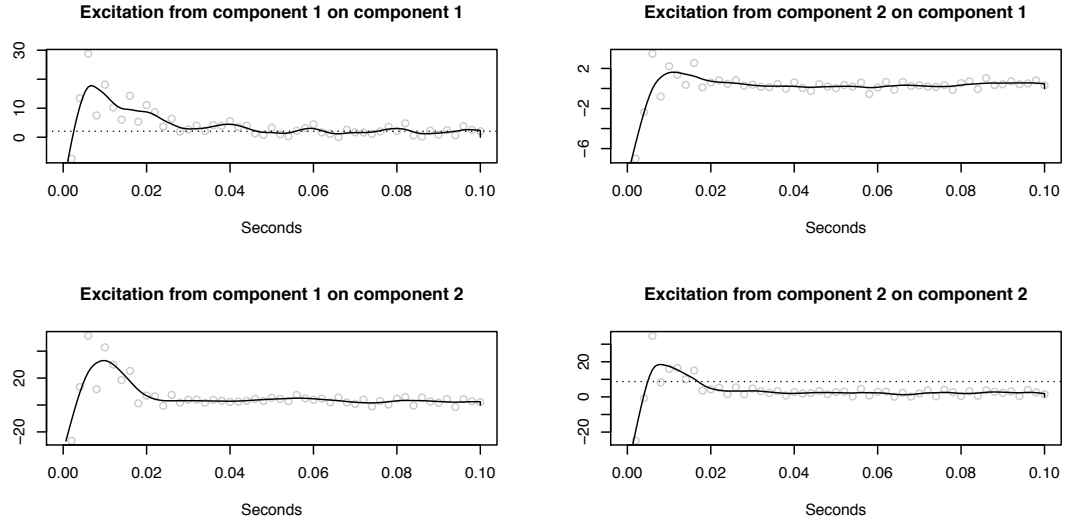
(b) After the rough analysis illustrated in Figure 9(a), we repeat the procedure for smaller support candidates and with respect to a much finer bin-size $\Delta = 0.01$ sec. We find an AIC-optimal support value of about 2.7 seconds. This value is stable over other choices of the bin-size. The attained minumum is remarkably clear-cut compared to the attained minimum in the simuation study illustrated in Figure 6.

(c) Bin-size analysis following the method from Section 4.2. The baseline estimates decrease in both components as the applied bin-sizes decrease. For Δ smaller than 0.01 sec, the decrease is of a lower magnitude than the 95%-confidence intervals. We conclude that, for $\Delta \leqslant 0.01$ sec, the bias of our estimation method becomes negligible.

Figure 9: Preliminary analysis for the bivariate data example $(\mathcal{T}, \mathcal{L})$ (trades/limit orders); see Section 5.2. Our nonparametric Hawkes estimator from Definition 11 depends on a support parameter s and a bin-size parameter Δ. Applying the selection methods from Section 4.1 and Section 4.2, we find that $s = 3$ sec and $\Delta = 0.01$ sec are reasonable choices.



(a) Bivariate fit with respect to bin size $\Delta = 0.01$ sec and support $s = 3$ sec. For the derivation of these estimation parameters; see Figure 9. Eyeball examination reveals local maxima in the lower panels at half seconds. $\log_{10}/\log_{10}$-plots of averaged estimates are given in Figure 11.



(b) We fit the Hawkes model to the same sample as in (a). This time however, we ignore the best support choice and set it naively to $s = 0.1$ sec only. In addition, we apply an extremely small bin size of $\Delta = 0.002$ sec. In the first milliseconds after each event, the results indicate an inhibitory effect; the Hawkes model does not allow for negative excitement. In the smoothed function-estimates, we detect a local maxima at 0.01 sec. For the estimated excitement function of the first component (the trades process) we observe further local maxima at multiples of 0.02 sec.

Figure 10: Exemplary biviariate Hawkes fits on a single 30 min-window; see Section 5.2. We apply two sets of estimation parameters (s, Δ). In the fitted bivariate process, the first component refers to the trade times $\mathcal{T}$ and the second component to the limit order arrivals and cancellations $\mathcal{L}$. The black solid lines are kernel-smoothed versions of the estimates; see the end of Section 3.2. The dotted lines in the diagonal plots refer to the fitted baseline-intensity components.

Choice of bin size

For a reasonable choice of the bin-size parameter Δ, we apply the method from Section 4.2. That is, we examine the impact of the bin-size choice on the estimation. We leave the support $s = 3\,\text{sec}$ fixed and, for different bin-size candidates Δ, we calculate the baseline-intensity estimate $\hat{\eta}_i^{(\Delta)}$, $i = 1, 2$, together with the corresponding confidence intervals; see (3.4) and (3.6) for the necessary calculations. We observe a monotone relation between the bin-size candidates and the corresponding baseline-estimates. However, for $\Delta \leqslant 0.01\,\text{sec}$, the differences of the estimates are of a lower order than their (estimated) confidence intervals; see Figure 9(c). So it is sensible to assume that, for this particular sample, the bias of our estimation method becomes negligible for bin-size choices of $\Delta \leqslant 0.01\,\text{sec}$.

Estimation results for single time window

From the bivariate event data set, we finally calculate the Hawkes estimator from Definition 11 with respect to support $s = 3\,\text{sec}$ and bin size $\Delta = 0.01\,\text{sec}$. Figure 10(a) summarizes the estimation results for this specific time thirty minute window.

The baseline intensity of the limit-order process $\mathcal{L}$ is about four times larger than the baseline intensity of trades process $\mathcal{T}$. In both processes, we observe a strong and quite similar self-excitement. The cross-excitement, however, is obviously directed: we observe a very strong cross-excitement from $\mathcal{T}$ on $\mathcal{L}$, but hardly any effect from $\mathcal{L}$ on $\mathcal{T}$. The estimated interactions can be summarized in the branching-matrix estimate

$$\begin{pmatrix} 0.62(\pm 0.04) & 0.03(\pm 0.01) \\ 0.55(\pm 0.06) & 0.54(\pm 0.03) \end{pmatrix}, \quad \text{i.e.,} \quad \text{`} \begin{pmatrix} \mathcal{T} \overset{0.62}{\rightsquigarrow} \mathcal{T} & \mathcal{L} \overset{0.03}{\rightsquigarrow} \mathcal{T} \\ \mathcal{T} \overset{0.55}{\rightsquigarrow} \mathcal{L} & \mathcal{L} \overset{0.54}{\rightsquigarrow} \mathcal{L} \end{pmatrix}\text{'}. \tag{5.1}$$

See Remark 12 for the calculation of the point estimates as well as the 95%-confidence bounds of the branching-matrix components. Also see the explanations after (2.5) for the interpretation of the branching-matrix that is indicated in the right matrix. If the underlying true excitement functions are heavy-tailed, then the estimates (5.1) may extremely depend on the choice of the support parameter s; see the simulation study described after (4.2). This means, that these kinds of estimates have to be interpreted (and communicated) together with the chosen value for s. In any case, the values and confidence bounds are a good description for the local excitement, i.e., for the influence on the intensities from the past s time units. The largest eigenvalue of matrix (5.1), i.e., the stability-criterion estimate, is 0.72. The strong asymmetry in (5.1) may be interpreted such that the trades cause the limit orders (and cancellations) and not vice versa. In further analysis, we found that the estimated branching-matrix, and in particular the asymmetric cross-excitement, is quite stable over all thirty minute windows of the regular trading hours (not illustrated). Clearly, the knowledge of the distribution of the entries in

(5.1) is attractive beyond confidence intervals: it allows testing for causal connections between event streams. This is particularly important in higher dimensional point-process networks; we demonstrate this possible application of our estimation method in Embrechts and Kirchner (2017)—with special hindsight on computational issues and implementation. We also apply this testing-idea to much larger data sets of limit order book data in Kirchner and Vetter (2017). In the cross-excitement from $\mathcal{T}$ on $\mathcal{L}$, we observe local maxima at half and whole seconds. This effect may have two causes: it reflects a preference either for absolute or for absolute round times. To put it differently: some of the order-sending algorithms that indeed react on trade events may have an implemented lag of half or full seconds.

In a second approach, we fit the Hawkes model to the same sample as above; see Figure 10(b). This time however, we ignore the best support choice and set it naively to $s = 0.1$ sec only. In addition, we apply an extremely small bin-size of $\Delta = 0.002$ sec. In the first milliseconds after each event, the results indicate an inhibitory effect; the standard linear Hawkes model does not allow for negative excitement because it could yield negative intensities with positive probability. In the smoothed function-estimate of the self-excitement of the first component (the trades process), we detect local maxima at 0.02 sec multiples. Also note that in this naive fit, the baseline-intensity estimates are much larger than in the first fit: these large values are a compensation for the too small support choice.

Naturally, the fitted Hawkes model is only completely specified when we smooth the results from the estimation method on the grid by some kind of smoothing mechanism that yields a function $\hat{H} : \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{2\times 2}$. We do this with a cubic smoothing spline method. Having thus completely specified the model, we apply a Kolmogoroff–Smirnov test on the transformed interarrival-times; see Section 4.3. The test rejects the fitted model for the 30 min-window. This is not surprising: given the very large sample-size, we are very likely to include 'abnormal' interarrival times that our model cannot catch; the Kolmogoroff–Smirnov test is particularly sensitive to such outliers. Dividing the 30 min-windows into smaller samples of 100 events yields plausible p-values (not illustrated). For further interpretation of the diagnostics; see the discussion in Section 5.3 below.

Estimation of whole data set

Next, we consider the whole data set; see Figure 8. More specifically, we consider thirteen 30 min-windows in the regular trading hours of the fourteen considered trading days. The support analysis of the windows yields AIC-optimal supports between 2 and 50 seconds, the majority (and the median) being approximately 10 seconds. We first calculate the Hawkes estimator from Definition 11 with respect to $s = 10$ sec and $\Delta = 0.01$ sec. We average the estimates over

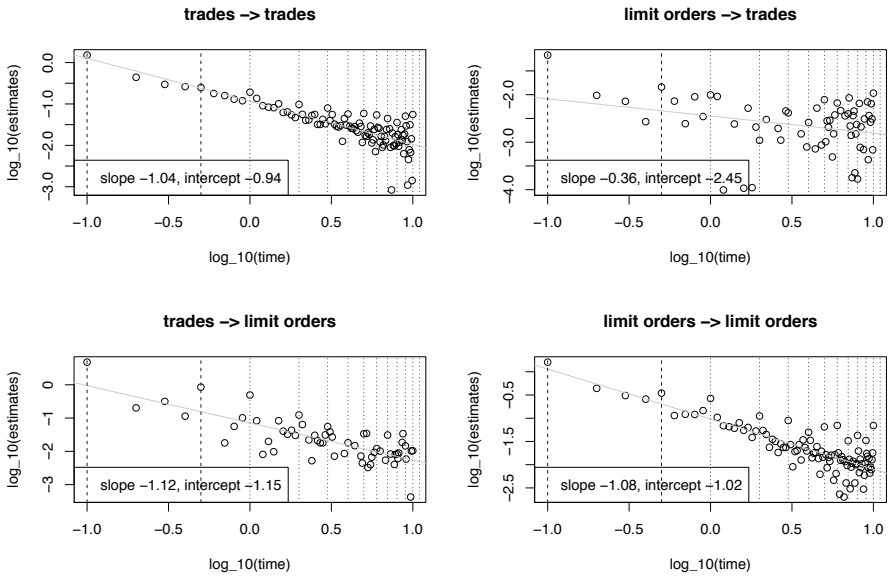


Figure 11: Log/log plots of the exitement estimates averaged over the regular trading hours of the three week data sample; see Section 5.2. The dotted vertical lines refer to seconds. Apart from the upper-right panel ($\mathcal{L} \rightsquigarrow \mathcal{T}$), the plots are very well linearly approximated. In the same three excitements we observe second periodicities (dotted lines) as well as peaks at 0.1 sec and 0.5 sec (dashed lines). Note that due the noise in the original estimates, we have some estimates < 0 (especially for large times); these estimates are ignored in the log-transformation. This introduces a bias for the slope estimate: the true decay is presumably faster than the slope estimates indicate.

all $13 \cdot 14 = 182$ time windows. The average branching matrix is

$$\begin{pmatrix} \mathcal{T} \overset{0.64}{\rightsquigarrow} \mathcal{T} & \mathcal{L} \overset{0.02}{\rightsquigarrow} \mathcal{T} \\ \mathcal{T} \overset{0.75}{\rightsquigarrow} \mathcal{L} & \mathcal{L} \overset{0.59}{\rightsquigarrow} \mathcal{L} \end{pmatrix}$$

which is quite similar as in the single-window analysis (5.1). In particular, we observe a similar asymmetry in the cross-excitements. The log/log-plots of the estimates indicate that the excitement functions decay with a power-law with decay parameters close to 1; see Figure 11. The log/log-plots of the results exhibit strong second-periodicities. Also note that there are local maxima in the excitements after 0.1 sec and 0.5 sec. We checked the absolute event-times in our data set for (absolute) round-time preferences. Statistical tests indicate such preferences. But we think that they are too weak to explain the strong periodicities in the excitement.

Then—just like for the single-window analysis above—we study the instantaneous excitement in more detail by setting s as small as 0.01 and Δ as small as 0.001 sec—which is the resolution of the original data. This very short excitement looks quite noisy (e.g. not monotone) on the one hand, on the other hand, it seems remarkably similar for all four excitements: we observe negative values up to 0.002 sec -0.003 sec, a maximum in 0.006 sec, and another local maximum at 0.01 sec. This delay in the excitements may be a manifestation of the *high-*

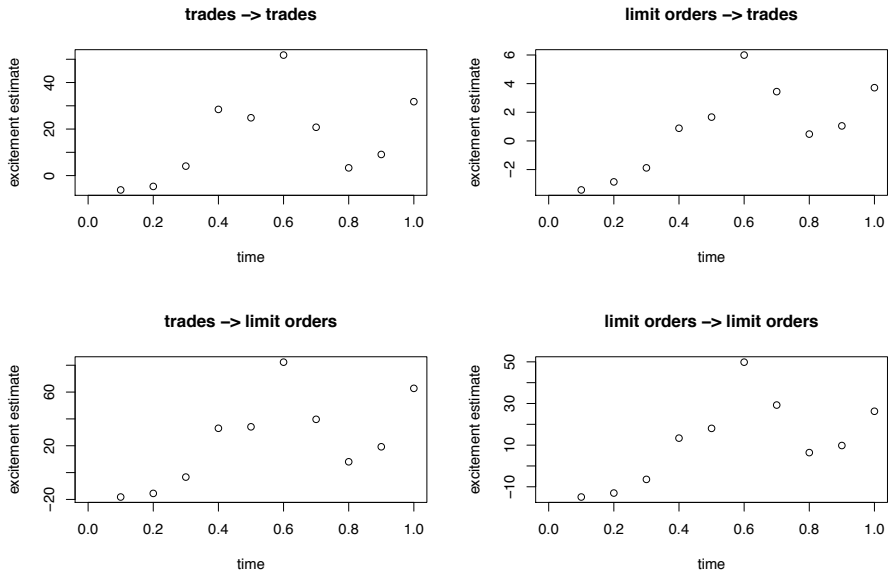


Figure 12: Average instantaneous excitement over the 182 considered 30min-windows. We fit the Hawkes model naively with a support parameter $s = 0.01$ sec and a super-small bin-size of 0.001 sec. Note that the resolution of our original data is milliseconds, so there is no data aggregation involved anymore. Consequently, it might be more appropriate to speak of an INAR(10) process rather than a Hawkes process. The shapes of the four estimated excitements are remarkably similar: there are negative values up to 0.003 sec, a maximum at 0.006 sec, and another local maximum at 0.01 sec. The excitement delay is presumably related to the *high-frequency cut-off* as discussed in Hardiman, S.J. et al. (2013).

frequency cut-off discussed in Hardiman, S.J. et al. (2013) or the *average latency* discussed in Bacry et al. (2014). We think that the observed inhibitory effects are statistical artifacts that are a compensation for some misspecification of the model. It would be interesting to examine this ultra-short behavior in more detail. Also note that when we get that close to the resolution of the data it might be more appropriate to apply the discrete-time INAR model directly instead of using the continuous time Hawkes model.

We also examine the development of the model over the day. To that aim, we illustrate the branching-coefficients estimates and the average baseline-intensity estimates over the fourteen 8:30am–9:00am time windows, over the fourteen 9:00am–9:30am time windows, and so on. The estimates are calculated as proposed before, that is, with respect to $s = 10$ sec and $\Delta = 0.01$ sec. The branching-coefficient estimates are quite stable over the regular trading hours—with outliers in the time windows right after the opening and right before the daily trading-stop (at 3:15pm). The average baseline-intensity exhibits the typical U-shape of the empirical intensity from Figure 8.

Note that—due to the power-law shape of the excitement—choosing a larger support-parameter will typically yield larger branching-coefficients. In this case, the branching-coefficient estimates from above are somewhat arbitrary (compared to the case of finite or exponentially

decaying excitement functions). Still they are meaningful *in combination with the applied* $s = 10\,\text{sec}$: an estimated branching-coefficient may be interpreted as the number of (direct) offspring of an event *in the next* 10 *seconds*; alternatively, it measures the influence of an event on the corresponding intensity component in the next 10 seconds.

Semiparametric fit of power-law parameters

An alternative way to infer the branching coefficients would be to conclude from Figure 12 that the model is governed by a parametric power-law function of the form $h(t) = \beta 1_{t \geq \varepsilon} t^{-\alpha}$ for some very small $\varepsilon > 0$ and $\alpha > 1$. If slope and intercept are estimated coefficients from the linear model explaining the $\log_{10}$-transformed estimates with the corresponding log-transformed times, then we have that $\alpha \approx -\text{slope}$ and $\beta \approx 10^{\text{intercept}}$. Consequently, we get

$$\int_0^\infty h(t)\mathrm{d}t = \frac{\beta}{\alpha - 1}\varepsilon^{1-\alpha} \approx -\frac{10^{\text{intercept}}}{1 + \text{slope}}\varepsilon^{1+\text{slope}}. \tag{5.2}$$

In principle, we can read-off the high-frequency cut-off ε from Figure 12. Say we set $\hat{\varepsilon} = 0.003$. If we plug in estimates for slope, intercept, and high-frequency cut-off, we obtain a semiparametric estimate for the branching coefficients. For the self-excitement of the limit orders in the lower right panel of Figure 11, for example, (5.2) yields a branching coefficient of approximately 1.9. This value indicates a supercritical self-excitement which would mean that the data-generating process is nonstationary and (in the long run) explosive. This result may indicate that the infinite extrapolation of the excitement is wrong. And even if the true underlying model were truly governed by a power-law excitement, there would be problems with the estimation approach derived from (5.2):

1. the (nonlinear) log transformation introduces (additional) bias;
2. the branching-coefficient approximation (5.2) is extremely sensitive on the value of ε, which itself is difficult to estimate;
3. some of the estimated excitement-values $\hat{h}_k, k = 1, \ldots, p$ might be negative due to noise—especially for large k. But then, these negative estimates get lost in the log-transform whereas their positive counterparts survive. This introduces an upward bias for large k and, consequently, the decay that we observe in the log/log representation is typically slower than the true decay.

In view of the above, the estimation approach based on (5.2) only makes sense for very large sample-sizes. In most cases, it is more sensible to estimate power-law parameters directly from original (untransformed) estimates via nonlinear least squares. Consider once more the self-excitement of the limit orders in the lower-right panel of Figure 11. Applying nonlinear least-squares optimization to the original estimates (e.g., with `nls()` in `R`) yields a decay-parameter

estimate of 1.25. This indicates a rather faster decay than the one resulting from the linear log/log-fit (1.08).

Power-law extrapolation vs. truncation

Next to the estimation issues discussed above, there are conceptual problems with extremely slowly decaying excitement-functions. We have already touched this issue in Section 4.1. In the following, we readdress the problem in our specific data-context: the AIC-optimal support-choice in Figure 9 is clear-cut. On the other hand, the linearity of the log/log-plot in Figure 11 is also convincing. However, the consequences of infinite excitements as in (4.2) or (5.2) with decay parameters of $\alpha = 1.1$ and lower are hard to interpret: in such a model, a significant part of the offspring happens days or weeks after the point of reference. In the case of (4.2), where most of the instantaneous excitement is shifted to the tail, such a low decay-parameter would mean that a large part of the offspring happens only after 30'000 years! These are brave extrapolations when the size of the data window is 30 min. So typically, one desires larger samples. Such long order-book histories, however, contain trading halts, night times, weekends, holidays, and so on. Including these obvious regime-switches in the estimation will typically yield excitement estimates that are even more heavy-tailed. In autoregressive fits, stationary models typically exhibit long-range dependence when calibrated to time windows where there are regime switches in the data; see Mikosch and Stărică (2000). That regime-switches have these effects in the point process context can easily be demonstrated in simulation, e.g., by calibrating a Hawkes model on data from a doubly stochastic Cox process as described in Section 4.3. Also note that other unobserved covariate-processes might have simililar effects. E.g., when we fit a univariate Hawkes model to data that is a margin of a biviariate Hawkes model, the fit typically also exhibits long-range dependence. These issues make observed long-range dependence as in Figure 11 hard to interpret. The interpretation of the results is finally a question of taste and, more importantly, a question of the application: in view of the argumentation in Section 4.1, the truncated model typically suffices if the goal is an algorithm that aims to calculate (as quickly as possible) the likelihood of a specific order-book event given the event-history (e.g., for the implementation of some strategy). Also, if we are mainly interested in the causal structure underlying data (which order flow affects which), the truncated models may suffice. On the other hand, if we are more interested in theoretical results such as connecting the market microstructure to coarser scale (price) processes, volatility estimation, and so on, then it is presumably more fertile to work with the extrapolation of the power-laws.

Comparison with earlier results

The most complete Hawkes-process based limit order book model as of now is Bacry et al.

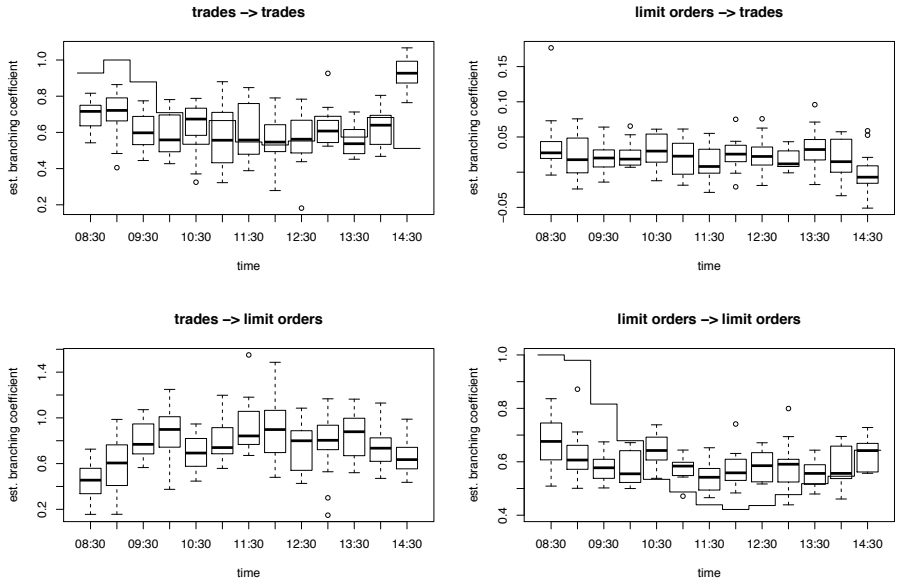


Figure 13: Comparison of the estimates over the trading day; see Section 5.2. For each of the fourteen sample days, we consider thirteen 30 min windows. The boxplots collect the branching-coefficient estimates (left scales); the baseline-intensity estimates are averaged over the corresponding time-windows (step functions). The branching coefficients are relatively stable over the trading day—with outliers at the opening and the close of the CME. The baseline intensity of the market orders reflects the U-shape of the average intensity from Figure 8.

(2014). Here, the authors distinguish the ask and the bid side of the order flow and also take price jumps and cancellations into account. The authors find that it is mainly the price jumps that drive the dynamics of the process. They also report a strong influence from trades on limit orders and less excitement vice versa, which supports our observations. Also the power-law shape with decay parameters around 1 of the excitements is reported in all papers where the Bacry–Muzy nonparametric method is applied to the order flow; see, e.g., Hardiman, S.J. et al. (2013) and Bacry et al. (2014). As to now, there seems to be no approach for the choice of the support parameter s, respectively, A in the BM-method. So the (relatively small) support estimator from our AIC-based selection cannot be compared with earlier results.

5.3 Interpretation of the estimation results

The interpretation of the estimation results from Section 5.2 (and of Hawkes fits in general) is not straightforward: observing the arrival of an order makes people (respectively, algorithms) send other orders. In this sense, we may expect some quite direct true excitement in LOB data. In the Hawkes modeling approach however, any fluctuation of exogenous processes that influence the observed event-process will also be detected as excitement. The past of the observed process then serves as a proxy for some unobserved covariate processes. E.g., the past price-jumps in Bacry et al. (2014) is reported to be the most important 'excitor' for the order flows.

These price jumps can be seen as a proxy for a change in the state of the book: updated—typically larger—volume at best bid and best ask, updated imbalance, and so on. Candidates for other covariate processes in our context are volume, arrival of orders away from the best bid or best ask price, spread, or even data from other assets such as options on S&P 500 E-mini futures. A most natural way to model this situation would be a joint multivariate Hawkes model. However, doing statistics with so little knowledge about the state (or even the dimensionality) of the process yields new problems. So the best way to get rid of artificial self-excitement in the Hawkes model is presumably to make the baseline intensity more flexible. For an example of such a Hawkes model with stochastic baseline-intensity; see Zhao (2012). To summarize: our estimation method can indeed detect self- or cross-excitement in data. However, we ought to be careful with interpretation of these terms.

Any Hawkes fit is meaningful and fertile despite the criticism above and despite the vanishing p-values in our application: plots of excitement estimates as in Figure 10 are visualizations of huge event-data sets in a compact and at the same time informative way. In that sense, any Hawkes fit—and our estimation method in particular—can be used as a graphical tool for exploratory event-stream analysis. Furthermore, even if the Hawkes model assumption may be completely wrong for the data-generating point process $\mathbf{N}$, an excitement-function estimate $\hat{h}_{ij}(\cdot)$ is still meaningful. It is an estimate for the best linear filter of '$\mathbb{E}\left[N_i(\mathrm{d}t)/\mathrm{d}t|\sigma\left(N_j(\{s\}), s < t\right)\right]$' which is a relevant quantity in all stationary models.

6 Conclusion

This paper demonstrates that applying methods from time series theory to the bin-count sequences of point process data yields a useful and intuitive nonparametric estimation method for the multivariate Hawkes process. The price for the fertile simplicity of the method is a bias due to various errors involved in the approximation. Simulation studies support that this bias can be controlled and that it is negligible for most practical means. The technique presented depends on the choice of the bin size and the assumed support of the excitement function(s). Methods for a sensible choice of these parameters are given. In any application, the robustness with respect to these choices ought to be studied. We treat computational issues in high-dimensional cases in Embrechts and Kirchner (2017). A larger limit-order book application is presented in Kirchner and Vetter (2017), where we consider an application of our method on Hawkes models with marks and with covariate-dependent baseline intensities.

Finally, note that in view of the analogy between discrete-time INAR(p) sequences and continuous-time Hawkes processes, analysts using the Hawkes model may consider to directly apply the INAR(p) model in the first place—as most event data live on relatively discrete time grids.

Acknowledgements

M.K. is indebted to Paul Embrechts for guidance and support during the preparation of the paper. The author acknowledges financial support from ETH RiskLab and the Swiss Finance Institute. Furthermore, M.K. thanks Robert Almgren for sharing his expertise on limit-order-book data, Marius Hofert as well as Martin Maechler (Bates and Maechler, 2015) for support with R, Valérie Chavez-Demoulin as well as Thibault Vatter for numerous comments on an earlier versions of the paper, and Rita Kirchner as well as Anne MacKay for help with the editing. Incorporating the comments of two anonymous referees made the paper more complete.

A Proofs

A.1 Proof of Proposition 6

First, we establish that $\mathbf{u}_n := \mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^p A_k \mathbf{X}_{n-k}$, $n \in \mathbb{Z}$, defines a white noise sequence. Stationarity of $(\mathbf{u}_n)$ follows from the stationarity of $(\mathbf{X}_n)$. For the sequel of the proof, fix any $n \in \mathbb{Z}$. Denote $\mathcal{F}_n := \sigma\{\mathbf{X}_k : k \leqslant n\}$. Note that $\mathbb{E}\left[A_k \circledast \mathbf{X}_{n-k} \,|\mathcal{F}_n\right] = A_k \mathbf{X}_n$, $A_k \in \mathbb{R}^{d\times d}_{\geqslant 0}$, and that ε_n is independent of $\mathcal{F}_n$. So we get

$$\mathbb{E}\left[\,\mathbf{u}_n \,|\mathcal{F}_{n-1}\right] = \mathbb{E}\left[\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k} + \varepsilon_n - \mathbf{a}_0 - \sum_{k=1}^\infty A_k \mathbf{X}_{n-k} \,|\mathcal{F}_{n-1}\right] = 0, \tag{A.1}$$

and, consequently, $\mathbb{E}\,\mathbf{u}_n = 0$, $n \in \mathbb{Z}$. For the autocovariances of the errors, note that, for $n' < n$ (and then, by symmetry, for $n' \neq n$),

$$\mathbb{E}\left[\,\mathbf{u}_n\,\mathbf{u}_{n'}\,\right] = \mathbb{E}\left[\,\mathbb{E}\left[\,\mathbf{u}_n\,\mathbf{u}_{n'}\,|\mathcal{F}_{n-1}\right]\right] = \mathbb{E}\left[\,\mathbf{u}_{n'}\,\underbrace{\mathbb{E}\left[\,\mathbf{u}_n\,|\mathcal{F}_{n-1}\right]}_{\overset{(A.1)}{=}0}\right] = 0.$$

Finally, we have that

$$\begin{aligned}\mathrm{Cov}(\mathbf{u}_n) &= \mathbb{E}\Big[\underbrace{\mathrm{Cov}\left(\,\mathbf{u}_n\,|\mathcal{F}_{n-1}\right)}_{=\mathrm{Cov}(\mathbf{X}_n\,|\mathcal{F}_{n-1})}\Big] + \mathrm{Cov}\left(\underbrace{\mathbb{E}\left[\,\mathbf{u}_n\,|\mathcal{F}_{n-1}\right]}_{\overset{(A.1)}{=}0}\right) = \mathrm{diag}\left(\,\mathbf{a}_0 + \sum_{k=1}^p A_k\,\mathbb{E}\left[\,\mathbf{X}_{n-k}\,\right]\right)\\ &= \mathrm{diag}\left(\,\mathbf{a}_0 + \sum_{k=1}^p A_k\Big(1_{d\times d} - \sum_{k=1}^p A_k\Big)^{-1}\mathbf{a}_0\right) = \mathrm{diag}\left(\Big(1_{d\times d} - \sum_{k=1}^p A_k\Big)^{-1}\mathbf{a}_0\right).\end{aligned}$$

□

A.2 Proof of Theorem 10

The first part of the proof largely depends on matrix manipulations. So it is important to remind the reader that all vectors are understood as column vectors. We rewrite the INAR(p) sequence $(\mathbf{X}_k) \subset \mathbb{N}_0^d$ as a standard multivariate linear autoregressive time series with white-noise error sequence $(\mathbf{u}_k)_{k\in\mathbb{Z}} := (\mathbf{X}_k - \mathbf{a}_0 + \sum_{l=1}^p A_l \mathbf{X}_{k-l})_{k\in\mathbb{Z}}$

$$\mathbf{X}_k = \mathbf{a}_0 + \sum_{l=1}^{p} A_l \mathbf{X}_{k-l} + \mathbf{u}_k, \quad k \in \mathbb{Z};$$

see Corollary 7. Then the distributional properties of the CLS-estimator are derived similarly as in Lütkepohl (2005), pages 70–75, where independent errors are assumed. In the following, let $\mathbf{Z} \in \mathbb{N}_0^{(dp+1)\times(n-p)}$ be the design matrix from the CLS Definition 8 with respect to the sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$. Furthermore, let $\mathbf{U} := (\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \ldots, \mathbf{u}_n) \in \mathbb{R}^{d\times(n-p)}$. Note that $\mathbf{Z}$ as well as $\mathbf{U}$ depend on n. We work under the assumption that

$$\frac{1}{n-p}\mathbf{Z}\mathbf{Z}^\top \xrightarrow{p} : \Gamma \in \mathbb{R}^{(dp+1)\times(dp+1)}, \quad n \longrightarrow \infty, \tag{A.2}$$

exists and is invertible. In addition, we use that, for $n \longrightarrow \infty$,

$$\frac{1}{\sqrt{n-p}}\mathrm{vec}\big(\mathbf{U}\mathbf{Z}^\top\big) \xrightarrow{\mathrm{d}} \mathcal{N}_{d^2p+d}\left(0_{d^2p+d}, \mathbb{E}\left[\Big(\mathbf{Z}_0 \otimes 1_{d\times d}\Big)\mathbf{u}_0\Big(\big(\mathbf{Z}_0 \otimes 1_{d\times d}\big)\mathbf{u}_0\Big)^\top\right]\right), \tag{A.3}$$

where $\mathbf{Z}_0 := \big(\mathbf{X}_{-1}^\top, \mathbf{X}_{-2}^\top, \ldots, \mathbf{X}_{-p}^\top, 1\big)^\top \in \mathbb{N}_0^{(pd+1)\times 1}$ has the same distribution as any of the columns of the design matrix $\mathbf{Z}$. We postpone the reasoning for (A.3) to the end of the proof. As a first step, weak consistency of $\hat{\mathbf{B}}^{(n)} \in \mathbb{R}^{d\times(dp+1)}$ is proven. To that aim, we will use that

$$\mathbf{Y} := (\mathbf{X}_{p+1}, \mathbf{X}_{p+1}, \ldots, \mathbf{X}_n) = \mathbf{B}\mathbf{Z} + \mathbf{U}\left(\in \mathbb{N}_0^{d\times(n-p)}\right); \tag{A.4}$$

see Definition 8.

$$\hat{\mathbf{B}}^{(n)} - \mathbf{B} = \mathbf{Y}\mathbf{Z}^\top\left(\mathbf{Z}\mathbf{Z}^\top\right)^{-1} - \mathbf{B} \overset{(A.4)}{=} (\mathbf{B}\mathbf{Z} + \mathbf{U})\mathbf{Z}^\top\left(\mathbf{Z}\mathbf{Z}^\top\right)^{-1} - \mathbf{B} = \mathbf{U}\mathbf{Z}^\top\left(\mathbf{Z}\mathbf{Z}^\top\right)^{-1}$$
$$= \frac{\mathbf{U}\mathbf{Z}^\top}{n-p}\left(\frac{\mathbf{Z}\mathbf{Z}^\top}{n-p}\right)^{-1}.$$

By (A.2), the second factor converges in probability to the constant matrix Γ. By (A.3), the first factor has the same asymptotic distribution as $\tilde{W}/\sqrt{n-p}$ where $\tilde{W}$ is a matrix consisting of jointly normally distributed entries not depending on n. So $\tilde{W}/\sqrt{n-p} \xrightarrow{p} 0_{d\times(dp+1)}$ and therefore $\hat{\mathbf{B}}^{(n)} - \mathbf{B} \xrightarrow{p} 0_{d\times(dp+1)}$. For establishing the asymptotic distribution, we treat the

difference of the estimated and true vectorized parameter-matrix in a similar way:

$$\begin{aligned}\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}(\mathbf{B}) = \operatorname{vec}\left(\hat{\mathbf{B}}^{(n)} - \mathbf{B}\right) &= \operatorname{vec}\left(\mathbf{U}\mathbf{Z}^\top\left(\mathbf{Z}\mathbf{Z}^\top\right)^{-1}\right) \\ &= \left(\left(\mathbf{Z}\mathbf{Z}^\top\right)^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\mathbf{U}\mathbf{Z}^\top\right) \\ &= \frac{1}{\sqrt{n-p}}\left(\left(\frac{\mathbf{Z}\mathbf{Z}^\top}{n-p}\right)^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\frac{\mathbf{U}\mathbf{Z}^\top}{\sqrt{n-p}}\right).\end{aligned} \tag{A.5}$$

In the third step of the calculation above we use that

$$\operatorname{vec}(AB) = \left(B^\top \otimes I\right)\operatorname{vec}(A), \tag{A.6}$$

for matrices A, B and identity matrix I such that the calculations are consistent dimensionwise; see A.12 in Lütkepohl (2005). It follows from (A.5) together with (A.2) that

$$\sqrt{n-p}\left(\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}(\mathbf{B})\right)$$

has the same asymptotic distribution as

$$\left(\Gamma^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\frac{\mathbf{U}\mathbf{Z}^\top}{\sqrt{n-p}}\right). \tag{A.7}$$

With (A.3), we then find that the asymptotic distribution of (A.7)—and therefore of

$$\sqrt{n-p}\left(\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}(\mathbf{B})\right)$$

—is centered normal with covariance matrix

$$\begin{aligned}&\left(\Gamma^{-1} \otimes 1_{d\times d}\right)\operatorname{cov}\left(\operatorname{dlim}_{n\to\infty}\operatorname{vec}\left(\frac{\mathbf{U}\mathbf{Z}^\top}{\sqrt{n-p}}\right)\right)\left(\Gamma^{-1} \otimes 1_{d\times d}\right) \\ &\overset{(A.3)}{=} \left(\Gamma^{-1} \otimes 1_{d\times d}\right)\mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_k\left(\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_k\right)^\top\right]\left(\Gamma^{-1} \otimes 1_{d\times d}\right).\end{aligned}$$

We still have to establish (A.3). To that aim, we rewrite the left-hand side of (A.3) as

$$\begin{aligned}
\frac{1}{\sqrt{n-p}}\mathrm{vec}\left(\mathbf{U}\,\mathbf{Z}^\top\right) &= \frac{1}{\sqrt{n-p}}\mathrm{vec}\left(\left(\sum_{j=1}^{n-p}\mathbf{Z}_{j,1}^\top\,\mathbf{U}_{\cdot,j},\ldots,\sum_{j=1}^{n-p}\mathbf{Z}_{j,dp+1}^\top\,\mathbf{U}_{\cdot,j}\right)\right)\\
&= \frac{1}{\sqrt{n-p}}\sum_{j=1}^{n-p}\mathrm{vec}\left(\left(\mathbf{Z}_{1,j}\,\mathbf{U}_{\cdot,j},\ldots,\mathbf{Z}_{dp+1,j}\,\mathbf{U}_{\cdot,j}\right)\right)\\
&= \frac{1}{\sqrt{n-p}}\sum_{j=1}^{n-p}\mathrm{vec}\left(\left(\mathbf{U}_{1,j},\ldots,\mathbf{U}_{d,j}\right)^\top\left(\mathbf{Z}_{1,j},\mathbf{Z}_{2,j},\ldots,\mathbf{Z}_{dp+1,j}\right)\right)\\
&= \frac{1}{\sqrt{n-p}}\sum_{k=p+1}^{n}\mathrm{vec}\left(\mathbf{u}_k\cdot\mathbf{Z}_k^\top\right),
\end{aligned}$$

where $\mathbf{Z}_k := \left(\mathbf{X}_{k-1}^\top,\mathbf{X}_{k-2}^\top,\ldots,\mathbf{X}_{k-p}^\top,1\right)^\top \in \mathbb{N}_0^{(pd+1)\times 1}$. Note that, for $k\in\{p+1,\ldots,n\}$, $\mathbf{Z}_k$ is the $(k-p)$-th column of the design matrix $\mathbf{Z}$. Now, let $\mathbf{w}_k := \mathrm{vec}\left(\mathbf{u}_k\cdot\mathbf{Z}_k^\top\right)\in\mathbb{R}^{pd^2+d}$, $k\in\mathbb{Z}$. We show that for the sequence $(\mathbf{w}_k)\subset\mathbb{R}^{pd^2+d}$, a central limit theorem for vector-valued martingale differences can be applied. Proposition 7.9 from Hamilton (1994) states that if $(\mathbf{w}_k)\subset\mathbb{R}^{\tilde{d}}$ is such that

(a) it defines a vector-valued martingale difference sequence, i.e., there is a filtration

$$(\mathcal{H}_k)_{k=p+1,p+2,\ldots,n}$$

such that $\mathbf{w}_k$ is $\mathcal{H}_k$-measurable and $\mathbb{E}\left[\mathbf{w}_k\,|\mathcal{H}_{k-1}\right]=0_{\tilde{d}},\ k\in\mathbb{Z}$,

(b) $\mathbb{E}\left[\mathbf{w}_k\,\mathbf{w}_k^\top\right]=:S\in\mathbb{R}^{\tilde{d}\times\tilde{d}}$ is a positive definite matrix independent of k,

(c) for all $k_1,k_2,k_3,k_4\in\mathbb{Z}$ and for all $i_1,\ldots,i_4\in\{1,2,\ldots,\tilde{d}\}$,

$$\mathbb{E}\left[\mathbf{w}_{k_1,i_1}\,\mathbf{w}_{k_2,i_2}\,\mathbf{w}_{k_3,i_3}\,\mathbf{w}_{k_4,i_4}\right]<\infty,$$

where $\mathbf{w}_{k,i}$ denotes the i-th component of $\mathbf{w}_k$, and

(d) $\sum_{k=p+1}^{n}\frac{1}{n-p}\,\mathbf{w}_k\,\mathbf{w}_k^\top\xrightarrow{p}S$,

then, for $n\longrightarrow\infty$, $1/\sqrt{n-p}\sum_{k=p+1}^{n}\mathbf{w}_k\xrightarrow{\mathrm{d}}\mathcal{N}_{\tilde{d}}(0_{\tilde{d}},S)$.

Proof of (a) Define the filtration $(\mathcal{H}_k)$ by setting

$$\mathcal{H}_k := \sigma\left(\left(\mathbf{u}_i,\mathbf{X}_{i-1},\mathbf{X}_{i-2},\ldots,\mathbf{X}_{i-p}\right):\ i\leqslant k\right),\ k\in\mathbb{Z}.$$

Then one can easily check that $\mathbf{w}_k=\mathrm{vec}\left(\mathbf{u}_k\cdot\mathbf{Z}_k^\top\right)$ is $\mathcal{H}_k$- measurable. It suffices to prove the martingale-difference property for the sequence $(\mathbf{u}_k)$ since $\mathbf{X}_{k'}$ for $k'<k$ and therefore $\mathbf{Z}_k$ are

$\mathcal{H}_{k-1}$-measurable. But then, because

$$\mathbb{E}\left[\mathbf{X}_k \middle| \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\varepsilon_k + \sum_{m=1}^{p} A_m \circledast \mathbf{X}_{k-m} \middle| \mathcal{H}_{k-1}\right] = \mathbf{a}_0 + \sum_{m=1}^{p} A_m \mathbf{X}_{k-m},$$

we obtain the martingale difference property:

$$\mathbb{E}\left[\mathbf{u}_k \middle| \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\mathbf{X}_k - \mathbf{a}_0 - \sum_{m=1}^{p} A_m \mathbf{X}_{k-m} \middle| \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\mathbf{X}_k \middle| \mathcal{H}_{k-1}\right] - \mathbf{a}_0 - \sum_{m=1}^{p} A_m \mathbf{X}_{k-m} = 0_d.$$

Proof of (b) Independency of k follows from stationarity of $(\mathbf{w}_k)$. Choose $k = 0$. We need to show that, for $b \in \mathbb{R}^{d(pd+1)} \setminus \{0_{d(pd+1)}\}$,

$$b^\top \mathbb{E}\left[\mathbf{w}_k \mathbf{w}_k^\top\right] b = \mathbb{E}\left[b^\top \mathbf{w}_0 \mathbf{w}_0^\top b^\top\right] = \mathrm{Var}\left(b^\top \mathbf{w}_0\right) > 0. \tag{A.8}$$

With (A.6), we find

$$\mathbf{w}_0 = \mathrm{vec}\left(\mathbf{u}_0 \cdot \mathbf{Z}_0^\top\right) = \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathrm{vec}\left(\mathbf{u}_0\right) = \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathbf{u}_0 \tag{A.9}$$

and therefore

$$\mathbb{E}\left[\mathbf{w}_0 \mathbf{w}_0^\top\right] = \mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathbf{u}_k \left(\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathbf{u}_0\right)^\top\right] = \mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathbf{u}_0 \mathbf{u}_0^\top \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\right].$$

To establish (A.8), we define the σ-algebra

$$\mathcal{F} := \sigma\left(\mathbf{X}_{-1}, \ldots, \mathbf{X}_{-p}, A_1 \circledast \mathbf{X}_{-1}, \ldots, A_p \circledast \mathbf{X}_{-p}\right).$$

Note that $\mathbf{Z}_0$ is $\mathcal{F}$-measurable and ε_0 is independent of $\mathcal{F}$. Using these facts when considering the expectation of the conditional variance of $b^\top \mathbf{w}_0$, we obtain

$$\begin{aligned}
\mathrm{Var}\left(b^\top \mathbf{w}_0\right) &= \mathbb{E}\left[\mathrm{Var}\left(b^\top \mathbf{w}_0 \,|\, \mathcal{F}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[b^\top \mathbf{w}_0 \,|\, \mathcal{F}\right]\right) \\
&\geqslant \mathbb{E}\left[\mathrm{Var}\left(b^\top \mathbf{w}_0 \,|\, \mathcal{F}\right)\right] \\
&\overset{(A.9)}{=} \mathbb{E}\left[\mathrm{Var}\left(b^\top \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathbf{u}_0 \,|\, \mathcal{F}\right)\right] \\
&= \mathbb{E}\left[b^\top \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right) \mathrm{Cov}\left(\mathbf{u}_0 \,|\, \mathcal{F}\right) \left(b^\top \left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\right)^\top\right].
\end{aligned} \tag{A.10}$$

Since

$$\mathbf{u}_0 = \mathbf{X}_0 - \mathbf{a}_0 - \sum_{i=1}^{p} A_i \mathbf{X}_{-i} = \varepsilon_0 + \sum_{i=1}^{p} A_i \circledast \mathbf{X}_{-i} - \mathbf{a}_0 - \sum_{i=1}^{p} A_i \mathbf{X}_{-i}, \tag{A.11}$$

the summand ε_0 is the only term that contributes to the conditional covariance matrix in (A.10)—the other summands in (A.11) are constant with respect to $\mathcal{F}$ and ε_0 is independent of $\mathcal{F}$. So we have $\operatorname{Cov}(\mathbf{u}_0 \,|\, \mathcal{F}) = \operatorname{Cov}(\varepsilon_0 \,|\, \mathcal{F}) = \operatorname{Cov}(\varepsilon_0) = \operatorname{diag}(\mathbf{a}_0)$ and continuing with (A.10) we find

$$\begin{aligned}
\operatorname{Var}\left(b^\top \mathbf{w}_0\right) &\geqslant \mathbb{E}\left[b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right) \operatorname{diag}(\mathbf{a}_0) \left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right)\right)^\top\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} \mathbf{a}_{0,i} \left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right)\right)_{1,i}^2\right] \\
&\geqslant \mathbf{a}_{0,i_0}\, \mathbb{E}\left[\left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right)\right)_{1,i_0}^2\right] > 0, \qquad \text{(A.12)}
\end{aligned}$$

where $i_0 \in \{1, 2, \ldots, d\}$ in (A.12) is chosen in such a way that $\mathbf{a}_{0,i_0} > 0$. (Remember that $\mathbf{a}_0 \neq 0_d$, by assumption.) The strict inequality in (A.12) follows because, for $j_0 \in \{1, 2, \ldots, np + d\}$ such that $b_{j_0} \neq 0$, we have that

$$\begin{aligned}
\mathbb{P}\left[\left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right)\right)_{1,i_0} \neq 0\right] = \mathbb{P}\left[b^\top \cdot \left(\mathbf{Z}_0^\top \otimes 1_{d\times d}\right)_{\cdot,i_0} \neq 0\right] &\geqslant \mathbb{P}\left[b_{j_0} \mathbf{X}_{k_0,l_0} \neq 0\right] = \mathbb{P}\left[\mathbf{X}_{k_0,l_0} \neq 0\right] \\
&> 0,
\end{aligned}$$

for some $k_0 \in \mathbb{Z}$ and some $l_0 \in \{1, 2, \ldots, d\}$ dependent on j_0. Note that $\mathbf{X}_{k,l}$ denotes the l-th component of $\mathbf{X}_k$. By stationarity, $k_0 \in \mathbb{Z}$ is irrelevant. And the case that $\mathbf{X}_{0,l_0} = 0$ *a.s.* for some $l_0 \in \{1, 2, \ldots, d\}$ we have excluded, so the strict inequality follows.

Proof of (c) Note that claim (c) follows if $\mathbb{E}\left[\mathbf{X}_{k_1,i_1} \cdots \mathbf{X}_{k_8,i_8}\right] < \infty$ for $k_1, \ldots, k_8 \in \mathbb{Z}$, $i_1, \ldots, i_8 \in \{1, 2, \ldots, d\}$. The boundedness of these expectations is established for the univariate case in Corollary 1 of Kirchner (2016). For the multivariate case, one can argue similarly via the existence of the moment generating function in a neighborhood of zero.

Proof of (d) We show that $\left(\mathbf{w}_k \mathbf{w}_k^{\mathrm{T}}\right)$ is ergodic. Then the claim of (d) follows with the Birkhoff–Khinchin Ergodic Theorem. The sequence $(\mathbf{X}_k)$ can be represented as margin of a pd-variate INAR(1) sequence $\left(\tilde{\mathbf{X}}_k\right)$; see Latour (1997). It is easily checked that the latter is an irreducible, aperiodic Markov chain on $\mathbb{N}_0^{pd}$. So $\left(\tilde{\mathbf{X}}_k\right)$ is ergodic; see Durrett (1996), page 338. As margins of ergodic processes are ergodic, $(\mathbf{X}_k)$ also is ergodic. As $\mathbf{w}_k$ can be written as a measurable function of the past of $(\mathbf{X}_k)$, $(\mathbf{w}_k)$ is also ergodic. Finally, $\left(\mathbf{w}_k \mathbf{w}_k^\top\right)$ is ergodic because it is a measurable transformation of the ergodic sequence $(\mathbf{w}_k)$.

□

Paper

D

Paul Embrechts, Matthias Kirchner

Hawkes graphs.

Theory of Probability and Its Applications,
62(1):163–193, 2017.

Hawkes graphs

Paul Embrechts, Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

This paper introduces the Hawkes skeleton and the Hawkes graph. These objects summarize the branching structure of a multivariate Hawkes point process in a compact, yet meaningful way. We demonstrate how graph-theoretic vocabulary ('ancestor sets', 'parent sets', 'connectivity', 'walks', 'walk weights', ...) is very convenient for the discussion of multivariate Hawkes processes. For example, we reformulate the classic eigenvalue-based subcriticality criterion of multitype branching processes in graph terms. Next to these more terminological contributions, we show how the graph view may be used for the specification and estimation of Hawkes models from large, multitype event streams. Based on earlier work, we give a nonparametric statistical procedure to estimate the Hawkes skeleton and the Hawkes graph from data. We show how the graph estimation may then be used for specifying and fitting parametric Hawkes models. Our estimation method avoids the a priori assumptions on the model from a straighforward MLE-approach and is numerically more flexible than the latter. Our method has two tuning parameters: one controlling numerical complexity, the other one controlling the sparseness of the estimated graph. A simulation study confirms that the presented procedure works as desired. We pay special attention to computational issues in the implementation. This makes our results applicable to high-dimensional event-stream data, such as dozens of event streams and thousands of events per component.

1 Introduction

This paper discusses the specification and estimation of multivariate Hawkes point process models from large, multitype event-stream datasets such as neural spike-trains, internet search-queries, or limit-order-book data in high-frequency finance. Our approach uses the notion of a

Hawkes skeleton and a Hawkes graph[1]. We demonstrate how these concepts are fertile beyond statistical estimation.

The Hawkes process was introduced in Hawkes (1971b,a) as a stationary point process on $\mathbb{R}$ whose points are assigned to a finite number of types. The (stochastic) intensity of a Hawkes process depends on the past of the process itself: given the occurrence of an event, the intensities—the expected mean number of events per time unit and event type—typically jump upwards and then decay. This structure can alternatively be represented as a multitype branching-process with immigration; see Hawkes and Oakes (1974). The crucial parameters of a Hawkes model are the *excitement functions* or, emphasizing the branching interpretation, the *reproduction intensities* that govern these self- and crosseffects. For a textbook reference that covers many aspects of the Hawkes process, see Daley and Vere-Jones (2009). Maximum likelihood estimation of Hawkes processes has been treated in Ozaki (1979) and Ogata (1988). Liniger (2009) deals especially with the construction of the multivariate and marked case; in particular, it formalizes a computationally beneficial recursive method for likelihood calculation in the exponential decay case.

In the present paper, we formally introduce the *Hawkes graph*. The Hawkes graph summarizes the branching structure of a multitype Hawkes point process as a directed graph with weighted vertices and edges. The vertices represent the possible event-types of the corresponding Hawkes process; an edge (i, j) denotes nonzero excitement from event-type i to event-type j. The vertex weights are the corresponding immigration intensities; the weight of an edge (i, j) is the expected number of type-j children events that an type-i event generates. The *Hawkes skeleton* is the Hawkes graph disregarding these vertex and edge weights. The network view on Hawkes processes has been considered in Zhou et al. (2013), Delattre et al. (2016), Bacry et al. (2015a), and Hall and Willett (2016). The graph terminology is convenient to describe many relevant aspects of multivariate Hawkes processes such as 'ancestor and parent sets', 'paths', 'path weights', 'feedback', 'cascades', or 'connectivity'. The graph representation of a Hawkes process also provides additional theoretical insight. For example, in Theorem 9, we give a graph-based criterion for subcriticality which is equivalent to the usual spectral-radius based criterion on the branching matrix. Furthermore, the graph approach turns out to be helpful for the estimation of multivariate Hawkes processes.

Concerning Hawkes process estimation, we see three main problems with the standard parametric likelihood approach. First of all, it uses many unjustified assumptions on the shape of the reproduction intensities. Secondly, the distribution of the MLE-estimator is (in general) not known. In particular, the likelihood approach does not provide tests to decide whether excite-

[1]Note that the term 'Hawkes graph' has already been introduced for the graph representation of a specific finite group; see Hawkes (1968). Neither the author of the latter paper, *T.* Hawkes, nor its content has anything to do with our notion of a Hawkes graph.

ment from one event type to another exists *at all*. Finally, there are numerical issues that make it difficult to apply MLE in a straightforward way with large, high-dimensional event-stream datasets.

Our approach leaves the choice of the excitement functions open to the very last. We apply an estimation procedure developed in Kirchner (2017a). This procedure is based on a limit-representation of the Hawkes process studied in Kirchner (2016): we discretize the original process and interpret it as an autoregressive model of bin-counts. The latter is statistically estimated using conditional least-squares. In this setup, the asymptotic distribution of the resulting estimators can be obtained. This opens the door to testing. Our procedure is numerically more robust than the standard MLE approach. However, for high-dimensional data our procedure cannot be applied in a straightforward manner either. This is why, in combination with the concept of a Hawkes skeleton and graph, we tackle the numerical difficulties by the following three-step algorithm:

1. Given a large multitype event-stream dataset, we first apply a specific testing scheme to decide whether there is *any* effect from a specific event type to any other event type. The test result yields the *Hawkes-skeleton estimate*. In this first step, we use a parameter allowing us to tune for a *very coarse discretization*; this keeps the computational complexity under control. Despite the resulting discretization error, this approach typically yields a *superset* of the true edge set. Under the paradigm that the graph of the true underlying multivariate Hawkes model is typically sparse, this estimated superset is still sparse.
2. In a second step, we estimate the *Hawkes graph given the skeleton estimate*. The Hawkes graph *quantifies* the remaining excitement effects. The sparseness of the estimated Hawkes-skeleton from (i) reduces the complexity of the estimation problem considerably: there are only few excitements left to estimate and there are fewer 'explanatory types' per event type, namely the estimated parent sets. Consequently, we may now choose a much finer discretization parameter and thus retrieve more precise edge and vertex weight estimates—including confidence intervals for all estimated values.
3. As a by-product, the calculations in (ii) yield estimates for the values of the nonzero excitement-functions on a finite equidistant grid. We exploit these estimation results graphically to choose appropriate parametric function-families. Finally, we fit the chosen parametric functions to the corresponding estimates by a non-linear least-squares method. This yields parameter estimates for parametric Hawkes models.

The multistep-procedure described above also works in a high-dimensional setting (such as dozens of event streams and thousands of events per component); the approach can be implemented in a straightforward way.

The paper is organized as follows: in Section 2, we give definitions and discuss graph attributes that are relevant for the description of multivariate Hawkes processes. In particular, we give results that clarify what kind of information on the Hawkes process a Hawkes graph encodes. In Section 3, we cite earlier results that allow for nonparametric estimation of Hawkes processes. We apply these methods to estimate the Hawkes skeleton and the Hawkes graph. Finally, we show how parametric families for the remaining nonzero reproduction intensities may be specified and calibrated. For an illustration of the new concepts introduced, we present a simulation study in Section 4. In Section 5, we conclude with directions for further research.

2 Definitions

In this section, we recall the branching construction of a multivariate Hawkes process as well as basic graph terminology. After this, we introduce the Hawkes skeleton as well as the Hawkes graph. The graph representation summarizes the branching structure of a Hawkes process in a compact and insightful manner.

2.1 Multivariate Hawkes processes

Throughout the paper, let $(\Omega, \mathbb{P}, \mathcal{F})$ be a complete probability space rich enough to carry all random variables involved. We give a constructive definition of the Hawkes process that emphasizes the branching structure. For a similar construction; see Hawkes and Oakes (1974) or Chapter 4 in Liniger (2009). The building blocks are Poisson random-measures on $\mathbb{R}$ endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

Definition 1. *Let $\lambda : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ be a locally integrable function. We say that M is a* Poisson random-measure *on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with* intensity function *λ whenever the following two conditions hold:*

1. *$M(B) \sim \text{Pois}\left(\int_B \lambda(s)\mathrm{d}s\right)$, $B \in \mathcal{B}(\mathbb{R})$.*
2. *If $B_1, B_2, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ with $B_i \cap B_j = \varnothing$, $i \neq j$, then $M(B_1), M(B_2), \ldots, M(B_n)$ are mutually independent.*

We write $M \sim \text{PRM}(\lambda \mathrm{d}s)$.

In the definition above we use the convention that $X \sim \text{Pois}(0) :\Leftrightarrow X \equiv 0$, a.s. and $X \sim \text{Pois}(\infty) :\Leftrightarrow X \equiv \infty$, a.s.

A multitype Hawkes process is a model for the occurrence of events on $\mathbb{R}$, where the events are assigned to a finite number of types. The different event-types are represented as (in general

dependent) random counting measures. For each event type, there is an immigration process. Each immigrant event independently generates a family. These families consist of cascades of Poisson random measures. A Hawkes process is the superposition of all such families. We formalize this construction in the definitions below. To emphasize the intuition behind the names of immigrants, generations, and families, we use the somewhat unusual letters **I**, **G**, and **F** for the corresponding processes.

Definition 2. *Let $d \in \mathbb{N}$ and $[d] := \{1, 2, \ldots, d\}$.*

1. *For $(i, j) \in [d]^2$, define* branching coefficients $a_{i,j} \geqslant 0$, displacement densities $w_{i,j}$ *supported on* $\mathbb{R}_{\geqslant 0}$, reproduction intensities $h_{i,j} := a_{i,j} w_{i,j}$, *and* reproduction processes $\xi_t^{(i,j)}(\cdot) := \xi^{(i,j)}(\cdot - t) \sim \mathrm{PRM}(h_{i,j}\mathrm{d}s)$, $t \in \mathbb{R}$, *mutually independent over* $(i, j, t) \in [d]^2 \times \mathbb{R}$.
2. *For $i_0 \in [d]$ and $g \in \mathbb{N}_0$, define the g-th* generation process *(generated by a type-i_0 event at time zero) as the d-tuple of random counting measures* $\mathbf{G}^{(i_0,g)} := \left(G_1^{(i_0,g)}, \ldots, G_d^{(i_0,g)}\right)$ *by*

$$
\begin{aligned}
G_j^{(i_0,0)}(B) &:= 1_{\{j=i_0\}} \delta_0(B), \quad B \in \mathcal{B}(\mathbb{R}),\ j \in [d], \\
G_j^{(i_0,g)}(B) &:= \sum_{i=1}^{d} \int_{\mathbb{R}} \xi_t^{(i,j)}(B) G_i^{(i_0,g-1)}(\mathrm{d}t), \quad B \in \mathcal{B}(\mathbb{R}),\ j \in [d],\ g \in \mathbb{N}. \qquad (2.1)
\end{aligned}
$$

3. *For $i_0 \in [d]$, define the* Hawkes family *(generated by a type-i_0 event at time zero) as the d-tuple of random counting measures*

$$
\mathbf{F}^{(i_0)} = \sum_{g \geqslant 0} \mathbf{G}^{(i_0,g)}.
$$

The branching structure of a Hawkes family is encoded in recursion (2.1). Note that the points of a Hawkes family actually form a *multitype branching random walk*; see Shi (2015). The following definition clarifies how the Hawkes family process is related to the prototypic branching process, the Galton–Watson process:

Definition 3. *For $i_0 \in [d]$, let $\mathbf{F}^{(i_0)}$ be a Hawkes family and let $\{\mathbf{G}^{(i_0,g)}\}_{g \in \mathbb{N}_0}$ be the corresponding generation processes constructed in Definition 2 above. For $g \in \mathbb{N}_0$, define*

$$
\mathbf{Y}_g^{(i_0)} := \left(Y_{g,1}^{(i_0)}, Y_{g,2}^{(i_0)}, \ldots, Y_{g,d}^{(i_0)}\right), \quad \text{where, for } j \in [d], \quad Y_{g,j}^{(i_0)} := G_j^{(i_0,g)}(\mathbb{R}).
$$

We call $(\mathbf{Y}_g^{(i_0)})_{g \in \mathbb{N}_0}$ the embedded generation process *of the Hawkes family $\mathbf{F}^{(i_0)}$.*

The embedded generation process $(\mathbf{Y}_g^{(i_0)})$ of a Hawkes family is a multitype Galton–Watson process. A multitype Galton–Watson process models the size of a population with individuals of d types, where each individual is alive during exactly one time unit; see Section 2.3 in

Haccou et al. (2005). The embedded generation process starts with a single type-i_0 individual in generation 0 and, for $g \in \mathbb{N}$, each type-i individual in generation $g-1$ gives offspring to $\text{Pois}(a_{i,j})$ $a_{i,j} = \int h_{i,j}\mathrm{d}t)$ type-j individuals in generation g. This is why $a_{i,j}$, $(i,j) \in [d]^2$, are called *branching coefficients* and why the matrix $A := (a_{i,j}) \in \mathbb{R}_{\geqslant 0}$ is called *branching matrix*.

Proposition 4. *Let A be the branching matrix of Hawkes families $\mathbf{F}^{(i_0)}$, $i_0 \in [d]$, respectively, of the corresponding embedded generation processes $(\mathbf{Y}_g^{(i_0)})$, $i_0 \in [d]$. Then we have that*

$$\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}) = \sum_{g \geqslant 0} \mathbb{E}\, Y_{g,j}^{(i_0)} < \infty, \quad (i_0, j) \in [d]^2, \tag{2.2}$$

if and only if the spectral radius of A is strictly less than 1. In this case, $(1_{d\times d} - A)$ is invertible and $(\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j)\in[d]^2} = (1_{d\times d} - A)^{-1}$.

Proof. Using

$$\mathbb{E}\, \mathbf{Y}_0^{(i_0)} = \mathbf{Y}_0^{(i_0)} = (0, \ldots, 0, \underbrace{1}_{i_0\text{-th entry}}, 0, \ldots, 0) \text{ and } \mathbb{E}\, \mathbf{Y}_g^{(i_0)} = \mathbb{E}\, \mathbf{Y}_{g-1}^{(i_0)} A,\ g \in \mathbb{N},\ i_0 \in [d],$$

it follows by induction that $(\mathbb{E}\, Y_{j,g}^{(i_0)})_{(i_0,j)\in[d]^2} = A^g$, $g \in \mathbb{N}_0$. By Fubini's theorem, we then get that $(\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j)\in[d]} = \sum_{g\geqslant 0} (\mathbb{E}\, Y_{g,j}^{(i_0)})_{(i_0,j)\in[d]} = \sum_{g\geqslant 0} A^g$. Given its entries are finite, the limit matrix $\sum_{g\geqslant 0} A^g$ is calculated like the limit of a real-valued converging geometric series. The equivalence in Proposition 4 follows from the fact that

$$\sum_{g=0}^{\infty} A^g \text{ converges} \quad \Leftrightarrow \quad \max\big\{|\lambda| :\ \lambda \text{ eigenvalue of } A\big\} < 1, \quad \text{for } A \in \mathbb{R}^{d\times d}. \tag{2.3}$$

A detailed proof for (2.3) can be found in Watson (2015). □

In particular, we get from Proposition 4 that a Hawkes family whose branching matrix satisfies (2.3) consists of an almost surely finite number of points.

Definition 5. *Let $\mathbf{I} = (I_1, I_2, \ldots, I_d)$ be a* Hawkes immigration process *with $I_{i_0} \sim PRM(\eta_{i_0}\mathrm{d}s)$, $i_0 \in [d]$, independent, where $\eta_{i_0} \geqslant 0$, $i_0 \in [d]$, are (constant)* immigration intensities. *Furthermore, let $\mathbf{F}_t^{(i_0)}(\cdot) := \mathbf{F}^{(i_0,t)}(\cdot - t)$, $t \in \mathbb{R}$, where $\mathbf{F}^{(i_0,t)}$, $t \in \mathbb{R}$, $i_0 \in [d]$, are independent copies of the generic Hawkes family processes $\mathbf{F}^{(i_0)}$ from Definition 2 above—also independent from the immigration process $\mathbf{I}$. Set*

$$\mathbf{N}(B) := \big(N_1(B), \ldots, N_d(B)\big) := \sum_{i_0=1}^{d} \int_{\mathbb{R}} \mathbf{F}_t^{(i_0)}(B)\, I_{i_0}(\mathrm{d}t), \quad B \in \mathcal{B}(\mathbb{R}).$$

*The d-tuple of random counting measures $\mathbf{N}$ is a d-*type Hawkes process. *If $N_i(\{T\}) = 1$, for some $i \in [d]$, we say that T is a* type-i event *or, synonymously, an* event in component *i.*

The Hawkes process **N** *is* subcritical *if the corresponding embedded generation processes are subcritical, i.e., if the spectral radius of their branching matrix is strictly smaller than 1.*

From Hawkes and Oakes (1974) we have that, in the subcritical case, a Hawkes process **N**, constructed as in Definitions 2 and 5, is a stationary solution to the system of implicit equations

$$\begin{aligned}\Lambda_j(t) &:= \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{E}\left[N_j\big((t, t+\delta]\big) \Big| \sigma\Big(\mathbf{N}\left((a,b]\right),\, a < b \leqslant t \Big) \right] \\ &= \eta_j + \sum_{i=1}^{d} \int_{-\infty}^{t} h_{i,j}(t-s) N_i\left(\mathrm{d}s\right), \quad t \in \mathbb{R},\ j \in [d]. \end{aligned} \tag{2.4}$$

We call $\mathbf{\Lambda}(t) := (\Lambda_1(t), \Lambda_2(t), \ldots, \Lambda_d(t))$ the *conditional intensity* of **N**. In terms of intensities, the value of a reproduction intensity at time t, $h_{i,j}(t)$, denotes the effect of an event $T^{(i)}$ in component i on the intensity of component j at time $T^{(i)} + t$.

Remark 6. In most work on Hawkes processes, including the original introductions (Hawkes, 1971b,a) and also including Kirchner (2017a), the function $h_{i,j}$ models the excitement *from component j on component i*. This somewhat counter-intuitive notation stems from the linear algebra used when writing (2.4) with matrix multiplication. In the present graph-driven work, '$a_{i,j}$', '$w_{i,j}$', '$h_{i,j}$', and '$(i,j) \in \mathcal{E}$' *all refer to the effect from component i on component j.*

2.2 Hawkes skeleton and Hawkes graph

We interpret the branching structure of the Hawkes process in terms of 'causality'. The overall goal of causality research is to describe dependencies in a directed manner—rather than applying commutative concepts such as correlation; see Pearl (2009) for a recent overview. The notion of causality is subtle. For Hawkes processes, however, the use of the term seems justified. Indeed, in the context of event streams, things cannot become much more 'causal' than in the recurrent parent/children relation of a branching process: if we delete an event in the branching construction from the definitions in Section 2.1 above, its offspring vanishes. So—without discussing causality formally—we postulate that given an event in component i, it directly *causes* $\mathrm{Pois}(a_{i,j})$ new events in component j. This makes the branching coefficient $a_{i,j}$ an obvious measure for the strength of the causal effect from component i on component j. Such causal effects are often represented as directed graphs. In the literature on causality, a graphical approach for modeling the interdependence of event streams can for instance be found in Meek (2014) or Gunawardana et al. (2014)—without any mentioning of 'Hawkes'. This shows how natural the definition of a Hawkes graph is. First, we introduce some general graph terminology:

Definition 7. *Let $d \in \mathbb{N}$ and $[d] = \{1, 2, \ldots, d\}$. A* directed graph *$\mathcal{G}$ is a 2-tuple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [d]$ is a set of* vertices *and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of* edges*. Given such a directed graph $\mathcal{G}$*

we introduce the following definitions:

i) Vertex i is a parent *of vertex j if $(i,j) \in \mathcal{E}$. We write* $\mathrm{PA}(j) := \{i : \ (i,j) \in \mathcal{E}\}$. *Vertex i is a* source vertex *if* $\mathrm{PA}(i) \backslash \{i\} = \varnothing$. *Vertex i is a* sink vertex *if* $\{j : \ (i,j) \in \mathcal{E}\} \backslash \{i\} = \varnothing$.

ii) For $g \in \mathbb{N}$, $(k_0, k_1, \ldots, k_g) \in \mathcal{V}^{g+1}$ is a walk *in $\mathcal{G}$ of length g from vertex i to vertex j if $k_0 = i, k_g = j$ and $(k_{l-1}, k_l) \in \mathcal{E}$, $l \in [g]$; $(k_0, k_1, \ldots, k_g) \in \mathcal{V}^{g+1}$ is a* closed walk *if it is a walk with $k_0 = k_g$. We denote the* set of walks in $\mathcal{G}$ from i to j with length $g \in \mathbb{N}$ *by* $\mathcal{W}_g^{(i,j)}$. *Furthermore, we set* $\mathcal{W}_0^{(i,j)} := \varnothing$ *if* $i \neq j$, $\mathcal{W}_0^{(i,j)} := \{(i)\}$ *if* $i = j$, $\mathcal{W}^{(i,j)} := \cup_{g \geqslant 0} \mathcal{W}_g^{(i,j)}$, *and* $\mathcal{W} := \cup_{(i,j) \in [d]^2} \mathcal{W}^{(i,j)}$.

iii) Vertex i is an ancestor *of j if there exists a walk of length $g \in \mathbb{N}$ from i to j. We denote the* ancestor set *of a vertex i in $\mathcal{G}$ by* $\mathrm{AN}(i)$.

iv) The vertices i and j are weakly connected *if $i = j$ or if there exists a set $\{(k_{l-1}, k_l),\ l = 1, \ldots, g : \ k_0 = i, k_g = j, (k_l, k_{l-1}) \in \mathcal{E}$ or $(k_{l-1}, k_l) \in \mathcal{E}\}$ for some $g \in \mathbb{N}$. The vertices i and j are* strongly connected *if the sets $\mathcal{W}^{i,j}$ and $\mathcal{W}^{j,i}$ are nonempty. A directed graph is* weakly (strongly) connected *if all pairs of its vertices are weakly (strongly) connected. A directed graph is* fully connected *if $(i,j) \in \mathcal{E}$, $(i,j) \in [d]^2$.*

In the remainder of the paper, we will simply write 'graph' for 'directed graph'. Also note that in our definition, a graph allows cycles and, in particular, self-loops. A vertex may or may not be an ancestor and, in particular, a parent of itself. Also note that any vertex i is always strongly connected to itself because $\{(i)\} \subset \mathcal{W}^{(i,i)}$, $i \in [d]$—no matter if i is contained in a closed walk or not. Consequently, the singleton graph is always strongly connected. However, it is only fully connected if is a self-loop. Next, we apply the graph terminology from Definition 7 to the Hawkes process:

Definition 8. *Let* $\mathbf{N}$ *be a d-type Hawkes process with immigration intensities $\eta_1, \eta_2, \ldots, \eta_d$ and branching coefficients $a_{i,j} (= \int h_{i,j}(t) \mathrm{d}t)$, $(i,j) \in [d]^2$; see Definitions 2 and 5. The* Hawkes graph skeleton $\mathcal{G}_{\mathbf{N}}^* = (\mathcal{V}_{\mathbf{N}}^*, \mathcal{E}_{\mathbf{N}}^*)$ *of* $\mathbf{N}$ *consists of a set of vertices $\mathcal{V}_{\mathbf{N}}^* = [d]$ and a set of edges*

$$\mathcal{E}_{\mathbf{N}}^* := \Big\{ (i,j) \in \mathcal{V}_{\mathbf{N}}^* \times \mathcal{V}_{\mathbf{N}}^* : \ a_{i,j} > 0 \Big\}.$$

For $j \in [d]$, we denote the parent, *respectively,* ancestor set *of j with respect to the Hawkes skeleton $\mathcal{G}_{\mathbf{N}}^*$ by* $\mathrm{PA}_{\mathbf{N}}(j)$ *and* $\mathrm{AN}_{\mathbf{N}}(j)$. *For the* Hawkes graph $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ *of* $\mathbf{N}$, *each vertex, respectively, edge of the corresponding Hawkes skeleton is supplied with a* vertex, *respectively, an* edge weight*:*

$$\begin{aligned} \mathcal{V}_{\mathbf{N}} &:= \Big\{ (j; \eta_j) : \quad j \in \mathcal{V}_{\mathbf{N}}^* \text{ and } \eta_j \text{ is the } j\text{-}th \text{ immigration intensity of } \mathbf{N} \Big\}, \\ \mathcal{E}_{\mathbf{N}} &:= \Big\{ (i,j; a_{i,j}) : \quad (i,j) \in \mathcal{E}_{\mathbf{N}}^* \text{ and } (a_{i,j})_{(i,j) \in [d]^2} \text{ is the branching matrix of } \mathbf{N} \Big\}. \end{aligned}$$

We call the branching matrix $A = (a_{i,j}) \in \mathbb{R}_{\geqslant 0}^{d\times d}$ *of* $\mathbf{N}$ *the* adjacency matrix *of* $\mathcal{G}_{\mathbf{N}}$.

i) A Hawkes graph $\mathcal{G}_{\mathbf{N}}$ *is* weakly, strongly, respectively, fully connected *if the corresponding skeleton* $\mathcal{G}_{\mathbf{N}}^*$ *is weakly, strongly, respectively, fully connected; see Definition 7.*

ii) Vertex $(j;\eta_j)$ *of a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$ *is a* source*, respectively,* sink vertex*, if it is a source, respectively, sink vertex in the corresponding skeleton* $\mathcal{G}_{\mathbf{N}}^*$*. Furthermore,* $(j;\eta_j)$ *is a* redundant vertex *if* $\eta_j = 0$ *and, in addition,* $\eta_i = 0$ *for all* $i \in \mathrm{AN}_{\mathbf{N}}(j)$.

iii) For any walk $w \in \mathcal{W}_{\mathcal{G}_{\mathbf{N}}}$ $(:= \mathcal{W}_{\mathcal{G}_{\mathbf{N}}^*})$ *in a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$*, we define the* walk weights

$$|w| = |(i_0, i_1, \ldots, i_g)| := \begin{cases} 1, & g = 0, \text{ and} \\ \prod_{l=1}^{g} a_{i_{l-1},i_l}, & g > 0, \end{cases}$$

where a_{i_{l-1},i_l}, $l = 1, 2, \ldots, g$, *are the edge weights from* $\mathcal{E}_{\mathbf{N}}$.

iv) A Hawkes graph is subcritical *if*

$$\sum_{w\in\mathcal{W}^{(i_0,i_0)}} |w| < \infty,\ i_0 \in [d], \text{ or, equivalently, } \sum_{\substack{w:\\ w \text{ closed walk in } \mathcal{G}_{\mathbf{N}}}} |w| < \infty. \tag{2.5}$$

Note that if a Hawkes graph vertex is redundant, then all its ancestors are also redundant. The notion of a subcritical Hawkes graph in Definition 8 iv) might ask for further explanation. The following theorem clarifies things:

Theorem 9. *Let* $\mathbf{N}$ *be a Hawkes process and let* $\mathcal{G}_{\mathbf{N}}$ *be the corresponding Hawkes graph. Then* $\mathbf{N}$ *is a subcritical Hawkes process (see Definition 5) if and only if* $\mathcal{G}_{\mathbf{N}}$ *is a subcritical Hawkes graph (see Definition 8).*

Proof. First, we prove that

$$\sum_{w\in\mathcal{W}^{(i_0,i_0)}} |w| < \infty,\ i_0 \in [d] \quad \Leftrightarrow \quad \sum_{w\in\mathcal{W}^{(i_0,j)}} |w| < \infty,\ (i_0, j) \in [d]^2. \tag{2.6}$$

'$\Leftarrow$' is trivial. We show '$\Rightarrow$' by induction over the graph size d: for $d = 1$, the implication is true. For $d > 1$, consider a graph with d vertices and assume that the left-hand side of (2.6) holds. Pick any $(i_0, j) \in [d]^2$. We split the possible paths from i_0 to j, $\mathcal{W}^{(i_0,j)}$, into *paths excluding* d, $\mathcal{W}^{(i_0,j)}_{\mathrm{excl}.d}$, and *paths including* d, $\mathcal{W}^{(i_0,j)}_{\mathrm{incl}.d}$:

$$\sum_{w\in\mathcal{W}^{(i_0,j)}} |w| = \sum_{w\in\mathcal{W}^{(i_0,j)}_{\mathrm{excl}.d}} |w| + \sum_{w\in\mathcal{W}^{(i_0,j)}_{\mathrm{incl}.d}} |w|. \tag{2.7}$$

The first sum is finite by the induction hypothesis. Now, assume the case that $i_0 \neq d$ and $j \neq d$. Every walk in the second sum of (2.7) may be (uniquely) split into the following five subwalks:

a d-avoiding walk w_1 from i_0 to some $i_1 \in \mathrm{PA}(d)$, a one-step walk (i_1, d), a walk $w_2 \in \mathcal{W}^{(d,d)}$, another one-step walk (d, j_1), with $d \in \mathrm{PA}(j_1)$, and finally some d-avoiding walk w_3 from j_1 to j. This yields

$$\begin{aligned}
&\sum_{w \in \mathcal{W}^{(i_0,j)}_{\text{incl.}d}} |w| \\
&\quad = \sum_{i_1 \in \mathrm{PA}(d)} \sum_{w_1 \in \mathcal{W}^{(i_0,i_1)}_{\text{excl.}d}} \sum_{w_2 \in \mathcal{W}^{(d,d)}} \sum_{j_1 : d \in \mathrm{PA}(j_1)} \sum_{w_3 \in \mathcal{W}^{(j_1,j)}_{\text{excl.}d}} |w_1|\, a_{i_1,d}\, |w_2|\, a_{d,j_1}\, |w_3| \\
&\quad \leqslant \sum_{i_1 \in \mathrm{PA}(d)} \sum_{j_1 : d \in \mathrm{PA}(j_1)} \max_{(i,j) \in [d]^2} a_{i,j}^2 \underbrace{\sum_{w_1 \in \mathcal{W}^{(i_0,i_1)}_{\text{excl.}d}} |w_1|}_{<\infty \text{ by ind. hyp.}} \underbrace{\sum_{w_2 \in \mathcal{W}^{(d,d)}} |w_2|}_{<\infty \text{ by assumption}} \underbrace{\sum_{w_3 \in \mathcal{W}^{(j_1,j)}_{\text{excl.}d}} |w_3|}_{<\infty \text{ by ind. hyp.}} < \infty.
\end{aligned}$$

Note that, by definition, $(i) \in \mathcal{W}^{(i,i)}$ and $|(i)| = 1$, $i \in [d]$, so that the calculation above also covers the cases $\mathrm{PA}(d) = \{i_0\}$ and $\mathrm{PA}(j) = \{d\}$ as well as d-including walks from i to j that touch d exactly once. If $i_0 = d$ or $j = d$, the splitting argument becomes even simpler; we do not give the details. We have proven the finiteness of the second sum in (2.7) and therefore (2.6).

Next, note that

$$\sum_{w \in \mathcal{W}^{(i_0,j)}} |w| = \sum_{g \geqslant 0} \sum_{w \in \mathcal{W}^{(i_0,j)}_g} |w| = \sum_{g \geqslant 0} \mathbb{E}\, Y^{(i_0)}_{g,j} = \mathbb{E}\, F^{(i_0)}_j(\mathbb{R}), \quad (i_0, j) \in [d]^2, \tag{2.8}$$

where $(\mathbf{Y}^{(i_0)}_g) = (Y^{(i_0)}_{g,1}, Y^{(i_0)}_{g,2}, \ldots, Y^{(i_0)}_{g,d})$ are the embedded generation processes of the generic family processes $\mathbf{F}^{(i_0)} = (F^{(i_0)}_1, \ldots, F^{(i_0)}_d)$ of $\mathbf{N}$; see Definition 3. Thus, (2.5) is a complicated way of saying that, for all $(i_0, j) \in [d]^2$, the expected total number of type-j offspring events of a type-i_0 event is finite, i.e., that $\mathbb{E}\, F^{(i_0)}_j(\mathbb{R}) < \infty$, $(i_0, j) \in [d]^2$. By Proposition 4, this in turn is equivalent to the spectral radius of the branching matrix being strictly less than 1—which is the original Hawkes-*process* subcriticality condition from Definition 5 □

Obviously, the Hawkes graph does not fully specify the corresponding Hawkes process; it only captures the structure of the embedded generation processes from Definition 3 together with the immigration intensities. Despite this simplification, the Hawkes graph gives relevant insight into the underlying Hawkes process—especially in the highdimensional case. For example, *connectivity and redundancy of vertices* are two graph-based concepts that become increasingly important the higher the dimension of the model considered is. If a Hawkes graph is not weakly connected, we may consider the *weakly connected* subgraphs separately and correspondingly split the original model into separate, lower-dimensional Hawkes processes. The notion of *redundant vertices* is important because, typically, we only want to consider 'acces-

sible' event types. *Sink* (*source*) *vertices* of a Hawkes graph correspond to Hawkes process components that only receive (give) excitement from (to) the system. The notion of *parent sets* is also helpful: e.g., for the marginal conditional intensity in (2.4), it is actually enough to sum over $i \in \mathrm{PA}(j)$ instead of $i \in [d]$ which may be computationally beneficial. The *ancestor sets* may be applied if we are only interested in modeling events of a particular type j. In this situation, it suffices to consider a Hawkes model for the event types in $\{j\} \cup \mathrm{AN}(j)$. Finally, we find the formulation of Hawkes graph *subcriticality* in (2.5) useful. It provides a more concrete meaning to the somewhat abstract eigenvalue-based criterion for the Hawkes process. E.g., (2.5) can be used when constructing subcritical Hawkes graphs, respectively, models. And—if a given graph is sparse and the closed walks are not too numerous—one can check subcriticality without even calculating any eigenvalue; see Section 4.1. Furthermore, in some cases, the *path weights* $|w|$ themselves might be worth calculating—even apart from criticality conditions; see the discussion in the proof of Theorem 9. Last but not least, the graph structure obviously allows for attractive self-explaining illustrations; see Figures 1 and 2. In the following proposition, we collect some specific graphical and statistical information that may be calculated from the adjacency matrix of a Hawkes graph:

Proposition 10. *For some $d \geqslant 2$, let $\mathbf{N}$ be a d-type subcritical Hawkes process. Furthermore, let $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ be the corresponding Hawkes graph with adjacancy matrix $A = (a_{i,j}) \in \mathbb{R}^{d\times d}_{\geqslant 0}$. Then we have that*

i) $a_{i,j} > 0 \quad \Leftrightarrow \quad i \in \mathrm{PA}_{\mathbf{N}}(j)$*;*

ii) $a_{i,j} = 0,\ j \in [d]\backslash\{i\} \quad \Leftrightarrow$ *vertex i is a sink vertex;*

iii) $a_{i,j} = 0,\ i \in [d]\backslash\{j\} \quad \Leftrightarrow$ *vertex j is a source vertex;*

iv) $(A^g)_{i,j} > 0 \quad \Leftrightarrow$ *there is a walk of length g from i to j;*

v) $(A^g)_{i,j} > 0$ *for some* $g \in [d] \quad \Leftrightarrow \quad i \in \mathrm{AN}(j)$*;*

vi) *for all* $(i,j) \in [d]^2$, $((A + A^\top)^g)_{i,j} > 0$ *for some* $g \in \{0\} \cup [d-1] \quad \Leftrightarrow$ *the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is weakly connected;*

vii) *for all* $(i,j) \in [d]^2$, $((A)^g)_{i,j} > 0$ *for some* $g \in \{0\} \cup [d-1] \quad \Leftrightarrow$ *the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is strongly connected;*

viii) $a_{i,j} > 0,\ (i,j) \in [d]^2 \quad \Leftrightarrow$ *the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is fully connected;*

The properties above can easily be checked. They may help to describe the relationships between Hawkes process components, respectively, Hawkes graph vertices. Two specific $\mathbb{R}^d_{\geqslant 0}$-vectors might be particularly meaningful statistical summaries of a Hawkes graph, respectively, Hawkes process:

Definition 11. *Let* **N** *be a subcritical d-type Hawkes process and let A be the adjacency matrix of the corresponding Hawkes graph* $\mathcal{G}_{\mathbf{N}}$. *Consider the limit matrix* $\mathbb{R}_{\geqslant 0}^{d\times d} \ni (e_{i,j}) := (1_{d\times d} - A)^{-1} = \sum_{g\geqslant 0} A^g \, (= (\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j)\in[d]^2})$ *from Proposition 4 and define*

$$c_{i_0} := \frac{\eta_{i_0} \sum_{j=1}^d e_{i_0,j}}{\sum_{i=1}^d \eta_i \sum_{j=1}^d e_{i,j}}, \quad i_0 \in [d], \quad \textit{and} \quad f_j := \frac{\eta_j e_{j,j}}{\sum_{i=1}^d \eta_i e_{i,j}}, \quad j \in [d].$$

We call $(c_{i_0})_{i\in[d]}$ *the* cascade coefficients *and* $(f_j)_{j\in[d]}$ *the* feedback coefficients.

One way of tuning a specific Hawkes graph may be achieved by 'switching-off' a selected vertex by forcing the corresponding immigration intensity to zero. The coefficients defined above summarize the effect of such a manipulation. In view of Proposition 4, we have the following interpretations. First of all, the *cascade coefficients* (c_i) are important from a *systemic* point of view. The cascade coefficient c_i measures the fraction of events in the system stemming from families with immigrated type-i ancestor. If $c_i > 1/d$, this indicates a relatively large impact of type-i events on the system. Secondly, the *feedback coefficients* (f_j) are more important from an *individual* point of view. They indicate how much of the total intensity that a vertex j *experiences* is due to its own immigration activity including the feedback it experiences by closed walks. We illustrate both concepts in Section 4.1.

3 Estimation

In this section, we give a summary of earlier work, where we introduced a nonparametric estimation procedure for the multivariate Hawkes process. Based on this approach, we introduce an estimation procedure for the Hawkes skeleton and the Hawkes graph. In particular, we clarify how one can bypass numerical problems in high-dimensional settings. Finally, we explain how one can use the results for completely specifying and estimating a parametric Hawkes model.

3.1 Earlier results

In (Kirchner, 2016), we showed that the distributions of the bin-count sequences of a Hawkes process can be approximated by the distribution of so called *integer-valued autoregressive time series* INAR(p). This approximation yields an estimation method for the Hawkes process: we fit the approximating model on observed bin-counts of point process data. The resulting estimates can be used as estimates of the Hawkes reproduction intensities on a finite and equidistant grid; see Kirchner (2017a). For illustration, consider a univariate Hawkes process N with reproduction intensity h and immigration intensity η. Given data from N in a time window $(0, T]$, $\Delta > 0$, small, bin counts $X_n^{(\Delta)} := N\big(((n-1)\Delta, n\Delta]\big)$, $k = 1, 2, \ldots, n := \lfloor T/\Delta \rfloor$, and some $p \in \mathbb{N}$, large,

we calculate

$$\left(\hat{\alpha}_0^{(\Delta)}, \hat{\alpha}_1^{(\Delta)}, \ldots, \hat{\alpha}_p^{(\Delta)}\right) := \operatorname{argmin}_{(\alpha_0^{(\Delta)}, \alpha_1^{(\Delta)}, \ldots, \alpha_p^{(\Delta)})} \sum_{k=p+1}^{n} \left(X_k^{(\Delta)} - \alpha_0^{(\Delta)} - \sum_{l=1}^{p} \alpha_l^{(\Delta)} X_{k-l}^{(\Delta)} \right)^2 . \quad (3.1)$$

Given (3.1), we estimate the reproduction-intensity values $h(k\Delta)$, $k = 1, 2, \ldots, p$, of N by $\hat{h}_k := \hat{\alpha}_k^{(\Delta)}/\Delta$ and the immigration intensity η by $\hat{\eta} := \hat{\alpha}_0^{(\Delta)}/\Delta$. The multivariate case is conceptually equivalent but somewhat cumbersome notationwise. Furthermore—due to the special distribution of the errors—the covariance matrix of the estimates is nonstandard. This is why we give all formulas in some detail. The following definitions and properties are taken from Kirchner (2017a)—modulo transposition as stated in Remark 6.

Definition 12. *Let* $\mathbf{N} = (N_1, N_2, \ldots, N_d)$ *be a subcritical d-type Hawkes process with immigration intensity* $\eta \in \mathbb{R}_{\geqslant 0}^d \setminus \{0_d\}$ *and reproduction intensities* $h_{i,j} : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, $(i, j) \in [d]^2$. *Let* $T > 0$ *and consider a sample of the process on the time interval* $(0, T]$. *For some* $\Delta > 0$, *construct the* $\mathbb{N}_0^d$*-valued* bin-count sequence *from this sample:*

$$\mathbf{X}_k^{(\Delta)} := \mathbf{N}\left(\big((k-1)\Delta, k\Delta\big]\right)^\top \in \mathbb{N}_0^{d\times 1}, \quad k = 1, 2, \ldots, n := \lfloor T/\Delta \rfloor . \quad (3.2)$$

Define the multivariate Hawkes estimator *with respect to some support* s, $\Delta < s < T$,

$$\widehat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Y} \quad \in \mathbb{R}^{(dp+1)\times d} . \quad (3.3)$$

Here,

$$\mathbf{Z}\left(\mathbf{X}_1^{(\Delta)}, \ldots, \mathbf{X}_n^{(\Delta)}\right) := \begin{pmatrix} (\mathbf{X}_p^{(\Delta)})^\top & (\mathbf{X}_{p-1}^{(\Delta)})^\top & \ldots & (\mathbf{X}_1^{(\Delta)})^\top & 1 \\ (\mathbf{X}_{p+1}^{(\Delta)})^\top & (\mathbf{X}_p^{(\Delta)})^\top & \ldots & (\mathbf{X}_2^{(\Delta)})^\top & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (\mathbf{X}_{n-1}^{(\Delta)})^\top & (\mathbf{X}_{n-2}^{(\Delta)})^\top & \ldots & (\mathbf{X}_{n-p}^{(\Delta)})^\top & 1 \end{pmatrix} \in \mathbb{R}^{(n-p)\times(dp+1)} \quad (3.4)$$

is the design matrix *and* $\mathbf{Y}\left(\mathbf{X}_1^{(\Delta)}, \ldots, \mathbf{X}_n^{(\Delta)}\right) := \left(\mathbf{X}_{p+1}^{(\Delta)}, \mathbf{X}_{p+2}^{(\Delta)}, \ldots, \mathbf{X}_n^{(\Delta)}\right)^\top \in \mathbb{R}^{(n-p)\times d}$ *with* $p := \lceil s/\Delta \rceil$.

For the following considerations, we drop the '(Δ, s)' superscript. Note that also the matrices $\mathbf{Z}$ and $\mathbf{Y}$ depend on Δ. Additional notation clarifies what the entries of the matrix $\widehat{\mathbf{H}}$ in (3.3)

actually estimate:

$$\begin{pmatrix} \widehat{H}_1 \\ \cdots \\ \widehat{H}_p \\ \hat{\eta} \end{pmatrix} := \widehat{\mathbf{H}} \in \mathbb{R}^{(dp+1)\times d}, \quad \text{where} \quad \widehat{H}_k := \begin{pmatrix} \hat{h}_{1,1}(k\Delta) & \hat{h}_{1,2}(k\Delta) & \dots & \hat{h}_{1,d}(k\Delta) \\ \hat{h}_{2,1}(k\Delta) & \hat{h}_{2,2}(k\Delta) & \dots & \hat{h}_{2,d}(k\Delta) \\ \dots & \dots & \dots & \dots \\ \hat{h}_{d,1}(k\Delta) & \hat{h}_{d,2}(k\Delta) & \dots & \hat{h}_{d,d}(k\Delta) \end{pmatrix}. \quad (3.5)$$

In Kirchner (2017a), we find that, for large T, small Δ and large p, the entries of $\widehat{\mathbf{H}}$ are approximately jointly normally distributed around the true values. Furthermore, the covariance matrix of $\text{vec}\left(\widehat{\mathbf{H}}^\top\right) \in \mathbb{R}^{d(dp+1)}$ ($\text{vec}(\cdot)$ stacks the columns of its argument) can be consistently estimated by

$$\widehat{S^2} := \frac{1}{\Delta^2}\left(\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right) \mathbf{W} \left(\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right) \in \mathbb{R}^{d(dp+1)\times d(dp+1)}. \quad (3.6)$$

Here, $\otimes$ denotes the Kronecker product, $\mathbf{Z}$ is the design matrix from (3.4) and $\mathbf{W} := \sum_{k=p+1}^{n} \mathbf{w}_k \mathbf{w}_k^\top \in \mathbb{R}^{d(dp+1)\times d(dp+1)}$, where, for $k = p+1, p+2, \dots, n$,

$$\begin{aligned} \mathbf{w}_k \;:=\; & \left(\left(\left(\mathbf{X}_{k-1}^{(\Delta)}\right)^\top, \left(\mathbf{X}_{k-2}^{(\Delta)}\right)^\top, \dots, \left(\mathbf{X}_{k-p}^{(\Delta)}\right)^\top, 1\right)^\top \otimes 1_{d\times d}\right) \\ & \cdot \left(\mathbf{X}_k^{(\Delta)} - \Delta\hat{\eta} - \sum_{l=1}^{p} \Delta \widehat{H}_l^\top \mathbf{X}_{k-l}^{(\Delta)}\right) \in \mathbb{R}^{d(dp+1)\times 1}. \end{aligned} \quad (3.7)$$

In Definition 12, we consider $\text{vec}(\mathbf{H}^\top)$ instead of $\text{vec}(\mathbf{H})$ in order to apply the results from Kirchner (2016) more directly; see Remark 6. We will discuss below how one retrieves specific values from the covariance matrix estimation in (3.6). The estimator from Definition 12 above depends on a support s, $0 < s << T$, and on a bin size Δ, $0 < \Delta \leqslant s$. Automatic methods for the choice of these estimation parameters are discussed in Kirchner (2016). In the present paper, we assume s given. Often, an upper bound for the support of the reproduction intensities can be guessed from the data context. The choice of Δ, however, will be crucial in high-dimensional settings. We will use it as a tuning parameter for controlling numerical complexity.

3.2 Estimation of the Hawkes skeleton

Our first goal is to identify the edges of the Hawkes skeleton from data; see Definition 8. The idea is simple: for $(i,j) \in [d]^2$, we estimate the edge weight $a_{i,j} = \int h_{i,j}(t)\mathrm{d}t$ by $\hat{a}_{i,j} := \Delta \sum_{k=1}^{p} \hat{h}_{i,j}(k\Delta)$; see (3.5) for the notation. Calculating the covariance estimate (3.6), we can check whether $\hat{a}_{i,j}$ is significantly larger than zero. If this is the case, we set $(i,j) \in \widehat{\mathcal{E}}^*$. In order to ease implementation, we explicitly give the necessary transformations for the estimates from Definition 12 and discuss numerical issues.

Definition 13. *Given d-type event-stream data on* $(0,T]$*, calculate the Hawkes estimator* $\mathbf{H}^{(\Delta_{skel},s)}$ *from Definition 12 with respect to some* s, $0 < s < T$, *and some* Δ_{skel}, $0 < \Delta_{skel} \leqslant s$. *For* $j \in [d]$, *let* $b_j \in \{0,1\}^{(dp+1)\times 1}$ *be column vectors with all entries zero but 1s at entries* $(k-1)d+j$, $k = 1,2,\ldots,p = \lceil s/\Delta_{skel} \rceil$. *Let* $B := (b_1, b_2, \ldots, b_d)^\top$, *and calculate*

$$(\hat{a}_{i,j})_{1\leqslant i,j\leqslant d} = \Delta_{skel} B \, \mathbf{H}^{(\Delta_{skel},s)} . \tag{3.8}$$

Fix $\alpha_{skel} \in (0,1)$ *and define the* Hawkes-skeleton estimator *as a graph* $\widehat{\mathcal{G}}^* := ([d], \widehat{\mathcal{E}}^*)$, *with*

$$\widehat{\mathcal{E}}^* := \Big\{ (i,j) \in [d]^2 : \ \hat{a}_{i,j} > \hat{\sigma}_{i,j} z^{-1}_{1-\alpha_{skel}} \Big\}. \tag{3.9}$$

Here, for $\beta \in (0,1)$, z_β^{-1} *denotes the* β*-quantile of a standard normal distribution. Efficient calculation of* $(\hat{\sigma}_{i,j})_{1\leqslant i,j\leqslant d}$ *will be given in Algorithm 14 below.*

The main point of this first estimation step is that we hope that the edge set $|\mathcal{E}^*|$ and, consequently $|\widehat{\mathcal{E}}^*|$ are typically much smaller than d^2, respectively, that $\mathrm{PA}_{\mathbf{N}}(j)$, $j \in [d]$, and, consequently, $\widehat{\mathrm{PA}}_{\mathbf{N}}(j)$, $j \in [d]$, are typically much smaller than d. If this is the case, the knowledge of the skeleton simplifies the estimation of the Hawkes graph considerably.

The role of Δ_{skel}

On the one hand, the smaller we choose the bin size Δ, the better the discrete approximation described in Section 3.1 works. On the other hand, the matrices involved in the calculation of the Hawkes estimator from Definition 12 become increasingly large when Δ decreases. More specifically, (3.3) involves the construction and multiplication of matrices with about ds/Δ rows and about T/Δ columns, where $T > 0$ denotes the sample window size, $d \in \mathbb{N}$ the number of event-types, and s, $\Delta \leqslant s << T$, the support parameter from Definition 12. Furthermore, we have to invert matrices of size $\lceil ds/\Delta \rceil \times \lceil ds/\Delta \rceil$. The crucial observation is that in the Hawkes-skeleton estimation, we may choose Δ_{skel} quite large for two reasons:

i) The test involved in (3.9) does not depend on Δ_{skel} too heavily. The false positive rate (that is, the probability of *including a false edge*) is well controlled by α_{skel}, because, under $H_0 : h_{i,j} \equiv 0$, discretizations as in (3.1) stay 'correct' even for very coarse Δ_{skel}; see (3.10) below. The false negative rate (probability of *missing a true edge*) naturally depends strongly on the true underlying edge weights. However, if there is truly considerable direct excitement from one component to another, then typically the effect from some bin to future bins will also be of some significance—which is exactly what our skeleton estimator tests. Our simulation study in Section 4.2 confirms these arguments.

ii) The actual *quantitative* estimation of the interactions between different event types will be performed in a second step when we consider the Hawkes *graph*. In this second step,

due to the (hoped-for) sparseness of the Hawkes skeleton, we are typically able to choose a much finer bin size Δ_{graph}. So we may ignore the bias stemming from a somewhat rough discretization in the first (skeleton-estimation) step.

By choosing $\Delta_{\text{skel}} = s/k$ for some small $k \in \mathbb{N}$ in the calculations of Definition 13 above, even Hawkes-skeleton estimates of very high-dimensional models (such as $d > 20$) become computationally tractable.

The role of α_{skel}

Note that under $H_0 : a_{i,j} \equiv 0$, we have that

$$\mathbb{P}_{H_0}[\hat{a}_{i,j} > \hat{\sigma}^2_{i,j} z^{-1}_{1-\alpha_{\text{skel}}}] \approx \alpha_{\text{skel}}. \tag{3.10}$$

Still, the parameter $\alpha_{\text{skel}} \in (0,1)$ should not so much be thought of as an actual significance level—due to the multiple testing setup over $(i,j) \in [d]^2$, and because of the dependence between the different edge tests. Despite this warning, note that in the simulation study from Section 4.2, the corresponding empirical false positive rates are very close to our (varying) choices of α_{skel}. In any case, α_{skel} is a flexible tuning parameter that allows for controlling the degree of sparseness in the estimated graph. A value of $\alpha_{\text{skel}} = 1$ will yield a fully connected estimated graph as Hawkes skeleton. When α_{skel} decreases, the skeleton estimate becomes sparser and sparser. For $\alpha_{\text{skel}} \geqslant 0.01$, we typically still *overestimate* the true edge set. In other words, for $j \in [d]$, we typically have that $\text{PA}_{\mathbf{N}}(j) \subset \widehat{\text{PA}}_{\mathbf{N}}(j)$ with high probability.

Variance estimate calculation

The most elaborate step from a computational point of view in Definition 12 is the calculation of the covariance estimator in (3.6). Here, we deal with matrices of size $\lceil d^2 s/\Delta \rceil \times \lceil d^2 s/\Delta \rceil$. Furthermore, we have to calculate approximately T/Δ vectors of size $d^2 s/\Delta$ and calculate and sum their crossproducts $\mathbf{w}_k \mathbf{w}_k^\top$. This is the numerical bottleneck of the procedure—in particular for high-dimensional setups. For the Hawkes-skeleton estimator from Definition 13, we simplify the calculation. First of all, we note that in the matrix $\widehat{S}^2$ from (3.6), we estimate many more covariance values than we actually need for the (marginal) distribution of the edge-weight estimates. After some linear algebra, we find that one can avoid the tedious computation of the $\mathbf{W}$ matrix from (3.6) by the following matrix manipulations.

Algorithm 14. *Let* $\mathbf{E} \in \{0,1\}^{d^2 \times (d^2 p + d)}$ *be a matrix with all entries zero but, for* $(i,j) = [d]^2$*, in row* $(i-1)d+j$ *we have 1s at entries* $(k-1)d^2+(i-1)d+j$*,* $k = 1, 2, \ldots, p$*. Let* $\mathbf{E}_{l,\cdot}$ *denote the l-th row of* $\mathbf{E}$*. With* $\widehat{S}^2$ *from* (3.6) *and for* $(i,j) \in [d]^2$*, we have that* $\hat{\sigma}^2_{i,j} := \Delta^2 \, \mathbf{E}^\top_{(i-1)d+j,\cdot} \, \widehat{S}^2 \, \mathbf{E}_{(i-1)d+j,\cdot}$

are the variance estimates for the $\hat{a}_{i,j}$ from (3.8)*. These estimates can be computed in the following way:*

i) Compute $\mathbf{E}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\mathbf{Z}^\top \otimes 1_{d\times d} \in \mathbb{R}^{d^2\times d(n-p)}$ and stack the rows of the result in a vector. Fill this vector row-wise in a $d^2(n-p)\times d$ matrix $\mathbf{C}$.

ii) Set $\mathbf{U} = (\mathbf{Y} - \Delta\,\mathbf{Z}\,\widehat{\mathbf{H}}) \in \mathbb{R}^{(n-p)\times d}$. Denoting $(U_{p+1}, U_{p+2}, \ldots, U_n)^\top := \mathbf{U}$, we now have that

$$U_k = \left(\mathbf{X}_k^{(\Delta)} - \Delta\hat{\eta} - \sum_{l=1}^{p} \Delta\widehat{H}_l^\top\, \mathbf{X}_{k-l}^{(\Delta)}\right), \quad k = p+1, p+2, \ldots, n.$$

Furthermore, let $\mathbf{U}^{(rep)} \in \mathbb{R}^{d^2(n-p)\times d}$ be a matrix consisting of d^2 repetitions of the $\mathbf{U}$ matrix stacked on top of each other.

iii) Multiply $\mathbf{C}$ from (i) pointwise with $\mathbf{U}^{(rep)}$ from (ii) and square the row sums of the resulting matrix. Row-wise fill the resulting vector into a $d^2\times(n-p)$ matrix and compute the row sums of this matrix.

iv) Row-wise fill the result from (iii) into a $d\times d$ matrix. This yields $\left(\hat{\sigma}^2_{i,j}\right)_{1\leqslant i,j\leqslant d}$.

3.3 Estimation of the Hawkes graph

Given an estimate $\widehat{\mathcal{G}}^*_{\mathbf{N}}$ of the Hawkes skeleton $\mathcal{G}^*_{\mathbf{N}}$ from Definition 13, we consider the estimation of the Hawkes graph $\mathcal{G}_{\mathbf{N}}$; see Definition 8. We aim to estimate vertex as well as edge weights, and to calculate corresponding confidence bounds for both. That is, after the more structural Hawkes-skeleton estimation from Section 3.2, we now *quantify* the various interactions between the observed event streams. Typically, after the skeleton estimation, we can reduce the effective dimensionality of the model considerably: in a first obvious step, we divide the skeleton $\widehat{\mathcal{G}}^*_{\mathbf{N}}$ into its weakly-connected subgraphs and treat them separately. In a second step, we identify $\widehat{\mathrm{PA}}_{\mathbf{N}}(j) := \{i \in \mathcal{V}_{\mathbf{N}} : (i,j) \in \widehat{\mathcal{E}}^*_{\mathbf{N}}\}$ for all $j \in \mathcal{V}_{\mathbf{N}}$. From the branching construction of a Hawkes process, respectively, of Hawkes families in Definitions 2 and 5, we have that any event in component j is either an immigrant stemming from a Poisson random measure with constant intensity η_j or has a direct explanation through an event in one of its parent components $\mathrm{PA}_{\mathbf{N}}(j)$. That is, in a multivariate version of (3.1), *it suffices to regress the bin-counts of component j on the bin-counts in* $\mathrm{PA}_{\mathbf{N}}(j)$. The constant term in this regression will represent the j-th immigration intensity. Considering only the parents instead of all of the d other components in the conditional-least-squares regression increases numerical efficiency and decreases estimation variance. In applications, however, we *do not know* the true parent set $\mathrm{PA}_{\mathbf{N}}(j)$. So, we have to substitute $\mathrm{PA}_{\mathbf{N}}$ with the estimate $\widehat{\mathrm{PA}}_{\mathbf{N}}$. As long as $\mathrm{PA}_{\mathbf{N}}(j) \subset \widehat{\mathrm{PA}}_{\mathbf{N}}(j)$ this is not an issue: from the branching construction, we have that the intensity at time t of component j, conditional on $\sigma(N_i(A) : A \in \mathcal{B}((-\infty,t]),\, i \in \mathrm{PA}_{\mathbf{N}}(j))$, is independent of the past of all other components

$\sigma(N_i(A) : A \in \mathcal{B}((-\infty, t]), i \notin \mathrm{PA}_{\mathbf{N}}(j))$. Consequently, additional vertices in the estimated parent sets do not introduce additional bias in this graph estimation. Apart from this restriction of the regression variables on (estimated) parent types, we apply the conditional-least-squares approach as in Definition 12. This time however, due to reduction of dimensionality, we will typically *be able to choose a much smaller bin size* Δ_{graph} than for the skeleton estimation before. To ease implementation, below we give convenient notations and the necessary calculations.

First, we drop the $\mathbf{N}$ subscript for the parent sets $\mathrm{PA}(j)$. Also, we write $\mathrm{PA}(j)$ instead of $\widehat{\mathrm{PA}}(j)$—keeping in mind that the first has to be substituted by the latter in most applications. For $k = 1, 2, \ldots, n$, $j \in [d]$ and some $0 < \Delta_{\mathrm{graph}} << \Delta_{\mathrm{skel}}$, let $\mathbf{X}_{k,j}^{(\Delta_{\mathrm{graph}})} := N_j\big(((k-1)\Delta_{\mathrm{graph}}, k\Delta_{\mathrm{graph}}]\big)$, $d_j := |\,\mathrm{PA}(j)|$, and

$$\mathbf{X}_{k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})} := \left(\mathbf{X}_{k,i_1}^{(\Delta_{\mathrm{graph}})}, \mathbf{X}_{k,i_2}^{(\Delta_{\mathrm{graph}})}, \ldots, \mathbf{X}_{k,i_{d_j}}^{(\Delta_{\mathrm{graph}})}\right)^\top. \tag{3.11}$$

In (3.11) and in what follows, we denote $\{i_1, i_2, \ldots, i_{d_j}\} := \mathrm{PA}(j)$ such that $i_1 < i_2 < \cdots < i_{d_j}$. The idea is to regress all the bin counts of all d event types separately on the past of their parents with Ansatz

$$\mathbb{E}\left[\mathbf{X}_{n,j}^{(\Delta_{\mathrm{graph}})} \middle| \mathbf{X}_{n-k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})},\, k = 1, 2, \ldots, p\right] = \alpha_{0,j}^{(\Delta_{\mathrm{graph}})} + \sum_{i \in \mathrm{PA}(j)} \sum_{k=1}^{p} \alpha_{k,i,j}^{(\Delta_{\mathrm{graph}})} \mathbf{X}_{n-k,i}^{(\Delta_{\mathrm{graph}})}, \quad j \in [d]. \tag{3.12}$$

Ansatz (3.12) should be compared with (3.1). Note that j itself may or may not be an element of $\mathrm{PA}(j)$.

Definition 15. *Let $\mathcal{G}_{\mathbf{N}}^*$ be a Hawkes skeleton (estimate) with respect to some d-type Hawkes process (data) $\mathbf{N}$. Given $d_j := |\,\mathrm{PA}(j)|$, $j \in [d]$, a bin size $\Delta_{graph} > 0$, a support s with $0 < \Delta_{graph} \leqslant s < T$, and $p := \lceil s/\Delta_{graph} \rceil$, calculate the conditional-least-squares estimates*

$$\widehat{\mathbf{H}}_j^{(\Delta_{graph}, s)} := \frac{1}{\Delta_{graph}} \left(\mathbf{Z}_j^\top \mathbf{Z}_j\right)^{-1} \mathbf{Z}_j^\top \mathbf{Y}_j \in \mathbb{R}^{(pd_j+1)\times 1}, \quad j \in [d_j], \tag{3.13}$$

with design matrices

$$\mathbf{Z}_j := \begin{pmatrix} (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{p-1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \\ (\mathbf{X}_{p+1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{2,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (\mathbf{X}_{n-1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{n-2,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{n-p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \end{pmatrix} \in \mathbb{N}_0^{(n-p)\times(pd_j+1)}, \quad j \in [d], \tag{3.14}$$

and vectors of responses

$$\mathbf{Y}_j := \left(\mathbf{X}_{p+1,j}^{(\Delta_{graph})}, \mathbf{X}_{p+2,j}^{(\Delta_{graph})}, \ldots, \mathbf{X}_{n,j}^{(\Delta_{graph})}\right)^\top \in \mathbb{N}_0^{(n-p)\times 1}, \quad j \in [d].$$

Given $\widehat{\mathbf{H}}_j^{(\Delta_{graph},s)}$, $j \in [d]$, *we define the* Hawkes-graph estimator $\widehat{G}_{\mathbf{N}} := (\widehat{\mathcal{V}}_{\mathbf{N}}, \widehat{\mathcal{E}}_{\mathbf{N}})$ *with* $\widehat{\mathcal{V}}_{\mathbf{N}} := \{(j; \hat{\eta}_j) : \ j \in [d]\}$ *and*

$$\widehat{\mathcal{E}}_{\mathbf{N}} := \bigcup_{j=1,\ldots,d} \left\{(i_l, j; \hat{a}_{i_l,j}) : \ \{i_1, \ldots, i_{d_j}\} = \mathrm{PA}(j), \ \hat{a}_{i_l,j} = b_{l,j}^\top \widehat{\mathbf{H}}_j^{(\Delta_{graph},s)}\right\}, \tag{3.15}$$

where, for $l \in [d_j]$, $b(l,j) \in \{0,1\}^{(d_j p+1)\times 1}$ *is a column vector with* 0*s in all components but 1s in components* $((k-1)d_j + l)$, $k = 1, 2, \ldots, p$. *Furthermore, for* $\alpha_{graph} \in (0,1)$, *we define the confidence intervals* $\left[\hat{\eta}_j \pm \hat{\sigma}_j z_{1-\alpha_{graph}}^{-1}\right)$, $j \in [d]$, *and, for* $i_l \in \mathrm{PA}_{\mathbf{N}}(j)$, $\left[\hat{a}_{i_l,j} \pm \hat{\sigma}_{i_l,j} z^{-1}(1-\alpha_{graph})\right)$. *We give the calculation of* $\hat{\sigma}_{i_l,j}$ *and* $\hat{\sigma}_j$ *in Algorithm 16, below.*

As before, additional notation clarifies what the entries of the matrices $\widehat{\mathbf{H}}_j^{(\Delta_{\mathrm{graph}},s)}$, $j \in [d]$, actually estimate:

$$\begin{pmatrix} \widehat{H}_{\mathrm{PA}(j),j}(\Delta_{\mathrm{graph}}) \\ \widehat{H}_{\mathrm{PA}(j),j}(2\Delta_{\mathrm{graph}}) \\ \ldots \\ \widehat{H}_{\mathrm{PA}(j),j}(p\Delta_{\mathrm{graph}}) \\ \hat{\eta}_j \end{pmatrix} := \widehat{\mathbf{H}}_j, \text{ with} \tag{3.16}$$

$$\widehat{H}_{\mathrm{PA}(j),j}(k\Delta_{\mathrm{graph}}) = \left(\hat{h}_{i_1,j}(k\Delta_{\mathrm{graph}}), \hat{h}_{i_2,j}(k\Delta_{\mathrm{graph}}), \ldots, \hat{h}_{i_{d_j},j}(k\Delta_{\mathrm{graph}})\right)^\top,$$

$k = 1, 2, \ldots, p$ and $\{i_1, i_2, \ldots, i_{d_j}\} = \mathrm{PA}(j)$. Finally, we provide efficient computations for the covariance estimates that are necessary for the confidence intervals around the estimated edge and vertex weights.

Algorithm 16. *Let* $j \in [d]$ *such that* $|\mathrm{PA}(j)| > 0$ *and let* $\{i_1, i_2, \ldots, i_{d_j}\} = \mathrm{PA}(j)$ *with* $i_1 < i_2 < \ldots, < i_{d_j}$. *For* (i_l, j), $l \in [d_j]$, *let* $e(i_l, j) \in \{0,1\}^{(d_j p+1)\times 1}$ *be a column vector with all entries 0, but 1s at components* $(k-1)d_j + (l-1)$, $k = 1, 2, \ldots, p$. *We compute* $\hat{\sigma}_{i_l,j}$ *in the following way:*

i) Compute $\mathbf{C}_{l,j} := e(i_l, j)^\top \left((\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j^\top\right) \in \mathbb{R}^{1\times(n-p)}$.

ii) Set $\mathbf{U}_j = (\mathbf{Y}_j - \Delta_{graph} \mathbf{Z}_j \widehat{\mathbf{H}}_j) \in \mathbb{R}^{(n-p)\times 1}$. *Denoting* $(U_{p+1,j}, U_{p+2,j}, \ldots, U_{n,j})^\top := \mathbf{U}_j$, *we have that*

$$U_{k,j} = \left(\mathbf{X}_{k,j}^{(\Delta_{graph})} - \Delta_{graph}\hat{\eta} - \sum_{m=1}^{p} \Delta_{graph} \widehat{H}_{\mathrm{PA}(j),j}^\top(m\Delta_{graph}) \, \mathbf{X}_{k-m,\mathrm{PA}(j)}^{(\Delta_{graph})}\right),$$

for $k = p+1, p+2, \ldots, n$.

iii) Pointwise multiply $\mathbf{C}_{l,j}$ *and* $\mathbf{U}_j$*. The sum of the squares of the result yields* $\hat{\sigma}^2_{i_l,j} \in \mathbb{R}_{\geqslant 0}$*.*

For the variance estimates corresponding to the j-th vertex weight, consider the last row of $\left((\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j\right) \in \mathbb{R}^{(d_j p+1)\times(n-p)}$*, multiply it pointwise with* $\mathbf{U}_j$ *from above, take the sum of squares of the results and multiply the result with* Δ^{-2}_{graph}*; this yields* $\hat{\sigma}^2_j$*.*

Remark 17. The bin size Δ_{graph} for the graph estimation in Definition 15 will typically be much smaller than the bin size Δ_{skel} for the skeleton estimation in Definition 13. After the graph estimation, one might again want to delete edges with edge-weight estimates non-significantly different from zero, or treat vertex-weight estimates, respectively, immigration intensities, that are not significantly different from zero as zero; see Figure 2. Also note that the latter could possibly be tested with a different significance parameter α_{vertex} than the significance parameter α_{graph} from the edge weight estimation. In any case, the resulting Hawkes-graph estimations ought to be checked for *redundant vertices*; see Definition 7. If the estimate has redundant vertices, the results are typically inconsistent with the data—as we typically observe data in all components. Therefore, if a fitted model has redundant vertices, we ought to increase α_{skel}, α_{graph}, and/or α_{vertex}. Thus, we obtain more estimated nonzero immigration intensities and/or larger estimated edge sets. We proceed with increasing the significance parameters until there are no redundancies left.

Given a Hawkes-graph estimate as in Definition 15, one may examine connectivity issues, path weights, graph distances, feedback and cascade coefficients, exploit graphical representations, etc.; see the example in Section 4.

3.4 Estimation of the reproduction intensities

For many applications, the results discussed above may already suffice. In other applications however, the graph estimation will only be a preliminary step and one would like to examine how the various excitements are distributed *over time*. In other words, one would like to explicitly estimate the displacement intensities or the reproduction intensities from Definition 2.

Parametric estimation

Given the Hawkes estimator from Definition 12, the Hawkes model is not yet completely specified. In particular, (3.13) only yields estimates of the reproduction intensities on a grid:

$$\left\{\left(k\Delta\right), \hat{h}_{i,j}(k\Delta)\right\}_{k=1,2,\ldots,p}, \quad i \in \widehat{\mathrm{PA}}(j),\ j \in [d]. \tag{3.17}$$

One obvious possibility to complete the model specification would be the application of any kind of smoothing method on (3.17). We want to consider another approach: we exploit (3.17)

graphically (examine log/log-plots, id/log-plots, check for local maxima, convex/concave regions, etc.) and identify appropriate parametric families. The parameters can then be fitted to the estimates (3.17) via non-linear least-squares (e.g., function `nls` in `R`):

Definition 18. *Consider a Hawkes-graph estimation as in Definition 15 with respect to some d-type event-stream data and a bin size $\Delta_{graph} > 0$. For $j \in [d]$ and $i \in \widehat{\mathrm{PA}}(j)$, let $w_{i,j}^{(\theta_{i,j})} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, $w_{i,j}^{(\theta_{i,j})}(t) = 0$, $t \leqslant 0$, be density families parametrized by $\theta_{i,j} \in \Theta_{i,j} \subset \mathbb{R}^{d_{i,j}}$. With the notation from* (3.17)*, let*

$$(\hat{a}_{i,j}, \hat{\theta}_{i,j}) := \mathrm{argmin}_{(a,\theta)\in\mathbb{R}_{\geqslant 0}\times\Theta_{i,j}} \sum_{k=1}^{p} \left(a w_{i,j}^{(\theta)}\left(k\Delta_{graph}\right) - \hat{h}_{i,j}(k\Delta_{graph}) \right)^2, \quad (i,j) \in \widehat{\mathcal{E}}^*, \tag{3.18}$$

and define the parametric reproduction-intensity estimates

$$\hat{h}_{i,j}^{(par)}(t) := \begin{cases} \hat{a}_{i,j} w_{i,j}^{(\hat{\theta}_{i,j})}(t), & (i,j) \in \widehat{\mathcal{E}}^*, \quad t \in \mathbb{R}, \\ 0, & (i,j) \notin \widehat{\mathcal{E}}^*, \quad t \in \mathbb{R}, \end{cases}$$

the parametric branching-matrix estimate

$$\widehat{A}^{(par)} := \left(\int \hat{h}_{i,j}^{(par)}(t) \mathrm{d}t \right)_{1\leqslant i,j\leqslant d},$$

and the parametric immigration-intensity estimates.

$$\hat{\eta}^{(par)} := \left(\hat{\eta}_1^{(par)}, \ldots, \hat{\eta}_d^{(par)} \right) := \lambda^{(emp)} \left(1_{d\times d} - \widehat{A}^{(par)} \right), \tag{3.19}$$

where $\lambda^{(emp)}$ denotes the observed empirical intensity $\lambda^{(emp)} := \mathbf{N}\left((0,T]\right)/T \in \mathbb{R}_{\geqslant 0}^{1\times d}$.

We illustrate this specification and estimation of a fully parametric multivariate Hawkes process in Figure 3. Here, we also see that the parameter estimates from (3.18) are symmetrically distributed around the true values. Even though the estimator calculations in Definition 18 stand at the end of a long chain of various discretizations and truncations, 'log-likelihood profile' confidence intervals (e.g., from `confint.nls` in `R`) give remarkably good coverage rates for the parameter estimates (not illustrated).

Remark 19. The definition of $\eta^{(\mathrm{par})}$ in (3.19) is motivated by the desirable equality

$$\eta^{(\mathrm{par})} \left(1_{d\times d} - (\widehat{A}^{(\mathrm{par})})^\top \right)^{-1} = \lambda^{(\mathrm{emp})}.$$

In other words, with this choice of $\hat{\eta}^{(\mathrm{par})}$, the observed unconditional intensity exactly equals the estimated unconditional intensity. This might be relevant in some applications (e.g., simulation

from a fitted model). Finally note that it might often be more efficient to consider weighted least squares in (3.18).

4 Example

We illustrate the concepts introduced in the previous sections with a ten-dimensional Hawkes model. We perform a simulation study and apply the estimation methods from Sections 3.2, 3.3, and 3.4 to the Hawkes skeleton, the Hawkes graph, and the reproduction-intensity parameters.

4.1 Example model

We consider a 10-type Hawkes process $\mathbf{N}$ as in Definition 5 with immigration intensities

$$\eta_i := \begin{cases} 1, & i \in \{1,7,10\}, \\ 0, & i \in \{2,3,4,5,6,8,9\}, \end{cases} \tag{4.1}$$

and reproduction intensities $h_{i,j}$, $(i,j) \in [10]^2$, defined, for $t \in \mathbb{R}$, by

$$h_{i,j}(t) := \begin{cases} 1.5\,\gamma(t), & (i,j) \in \{(1,2),(2,4),(8,9)\}, \\ 1_{t\in[1,2]}0.5, & (i,j) \in \{(1,1),(2,3),(3,5),(4,3),(4,5),(4,6),(5,3),(7,8),(9,7)\}, \\ 1_{t\in[1,2]}0.1, & (i,j) = (5,7), \\ 0, & \text{else.} \end{cases} \tag{4.2}$$

Here, γ denotes a Gamma density with shape parameter 6 and rate parameter 4, i.e., $\gamma(t) = 1_{t\geqslant 0}t^5 \exp\{-4t\}(4^6)/(5!)$. In Hawkes graph terminology, we have 13 edges supplied with three different kinds of edge weights: a *heavy weight* (1.5) for three edges, a *light weight* (0.5) for seven edges, and one edge with a *super-light weight* (0.1). An illustration of the corresponding graph $\mathcal{G}_{\mathbf{N}}$ is much more meaningful than (4.2); see the left graph in Figure 1. From this figure, the various direct and indirect dependencies can be read off instantaniously; only the large nodes have nonzero immigration intensity; a fat edge corresponds to an edge weight of 1.5; a thin edge corresponds to an edge weight of 0.5; the dashed line corresponds to the super-light edge weight 0.1. We examine the Hawkes-graph properties introduced in Definitions 8 and 11:

Redundancy The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ has no *redundant vertices*: all small vertices have a large vertex as one of their ancestors. If vertex 1 were small, the vertices $1,2,3,4,5$ and 6 would be

redundant as they could not generate events.

Connectivity The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is *not weakly connected*. The graph can be divided in two separate weakly-connected Hawkes subgraphs with vertex sets $\{1,2,3,4,5,6,7,8,9\}$, and $\{10\}$. Deleting edge $(5,7;0.1)$ would yield three separate weakly-connected Hawkes subgraphs.

Criticality The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is *subcritical*: all vertices but vertex 10 are part of closed walks. It suffices to check criterion (2.5) for vertices $i_0 \in \{1,2,3,7\}$. For vertex 1, we find that

$$\mathcal{W}_g^{(1,1)} = \{(\underbrace{1,1,\ldots,1}_{g+1 \text{ times}})\},\ g \in \mathbb{N}, \quad \text{and} \quad |(\underbrace{1,1,\ldots,1}_{g+1 \text{ times}})| = 0.5^g,\ g \in \mathbb{N}\,.$$

Consequently, $\sum_{g=1}^{\infty} \sum_{w_g \in \mathcal{W}_g^{(1,1)}} |w_g| = \sum_{g=1}^{\infty} 0.5^g < \infty$. For vertex 2, we find that

$$\mathcal{W}_1^{(2,2)} = \mathcal{W}_2^{(2,2)} = \varnothing,\ \mathcal{W}_3^{(2,2)} = \{(2,4,6,2)\},\ \mathcal{W}_4^{(2,2)} = \mathcal{W}_5^{(2,2)} = \varnothing,\ \mathcal{W}_6^{(2,2)} = \{(2,4,6,2,4,6,2)\},\ldots$$

With $|(2,4,6,2)| = 1.5 \cdot 0.5 \cdot 0.5 = 0.375$, $|(2,4,6,2,4,6,2)| = 0.375^2$, ..., criterion (2.5) again follows. For vertices 3 and 7, one argues analogously. In other words, as long as closed walks do not overlap, we can construct large subcritical Hawkes graphs without calculating any eigenvalues. When closed walks overlap, the underlying combinatorics typically become too involved as to proceed in this manner. In this case one could calculate the spectral radius of the adjacency matrix of the involved edges only. For example, if we wanted to introduce another edge $(9,5;a_{9,5})$ in model (4.2), respectively, Figure 1, we would have to calculate the spectral radius of the adjacency matrix corresponding to the Hawkes (sub-)graph with edges

$$\big\{(3,5;0.5),(5,3;0.5),(5,7;0.1),(7,8;0.5),(8,9;0.5),(9,5;a_{9,5}),(9,7;0.5)\big\};$$

see Theorem 9.

Cascade and feedback coefficients We calculate the coefficients from Definition 11 with respect to the example model; see Table 1. The cascade and feedback coefficients summarize the impact of the driving vertices 1, 4 and 10 (that is, of the vertices with nonzero vertex weights) on the process. The *cascade coefficients* measure the impact of each vertex on the whole system. In our example, the immigrants in the first vertex together with the cascades that they trigger are responsible for about 82% of all events that occur in the system. The *feedback coefficients* measure the impact of each vertex on itself. In our example, for vertex 8 this means that 76% of its activity are explained by its own immigration activity and by the feedback loops that the type-8 immigrants trigger via closed walks. Vertex 1 is only excited by its own activity. For

vertex 10 the feedback coefficient is also equal 1—albeit there is no true feedback involved. Still, its intensity would decrease by 100% if it were switched off.

Table 1: Cascade and feedback coefficients

	1	2	3	4	5	6	7	8	9	10
cascade.coefficients	0.82	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.04
feedback.coefficients	1.00	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.00	1.00

4.2 Simulation study

We simulate $n_{\text{sim}} = 1000$ realizations of the Hawkes process $\mathbf{N}$ from Section 4.1. We use the branching construction from Definitions 2 and 5 as simulation algorithm. In each realization, we simulate a time window of 500 time units. This typically yields between 500 and 2000 events per component. Given each of these realized event streams, we calculate the Hawkes-skeleton estimator from Definition 13—with respect to different values of Δ_{skel} and α_{skel}. Given these skeleton estimates, we calculate the Hawkes-graph estimator from Definition 15—including confidence bounds for all vertex and edge weights. Finally, we analyze the scatterplots for branching-intensity estimates, choose parametric function families, and fit the parameters on the estimates by nonlinear least squares. Figures 1 and 2 illustrate the procedure.

Hawkes-skeleton estimation

We fix $s = 5$ and, for each simulated event-stream, we calculate the Hawkes-skeleton estimates from Definition 13 with respect to this support parameter s, bin sizes $\Delta_{\text{skel}} \in \{0.2, 0.5, 1, 2\}$, and various sparseness parameters $\alpha_{\text{skel}} \in \{0.005, 0.01, 0.05, 0.1, 0.25\}$. We denote the estimated edge sets by $\{\widehat{\mathcal{E}}^*(k)\}_{k=1,2,\ldots,n_{\text{sim}}}$ and the true edge set by $\mathcal{E}^*$. Using this notation, we summarize the results of the simulation study in Tables 2, 3, 4, and 5 with the following statistics:

i) *nedges*: average size of estimated edge-sets (true number is 13), that is, $\sum_{k=1}^{n_{\text{sim}}} |\mathcal{E}^*(k)|/n_{\text{sim}}$.

Table 2: $\Delta_{\text{skel}} = 0.2$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.324	0.902	1.000	0.956	0.121	0.993
0.010	13.066	0.917	1.000	0.970	0.190	0.987
0.050	17.296	0.946	1.000	0.990	0.379	0.942
0.100	21.995	0.959	1.000	0.995	0.507	0.890
0.250	35.015	0.979	1.000	0.999	0.739	0.744

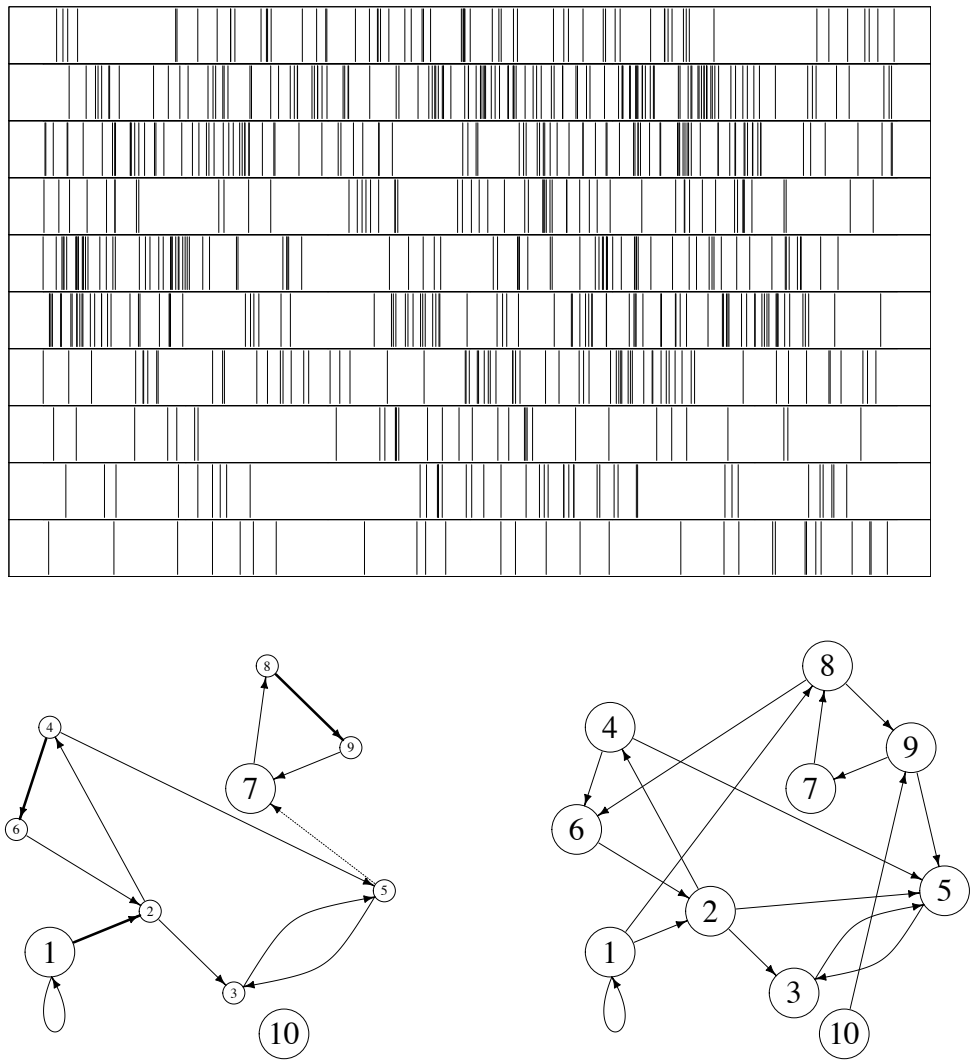


Figure 1: Hawkes process simulation, Hawkes graph, and estimated Hawkes skeleton. The left graph represents the Hawkes graph corresponding to the Hawkes process example from Section 4.1; the graph is a summary of the immigration and branching structure of the model: edges from one vertex to another vertex denote nonzero reproduction intensities, respectively, excitement. Fat edges refer to heavy excitement (1.5 expected children events in branching construction); thin edges to small excitement (0.5 expected children) and the dotted line refers to a very small excitement (0.1 expected children); see (4.2). Large vertices correspond to nonzero immigration-intensities ($= 1$) and small vertices to the zero-immigration vertices; see (4.1). The barcode plots illustrate a 30 time-units window of a simulated realization of the model (after some burn-in): we observe events of ten types, respectively, in ten components. One goal of our paper is to retrieve the graph on the left from such a realization. As a first step towards this aim, we calculate the Hawkes-skeleton estimate from Definition 15 with respect to a coarse bin size $\Delta_{\text{skel}} = 1$ and a sparseness parameter $\alpha_{\text{skel}} = 0.05$. The right graph illustrates such an estimate. This skeleton will be used in a second step to retrieve the Hawkes-graph estimate; see Figure 2. Comparing the skeleton with the true graph on the right, we see that we catch twelve of the thirteen true edges. We miss edge $(5, 7)$. Furthermore, the estimate introduces five additional wrong edges $(1, 8)$, $(2, 5)$,$(8, 6)$, $(9, 5)$, and $(10, 9)$. The three crucial points are: (i) These five false-positive edges *do not introduce additional bias* in the graph estimation. (ii) Due to the coarse Δ_{skel}-value, the calculation of the skeleton estimate is computationally simple. (iii) The resulting skeleton estimate is nearly as sparse as the true skeleton. This considerably reduces the complexity of the graph estimation (with a very fine Δ_{graph}-parameter). See Figure 2, for the Hawkes-graph estimation with respect to the skeleton estimate from above.

Table 3: $\Delta_{\text{skel}} = 0.5$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.353	0.902	1.000	0.957	0.120	0.993
0.010	13.118	0.917	1.000	0.971	0.179	0.986
0.050	17.255	0.945	1.000	0.990	0.375	0.943
0.100	21.952	0.959	1.000	0.995	0.514	0.891
0.250	34.805	0.980	1.000	0.999	0.745	0.746

Table 4: $\Delta_{\text{skel}} = 1$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.476	0.910	1.000	0.967	0.129	0.993
0.010	13.171	0.921	1.000	0.977	0.178	0.986
0.050	17.264	0.949	1.000	0.993	0.400	0.943
0.100	21.806	0.962	1.000	0.997	0.535	0.893
0.250	34.465	0.979	1.000	0.999	0.730	0.750

Table 5: $\Delta_{\text{skel}} = 2$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.244	0.810	1.000	0.828	0.074	0.980
0.010	13.680	0.846	1.000	0.876	0.119	0.969
0.050	19.709	0.913	1.000	0.957	0.262	0.910
0.100	25.065	0.936	1.000	0.978	0.369	0.852
0.250	38.186	0.966	1.000	0.994	0.605	0.705

ii) *total*: fraction of correctly included edges, i.e, of pairs $(i,j) \in \widehat{\mathcal{E}}^*_{\mathbf{N}}(k)$ such that $(i,j) \in \mathcal{E}_{\mathbf{N}}$:

$$\frac{\sum_{k=1}^{n_{\text{sim}}} \sum_{(i,j)\in\mathcal{E}^*} 1_{\{(i,j)\in\widehat{\mathcal{E}}^*(k)\}}}{n_{\text{sim}}|\mathcal{E}^*|}.$$

Note that $1 - \textit{total}$ is the *false-negative rate*.

iii) *heavy/light/super.light*: more detailed version of ii) above; fractions of correctly estimated edges with heavy (1.5), light (0.5) and super-light (0.1) edge weights.

iv) *zero*: fraction of correctly excluded edges, i.e., of pairs $(i,j) \notin \widehat{\mathcal{E}}^*_{\mathbf{N}}(k)$ such that $(i,j) \notin \mathcal{E}_{\mathbf{N}}$:

$$\frac{\sum_{k=1}^{n_{\text{sim}}} \sum_{(i,j)\notin\mathcal{E}^*} 1_{\{(i,j)\notin\widehat{\mathcal{E}}^*(k)\}}}{n_{\text{sim}}\big(d^2 - |\mathcal{E}^*|\big)}.$$

Note that $1 - \textit{zero}$ is the *false-positive rate*.

First, we discuss the estimations with respect to bin size $\Delta_{\text{skel}} = 0.2$; see Table 2. We note from the last column, *zero*, that the false-positive rate is indeed very close to the value of the chosen theoretical significance level α_{skel}. Going back to Definition 13, we see that the larger α_{skel}, the more edges are included in the Hawkes-skeleton estimation. This is reflected in all of the columns. However, even for very small α_{skel}, we detect *all* of the edges with a heavy edge weight and most of the edges with light edge weight. The edge $(5,7)$ with the super-light weight (0.1) is obviously a hard-to-detect alternative to the zero hypothesis. Note that Tables 3, 4, and 5 look roughly the same as Table 2 one above—though the estimates were calculated with respect to completely different bin sizes Δ_{skel}. So, in this first estimation step, we may use a very coarse bin size Δ_{skel}. This makes the calculations underlying the skeleton estimation feasible even for much higher dimensions.

The main purpose of the skeleton estimation is to lay the ground for the graph estimation which itself depends on a given estimated skeleton; see Definition 15. Missing edges in the skeleton estimate will typically introduce a bias for the graph-weight estimates. We therefore want to keep the false-negative rate ($= 1 - \textit{total}$) in the skeleton estimation very small. As a consequence, we need α_{skel} large to include more edges. Note that false-positive edges do *not* add additional bias in the graph estimation; see Section 3.3. So the increase of the false-positive rate (that is, the decrease in the *zero*-column) does not prevent us from increasing the α_{skel}-parameter. Note, however, that the whole reason for the two-step estimation procedure is that in the first step we want to take advantage of the sparseness of the underlying true Hawkes graph and *reduce* the complexity of the a priori fully connected network. Too many additional false-positive edges would hamper this advantage. In this sense, not only Δ_{skel} but also α_{skel} can be understood as a parameter controlling the numerical complexity of the method: the smaller

Table 6: $\Delta_{\text{graph}} = 0.1$ and $\alpha_{\text{graph}} = 0.05$

applied.skeleton	vertex.weight.coverage	edge.weight.coverage
alpha.skel = 0.005	0.859	0.907
alpha.skel = 0.01	0.867	0.904
alpha.skel = 0.05	0.896	0.893
alpha.skel = 0.1	0.907	0.900
alpha.skel = 0.25	0.915	0.932
true skeleton	0.947	0.943

α_{skel}, the sparser the estimated skeleton, the less complex the computations for the Hawkes-graph estimate from Definition 15. We see in our tables that, for all choices of Δ_{skel} and all values of α_{skel}, we typically catch all the true edges, i.e., the false-negative rate is really small. In the next section, we will see that the graph estimates are not dramatically sensitive to the α_{skel} parameter in the skeleton estimation.

Hawkes-graph estimation

In a further step, we quantify the estimated excitements. That is, given a Hawkes skeleton, we estimate the corresponding graph as in Definition 15; see Figure 2. We do this both with respect to the true skeleton and with respect to the estimated skeletons from the first estimation step. For comparison, we apply skeletons that were estimated with respect to different α_{skel}-parameters. However, we only consider the skeletons that were estimated with respect to the (rough) bin size $\Delta_{\text{skel}} = 1$. As opposed to the skeleton estimation, we may now use a much smaller bin size $\Delta_{\text{graph}} = 0.1$ for the graph estimation. In the present example, this is approximately the lower bin-size bound for tolerable computing time for the simulation study using a 2.3 GHz Intel Core processor (about 10sec for each of the estimations, no parallelization). Furthermore, we apply $s = 5$ and $\alpha_{\text{graph}} = 0.05$ in the calculation. For each simulation, we also calculate the confidence bounds for all vertex and edge weights from Definition 15. Table 6 reports the coverage rates.

The coverage rates of the graph estimations that were calculated with respect to the true underlying skeleton correspond well with the significance parameter $\alpha_{\text{graph}} = 0.05$. Naturally, the coverage rates for the estimates with respect to the estimated skeleton are smaller: as soon as the estimated skeleton misses an edge (e.g., the super-light edge $(5, 7; 0.1)$), the model calibration balances this missing possibility of excitement by increased baseline intensities or increased edge weights. The larger α_{skel}, the lower the probabilty of missing an edge, the better the coverage rates. Note, however, that at the same time, the corresponding skeleton estimate becomes increasingly dense and with it the graph estimation becomes increasingly time-consuming.

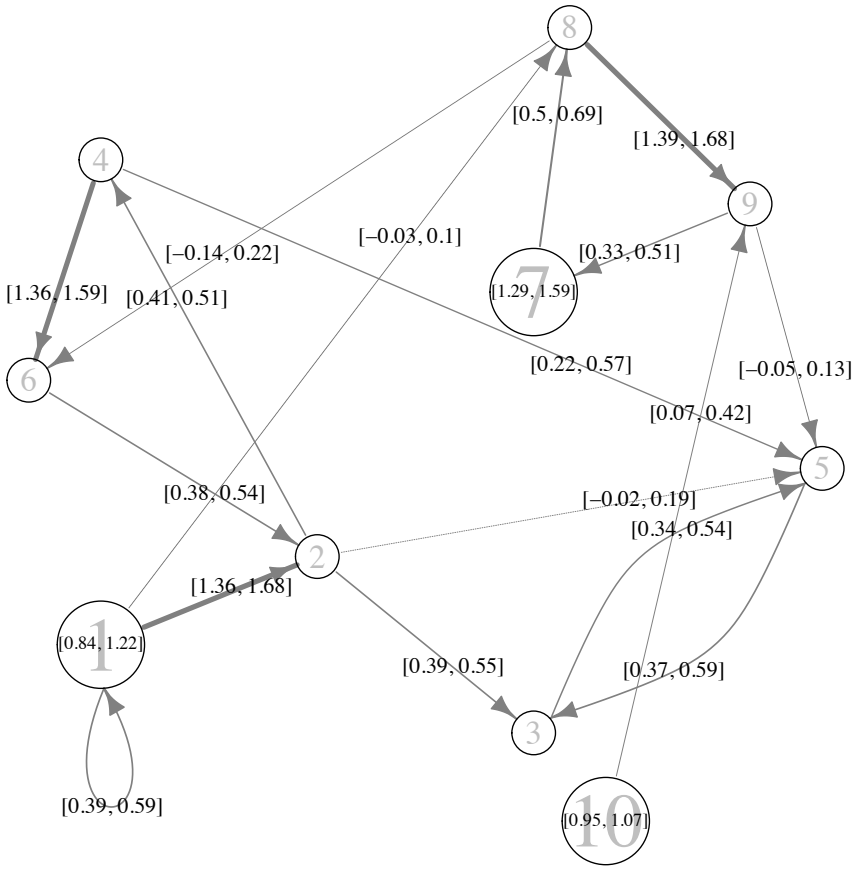


Figure 2: Hawkes-graph estimation. Given a single simulation of length $T = 1000$ from the example Hawkes process in Section 4.1, we calculate the Hawkes-graph estimator from Definition 15 with respect to the Hawkes-skeleton estimation from Figure 1; we apply a bin size $\Delta_{\text{graph}} = 0.025$ and a significance parameter $\alpha_{\text{graph}} = 0.05$. This calculation allows us to supply each vertex and each node from this estimated skeleton with confidence intervals for their weights in the corresponding Hawkes graph. The edge widths in the illustration are chosen proportional to the estimated edge weights. Estimated edge weights that are not significantly larger than zero are illustrated as a dashed edge. Similarly, vertices where the confidence interval for the vertex weight contains 0 are plotted as smaller circles—the corresponding confidence bounds are left away in this latter case. Comparing the results with the true Hawkes graph in Figure 1, respectively, with the Hawkes process parametrization in (4.1) and (4.2), we see that for all correct edges, the true weights are covered by the confidence intervals. And for the wrong, additional edges from the skeleton estimation $(1, 8)$, $(2, 5)$,$(8, 6)$, and $(9, 5)$, we see that their weights are not significantly different from zero ($\alpha_{\text{graph}} = 0.05$). The estimated edge weight for the wrong edge $(10, 9)$ is significantly larger than zero but still small. All true vertex weights but the weight of vertex 7 are also covered by the confidence intervals. The weight of vertex 7 is overestimated because we missed the (light) edge $(5, 7; 0.1)$ in the skeleton estimation; this missing explanatory variable for the events in component 7 is compensated by an extra large vertex weight in the graph estimation. Deleting all insignificant (in figure dashed) edges and setting the vertex weight of the insignificant (in figure small) vertex-weights to zero, we recover the original underlying graph almost perfectly.

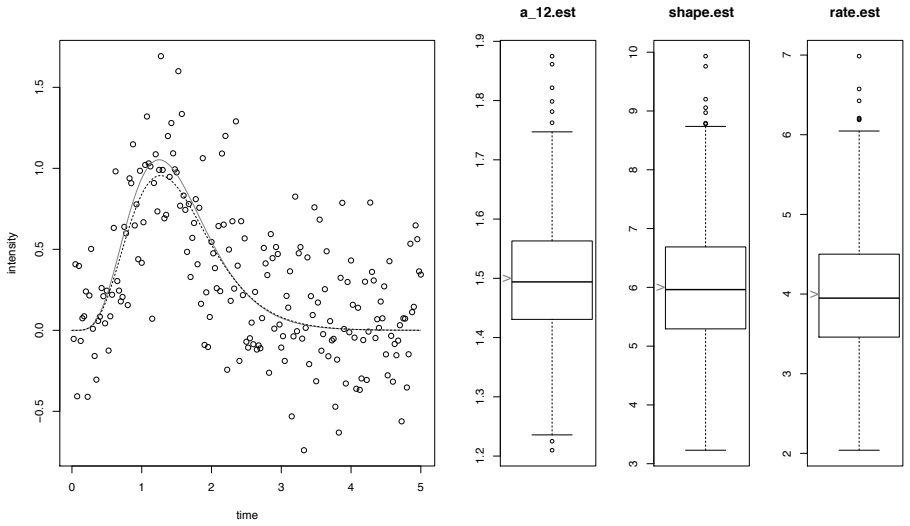


Figure 3: Parametric estimation. *Left:* From a single realization of the example Hawkes model from Section 4.1 with length $T = 1000$, we calculate Hawkes-skeleton and Hawkes-graph estimates from Definitions 13 and 15; see Figures 1 and 2. As a by-product of these calculations, we retrieve pointwise estimates (circles) for the values of the reproduction intensity $h_{1,2}$ on an equidistant grid; see (3.16). From these estimates, one may guess that $h_{1,2}(t) = a_{1,2}\gamma(t)$, where γ is a Gamma density depending on a shape and on a rate parameter. We fit the three parameters by nonlinear least squares as described in Section 3.4. The dotted black line refers to the corresponding estimated parametric function. It catches the true underlying function (grey solid line) quite well; see (4.2). *Right:* We apply this parametric estimation of $h_{1,2}$ on 1000 independent realizations of length $T = 500$. The boxplots collect the parameter estimates for each of the 1000 estimations of the simulation study. The grey marks refer to the corresponding true values. Eyeball-examination shows that the estimates are remarkably symmetric distributed and unbiased. QQ-plots (not illustrated) support asymptotic normality.

Parametric reproduction intensity estimation

Finally, we check how the various excitements are distributed over time. As examples, we examine the reproduction intensity $h_{1,2}$. From the calculation of the Hawkes-graph estimate, we retrieve estimates of the reproduction intensity values on an equidistant grid; see (3.16). Based on the scatter plots of these estimates, we choose appropriate parametrized function families. Given such parametric functions, the parameters are fit to the pointwise estimates via nonlinear least squares; see Figure 3. QQ-plots (not included) support asymptotic normality for the parameter estimates.

5 Conclusion

The Hawkes graph and the Hawkes skeleton describe the immigration and branching structure of a Hawkes process in a graph-theoretical framework. We demonstrate how graph terminology can be very useful for multivariate Hawkes processes. Combining the new concepts with an estimation procedure from earlier work, we develop a statistical estimation method for the

Hawkes skeleton and the Hawkes graph. The key idea is that in a preliminary step we only test if there is *at all* excitement from any vertex to another vertex. We show that this first step is relatively simple to implement. The knowledge of the Hawkes skeleton makes the second step, the estimation of the Hawkes graph, much more efficient—both from a computational and statistical point of view. The simulation study shows that the procedure works as desired. As long as the true underlying graph is sparse (e.g., if the typical number of parents of a node is not larger than 5 and does not depend on the dimension of the process) the approach may be applied in even higher-dimensional situations. In any case, the method may be a useful tool for preliminary analysis when examining large multi-type event-stream data in the Hawkes framework.

It might be worthwile to study the distributional properties of the parameter estimates from Section 3.4 in more detail. Also note that the graph representation would also apply for discrete-time event-stream models, i.e., for multivariate time series of counts. More specifically, the present paper could have been developed in complete analogy for multivariate integer-valued autoregressive time series (INAR(∞)) which can be interpreted as discrete-time versions of the Hawkes process; see Kirchner (2016). In this latter case, all results that we apply in our paper would be valid without taking any discretization error into account. In any case, when applied to real data, the discretization error is *not* the major drawback of our method: our method does indeed solve the important problem of how to decide whether an edge between two components exists at all. But for the specification of a Hawkes process we need to solve another—more important—issue. We want to be able to decide whether we observe a *complete* Hawkes graph or whether our data lack some non-redundant vertices! In particular, the method presented will also yield reasonable results for data stemming from models with no or less underlying 'causality'. The seeming excitement can then be explained by a confounding factor that we do not observe (and ignore). We believe, in view of the widespread interpretation of the Hawkes model as a causal model (an interpretation we share), it would be of utmost importance to derive tests for the presence of such hidden confounding factors in the event-stream context.

Acknowledgements

This research has been supported by the ETH RiskLab and the Swiss Finance Institute. The authors wish to express their gratitude to all the R-programmers providing and maintaining powerful statistical software. For our work, the igraph package (Csardi and Nepusz, 2006) and the Matrix package (Bates and Maechler, 2015) have been particularly helpful. We thank Vladimir Ulyanov for his comments on an earlier version of the paper which helped to improve the presentation. We also thank Philippe Deprez for a fertile discussion about Theorem 9.

Paper

E

Matthias Kirchner, Silvan Vetter.

Hawkes model specification for limit order books.

Submitted.

Hawkes model specification for limit order books

M. Kirchner and S. Vetter

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

This paper discusses Hawkes modeling of order arrivals in limit order books. We model the flow of market orders, limit orders, and cancelations by a self- and crossexciting multivariate marked Hawkes process with state-dependent baseline intensities. The marks carry the order sizes and the state of the book is summarized by the 'limit-order-book imbalance'. We specify the model very carefully—with few a priori assumptions: we select the non-zero excitements (the 'Hawkes skeleton'), the shape of the decay kernels, and the shape of the impact functions in a nonparametric manner. Furthermore, we show that our data exhibit perfect bid–ask symmetry. We observe that the imbalance of the order book explains the probability for a bid (ask) market order—given the occurrence of a market order—in a perfectly linear manner. Thus, we include a term involving the imbalance in the baseline intensity of the process. We calibrate the specified parametric model by maximum likelihood estimation and discuss the results. Finally, we apply the fitted model in order to estimate the conditional distribution of the next order type. This opens the door to order-type prediction.

1 Introduction

A limit order book (LOB) is a double auction system that organizes the market at the important stock exchanges; see Gould et al. (2013). We consider two NASDAQ datasets of Intel and Microsoft stocks covering three months. We model the arrivals of market orders, limit orders, and cancelations at best bid price and best ask price with marked Hawkes point processes. More specifically, in our model, each LOB event potentially excites new orders and cancelations. The strength of these excitements is governed by *branching coefficients*, the persistence of the excitements over time is governed by *decay kernels*, and the influence of the order sizes on the

excitements is governed by *impact functions*. We summarize the state of the book by the *order-book imbalance* and include it in our model by letting the *baseline intensities* of the Hawkes process depend on it. This makes our approach a mélange of a purely state-dependent model (as, e.g., in Cont and de Larrard (2013)) and a purely past-dependent model (as, e.g., in Bacry et al. (2014)). We pay special attention to the specification of a parametric model. Using methodology from Kirchner (2017a), Embrechts and Kirchner (2017), and Bacry et al. (2015a), we select the set of non-zero excitements ('Hawkes skeleton'), the shape of the decay kernels, as well as the shape of the impact functions nonparametrically. In addition, we conclude from the nonparametric estimates that we may assume bid–ask symmetry for all model parameters. Finally, we fit the parametric model to the data. The results are in line with descriptive analysis and nearly equivalent for the two considered datasets. We explain how the model can be used for order-type prediction. We find that this procedure predicts the arrival of certain orders rather well.

1.1 Main contributions

The main aim of this paper is to show how to carefully select a parametric multivariate marked Hawkes model—without too many a priori assumptions. The model specification involves identification of ...

1. ... zero-excitements: we apply the method from Embrechts and Kirchner (2017) to retrieve the 'Hawkes skeleton' and the 'Hawkes graph' from our data. We find that the market orders are the driving processes in the order book: they are hardly affected by other LOB events; see Figure 4.
2. ... decay kernels: with the nonparametric estimation method from Kirchner (2017a), we find that the non-zero excitements exhibit an overall power-law behavior.
3. ... impact functions: following an idea from Bacry et al. (2014), we extend our method from Kirchner (2017a) to the marked case; see Section 2.5. It turns out that linear impact functions are a good choice. That is, a market order of 1000 lots excites the incoming of limit orders ten times more than a market order of 100 lots.
4. ... bid–ask symmetry: descriptive analysis as well as all (non-parametric) estimation results exhibit clear symmetry. This halves the number of parameters, and consequently also standard deviation.
5. ... state dependence: this is a result of our descriptive analysis. We observe that the LOB imbalance almost perfectly describes on which side of the LOB an order occurs; see Figure 3.

As an application of the specified and calibrated model, we estimate the conditional distribution of the next order type. This works remarkably well and opens the door to order-type prediction.

1.2 Relevant literature

Gould et al. (2013) gives an overview for the modeling of limit order books. For the applied nonparametric estimation method of multivariate Hawkes processes, see Kirchner (2017a). For an application to the inference of the Hawkes skeleton and the Hawkes graph from data, see Embrechts and Kirchner (2017). For the notation and MLE estimation of marked Hawkes processes, we refer to Liniger (2009). We use a formalization of the limit order book similar to Cont and de Larrard (2013). For more information on Hawkes modeling of limit order books, see Bacry and Muzy (2015) and the references therein.

1.3 Structure

In Section 2, we introduce multivariate marked Hawkes processes and discuss their specification and estimation. In Section 3, we formalize limit order books and give a descriptive analysis of our datasets. In Section 4, we specify a parametric model for order arrivals. In Section 5, we present MLE calibration of the model and possible economic interpretations of the results. In Section 6, we illustrate how the calibrated model can be used to predict the next order type. In Section 7, we summarize our results, discuss specific problems, and propose possible extensions and applications.

2 Methodology

2.1 Definitions and terminology

Marked Hawkes processes are self- and crossexciting point processes; their intensity depends on the past of the process itself, see Hawkes (1971b). As such, they can be interpreted as point process counterparts to autoregressive time series in discrete time; see Kirchner (2016). For $d \in \mathbb{N}$, a d-type (or d-variate) marked Hawkes process is a simple point process on $\mathbb{R}$, where, for $k \in \mathbb{Z}$, the k-th event is described by the triple $(T_k, L_k, Z_k^{(L_k)})$. Here, $T_k \in \mathbb{R}$ is the *arrival time* (we assume that $T_k < T_{k+1}$), $L_k \in [d] := \{1, \dots, d\}$ is the *type*, and $Z_k^{(L_k)} \in \mathbb{R}$ is the *mark value* of the k-th event. We assume that $\{Z_k^{(j)},\ k \in \mathbb{Z},\ j \in [d]\}$ are independent random variables with $Z_k^{(j)} \sim F_j$, $k \in \mathbb{Z}$, $j \in [d]$, where $\{F_j\}_{j\in[d]}$ are absolutely continuous *mark distributions* with densities $\{f_j\}_{j\in[d]}$. We set $N_j(A \times B) := \#\{T_k \in A :\ L_k = j, Z_k^{(j)} \in B\},\ j \in [d], A, B \in \mathcal{B}(\mathbb{R})$ and

$\mathbf{N} = (N_1, N_2, \ldots, N_d)$. Furthermore, the *history* of $\mathbf{N}$ is

$$\mathcal{H}_t^{(\mathbf{N})} := \sigma\Big(N_j\big(A \times B\big) : A \in \mathcal{B}\big((-\infty, t]\big),\ B \in \mathcal{B}(\mathbb{R}),\ j \in [d]\Big), \quad t \in \mathbb{R}.$$

The *conditional intensity* of N_j is

$$\lambda_j^{(\mathbf{N})}(t) = \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{E}\Big[N_j\big((t, t+\delta] \times \mathbb{R}\big) \,\Big|\, \mathcal{H}_t^{(\mathbf{N})}\Big], \quad j \in [d], \quad t \in \mathbb{R}.$$

A *d-type marked Hawkes process* $\mathbf{N} = (N_1, N_2, \ldots, N_d)$ solves the family of equations

$$\lambda_j^{(\mathbf{N})}(t) = \eta_j + \sum_{i=1}^{d} \int_{(-\infty,t)\times\mathbb{R}} m_{i,j} w_{i,j}(t-s)\, g_{i,j}(z)\ N_i(ds \times dz), \quad j \in [d], t \in \mathbb{R}. \tag{2.1}$$

For $j \in [d]$, $\eta_j \geqslant 0$ are the *baseline intensities*. For $(i,j) \in [d]^2$, $m_{i,j} \geqslant 0$ are the *branching coefficients*, $w_{i,j} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, $w_{i,j}(t) = 0$, $t \leqslant 0$, are the *decay kernels*, and $g_{i,j} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ are the *impact functions*. Furthermore, we define the *excitement functions* $h_{i,j} := m_{i,j} w_{i,j}$, $(i,j) \in [d]^2$. Decay kernels and the impact functions are normalized by the conditions

$$\int w_{i,j}(t)dt = 1 \quad \text{and} \quad \mathbb{E}g_{i,j}(Z_0^{(i)}) = 1, \quad (i,j) \in [d]^2. \tag{2.2}$$

These normalizations make the model identifiable. Note that, in general, the second condition in (2.2) is also a moment condition on the mark distributions F_i, $i \in [d]$. Obviously, the defining equations (2.1) are self-referencing: the process is defined via its conditional intensity and the conditional intensity is defined via the process. So this is an implicit definition. It is comparable to the implicit definition of autoregressive time series. It is non-trivial, but well-known, that a solution for (2.1) exists if the spectral radius of the *branching matrix* $M = (m_{i,j})_{1 \leqslant i,j \leqslant d} \in \mathbb{R}_{\geqslant 0}^{d \times d}$ is strictly less than one; see Liniger (2009). This solution specifies a unique point process distribution, a Hawkes process.

2.2 MLE estimation

We consider parametric families of decay kernels and impact functions, $\{w_{i,j}^{(\vartheta_{i,j})}\}$ and $\{g_{i,j}^{(\rho_{i,j})}\}$ as well as mark-distribution densities $\{f_i^{(\rho_i)}\}$. From these parametrizations we obtain a parametric Hawkes family

$$\Big\{N^{(\theta)} :\ \theta = \big((\eta_j), (m_{i,j}), (\vartheta_{i,j}), (\rho_{i,j}), (\rho_i)\big)\Big\},$$

with obvious restrictions on the parameter space. The parameter-vector θ is a concatenation of the baseline intensities, the branching coefficients as well as the parameters for decay kernels, impact functions, and mark densities. For $j \in [d]$, denote by $\theta_j \subset \theta$ the subset of parameters

that the j-th conditional-intensity component, $\lambda_j^{(\mathbf{N})}$, depends on; see (2.1). At this point, we drop the superscript $(\mathbf{N})$ and write $\lambda_j^{(\theta_j)}$ instead. Note that $\theta_i \cap \theta_j = \varnothing,\ i \neq j$. The (approximative) *log likelihood* of θ with respect to a sample of N on $[T_*, T^*]$, $T_* < T^*$, then reads

$$l_{[T_*,T^*]}(\theta) = \sum_{j=1}^{d} \int_{[T_*,T^*]\times\mathbb{R}} \log \lambda_j^{(\theta_j)}(t)\, N_j(dt \times dz) - \sum_{j=1}^{d} \int_{T_*}^{T^*} \lambda_j^{(\theta_j)}(t) dt + \sum_{j=1}^{d} \int_{\mathbb{R}} \log f_j^{(\theta_j)}(z)\, N_j([T_*, T^*] \times dz). \tag{2.3}$$

In general, the conditional intensities $\lambda_j^{(\theta_j)}$, $j \in [d]$, (and hence also the likelihood $l_{[T_*,T^*]}$) depend on values of $\mathbf{N}$ before time T_*; see (2.1). Here, we have no data. This is why we call the likelihood approximative. To make up for this approximation, one could add a compensation to the likelihood part (by approximating the past of the process before T_* by its first moment measure which in turn depends on the parameters). This would correct part of the edge effects. However, it also increases computational complexity in the optimization. As we are more interested in local modeling, we ignore these edge effects. Furthermore, we do not consider the specification and estimation of the mark distribution; we treat the mark sizes as given throughout. So the last summand in (2.3) is irrelevant for us. Finally, note that for the sake of optimization, we can split (2.3) into d summands as the conditional intensities $\lambda_i^{(\theta_i)}$ depend on disjoint sets of parameters. For this representation (and for implementation), it is convenient to split the data into the different types, that is,

$$\dot{\bigcup}_{j\in[d]} \big\{(T_k^{(j)}, Z_k^{(j)})\big\}_{k\in[n_j]} := \big\{(T_k, Z_k^{(L_k)})\big\}_{k\in[n]},$$

with $n_j := N_j\big([T_*, T^*] \times \mathbb{R}\big)$, $j \in [d]$, and $n := \sum_{j\in[d]} n_j$. Summarizing, the log likelihoods we consider are

$$l^{(j)}_{[T_*,T^*]}(\theta_j) = \sum_{k=1}^{n_j} \log \tilde{\lambda}_j^{(\theta_j)}(T_k^{(i)}) - \int_{T_*}^{T^*} \tilde{\lambda}_j^{(\theta_j)}(t)\mathrm{d}t, \quad j \in [d], \tag{2.4}$$

where

$$\tilde{\lambda}_j^{(\theta_j)}(t) := \eta_j + \sum_{i\in[d]} \sum_{T_k^{(i)}\in[T_*,t)} m_{i,j} g_{i,j}^{(\rho_{i,j})}\big(Z_k^{(i)}\big) w_{i,j}^{(\vartheta_{i,j})}(t - T_k^{(i)}), \quad j \in [d], \tag{2.5}$$

with $\theta_j := \big(\eta_j, (m_{i,j})_{i\in[d]}, (\vartheta_{i,j})_{i\in[d]}, (\rho_{i,j})_{i\in[d]}\big)$, $j \in [d]$. Note that the number of parameters quickly becomes large in d: there are d baseline intensities and d^2 branching coefficients. In addition, each of the d^2 impact and decay functions typically has at least one free parameter. That is, there are at least $3d^2 + d$ parameters which we have to estimate. Also note that it is not a priori clear how parametric families for the impact and decay functions ought to be chosen.

2.3 Nonparametric estimation

In Kirchner (2017a), we introduce a nonparametric estimation procedure for multivariate Hawkes processes without marks. The method depends on discretization. The theoretical basis for discrete approximation in a Hawkes setup is a convergence theorem derived in Kirchner (2016) (univariate case, autoregressive view) and in Kirchner (2017b) (multitype case, branching view). We divide the time line into bins of size $\Delta > 0$. Given an unmarked Hawkes process $\mathbf{N} = (N_1, N_2, \ldots, N_d)$, we count the number of events in each bin (for each event type) and obtain an $\mathbb{N}_0^d$-valued sequence $(\mathbf{X}_n^{(\Delta)})_{n\in\mathbb{Z}}$, where $\mathbf{X}_n^{(\Delta)} := (X_{n,1}^{(\Delta)}, X_{n,2}^{(\Delta)}, \ldots, X_{n,d}^{(\Delta)})$ with

$$X_{n,j}^{(\Delta)} := N_j\Big(\big((n-1)\Delta, n\Delta\big]\Big), \quad j \in [d],\, n \in \mathbb{Z}. \tag{2.6}$$

For small $\Delta > 0$ and large $p \in \mathbb{N}$, we obtain

$$\mathbb{E}\left[X_{n,j}^{(\Delta)} | \sigma\big(\mathbf{X}_{n-k,j}^{(\Delta)} : k \in \mathbb{N}\big)\right] \approx \Delta\eta_j + \sum_{i=1}^{d}\sum_{k=1}^{p} \Delta h_{i,j}(k\Delta) X_{n-k,i}^{(\Delta)}, \quad j \in [d],\, n \in \mathbb{Z}. \tag{2.7}$$

Treating approximation (2.7) as an equality, we apply conditional least-squares optimization for the coefficients. After renormalization of the results, we obtain the *multivariate Hawkes estimator* with respect to the corresponding support s and bin size Δ; see Kirchner (2017a) for details. The multivariate Hawkes estimator collects estimates for η_j, $j \in [d]$, and $h_{i,j}(k\Delta)$, $k = 1, 2, \ldots, p$, $(i, j) \in [d]^2$. In Kirchner (2017a), we discuss the selection of Δ and s. We also prove asymptotic normality and derive consistent covariance estimates. This opens the door to testing. In the present paper, we extend the method to the marked case.

2.4 Hawkes graphs and skeletons

Embrechts and Kirchner (2017) introduces the *Hawkes graph*—a compact, yet meaningful summary of an unmarked d-type Hawkes process. The d vertices of the Hawkes graph correspond to the possible event types. There is an edge from vertex i to vertex j if and only if $m_{i,j} > 0$ for the corresponding Hawkes process. That is, each event type j is only excited by the types corresponding to the parents of vertex j in the Hawkes graph. Note that a vertex can be its own parent. In addition, vertices and edges are supplied with *weights*. The weight of vertex i corresponds to the baseline intensity $\eta_i\,(\geqslant 0)$; the weight of edge (i, j) corresponds to the branching coefficient $m_{i,j}$. Studying the Hawkes graph (walks, walk weights, in and out degrees, feedback loops, etc.) gives complementary insight into the process. For any estimation method, the knowledge of the *Hawkes skeleton*, that is, the Hawkes graph without weights, is very helpful; it potentially reduces the dimensionality of the optimization problem significantly. In Embrechts and Kirchner (2017), we propose to infer the Hawkes skeleton from data with

the nonparametric method introduced in Section 2.3: we calculate estimates for $m_{i,j}$ and the corresponding variance estimates from the time series estimates derived from (2.7), that is, the unmarked case. Then we test for all pairs $(i, j) \in [d]^2$ whether the estimated branching coefficient $\hat{m}_{i,j}$ is significantly larger than zero. If so, we introduce the edge (i, j) into the estimated skeleton. Under the paradigm that the underlying true skeleton is sparse, the skeleton estimate will also be sparse. We explain (and confirm in simulations) that for estimation of the skeleton, Δ can be chosen rather coarse. Therefore, the discretization parameter Δ can be used for controlling computational complexity. Note that the significance parameter $\alpha \in (0, 1)$ for the edge selection controls sparseness of the graph. More specifically, α gives the fraction of falsely introduced edges. Such redundant edges do not introduce additional bias in the final model estimation. So the typically 'overestimated'—though still sparse—estimated skeleton may be used as a basis for further nonparametric or parametric analysis of the process. Note that when we again estimate the Hawkes graph (including the edge weights) nonparametrically from (2.7), we may typically apply a much smaller bin size—given a relatively sparse Hawkes-skeleton estimate.

2.5 Nonparametric estimation for the marked case

For our case study, we extend the estimation method from Section 2.3 to the marked case: we discretize both time and mark space. For each event type $i \in [d]$, we partition the range of mark sizes $Z_k^{(i)}$ into $q_i \in \mathbb{N}$ disjoint *mark regimes* $I_l^{(i)} \subset \mathbb{R}$, $l = 1, \ldots, q_i$, with $\dot{\cup}_{l=1}^{q_i} I_l^{(i)} = \text{range}(Z_k^{(i)})$, $i \in [d]$. We write $p_l^{(i)} := \mathbb{P}\left[Z_0^{(i)} \in I_l^{(i)}\right]$, $l = 1, 2, \ldots, q_i$, $i \in [d]$, for the mark-regime probabilities and $\hat{p}_l^{(i)}$ for their empirical counterparts. The approximative model equation from (2.7) now reads, for $j \in [d]$,

$$\mathbb{E}\left[X_{n,j}^{(\Delta)} \middle| \sigma\big(\mathbf{X}_{n-k}^{(\Delta)} : k \in \mathbb{N}\big)\right] \approx \Delta\eta_j + \sum_{i=1}^{d}\sum_{k=1}^{p}\sum_{l=1}^{q} \Delta h_{i,j}(k\Delta) g_{i,j}^{(l)} \tilde{X}_{n-k,i,l}^{(\Delta)}, \quad n \in \mathbb{Z}, \tag{2.8}$$

where

$$\tilde{X}_{n,i,l}^{(\Delta)} := \#\Big\{k : T_k \in \big((n-1)\Delta, n\Delta\big],\, L_k = i,\, Z_k^{(L_k)} \in I_l^{(i)}\Big\}, \quad n \in \mathbb{Z},$$

that is, $\tilde{X}_{n,i,l}^{(\Delta)}$ counts the number of type-i events in the n-th time bin with mark value lying in the l-th regime. Note that $\sum_{l=1}^{q_i} \tilde{X}_{n,i,l}^{(\Delta)} = X_{n,i}^{(\Delta)}$, $i \in [d]$, $n \in \mathbb{Z}$. We calculate estimates for constant-part and linear coefficients in (2.8) by conditional least-squares estimation—in analogy to (2.7). Thus, we obtain estimates $\hat{\pi}_{i,j,k,l}$ for the coefficients $\pi_{i,j,k,l} := h_{i,j}(k\Delta) g_{i,j}^{(l)}$ in a straightforward way. Together with the normalizing conditions (2.2), we derive estimates $\hat{h}_{i,j,k}^{(\Delta)}$ for $h_{i,j}(k\Delta)$,

estimates $\hat{w}^{(\Delta)}_{i,j,k}$ for $w_{i,j}(k\Delta)$ and estimates $\hat{g}_{i,j,l}$ for $g_{i,j}(\mathbb{E}\,1_{Z^{(i)}_k \in I^i_l} Z^{(i)}_k)$: for $(i,j) \in [d]^2$, we set

$$\hat{h}^{(\Delta)}_{i,j,k} := \hat{h}^{(\Delta)}_{i,j,k} \underbrace{\sum_{l=1}^{q_i} \hat{p}^{(i)}_l \hat{g}_{i,j,l}}_{\overset{(2.2)}{:=}1,} = \sum_{l=1}^{q_i} \hat{p}^{(i)}_l \hat{\pi}_{i,j,k,l}, \quad k = 1,2,\dots,p, \tag{2.9}$$

$$\hat{m}_{i,j} := \sum_{k=1}^{p} \Delta \hat{h}_{i,j}(k\Delta), \tag{2.10}$$

$$\hat{g}_{i,j,l} := \underbrace{\frac{1}{\hat{m}_{i,j}} \sum_{k=1}^{p} \Delta \hat{h}_{i,j}(k\Delta)}_{\overset{(2.10)}{:=}1} \hat{g}_{i,j,l} = \frac{\Delta}{\hat{m}_{i,j}} \sum_{k=1}^{p} \hat{\pi}_{i,j,k,l}, \quad l = 1,2,\dots,q_i, \text{ and} \tag{2.11}$$

$$\hat{w}^{(\Delta)}_{i,j,k} := \frac{\hat{h}_{i,j}(k\Delta)}{\hat{m}_{i,j}}, \quad k = 1,2,\dots,p. \tag{2.12}$$

For the case study in this paper, we choose $q_i \equiv q \in \mathbb{N}$, $i \in [d]$, and $I^{(i)}_l$, $l = 1,2,\dots,q$, such that $\hat{p}^{(i)}_l \approx q^{-1}$. Thus, the variance for the estimates of $g^{(l)}_{i,j}$ is more or less equally distributed over $l = 1,2,\dots,q$. Also note that in (2.8), i only runs over $i \in \mathrm{PA}(j)$, where $\mathrm{PA}(j) := \{i : m_{i,j} > 0\}$, $j \in [d]$, are the 'parent types' of type i. Therefore, if the Hawkes skeleton (estimate) is given, the design matrix for conditional-least squares estimation becomes smaller accordingly.

3 Data

In this section, we introduce limit-order-book (LOB) terminology. Moreover, we provide a descriptive analysis of the data examined.

3.1 Limit order books

An LOB is a double auction system that allows market participants to sell and buy stocks by posting orders and, if necessary, canceling them. On the *bid side* of the LOB, buy offers are recorded. On the *ask side*, sell offers are recorded. In the LOB, market participants are allowed to post offers at any price on an (equidistant) grid. The grid size is specified by the exchange and is called the *(price) tick*. Typically, orders for many levels of bid and ask prices exist. The *best bid (price)* and *best ask (price)* indicate the best prices at which one can immediately sell (best bid price) and buy (best ask price) the asset. The difference between best ask price and best bid price is called *(bid–ask) spread*. Every order has a volume, the *order size*. It denotes the quantity of assets the market participant wants to buy, sell, or cancel. The *lot size* denotes

the minimum number of shares one can post as limit or market order. Thus, order sizes are always given in multiples of the lot size. Roughly speaking, market participants have six types of actions they can pursue:

- *Market orders*: send a sell order at the best bid price or send a buy order at the best ask price. Market orders immediately inititiate trades.
- *Limit orders*: send a limit order. These can either be *limit bid orders* (offer to buy stocks for a price less than the best ask price) or *limit ask orders* (offer to sell stocks for a price higher than the best bid price).
- *Cancelations*: cancel a previously sent bid or ask limit order.

Whenever all limit orders at the best bid or best ask price get matched with incoming market orders (or are cancelled), the best bid or best ask price jumps to the next price level with existing limit orders. Note that, in our terminology, all orders which include the word 'bid' are *bid-side events*: in other words, limit bid orders are buy orders, whereas *market bid orders are sell orders*. These get matched with limit orders at the best bid, hence the terminology.

3.2 Stylized order book

For simplicity, we only keep track of orders and cancelations which arrive at the best bid and the best ask price. We introduce abbreviations for the six LOB events we consider:

- Market bid orders at best bid (MB) (= sell market order).
- Limit bid orders at best bid (LB).
- Cancelations at best bid (CB).
- Market ask orders at best ask (MA) (= buy market order).
- Limit ask orders at best ask (LA).
- Cancelations at best ask (CA).

We call these six event types *orders* (although—strictly speaking—a cancelation is not an 'order'). We denote the set of all possible orders by $O := \{\mathrm{MB}, \mathrm{LB}, \mathrm{CB}, \mathrm{MA}, \mathrm{LA}, \mathrm{CA}\}$. If we refer to order types on the bid *or* the ask side, we speak of *market orders* (MO), *limit orders* (LO), and *cancelation orders* (CO). We introduce some more LOB notation:

- $Q^{(b)}(t)$ and $Q^{(a)}(t)$, $t \in \mathbb{R}$: number of limit orders at the best bid and at the best ask at time t.

- $I(t)$, $t \in \mathbb{R}$: *LOB imbalance* at time t; we define it as

$$I(t) = \frac{Q^{(a)}(t) - Q^{(b)}(t)}{Q^{(b)}(t) + Q^{(a)}(t)} \big(\in [-1, 1] \big), \quad t \in \mathbb{R}. \tag{3.1}$$

- $\delta > 0$: *tick size*; denotes the minimal price increment possible in the LOB.
- $\sigma \in \mathbb{N}$: *lot size*; denotes the minimal number of shares one can post as limit or market order. Note that order sizes are always given in multiples of the lot size.

We denote an LOB event as $(T_k, O_k, V_k^{(O_k)})$, where $T_k \in [T_*, T^*]$ is the time stamp, $O_k \in \{LB, LA, MB, MA, CB, CA\}$ gives the order type, and $V_k^{(O_k)} \in \sigma\mathbb{N}$ denotes the corresponding order size or volume.

3.3 Descriptive analysis

We provide an overview of the LOB datasets under examination. We use NASDAQ data bought from LOBSTER (`https://lobsterdata.com` made by frischedaten UG (haftungsbeschraenkt)). We consider two stocks: Intel (INTC15) and Microsoft (MSFT15). We choose these two assets because they are 'large-tick assets'. That is, the price tick δ is large compared to the price level of the stock. These kinds of assets have several convenient features: the spread is nearly constant to one price tick all over the data and a large part of the LOB volume is concentrated at best bid and best ask. Furthermore, there are no preferences for round prices (that is, the set of observed prices modulo 1\$, respectively, modulo 0.1\$ is uniform over $(0, 1)$, respectively $(0, 0.1)$. For both assets, we consider a 3-month time window (September, October, November in 2015) which means 61 trading days. The data are provided in two files, an *order-book file* and a *message-book file*. The order-book files provide the values of $Q_b(t)$ and $Q_a(t)$ whenever one of the queues changes. The message-book files record what kind of LOB event happened with a time resolution of 10^{-9} sec. For our analysis, we derive the order streams from the message-book data (only at best bid and best ask). The order-book file will be used to calculate the imbalance. The arrival intensity of orders exhibits the typical U-shape with many order-book events at the beginning and end of the trading day and much calmer periods midday for both assets; see Figure 1. The number of events per order type and the corresponding volumes over one trading day is given in Figure 2. We observe that a large majority of the LOB events consists of limit orders and cancelations. Market orders are much rarer. Order sizes are clustered at multiples of 100 for smaller orders and at multiples of 1000 for larger orders. Only few orders attain sizes outside this discrete grid. Most volume sequences are autocorrelated (not illustrated). Empirical transition probabilities between all order types are displayed in Table 1. In this table, we observe that limit orders are often followed by cancelations on the same side of the book. Market orders are most often followed by limit orders on the other side of the

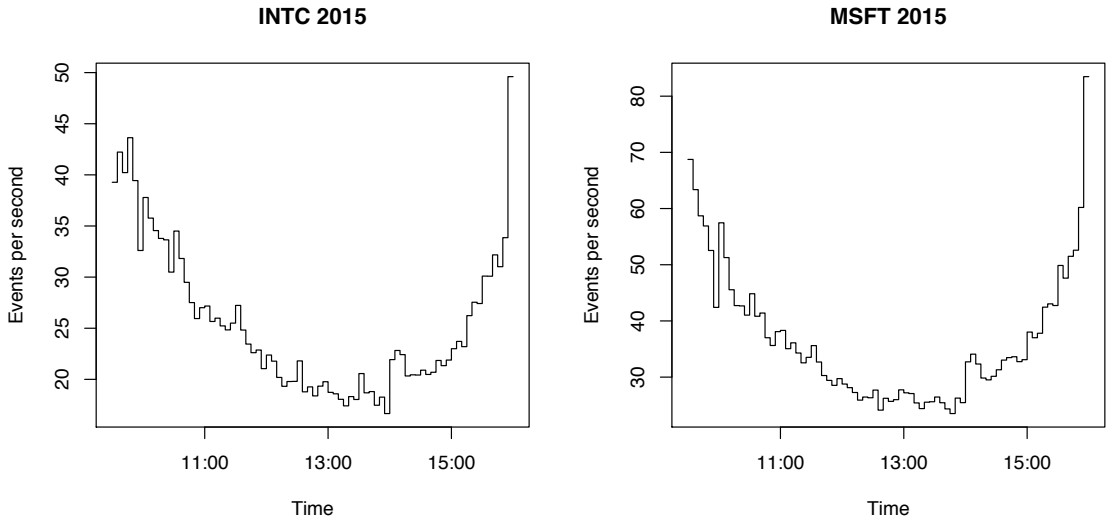


Figure 1: Empirical intensities: for each of the 61 trading days in our datasets, we calculate the average number of order book events per second in five minute windows. The figures illustrate the averages over the corresponding time windows. The U-shapes are obvious. We also note some patterns that do not seem to be coincidental such as the sudden drop at 10am or the sudden increase at 2pm.

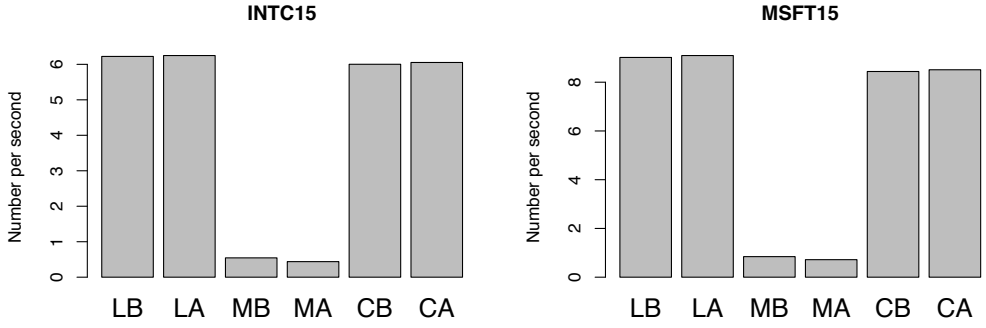


Figure 2: Composition of the order streams: there are far more limit orders and cancelations than market orders.

book. The bid–ask symmetry in the empirical transition probabilities is obvious. In addition, the values are quite similar for both INTC15 and MSFT15.

3.4 Imbalance

We summarize the state of an LOB at time $t \in \mathbb{R}$ by its *imbalance* $I(t)$; see (3.1). Note that $I(t) \in [-1, 1]$. If $I(t)$ is negative, then there are more limit orders at the best bid than at the best ask; if $I(t)$ is positive then there are more limit orders at the best ask than at best bid. An imbalance close to -1 means that relatively many limit orders are present at the best bid price, but relatively few at the best ask price. Consequently, it is likely that the price will

Table 1: Unconditional empirical transition probabilities: the first values refer to INTC15, the values in brackets to MSFT15. Note that the results are almost equal for both stocks. Also note that the diagonal and crossdiagonal 3×3 matrices are very similar. In other words, the transitions are bid–ask symmetric. The high values on the diagonals are presumably owed to order splits.

	MB	LB	CB	MA	LA	CA
MB	0.61 (0.61)	0.05 (0.05)	0.05 (0.05)	0.01 (0.01)	0.21 (0.21)	0.07 (0.06)
LB	0.01 (0.01)	0.40 (0.39)	0.31 (0.31)	0.01 (0.01)	0.10 (0.11)	0.17 (0.18)
CB	0.01 (0.01)	0.32 (0.33)	0.35 (0.32)	0.01 (0.01)	0.17 (0.19)	0.14 (0.14)
MA	0.02 (0.02)	0.21 (0.21)	0.06 (0.05)	0.67 (0.67)	0.01 (0.01)	0.03 (0.04)
LA	0.01 (0.01)	0.10 (0.10)	0.17 (0.18)	0.00 (0.00)	0.40 (0.38)	0.32 (0.31)
CA	0.01 (0.01)	0.17 (0.19)	0.14 (0.14)	0.00 (0.01)	0.32 (0.33)	0.35 (0.32)

increase in the near future after the remaining few limit orders on the ask side get matched with market orders or get cancelled. Traders therefore have an incentive to send their market ask order (MA) before the (best ask) price increases and thus 'earn the spread'. This explains why the relative frequency of market *ask* orders—given a market order—is almost equal to 1 for imbalance values close to -1 and then decreases to 0 with increasing imbalance as illustrated in Figure 3. The figure also shows that this dependence exhibits a stretched inverted S-shape that can be approximated by a linear function. More specifically, we observe the remarkably simple empirical relation

$$\mathbb{P}_{\text{emp}}\left[\text{MA at time } t \mid \text{MO at time } t\right] \approx \frac{1 - I(t)}{2}, \quad t \in \mathbb{R}. \tag{3.2}$$

Hence, given that we observe a market order, the state of the imbalance explains the probability that this market order arrives at the ask side. Note that, for market bid orders, the relation reads $\frac{1+I(t)}{2}$. We find a similar effect for cancelations and limit orders—though not as clear-cut as for market orders. The slope is positive for market bid orders, for cancelation on the bid sid, and for limit ask orders; the slope is negative for the other orders. In any case, we conclude that the imbalance is a variable that we ought to include into our model. Furthermore, it seems to be reasonable to include the imbalance in a *linear* manner.

4 Model specification

The goal of this section is to specify a parametric model for our LOB data in a reasonable way, that is, without postulating too many a priori assumptions. We motivate the use of a self- and crossexciting Hawkes model for this kind of dataset. As a first step towards such a model, we infer a preliminary Hawkes skeleton of the six-type order event stream—assuming constant impact functions and constant baseline intensities—using the nonparametric method explained in Section 2.4. Given this preliminary skeleton, we estimate the edge weights and give a prelimary nonparametric analysis of the excitement functions. This analysis yields a

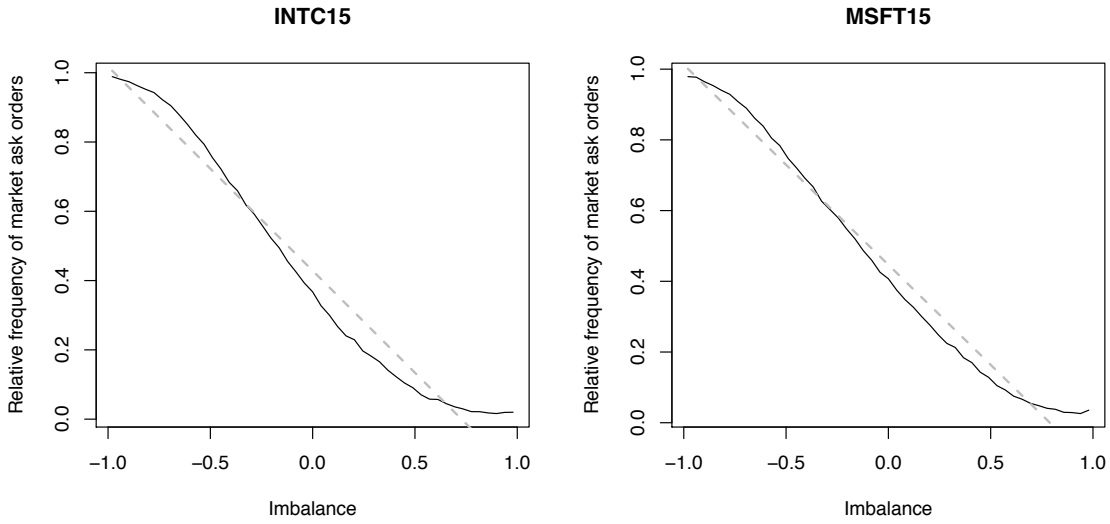


Figure 3: Relative frequency of market ask orders in terms of all market orders plotted against the imbalance for MSFT: we observe a stretched inverted S-shape. The dashed line refers to a linear regression.

nonparametric Hawkes-graph estimate and, in particular, the final model skeleton with respect to which all further calculation is done. In a second step, we also consider the volume of the orders and examine the shape of the impact and the excitement functions using the bin-count method for the marked case as explained in Section 2.5. Finally, we include bid–ask symmetry considerations as well as state dependence (via the LOB imbalance) and formulate a corresponding fully parametric Hawkes model.

4.1 Motivation of Hawkes model

We found that the (marginal) distribution of many interarrival times in the data can adequately be described by Gamma distributions (not illustrated). However, one can also show that the interarrival-time sequences are strongly autocorrelated. Consequently, a pure Gamma renewal process is not a good model. It cannot explain the clustering of events or the longer 'calm' periods that we observe in the data. This makes a Hawkes process a more plausible model for the LOB event streams than a pure renewal model. Another benchmark model is the class of doubly stochastic Poisson processes, where the intensity depends on the state of the book. E.g., in Cont and de Larrard (2013), the arrival intensity of the various orders is explained by the size of the bid and the ask queue. Clearly, this kind of model can reproduce clustering of events. However, pure state dependence cannot (or only very indirectly) capture the obvious excitements between some order types. For example, limit orders are often followed by cancelations, market orders by limit orders, etc.; see Table 1. However, we have shown in Section 3.4 that the imbalance contains information that is 'directly' relevant for the order flow. By letting the

Table 2: Results of the skeleton analysis described in Section 2.4 for the INTC15 data: for each edge, we give the average p-value as well as the relative frequency of the inclusion in the estimated skeleton with respect to a significance level of 0.05 (of over 2000 considered 15 min windows). Both tables are remarkably bid–ask symmetric; that is, the diagonal and crossdiagonal 3×3 submatrices are very similar. On the basis of these tables, we choose a preliminary skeleton that is the base for the Hawkes-graph analysis; see Figure 4.

	MB	LB	CB	MA	LA	CA
MB	0.19 (0.41)	0.30 (0.29)	0.29 (0.31)	0.43 (0.11)	**0.04** (0.83)	**0.06** (0.76)
LB	0.84 (0.00)	0.14 (0.57)	**0.05** (0.83)	0.51 (0.05)	0.72 (0.03)	0.27 (0.37)
CB	**0.11** (0.53)	0.83 (0.02)	0.71 (0.05)	0.60 (0.03)	**0.11** (0.59)	0.42 (0.15)
MA	0.41 (0.11)	**0.04** (0.85)	**0.06** (0.75)	0.26 (30)	0.45 (0.15)	0.38 (0.20)
LA	0.58 (0.03)	0.76 (0.02)	0.30 (0.33)	0.86 (0.00)	0.15 (0.57)	**0.06** (0.79)
CA	0.50 (0.08)	**0.09** (0.67)	0.38 (0.18)	**0.10** (0.61)	0.82 (0.02)	0.70 (0.05)

baseline intensities of our model depend on the value of the imbalance, we incorporate state dependence to the past dependence of the Hawkes model. We thus work with a state-dependent multivariate Hawkes process.

4.2 Hawkes-skeleton and Hawkes-graph estimation

As a first step, we infer the Hawkes skeleton and the Hawkes graph from the data; see Section 2.4. The results will later be used for the parametric MLE estimation. However, they are interesting in their own right. They give a (preliminary) graphical summary of the relations between the six order streams; see Figure 4. The six considered order types are represented by vertices and the non-zero excitements by edges of the skeleton. Note that we ignore the volume of the orders as well as the LOB imbalance at this point. We split the data in 15 min windows—to make the (otherwise constant) baseline intensities more flexible. For each of the 2084 15 min time windows, we estimate the Hawkes skeleton with respect to a truncation parameter $s = 5 \cdot 10^{-1}$ sec and a (relatively coarse) bin size $\Delta_{\text{coarse}} = 10^{-2}$ sec. This is a quite 'local' analysis in order to keep computing time tolerable. However, working with (much larger) single windows and larger s, we found that the size of s is not crucial for detection of excitement. In the final MLE model calibration, we will drop this truncation anyway. We estimate the Hawkes skeleton as explained in Section 2.4. That is, for each $(i, j) \in [d]^2$, we calculate p-values with respect to the null hypothesis that there is no excitement from i to j; see Embrechts and Kirchner (2017) for the necessary calculations. In Table 2, we give these average p-values for the INTC15 data. For each edge, we also give the relative frequency of the corresponding p-values less than 0.05. As for the order-type transitions (Table 1), the table is remarkably bid–ask symmetric. Moreover, we found that the corresponding table for the MSFT15 data is nearly equivalent (not illustrated). Based on Table 2, we pick a preliminary edge set, i.e., skeleton. Note that including redundant edges ('false positive edges') will yield larger variance and computation times

whereas excluding true edges ('false negative edges') will yield bias. In view of this observation, we pick too many edges rather than too few—keeping in mind that the actual effect will be quantified only in a next step. These considerations yield the *preliminary Hawkes-skeleton estimate*

$$\begin{aligned}\hat{\mathcal{E}}_{\text{prelim}} = \big\{&(\text{MB} \to \text{LA}), (\text{MB} \to \text{CA}), (\text{LB} \to \text{LB}), (\text{LB} \to \text{CB}), (\text{CB} \to \text{LA}), (\text{CB} \to \text{MA}) \\ &(\text{MA} \to \text{LB}), (\text{MA} \to \text{CB}), (\text{LA} \to \text{LA}), (\text{LA} \to \text{CA}), (\text{CA} \to \text{LB}), (\text{CA} \to \text{MB})\big\}\end{aligned} \tag{4.1}$$

Note that only relatively few of the estimated window-wise skeletons (not illustrated) contain edges not selected in $\hat{\mathcal{E}}_{\text{prelim}}$. In this sense, the preliminary skeleton *overestimates* the true skeletons. Next, we estimate the Hawkes-graph weights with respect to the preliminary skeleton from (4.1). To that aim, we may choose a finer bin size $\Delta_{\text{fine}} = 10^{-4}$ as the computations become less involved. In Figure 4, we illustrate both: the prelimary skeleton as well as the corresponding average Hawkes graph for the INTC15 data. Thick arrows indicate strong excitement, whereas thin arrows represent weak excitement for the corresponding edges. We list the main characteristics of our Hawkes-skeleton and Hawkes-graph analysis:

1. Hawkes skeleton and Hawkes graph estimates are bid–ask symmetric and nearly equivalent for both the INTC15 and the MSFT15 data.
2. The heaviest edges are $(\text{MB} \to \text{LA})$ and $(\text{MA} \to \text{LB})$. They are also the most consistent edges over the various skeleton estimations; see Table 2. (Remember that, by our definition, market bid orders are sell orders, that is, market orders which get matched with a limit order on the bid side as described in Section 3.1.)
3. Market orders are the drivers of the process in that they excite limit orders and cancelations on the other side of the book, but get hardly excited by any other order type. This makes sense from an (naive) economic point of view: if nobody actually buys or sells, the market comes to a standstill.
4. There are several loops, for example $(\ldots \to \text{LB} \to \text{CB} \to \text{LA} \to \text{CA} \to \text{LB} \to \ldots)$. The excitements $(\text{LB} \to \text{CB})$ and $(\text{LA} \to \text{CA})$ can presumably be explained by traders who send limit orders and delete them instantaneously. The edges $(\text{CB} \to \text{LA})$ and $(\text{CA} \to \text{LB})$ are harder to interpret. Maybe this loop catches the 'searching for liquidity' of algorithms that comb in turns through either side of the book.
5. There are quite heavy edges $(\text{MB} \to \text{CA})$ and $(\text{MA} \to \text{CB})$. These excitements are hard to explain. Maybe, we observe an aggregation of excitement from market orders to limit orders on the other side of the book and from these limit orders to cancelations on the same side of the book.

6. There are four edges in the Hawkes skeleton which do not seem to be significant anymore in the graph estimations. Namely, $\big\{(\text{LB} \to \text{LB}; 0.011), (\text{LA}, \text{LA}; 0.012), (\text{CB} \to \text{MA}; 0.020), (\text{CA}, \text{MB}; 0.018)\big\}$. Here, the number denotes the average of the estimated branching coefficients/edge weights. These are much smaller than for the remaining edges in the graph. We will ignore these four edges in the final skeleton.

Note that we repeated these Hawkes-skeleton and Hawkes-graph estimations over many values of Δ, s, and various window sizes. We found that Hawkes-skeleton and Hawkes-graph estimates are remarkably stable with respect to these choices. Summarizing the considerations from above, for all following estimations, we work with the *final Hawkes-skeleton estimate*

$$\begin{aligned}\hat{\mathcal{E}}_{\text{final}} = \big\{&(\text{MB} \to \text{LA}), (\text{MB} \to \text{CA}), (\text{LB} \to \text{CB}), (\text{CB} \to \text{LA}) \\ &(\text{MA} \to \text{LB}), (\text{MA} \to \text{CB}), (\text{LA} \to \text{CA}), (\text{CA} \to \text{LB})\big\}. \end{aligned} \tag{4.2}$$

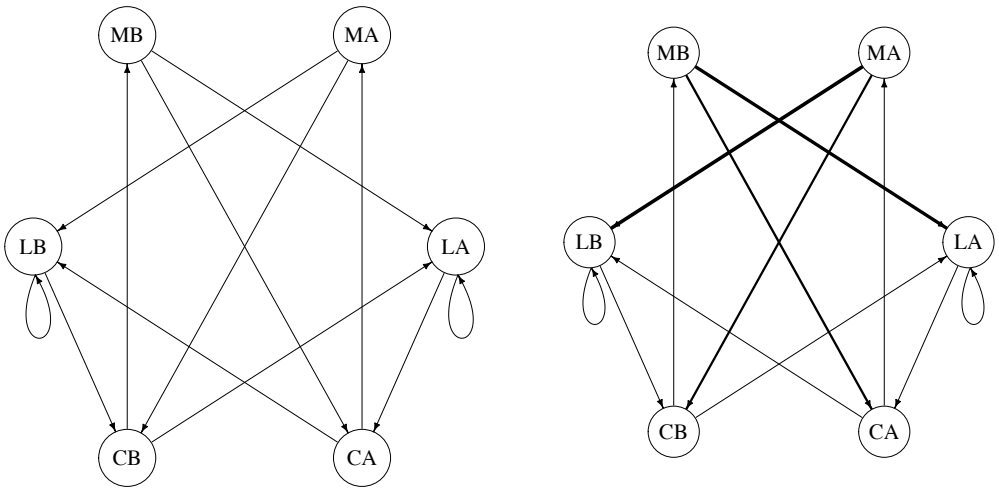


Figure 4: Preliminary Hawkes-skeleton estimate (left) and Hawkes-graph estimate (right): the Hawkes-skeleton estimate is based on the average skeleton estimates of the INTC15 data from Table 2.

4.3 Choice of excitement and impact functions

We remind the reader of the Hawkes terminology from Section 2.1: the strength of the excitement from type-i events to type-j events is measured by the branching coefficient $m_{i,j}$. The excitement decreases over time according to the decay kernel $w_{i,j}$. Also remember that we call

$h_{i,j} := m_{i,j} w_{i,j}$ excitement function. The impact of the mark $Z_k^{(i)}$ attached to the k-th type-i event boosts the excitement by another scaling factor $g_{i,j}(Z_k^{(i)})$, where $g_{i,j}$ is the impact function. Also note that we only have to estimate $h_{i,j}$ and $g_{i,j}$ for $(i,j) \in \mathcal{E}_{\text{final}}$, that is, an edge in the final skeleton from (4.2). This decreases both variance and computation time of the estimation. In a next step, we calculate nonparametric pointwise estimates for the excitement and impact functions as described in Section 2.5. From these pointwise estimates, we will specify the shape of the eight excitement functions, respectively, decay kernels as well as the eight impact functions in our final parametric model. In this section, we apply $\Delta_{\text{fine}} = 10^{-4}$ sec and $s = 10^{-1}$ sec for all calculations.

Excitement functions

From (2.9), we obtain $p = \lceil s/\Delta_{\text{fine}} \rceil = 10^3$ estimates for all time windows and for all the excitement functions evaluated at $\Delta_{\text{fine}}, 2\Delta_{\text{fine}}, \ldots, p\Delta_{\text{fine}}$. We plot the average of these estimates against the time grid to analyze the excitement shape. Two types of decay kernels are obvious candidates: exponential decay and power decay. For each of the eight considered edges, log-log plots exhibit linearity. We hence choose power-law decay kernels for our model. Furthermore, the plots are bid–ask symmetric.

Impact functions

We calculate (2.11) for all time windows in order to examine the impact functions of the order arrival process. We partition the range of order volumes into $q_i \equiv q = 5$ mark regimes $(I_l^{(i)})_{l=1,\ldots,5}$, $i \in [d]$, namely

$$(0, 100], (100, 300], (300, 500], (500, 700], (700, \infty).$$

For $(i,j) \in \mathcal{E}_{\text{skel}}$, we plot the average of the estimates $\hat{g}_{i,j,l}$, $l = 1, 2, \ldots, 5$, against the empirical values of $\mathbb{E}\left[1_{Z_k^{(i)} \in I_l^{(i)}} Z_k^{(i)}\right]$, $l = 1, 2, \ldots, 5$ (not illustrated). We observe that the estimated impact functions for all edges increase and may well be approximated by a linear function. Thus, we choose a linear impact function $g(x) = ax$ for the excitements. Note that in this case, there are no parameters left to estimate for these linear impact functions g. Indeed, the condition $\mathbb{E}g(Z) = 1$ from (2.2) together with the linearity of the expected value yields $a = (\mathbb{E}Z)^{-1}$, where Z is a generic mark of a market order. A linearly increasing impact function means that large orders cause more excitement than small ones. This seems to be reasonable.

4.4 Symmetry

We observe a clear-cut bid–ask symmetry in the Hawkes-skeleton and Hawkes-graph plots of Figure 4 as well as in the excitement and impact estimates described in Section 4.3 Furthermore, Table 2 for the average p-values, the imbalance dependence in Figure 3 and the empirical order-type transition probabilities in Table 1 exhibit bid–ask symmetry. Therefore, it seems well justified to consider a bid–ask symmetric model. Thus, we assume symmetry between the bid and the ask side in all parameters for the final calibration in Section 5. For example, we will assume that the excitement and impact functions for $(\mathrm{MB} \to \mathrm{LA})$ are the same as for $(\mathrm{MA} \to \mathrm{LB})$. The symmetry assumption halves the number of parameters and accordingly halves standard deviation for the remaining estimates.

4.5 State dependence

We have observed in Section 3.4 that the relative frequency of market ask orders (MA) can be approximately explained by the relation

$$\mathbb{P}_{\mathrm{emp}}\big[\text{MA at time } t \,\big|\, \text{MO at time } t\big] \approx \frac{1 - I(t)}{2}, \quad t \in \mathbb{R}, \tag{4.3}$$

where $I(t)$ is the imbalance process from (3.1). To incorporate this relation into our model, we include the imbalance for the modeling of market orders by letting the baseline intensities depend on the LOB imbalance in a linear and bid–ask symmetric manner: when the imbalance is close to -1, more ask market orders arrive; when it is close to 1, more market bid orders arrive. Accordingly, we make the baseline intensities $\eta_{\mathrm{MB}}(t)$ ($\eta_{\mathrm{MA}}(t)$) of market bid (ask) orders state dependent by setting

$$\eta_{\mathrm{MB}}(t) := \eta_{\mathrm{MO}}\big(1 + I(t)\big) \quad \text{and} \quad \eta_{\mathrm{MA}}(t) := \eta_{\mathrm{MO}}\big(1 - I(t)\big), \tag{4.4}$$

where $t \in \mathbb{R}$ and $\eta_{\mathrm{MO}} \geqslant 0$ is the *average baseline intensity*. Note that (4.4) is consistent with (4.3). We include the imbalance in the baseline intensities of limit orders and cancelations in a similar manner. Our descriptive analysis shows that in those cases, the imbalance dependence is not as clear-cut as for market orders (not illustrated). In the model, however, the imbalance dependence will be distorted by the additional excitement that limit orders and cancelations experience.

4.6 Model formulation

From the previous sections, we have identified a set of non-zero edges in the Hawkes skeleton (4.2), power-law decay of the excitement functions (and thus for the decay kernels), linearity of

impact functions, a general bid–ask symmetry of the model, as well as linear dependence of all conditional intensities on the imbalance. We finally summarize these findings in a parametric Hawkes model. At this point, we drop the general '(i, j)' notation. Instead, we explicitly name the type labels after the order types to simplify tractability and implementation. For $O \in \mathcal{O} := \{\mathrm{MB}, \mathrm{MA}, \mathrm{LB}, \mathrm{LA}, \mathrm{CB}, \mathrm{CB}\}$ and $k \in \mathbb{Z}$, let $(T_k^{(O)}, V_k^{(O)})$ denote the k-th type-O order-book event with time stamp $T_k^{(O)}$ and volume $V_k^{(O)}$, and let $I(t) \in [-1, 1]$ denote the imbalance at time t as in (3.1). (We assume the random variables $I(t)$ are $\mathcal{H}_t$-measurable for $t \in \mathbb{R}$.) We formulate the final parametric model:

Parametric model

For the conditional intensities at time $t \in \mathbb{R}$, we set

$$
\begin{aligned}
\lambda_{\mathrm{MB}}^{(\theta_{MB})}(t) &= \eta_{MB}(1 + I(t-)),\\
\lambda_{\mathrm{LB}}^{(\theta_{LB})}(t) &= \eta_{LB}\big(1 - I(t-)\big) + \sum_{O\in\{\mathrm{MA},\mathrm{CA}\}} m_{O,\mathrm{LB}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{LB}})}(t - T_k^{(O)}),\\
\lambda_{\mathrm{CB}}^{(\theta_{\mathrm{CB}})}(t) &= \eta_{\mathrm{CB}}\big(1 + I(t-)\big) + \sum_{O\in\{\mathrm{LB},\mathrm{MA}\}} m_{O,\mathrm{CB}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{CB}})}(t - T_k^{(O)}),\\
\lambda_{\mathrm{MA}}^{(\theta_{MA})}(t) &= \eta_{MA}(1 - I(t-))\\
\lambda_{\mathrm{LA}}^{(\theta_{LA})}(t) &= \eta_{LA}\big(1 + I(t-)\big) + \sum_{O\in\{\mathrm{MB},\mathrm{CB}\}} m_{O,\mathrm{LA}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{LA}})}(t - T_k^{(O)}), \text{ and}\\
\lambda_{\mathrm{CA}}^{(\vartheta_{\mathrm{CA}})}(t) &= \eta_{\mathrm{CA}}\big(1 - I(t-)\big) + \sum_{O\in\{\mathrm{LA},\mathrm{MB}\}} m_{O,\mathrm{CA}} \sum_{k\in\mathbb{Z}} \frac{V_k^{(O)}}{\mathbb{E}\, V_k^{(O)}} w^{(\vartheta_{O,\mathrm{CA}})}(t - T_k^{(O)}).
\end{aligned}
$$

The decay kernels are parametrized by $\vartheta = (\alpha, x_m) \in (0, \infty)^2$, where

$$
w^{(\vartheta)}(t) := \begin{cases} \alpha x_m^{\alpha}(x_m + t)^{-(1+\alpha)}, & t > 0,\\ 0, & \text{else.} \end{cases} \tag{4.5}
$$

Furthermore, all parameters are bid–ask symmetric, that is,

$$\begin{aligned}
\theta_{\mathrm{MB}} &= (\eta_{\mathrm{MB}}) = (\eta_{\mathrm{MA}}) = \theta_{\mathrm{MA}},\\
\theta_{\mathrm{LB}} &= (\eta_{\mathrm{LB}}, m_{\mathrm{MA,LB}}, m_{\mathrm{CA,LB}}, \vartheta_{\mathrm{MA,LB}}, \vartheta_{\mathrm{CA,LB}})\\
&= (\eta_{\mathrm{LA}}, m_{\mathrm{MB,LA}}, m_{\mathrm{CB,LA}}, \vartheta_{\mathrm{MB,LA}}, \vartheta_{\mathrm{CB,LA}}) = \theta_{\mathrm{LA}},\\
\theta_{\mathrm{CB}} &= (\eta_{\mathrm{CB}}, m_{\mathrm{LB,CB}}, m_{\mathrm{MA,CB}}, \vartheta_{\mathrm{LB,CB}}, \vartheta_{\mathrm{MA,CB}}), \text{ and}\\
&= (\eta_{\mathrm{CA}}, m_{\mathrm{LA,CA}}, m_{\mathrm{MB,CA}}, \vartheta_{\mathrm{LA,CA}}, \vartheta_{\mathrm{MB,CA}}) = \theta_{\mathrm{CA}}.
\end{aligned}$$

Note that in the implementation, we have to make sure that $w^{(\vartheta_{o_1,o_2})}(0) = 0$. For example, in R: `dpareto(`$x_m$`, shape =` α`, scale =` x_m`)` > 0—so this has to be modified. Similarly, it is important that in the likelihood, the imbalance relevant for the conditional intensity at an event time $T_k^{(j)}$ is $I(T_k^{(j)}-)$. That is, we consider the imbalance just before the considered event, not the imbalance that takes the event into account. In total, the model is described by 15 parameters which is still quite a lot. However, this number of parameters is relatively low in comparison to $6^2(1+2+2)+6 = 136$ potential parameters—working with the fully connected Hawkes skeleton with 36 edges, each supplied with one branching parameter, two decay parameters, and two impact parameters. One might even think about choosing the scale parameters x_m fixed. But we found that the estimates of the shape parameters depend quite heavily on this choice; so we prefer to work with free scale parameters.

5 Model calibration

For the six parameter vectors $\theta_{\mathrm{MB}}, \theta_{\mathrm{MA}} \in \mathbb{R}_{>0}$ and $\theta_{\mathrm{LB}}, \theta_{\mathrm{CB}}, \theta_{\mathrm{LA}}, \theta_{\mathrm{CA}} \in \mathbb{R}^7_{>0}$, we calculate the MLE estimates as explained in Section 2.2 (with the obvious extensions to state-dependent baseline intensities, and substituting $\mathbb{E}\,Z^{(O)}$, $O \in \mathcal{O}$, with the average of the observed order sizes in question). Note that we optimize the likelihoods corresponding to the six order types *separately*. As explained, we assume that these vectors are pairwise equal (by the symmetry assumption). The average of these estimation pairs yields the final estimate. For example, we estimate θ_{MB} and θ_{MA} separately. For the final estimate of θ_{MO} $(= \theta_{\mathrm{MA}} = \theta_{\mathrm{MB}})$, we set $\hat{\theta}_{\mathrm{MO}} := (\hat{\theta}_{\mathrm{MB}} + \hat{\theta}_{\mathrm{MA}})/2$. We treat the other parameter vectors in the same way. We apply this procedure to data windows containing exactly 1000 events—the actual length of these windows ranging from a few seconds to several minutes. In this manner, the calibration of the model on *one* of the 61 considered trading days takes five computing hours (parallelized on 24 cores). This is why we do not consider more events per estimation window. For some of these 1000-events samples the likelihood-optimization algorithm does not converge. We omit these estimates

(about 4% of all estimation results). In a first step, we average over all of these estimations. Note that we take all averages in a time-weighted manner—otherwise we would overweight very active periods. In a second step, we average the estimation results separately over all 9:30h–9:45h windows, over all 9:45h–10:00h windows, etc. This allows us to examine the dynamics of the model over a typical trading day. We only present the results for the MSFT15 dataset; we found that the results for the INTC15 dataset are quasi equivalent.

5.1 Average calibration

We present the overall average estimation results. Tables 3 and 4 give the estimated parameter values; Figure 5 illustrates the estimated average Hawkes graph derived from the MLE estimates. Obviously, the excitements (= branching coefficients) from market orders to limit orders and cancelations on the other side of the book are by far the most important. The excitements from cancelations to limit orders and vice versa are much smaller, though significantly larger than zero. Furthermore, they are of exactly the same order. The spectral radius of the average branching-matrix estimate is approximately 0.5. The estimated shape parameters of the decay kernels are quite large, indicating a 'fast power-law' decay. In particular, the shape parameters are all larger than 1. So we may consider '$x_m/(\alpha - 1)$', the expected value of the distribution defined by the estimated decay kernels in (4.5). These expected values measure the persistence of the excitement over time. We consider these *mean excitement times* more meaningful than the values of the scale-parameter estimates. The estimated mean excitement times are all fractions of a second. The shortest mean excitement time corresponds to the excitement from limit orders to cancelations: together with the estimated value of the corresponding branching-coefficient estimated, 0.50, this can be interpreted that every other limit order is canceled an instant after its submission.

Our results for the spectral radius and for the shape of the power-law decay are quite different from the results in the relevant literature with values of $\alpha \in (0, 1)$ and the spectral radius near the critical case 1; see Hardiman, S.J. et al. (2013) or Bacry et al. (2014). We suppose that the inclusion of order size and, even more important, state dependence is the reason for these somewhat diverging results. To put it differently, the often observed near criticality and long-range dependence property of Hawkes-model fits may presumably be explained by ignoring relevant covariates in the model. This lack of explanatory variables is compensated by large branching coefficients and very heavy-tailed decay kernels of infinite expectations. In this context, also note that we have fitted the model on quite small time windows for computational reasons. Calibrations on single larger windows did not change the results significantly. In any case, the literature referred to above says that the decay shape parameter is also very small for very small waiting times. So small shape parameters would be detected in the short time windows.

Table 3 provides the average baseline-intensitiy estimates for market orders, limit orders

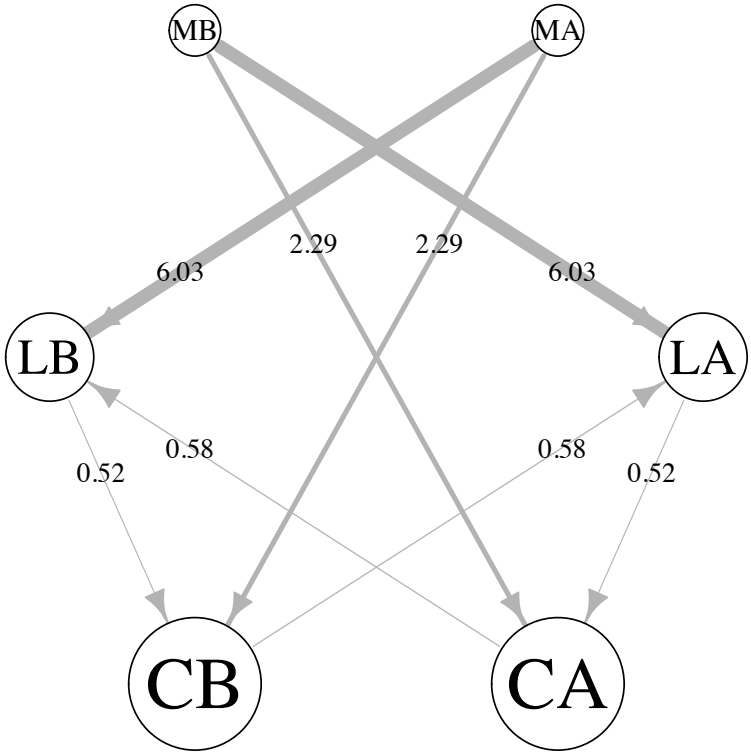


Figure 5: Average Hawkes graph of the MSFT15 dataset based on the average MLE estimates. The vertex diameters are proportional to the vertex weights, that is, to the corresponding baseline-intensity estimates.

and cancelations. The baseline intensities are of surprisingly similar order for all event types—though there are far fewer market orders than limit orders or cancelations in the data. This is reflected by the fact that every market bid order of average size (in this model) triggers

$$\sum_{n\geqslant 0}\sum_{O\in\mathcal{O}}^{6}(\hat{M}^n)_{\mathrm{MB},O} = \sum_{O\in\mathcal{O}}\left((1_{6\times 6}-\hat{M})^{-1}\right)_{\mathrm{MB},O} \approx 17 \tag{5.1}$$

subsequent orders (counting also the triggering market order)—when $\hat{M} := (\hat{m}_{i,j})_{(i,j)\in[d]}$ denotes the estimated branching matrix. We compare the influence of specific order types on the whole system by calculating the so called *cascade coefficients*; see Definition 7 in Embrechts and Kirchner (2017). That is, for *each* order type, we first calculate its influence on the system as in (5.1) and then compare the resulting numbers (weighted by the individual activity of the order type in question, that is, weighted by the corresponding baseline intensities). We give the resulting estimated values in Table 3. E.g., for market bid orders, the estimated cascade coefficient is 0.33. The interpretation is that if we suppressed all market bid orders, then, in the model, the total activity of the six considered order streams would be reduced by 43%. In particular, if we suppressed *all* market orders, the activity of the LOB would decrease by 66%. This strong decrease seems to make sense: if nobody actually buys or sells anything, there is no use in sending offers to buy or sell and the market eventually 'stops'.

	$\hat{\eta}$	cascade-coefficient estimate
MB	1.06 (0.48)	0.33 (0.01)
LB	2.1 (0.32)	0.08 (0.01)
CB	2.5 (0.53)	0.09 (0.01)

Table 3: Average baseline-intensity and cascade-coefficient estimates (standard deviations in brackets). The relatively large standard deviations reflect the variability of the parameter values over the trading day.

	$\hat{m}$	$\hat{\alpha}$	$\hat{x}_m$	mean-displacement estimate
MB → LA	5.33 (0.54)	6.42 (0.69)	0.13 (0.02)	0.026 (0.007)
MB → CA	2.38 (0.16)	4.71 (0.42)	0.6 (0.08)	0.165 (0.038)
LB → CB	0.5 (0.02)	18.95 (2.4)	0.02 (0.01)	0.001 (0.007)
CB → LA	0.53 (0.08)	6.33 (1.17)	0.2 (0.14)	0.054 (0.095)

Table 4: Average excitement-parameter estimates (standard deviations in brackets). Note that most of the standard deviations in this table are significantly smaller than the ones for the baseline-intensity estimates in Table 3. This reflects the fact that the excitement parameters are more stable over a trading day than the baseline intensities.

5.2 Parametric representation of a trading day

Instead of averaging over all estimates, we now average over the trading days separately for all 15 min data windows, that is, for 9:30am–9:45am, all 9:45am–10:00am, etc. This procedure

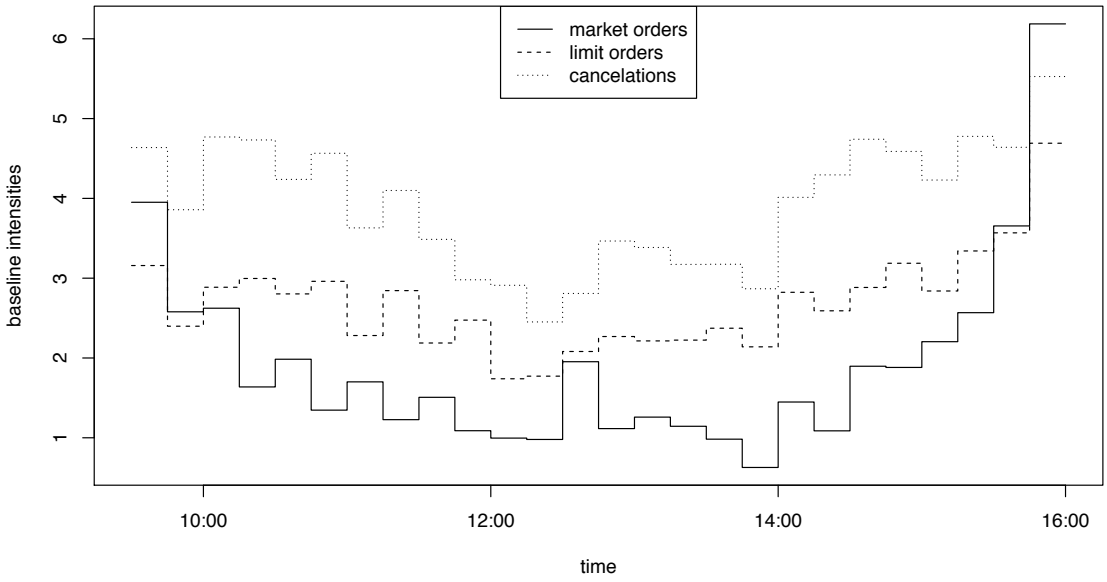


Figure 6: Baseline-intensity estimates for an average trading day.

yields 26 estimates for each considered parameter. We illustrate the results in Figures 6, 7, and 8.

Average baseline intensities

The estimates for the average baseline intensities are displayed in Figure 6. For all order types, we observe the typical U-shape already noticed for the descriptive order-arrival intensity.

Branching coefficients

Figure 7 illustrates the development of the branching-coefficient estimates over the trading day as well as the corresponding spectral-radius estimates. For nearly all time windows the model calibration is bounded away from criticality, that is, the spectral radius is always significantly less than 1. For the heaviest edges from market to limit orders and from limit orders to cancelations, the estimated coefficients exhibit a slight inverted U-shape whereas the estimates corresponding to edges from limit orders to cancelations and from cancelations to limit orders exhibit U-shapes. Thus, near the opening and the closing, cancelations and limit orders 'react' (a bit) less sensitive on market orders than during the day. On the other hand, the excitements between limit orders and cancelations are heavier during the trading day. In any case, between 11am and 15pm, all branching-coefficient estimates are roughly constant over the different time windows.

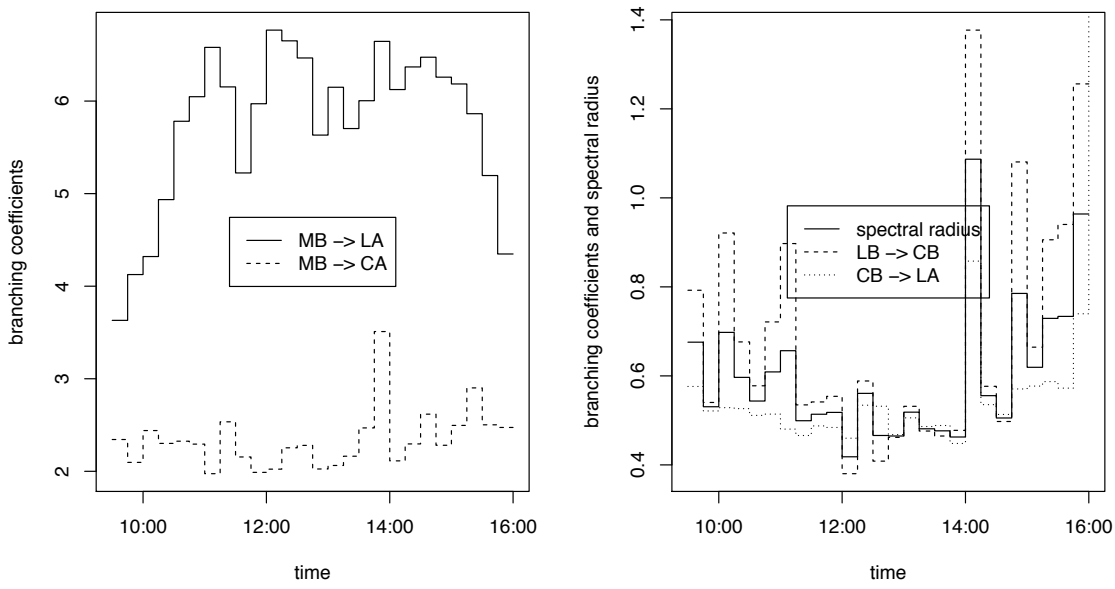


Figure 7: Branching-coefficient and spectral-radius estimates for an average trading day.

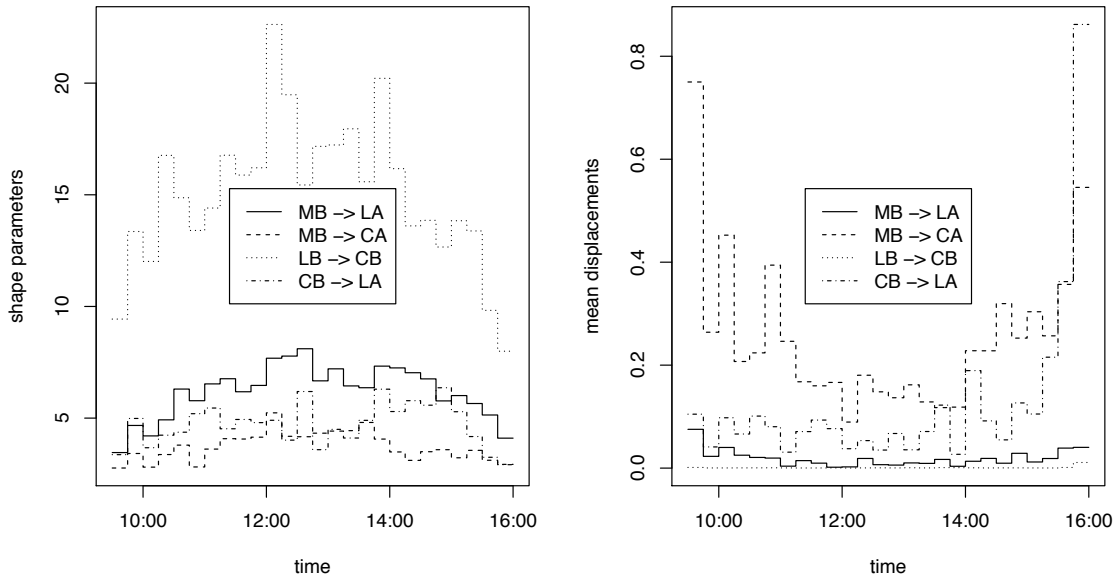


Figure 8: Decay-parameter estimates over an average day: rather than giving the estimated scale parameters 'x_m', we give the estimates for $x_m/(\alpha - 1)$, that is, the mean excitement-time in seconds.

Decay parameters

Figure 8 illustrates the development of the decay-parameter estimates of an average trading day. We observe an inverted U-shape for the shape parameters and a U-shape for the mean excitement-times (= 'scale / (shape - 1)). So, though the market is overall more active near opening and closing, the 'reaction times' seem to become quicker during the trading day and slow down before closing. Throughout, we observe the smallest mean excitement-time estimates for the edges from limit orders to cancelations.

6 Application: order-type prediction

We may use our calibrated order-arrival model to develop a predictor for the next order-book event. E.g., one can consider predictions on whether the next order arrives on the bid or ask side, predictions on the kind of the next order (market order, limit order, or cancelation), or predictions on the next actual order type $O := \{\mathrm{MB}, \mathrm{LB}, \mathrm{CB}, \mathrm{MA}, \mathrm{LA}, \mathrm{CA}\}$. For all cases, one uses that the conditional order-type distribution of the model reads

$$p_t(A) := \mathbb{P}\big[\text{Order type } O \in A \text{ at time } t \,\big|\, \text{Order at time } t\big] = \frac{\sum_{O \in A} \lambda_O(t-)}{\sum_{\tilde{O} \in \mathcal{O}} \lambda_{\tilde{O}}(t-)}, \quad A \subset \mathcal{O}. \tag{6.1}$$

Substituting the conditional intensities with the corresponding (local) estimates from our calibrated model yields a conditional order-type probability $\hat{p}_t(A)$. We predict A if $\hat{p}_A(t)$ is larger than a given threshold α. This procedure works especially well for limit orders. Here—that is, for $A = \{\mathrm{LB}\}$ or $A = \{\mathrm{LA}\}$ and $\alpha = 0.5$—we found that we get about 50% false positive predictions and less than 10% false negative predictions. The model predicts bid and ask cancelations similarly well. There are hardly any moments in time, where $p_{\{\mathrm{MB}\}}(t)$ or $p_{\{\mathrm{MA}\}}(t)$ is larger than 0.5; so prediction for market orders makes no sense in this procedure. Naturally, the performance could be optimized by using a local (maybe even non-symmetric) model.

7 Discussion

We specified a fully parametric Hawkes-process based model for order arrivals in limit order books. The various results give a compact, yet meaningful summary of the large dataset. Furthermore, the model turns out to be useful for order-type prediction, if desired. We use the LOB imbalance as an explanatory variable for the determination of the side of the book on which a market order arrives. Thus, we add state-dependence to the otherwise purely autoregressive structure of the model. Working with impact functions includes the intuition that large orders lead to higher numbers of subsequent orders than small ones. Because of the linear choice of

impact functions, these additional explanatory variables do not increase model complexity (in terms of number of parameters). The final calibration of the model to LOB data is in line with descriptive analysis; the calibrated model seems reasonable from an economic perspective. E.g., market orders are identified as the main driving force behind the order arrival process. They excite limit orders and cancelations, but get hardly excited by any other orders themselves. If we could suppress market orders, the market (model) would eventually nearly stop.

Clearly, our analysis depends on many choices, such as the window size for the MLE, or the discretization and truncation parameters for the nonparametric estimation method. Due to the large dataset (and the large number of choices), we were not able to give a more systematic sensitivity analysis. Note, however, that we experimented quite freely with different parameters and window sizes on single time windows. We found that the results do not change significantly: the selected non-zero edges of the estimated Hawkes skeletons are stable, and the Hawkes-graph weights are stable in a sense that the order of magnitude is hardly affected. The same is true for the other parameters. Furthermore, the results are almost equivalent over the two examined datasets.

Note that the specified model from Section 4.6 is incomplete: we did not consider the modeling of the order sizes, but treated the volumes as given. As subsequent order sizes are strongly correlated, this would ask for separate modeling. Furthermore, if we wanted to simulate from the model, we would also have to keep track of the actual number of orders at the best bid and best ask to be able to calculate the order book imbalance. Then, however, after a price jump, one would also have to model the size of the queues at the new best bid, respectively, new best ask. In view of these complications, we have restricted ourselves to the case where we treat the marks as given quantities.

One could incorporate price jumps into our model, e.g., by including the event streams of bid- and ask-price jumps. These would be extremely rare events compared to the order flow. It would be interesting to examine the relevant in- and outgoing edges of the 'price-jump vertices' of the corresponding Hawkes graph. It seems obvious that the price-jump intensities will be intimately connected with the orderbook imbalance.

The descriptive analysis and the model specification in this paper are carried out for large-tick assets, where one price tick is relatively large compared to the price of the asset. If we want to extend the analysis to small-tick assets, we will have to reconsider the model. E.g., in the large-tick case, the spread is often larger than one tick. A possible way to deal with this situation would be to scale the imbalance by the spread size as follows

$$\tilde{I}(t) := (P^{(a)}(t) - P^{(b)}(t)) \frac{Q^{(a)}(t) - Q^{(b)}(t)}{Q^{(b)}(t) + Q^{(a)}(t)} \quad (\in \mathbb{R}), \tag{7.1}$$

and then use this scaled version as an explanatory variable for the baseline intensities.

It might be tempting to use the predictive power of our model for building high-frequency-trading strategies. This has to be handled with care as any intervention will presumably change the stochastic rules of the LOB. Here, a promising approach would be to analyze labelled datasets that allow identification of the agents involved. This would open the door to statistical analysis that takes interventions into account.

Paper

F

Matthias Kirchner.

A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation.

Submitted.

A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation

M. Kirchner and A. Bercher

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

In earlier work (Kirchner (2017a)), we introduced a nonparametric estimation method for the Hawkes point process. In this paper, we present a simulation study that compares this specific nonparametric method to maximum-likelihood estimation. We find that the standard deviations of both estimation methods decrease as power-laws in the sample size. Moreover, the standard deviations are proportional. E.g., for a specific Hawkes model, the standard deviation of the branching-coefficient estimate is roughly 20% larger than for MLE—over all sample sizes considered. This factor becomes smaller when the true underlying branching coefficient becomes larger. In terms of runtime, our method clearly outperforms MLE. The present bias of our method can be well explained and controlled. As an incidental finding, we see that also MLE estimates seem to be significantly biased when the underlying Hawkes model is near criticality. This asks for a more rigorous analysis of the Hawkes likelihood and its optimization.

1 Introduction

The Hawkes point process has been introduced in Hawkes (1971b,a); Hawkes and Oakes (1974); it is a model for (possibly multitype) event streams. For a comprehensive discussion, see Liniger (2009). The monograph Daley and Vere-Jones (2009) also discusses many properties of the Hawkes process. Today, large event-stream datasets are omnipresent, with the Hawkes process as a popular modeling tool—typical applications can be found in high-frequency trading, neurology, or internet traffic. For such large datasets, however, maximum likelihood estimation (MLE) for calibration of a Hawkes process can not be applied in a straightforward manner because the evaluation of the likelihood quickly becomes numerically complex. As possible solutions, one then relies on E–M algorithms or one considers a special Hawkes model that allows

recursive calculation for the likelihood. At the same time, these large datasets open the door to nonparametric estimation methods. In Kirchner (2017a), we introduce a simple nonparametric estimation procedure (NPE) for multivariate Hawkes point processes. Our method consists of three steps:

1. *Discretization*: given a point process sample, divide the time-line into bins and calculate the 'bin-count sequence'.
2. *Optimization*: fit a linear autoregressive model on the bin-count sequence via conditional-least-squares.
3. *Normalization*: retranslate the resulting autoregression estimates to parameters of the Hawkes model.

In the paper Kirchner (2017a), we performed a bivariate simulation study on the method to study bias, asymptotic normality, and validity of variance estimates. In Embrechts and Kirchner (2017), we present a simulation study on a 10-dimensional Hawkes process that concentrates on the application of our method on the detection of *significant* excitement from one event stream to another. In the present paper, we aim to close a gap by comparing NPE to the MLE benchmark with respect to different Hawkes models. As a third method, we also consider a semiparametric approach (SPE). For the latter, we add information on the parametric model to the NPE method. In our study, we only consider univariate Hawkes processes starting at zero. We found that otherwise

- the influence of the numerical optimization of the likelihood (like choice of optimization method, choice of starting values, choice of optimization set boundaries) and
- edge effects (choice of burn-in times, approximation of likelihood)

have strong effects on the estimation results; see Section 4.2.

We find that for all methods, standard deviations and root mean squared errors decrease as a power-law in the sample size. Moreover, the standard deviations of the three examined methods are roughly proportional. E.g., for the branching coefficient estimate, the standard deviation of NPE is approximately 20% times larger than for MLE (over all considered sample sizes). In addition, we find that the effect of increasing the sample size on computation time is much larger for MLE. From the analysis, we conclude that for smaller sample sizes (e.g., less than 200 observed points), MLE is clearly the most suitable method. However, we have to keep in mind that the comparison is not fair: MLE assumes perfect a-priori information on the underlying parametric model whereas NPE works with zero information. This imbalance of pre-knowledge makes the good performance of NPE even more interesting. Another important and somewhat surprising finding is that SPE (i.e. NPE, fed with additional information) *increases variance*.

We conclude that SPE is not necessarily the best way to further develop our NPE method.
Our paper is organized as follows: in the second section, we introduce notation and recall the definition of Hawkes processes and their estimation. In the third section, the main part of the paper, we present set-up and results of the simulation study. In the fourth section, we conclude with implications of our findings on the estimation of Hawkes processes.

2 Material and methods

2.1 Hawkes processes

A *point process* N is a measurable mapping from some probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ into $(M_p, \mathcal{M}_p)$, where M_p denotes the space of locally-finite counting measures on the bounded Borel sets $\mathcal{B}_b(\mathbb{R})$ in $\mathbb{R}$, and

$$\mathcal{M}_p := \sigma\Big(\big\{m \in M_p : \ m(A) = n\big\} : \ A \in \mathcal{B}_b(\mathbb{R}), n \in \mathbb{N}_0\Big).$$

For any point process $N : (\Omega, \mathcal{F}, \mathbb{Q}) \to (M_p, \mathcal{M}_p)$, we set

$$\mathcal{H}_t^{(N)} := \sigma\Big(\big\{\omega \in \Omega : N(\omega)(A) = k\big\} : \ A \in \mathcal{B}_b\big((-\infty, t]\big), \, k \in \mathbb{N}\Big), \quad t \in \mathbb{R}\,.$$

Note that, $\mathcal{H}_t^{(N)} \subset \mathcal{F}$, $t \in \mathbb{R}$. We say N is simple if $\mathbb{Q}\big[N(\{t\}) \in \{0,1\}, \, t \in \mathbb{R}\,\big] = 1$. For $a < b$, we write

$$\int_a^b f(t) N(\mathrm{d}t) := \sum_{\substack{a<t\leqslant b:\\ 0<N(\{t\})}} f(t) N(\{t\}) \quad (\leqslant \infty).$$

In this paper, we only consider Hawkes processes that are void on the negative halfline: a *Hawkes process* is a simple point process N with $N\big((-\infty, 0]\big) = 0$ and

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}\Big[N\big((t, t+\delta]\big)\Big|\mathcal{H}_t^{(N)}\Big]}{\delta} = \eta + \int_0^t m w(t-s) N(\mathrm{d}s), \quad t > 0. \tag{2.1}$$

Here, $w : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, such that $\int w(t)\mathrm{d}t = 1$ and $w(t) = 0$, $t \leqslant 0$, is the *displacement density* and m, $0 \leqslant m < 1$, is the *branching coefficient*. Furthermore, $\eta > 0$ is the (constant) *immigration intensity*. This terminology stems from the fact that Hawkes processes can be represented as branching random walks with immigration; see Hawkes and Oakes (1974) and Shi (2015). One can show that (2.1) above determines a unique distribution on $(M_p, \mathcal{M}_p)$.

2.2 Maximum likelihood estimation

We consider a realization $N_{[0,T]}$ of a point process N on a time window $[0,T]$. The *log-likelihood* of a parametric Hawkes model (η, m, θ), that is, with immigration intensity η, branching coefficient m and parametric displacement density $w_\theta \in (w_{\theta'})_{\theta' \in \Theta}$ with respect to the realization $N_{[0,T]}$ is

$$l^{(N_{[0,T]})}(\eta, m, \theta) = \int_0^T \log\left(\lambda_{\eta,m,\theta}^{(N_{[0,T]})}(t)\right) N_{[0,T]}(\mathrm{d}t) - \int_0^T \lambda_{\eta,m,\theta}^{(N_{[0,T]})}(s)\mathrm{d}s, \tag{2.2}$$

where $\lambda_{\eta,m,\theta}^{(N_{[0,T]})}(t) = \eta + \int_0^t m w_\theta(t-s) N_{[0,T]}(\mathrm{d}s)$. Remember that we assume Hawkes processes have no mass on $(-\infty, 0]$. Also note that we assume $w(0) = 0$ for displacement densities w. We set

$$(\hat{\eta}^{(\mathrm{MLE})}, \hat{m}^{(\mathrm{MLE})}, \hat{\theta}^{(\mathrm{MLE})}) := \mathrm{argmax}_{(\eta,m,\theta) \in \mathbb{R}_{\geqslant 0} \times [0,1) \times \Theta} \, l^{(N)}(\eta, m, \theta). \tag{2.3}$$

Note that the complexity of the likelihood (2.2) grows quadratically in the number of observed points. We calculate (2.3) with the `R` function `optim` applying the method '`L-BFGS-B`'. The algorithm allows bounded optimization regions and is described in Byrd et al. (1995).

2.3 Nonparametric estimation

Procedure

The estimation procedure from Kirchner (2017a) consists of three steps: discretization, optimization, and normalization. Let $N_{[0,T]}$ be a realization of a Hawkes process N on a time window $[0,T]$ with immigration intensity η, branching coefficient m and displacement density w. Fix *estimation parameters* Δ, $0 < \Delta << T$, Δ 'small', as well as $p \in \mathbb{N}$, $p\Delta$ 'large'.

1. *Discretization:* calculate the bin-count sequence $(X_n^{(\Delta)})$, that is,

$$X_n^{(\Delta)} := N_{[0,T]}\Big(\big((n-1)\Delta, n\Delta\big]\Big), \quad n = 1, 2, \ldots, \lfloor T/\Delta \rfloor. \tag{2.4}$$

2. *Optimization:* set

$$(\hat{\alpha}_0^{(\Delta)}, \hat{\alpha}_1^{(\Delta)}, \ldots, \hat{\alpha}_p^{(\Delta)}) := \mathrm{argmin}_{(\alpha_k) \in \mathbb{R}^{p+1}} \sum_{n=p+1}^{\lfloor T/\Delta \rfloor} \Big(X_n^{(\Delta)} - \alpha_0 - \sum_{k=1}^{p} \alpha_k X_{n-k}^{(\Delta)}\Big)^2. \tag{2.5}$$

3. *Normalization:*

$$\hat{\eta}^{(\mathrm{NPE})} := \frac{\hat{\alpha}_0^{(\Delta)}}{\Delta}, \quad \hat{m}^{(\mathrm{NPE})} := \sum_{k=1}^{p} \hat{\alpha}_k^{(\Delta)}, \quad \hat{w}_k^{(\Delta)} := \frac{\hat{\alpha}_k^{(\Delta)}}{\Delta \hat{m}^{(\mathrm{NPE})}},\ k = 1, 2, \ldots, p. \tag{2.6}$$

Note that $\hat{w}_k^{(\Delta)}$ is an estimate for $w(k\Delta)$, respectively, for $w((k-0.5)\Delta)$—taking a discretization correction into account. Furthermore, if we assume that the data-generating displacement density w lies in a parametric density family $(w_{\theta'})_{\theta' \in \Theta}$, we estimate the unknown parameter θ from

$$\hat{\theta}^{(\mathrm{NPE})} := \mathrm{argmin}_{\theta \in \Theta} \sum_{k=1}^{p} \Big(w_\theta\big((k-0.5)\Delta\big) - \hat{w}_k^{(\Delta)} \Big)^2. \tag{2.7}$$

For optimization in (2.7), we apply the R-function `nls` with an algorithm from the `port` library. Least-squares optimization in (2.5) obviously results in matrix inversion. For small Δ, the design matrix of the problem becomes very sparse. Specialized software—such as the R-package 'Matrix'—makes construction and manipulation of such sparse matrices numerically efficient; see Bates and Maechler (2015). The estimates obtained in (2.6) and (2.7) depend on the choice of the estimation parameters Δ and p. In Kirchner (2017a), we discuss methods for the optimal choice of Δ and p. Throughout the present study, we keep the parameters fixed at $\Delta = 10^{-2}$ and $p = 400$.

Error analysis

Equation (2.5) assumes an obviously erroneous (approximative) model for the bin counts. The wrong model assumptions will typically cause the estimates to be biased. Therefore, it is desirable to have a measurement of 'how erroneous' the approximative model is. First of all, we decompose and quantify the error in three parts: Given a Hawkes model (η, m, w) and estimation parameters Δ and p, we define

- *the discretization error*

$$e_1 := m \left| \int_0^{p\Delta} w(t)\mathrm{d}t - \sum_{k=1}^{p} \Delta w(k\Delta) \right| \tag{2.8}$$

  that results from the discrete sampling of the reproduction density in (2.5),

- the *cut-off error*

$$e_2 := \frac{m\eta}{1-m} \int_{p\Delta}^{\infty} w(t)\mathrm{d}t \tag{2.9}$$

that results from ignoring the influence of the bin counts $X_{n-k}^{(\Delta)}$, $k > p$, on the counts of bin n in (2.5), and

- the *distributional error*

$$e_3 := \frac{m\eta}{1-m}\frac{\Delta}{2}w(0+) \tag{2.10}$$

that results from ignoring (typically) important explanatory variables in (2.5).

Below, we clarify the rationale behind (2.10): suppose we observe two events in a bin ('$X_k^{(\Delta)} = 2$'). In the original Hawkes model, the first event typically increases the conditional intensity; it may well be that the second event is explained by this increase. However, in the approximating equation (2.5), we ignore this possibility and aim to explain both events by events in *earlier* bins or by the constant term. To quantify the distributional error, we approximate the contributions to the conditional intensity that are ignored in (2.5), averaged over some bin k:

$$\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta}\int_{(k-1)\Delta}^{t} mw(t-s)N(\mathrm{d}s)\mathrm{d}t \approx \frac{1}{2}mw(0+)X_k^{(\Delta)}.$$

Taking expectations (and assuming stationarity for the sake of simplicity), yields (2.10). Note that all errors are scaled by m: decreasing m decreases the influence of the errors. It is clear that all errors can be made arbitrarily small by letting $\Delta \downarrow 0$ and $p\Delta \uparrow \infty$. Finally note that the errors can only be calculated when the true model is given (or estimated). We find that the cut-off (e_2) and the distributional error (e_3) that stem from ignoring explanatory variables in (2.5) are typically compensated by overestimation of the baseline intensity. That is, we typically have that $\mathbb{E}\,\hat{\eta}^{(\mathrm{NPE})} \approx \eta + e_1 + e_2$.

2.4 Semiparametric estimation

We also consider a semiparametric estimation method, SPE, where we fit a parametric Hawkes model to the NPE results from Section 2.3: given a point-process sample $N_{[0,T]}$ as described above (2.4), we first calculate the nonparametric estimates $(\hat{\alpha}_k^{(\Delta)})_{k=1,\dots,p}$ from (2.5). As in (2.7), we assume that the underlying Hawkes displacement density belongs to a parametric family $(w_\theta)_{\theta\in\Theta}$. This time, however, we fit *all* parameters anew in the light of this additional information: namely, we set

$$\big(\hat{m}^{(\mathrm{SPE})}, \hat{\theta}^{(\mathrm{SPE})}\big) := \operatorname{argmin}_{(m,\theta)\in[0,1]\times\Theta} \sum_{k=1}^{p}\Big(\hat{\alpha}_k^{(\Delta)} - mw_\theta\big((k-0.5)\Delta\big)\Big)^2 \tag{2.11}$$

and we adapt the immigration-intensity estimate accordingly:

$$\hat{\eta}^{(\mathrm{SPE})} := N\big([0,T]\big)\big(1-\hat{m}^{(\mathrm{SPE})}\big). \tag{2.12}$$

Note that $\hat{\eta}^{(\mathrm{SPE})}$ in (2.12) is defined in such a way that the number of observed points corresponds approximately to the expected number of points of the estimated model. For optimization in (2.11), we apply the `R`-function `nls` with the algorithm from the `port` library. Note that the optimization problem in (2.11) is typically far less complex than the likelihood optimization in (2.3).

3 Simulation study

3.1 Setup

We consider four different Hawkes models, namely

- subcritical with exponential decay,
- nearly critical with exponential decay,
- subcritical with power-law decay, and
- nearly critical with power-law decay.

More explicitly, we vary the branching coefficient m over $\{0.4, 0.95\}$. The terms *subcritical*, *nearly critical* refer to cases $m = 0.4$, $m = 0.95$, respectively. Furthermore, we consider two different decay shapes for the displacement densities: the terms *exponential decay*, *power-law decay* refer to displacement densities of the form $w_\theta(t) = 1_{t>0}\theta\exp(-\theta t)$, $w_\theta(t) = 1_{t>0}\theta(1+t)^{-(1+\theta)}$, respectively. In both decay cases, the true underlying parameter θ is equal to 1.5. For all four models, we set the immigration intensity $\eta := 1-m$. Thus, the number of points in $[0,T]$ will be approximately T. For the main study, we consider $T = 500$. As a second step, we also vary T over $[200, 2'000]$. We simulate $nsim = 1'000$ times from each model. For each simulation, we calculate the estimates for the three methods MLE, NPE, and SPE. From these collections of estimates, we calculate mean, standard deviation, and root mean squared error.

For the estimation parameters of the nonparametric method, we choose $\Delta = 10^{-2}$ and $p = 4/\Delta = 400$. To make the results comparable, we keep Δ and p fixed. In particular, this allows us to study the effect of a large cut-off error in the critical power-law case; see the discussion on the cut-off error in Section 3.2 as well as Table 1. The bin-size $\Delta = 0.01$ seems small enough.

Table 1: Overview for the error measurements related to the NPE-method; see Section 3.2. The NPE-method depends on the choice of discretization parameter Δ and a cut-off parameter s. The three errors below indicate how the choices contribute to potential bias of the estimates.

	error		
model	discretization e.	cut-off e.	distributional e.
subcritical, exponential decay	0.003	0.001	0.003
nearly critical, exponential decay	0.007	0.002	0.007
subcritical, power-law decay	0.004	0.036	0.003
nearly critical, power-law decay	0.007	0.085	0.007

3.2 Results

We collect the distributional results of the study in Tables 2 and 3. In Figure 1, we compare the performance of the estimators relative to the sample size. In Figure 2, we compare the runtime of the different methods.

Bias

We collect the three errors with respect to the four models considered in Table 1. Comparing these errors with Tables 2 and 3, we see that the sum of cut-off error (e_2) and distributional error (e_3) yields approximately the bias of $\hat{\eta}^{(\mathrm{SPE})}$—as expected. The cut-off error is naturally larger for the power-law decay. Accordingly, in the power-law decay case, we underestimate m and overestimate η quite heavily. Clearly, this bias is mainly due to the cut-off error; it can be adjusted by applying a larger parameter p in the estimation procedure. For the SPE-estimates, a part of this bias vanishes—maybe due to some kind of smoothing effect. Note that t-tests clearly reject the null-hypothesis of the MLE estimates being unbiased (not illustrated). That is, MLE-estimates of Hawkes parameters *are* considerably biased—even for our rather large sample sizes (ca. 500 points).

Variance

As expected, standard deviation for MLE is smaller than for NPE and SPE; see Tables 2 and 3 and Figure 1. Remember that for SPE, we apply the parametric model information on the NPE estimates. Surprisingly, this additional information does not yield a lower variance for the estimate of m. All in all, we observe that '$\mathrm{sd}(\mathrm{MLE}) < \mathrm{sd}(\mathrm{NPE}) < \mathrm{sd}(\mathrm{SPE})$'—for all models and all parameters (with few less important exceptions). In the exponential, nearly critical case, we perform simulations for different sample-window sizes $T \in \{200, 300, 400, 500, 1000, 2000\}$. In Figure 1, we find that for all methods and all parameters, standard deviation decreases as a power-law in the sample size. In the same figure, we see that the differences of the logarithmic standard deviations of the methods stay roughly constant over different sample window sizes

T. That is, the standard deviations are roughly proportional. We find that, independently of the sample window size,

$$\frac{\mathrm{sd}(\hat{\eta}^{(NPE)})}{\mathrm{sd}(\hat{\eta}^{(MLE)})} \approx 2.5, \quad \frac{\mathrm{sd}(\hat{m}^{(NPE)})}{\mathrm{sd}(\hat{m}^{(MLE)})} \approx 1.2, \quad \frac{\mathrm{sd}(\hat{\theta}^{(NPE)})}{\mathrm{sd}(\hat{\theta}^{(MLE)})} \approx 1.9. \tag{3.1}$$

Figure 1 as well as the numbers in (3.1) above refer to the exponential, nearly critical model. We also performed the same analysis for the three other models. We always find proportionality of standard deviations. However, the proportionality factors depend on the underlying models. The factors become smaller when the underlying branching-coefficient m increases (not illustrated). Note that for all models, we have that $sd(\hat{m}^{(NPE)})/sd(\hat{m}^{(MLE)})$ is the smallest of the three parameters estimates. Also note that we have shown already in paper Kirchner (2017a), that the variances of the aggregated parameters are hardly affected by the choice of the bin-size Δ.

Runtime

In the exponential, nearly critical case, we perform simulations for different sample-window sizes $T \in \{100, 200, 300, 400, 500, 1000, 2000\}$. Figure 2 shows how MLE runtime exhibits polynomial growth whereas SPE runtime growth is linear. The runtime differences increase dramatically in the multivariate case (not illustrated). For MLE, we could simplify calculations by applying a cut-off in (2.2). However, such cut-offs do not improve the MLE runtime significantly (not illustrated). Furthermore, such a cut-off would sacrifice MLE's better performance in terms of bias. In the (very special) exponential decay case, we could also calculate the likelihood values at the observation points recursively. Finally note that for our SPE method, we are numerically more flexible as we may not only vary the cut-off $s = p\Delta$ but also the discretization parameter Δ.

Estimation parameters

We apply $\Delta = 0.01$ and $p = 4/0.01 = 400$ throughout to make the results comparable. For the critical power-law case, a larger p would be preferable; see the discussion on the cut-off error as well as Table 1. Increasing p decreases the bias. With respect to the discretization and distributional error, $\Delta = 0.01$ seems small enough.

Table 2: Results of the simulation study for exponential decay; we consider a subcritical regime ($m = 0.4$, upper three rows), as well as a nearly critical regime ($m = 0.95$, lower three rows). The sample window size is $T = 500$. Clearly, in terms of root mean squared error, MLE performs best—for all parameters and both models. For the nearly critical model, the effect of the cut-off error on the NPE and the SPE is eye-catching: we underestimate m and overestimate η. Note, however, that also MLE clearly underestimates m.

	true	mean			sd			rmse		
		MLE	NPE	SPE	MLE	NPE	SPE	MLE	NPE	SPE
η	0.60	0.607	0.619	0.614	0.061	0.074	0.068	0.062	0.076	0.069
m	0.40	0.391	0.383	0.387	0.058	0.069	0.063	0.058	0.071	0.064
θ	1.50	1.572	1.604	1.617	0.396	0.437	0.516	0.402	0.449	0.528
η	0.05	0.051	0.081	0.079	0.012	0.028	0.072	0.012	0.041	0.077
m	0.95	0.903	0.870	0.872	0.069	0.080	0.110	0.083	0.114	0.134
θ	1.50	1.523	1.589	1.665	0.213	0.376	0.706	0.215	0.387	0.725

Table 3: Results of the simulation study for power-law decay; we consider a subcritical regime ($m = 0.4$, upper three rows) as well as a nearly critical regime ($m = 0.95$, lower three rows). For both models, NPE performs better than SPE—despite the SPE-advantage of additional parametric information. As in Table 2, we see that also MLE underestimates the nearly critical branching coefficient m considerably. This is hard to explain (and hard to control) whereas the bias of NPE and SPE can be explained by the cut-off error. Also note that the cut-off reduces the standard deviation of the estimates: for the subcritical model, NPE has an even lower variance than MLE.

	true	mean			sd			rmse		
		MLE	NPE	SPE	MLE	NPE	SPE	MLE	NPE	SPE
η	0.60	0.616	0.655	0.599	0.085	0.078	0.126	0.086	0.096	0.126
m	0.40	0.382	0.344	0.401	0.083	0.071	0.124	0.085	0.090	0.124
θ	1.50	1.737	1.883	1.763	0.622	0.638	0.964	0.665	0.744	0.998
η	0.05	0.053	0.104	0.063	0.016	0.040	0.074	0.017	0.068	0.075
m	0.95	0.902	0.836	0.894	0.070	0.084	0.118	0.085	0.142	0.131
θ	1.50	1.544	1.661	1.611	0.205	0.356	0.607	0.210	0.391	0.617

4 Discussion

4.1 Implications of the results

In terms of runtime, NPE as well as SPE are much faster than MLE. In addition, note that in our nonparametric methods, we have a tool to adjust the complexity of calculations to computational power available: we decrease complexity by picking a coarser Δ—at the price of increasing the bias; see Kirchner (2016). Clearly, these problems will be even more accentuated in the multivariate case. As a side result, we find that, in our set-up, MLE Hawkes parameter estimates seem to be quite heavily biased for nearly critical models. This asks for further examinations of the Hawkes likelihood and its optimization. So, in the light of our study, we propose the following approach to Hawkes process estimation for large sample sizes like 500 points (per point type) and more:

1. Calculate NPE estimates—with the most conservative choices for Δ and p that are computationally acceptable.
2. On the basis of the results, choose a parametric model (in the multivariate case, this choice includes fixing the 'zero edges', that is, choosing the 'Hawkes skeleton'; see Embrechts and Kirchner (2017)).
3. Estimate the parameters by MLE—possibly using an E–M algorithm. If this task is still too complex, we propose to smoothen the SPE-estimates as in (2.11).

We see that in terms of variance, NPE compares favorably with MLE; the methods yield comparable results. In terms of bias, we see that MLE is also not immaculate. The bias of the NPE-estimates can be controlled; it is less obvious how this can be achieved for MLE. Whereas NPE performs very well compared to MLE, this comparison is not really fair: indeed for NPE we use a-priori information of the shape of the true underlying displacement density. In real applications this information is seldomly available.

4.2 Multivariate case

Multivariate studies and the present univariate study bring similar results. However, we found that the optimization of the likelihood is so involved in the multivariate case that numerical effects dwarf statistical properties. To study the role and effects of the most efficient optimization methods (recursive likelihood calculation, E–M algorithms, cutting-off the displacement densities, ...) would be a research project of its own. In any case, these issues again highlight one advantage of the NPE-method over the MLE approach in the context of large datasets: NPE is easier to implement and faster. For the purpose of this paper, where we use MLE as a benchmark model, and in order to achieve readability, we forgo a study of the multivariate case.

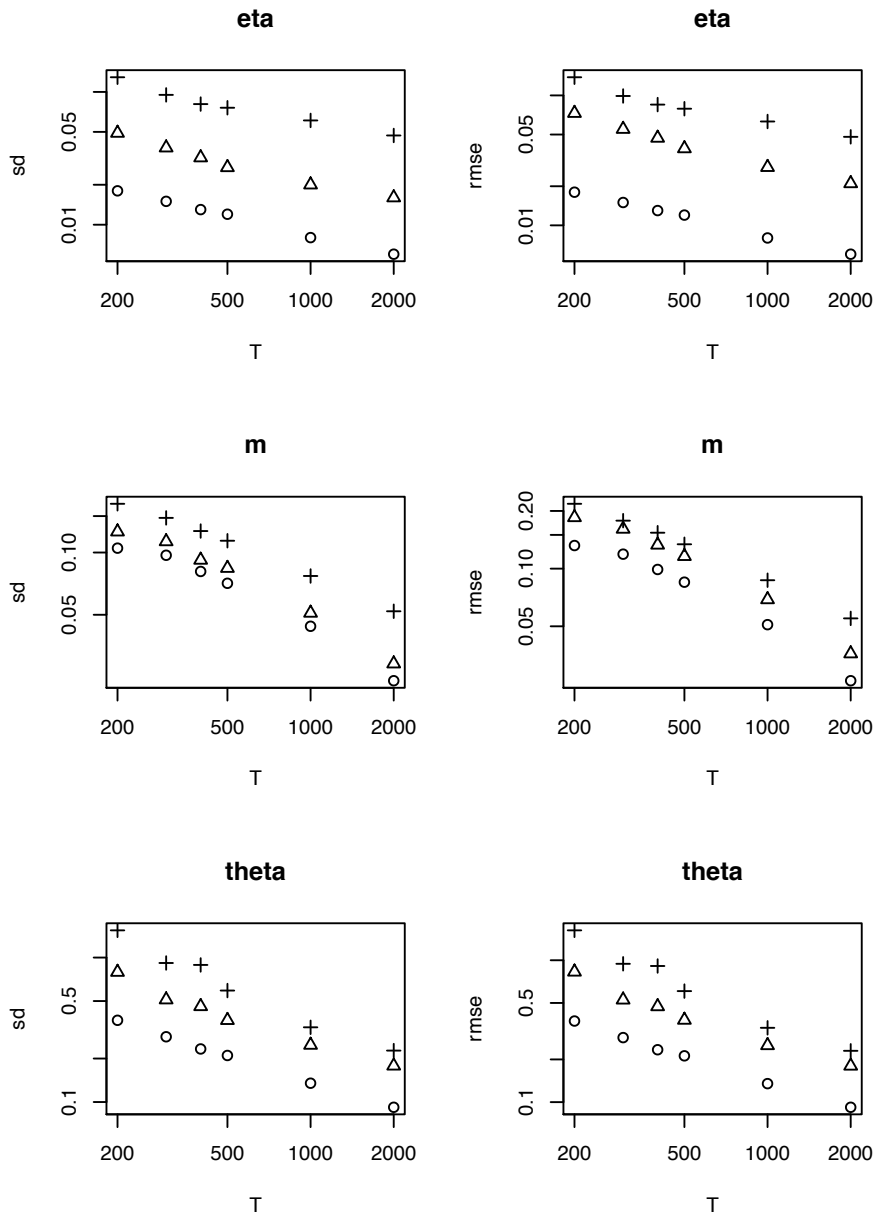


Figure 1: Comparison of sample size and estimator-performance for the nearly critical model ($m = 0.95$) with exponential decay: the panels in the left column indicate standard deviations (of 1000 samples), the panels in the right column indicate root mean squared errors. Circles refer to MLE, triangles to NPE and crosses to SPE. Note the log/log scales. The plots indicate power-law decays of the standard deviation. We also see that the values are roughly proportional.

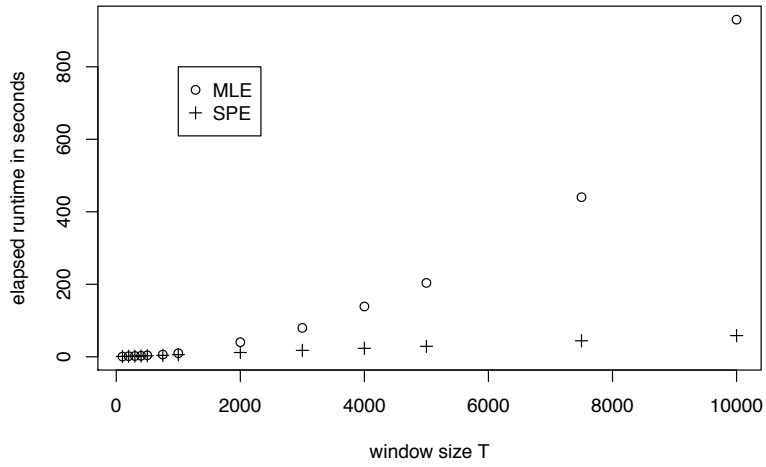


Figure 2: Runtime over different sample window sizes. We measure the elapsed time for a single realization. For this analysis, we only consider the exponential subcritical case in order to keep the variance of the realizations low and make the runtimes representative. The runtimes of SPE nearly coincide with those of NPE so we do not illustrate them. We observe mild linear growth for SPE and polynomial growth for MLE.

Acknowledgements

We gratefully acknowledge Paul Embrechts for guidance and support during the preparation of the paper. This simulation study was suggested by Valérie Chavez-Demoulin. We thank her for her advice.

Paper

G

Matthias Kirchner.

A note on critical Hawkes processes.

Working paper.

A note on critical Hawkes processes

Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich,
8092 Zurich, Switzerland.

Abstract

Let F be a distribution function on $\mathbb{R}$ with $F(0) = 0$ and density f. Let $\tilde{F}$ be the distribution function of $X_1 - X_2$, $X_i \sim F$, $i = 1, 2$, iid. We show that for a critical Hawkes process with displacement density (= 'excitement function' = 'decay kernel') f, the random walk induced by $\tilde{F}$ is necessarily transient. Our conjecture is that this condition is also sufficient for existence of a critical Hawkes process. Our train of thought relies on the interpretation of critical Hawkes processes as cluster-invariant point processes. From this property, we identify the law of critical Hawkes processes as a limit of independent cluster operations. We establish uniqueness, stationarity, and infinite divisibility. Furthermore, we provide various constructions: a Poisson embedding, a representation as Hawkes process with renewal immigration, and a backward construction yielding a Palm version of the critical Hawkes process. We give specific examples of the constructions, where F is regularly varying with tail index $\alpha \in (0, 0.5)$. Finally, we propose to encode the genealogical structure of a critical Hawkes process with Kesten (size-biased) trees. The presented methods lay the grounds for the open discussion of multitype critical Hawkes processes as well as of critical integer-valued autoregressive time series.

Keywords: Hawkes process, critical cluster field, cluster invariance, branching random walk, renewal theory, regular variation, Kesten tree.

1 Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{X} \in \mathcal{B}(\mathbb{R}^d)$ for some d $\in \mathbb{N}$, and let $\mathcal{B}(\mathcal{X})$ ($\mathcal{B}_b(\mathcal{X})$) denote the (bounded) Borel sets of $\mathcal{X}$. A *point process on* $\mathcal{X}$ is a measurable mapping $N : (\Omega, \mathcal{F}) \to (M_p, \mathcal{M}_p)$, where M_p denotes the space of locally finite counting measures on $\mathcal{X}$ and $\mathcal{M}_p$ is the smallest σ-algebra of M_p such that all mappings $M_p \ni m \mapsto m(B)$, $B \in \mathcal{B}_b(\mathcal{X})$, are measurable. For any point process N on $\mathbb{R}$, set $\mathcal{F} \supset \mathcal{H}_x^N := \sigma\big(\{\omega \in \Omega : N(\omega)(B) = n\} :$

$n \in \mathbb{N}_0, B \in \mathcal{B}_b((-\infty, x])\big)$, $x \in \mathbb{R}$. The filtration $(\mathcal{H}_x^N)$ is the *history of* N. We say that $\lambda(\cdot)$ is an $(\mathcal{H}_x^N)$*-intensity* of a point process N if $\lambda(\cdot)$ is adapted to $(\mathcal{H}_x^N)$, is almost surely locally integrable, and

$$\mathbb{E}\left[N\big((a,b]\big)1_A\right] = \mathbb{E}\left[\int_{(a,b]} \lambda(x)\mathrm{d}x 1_A\right], \quad a < b, A \in \mathcal{H}_a^N. \tag{1.1}$$

If $\mathbb{P}[\cap_{k=1}^n \{N(B_k) = n_k\}] = \mathbb{P}[\cap_{k=1}^n \{N(B_k + x) = n_k\}]$, $x \in \mathbb{R}$, $n_k \in \mathbb{N}$, $B_k \in \mathcal{B}(\mathbb{R}), k = 1, 2, \ldots, n \in \mathbb{N}$, then N is *stationary*. If $\mathbb{P}[\cap_{t\in\mathbb{R}} \{N(\{t\}) \leqslant 1\}] = 1$, then N is *simple*. For a simple and stationary point process N, we have that $\mathbb{E}\,N(B) = \lambda \int_B \mathrm{d}x$ for some $\lambda \in [0, \infty]$. We call λ *(average) intensity* and $\mathbb{E}\,N(\cdot)$ *mean measure* of N. A *subcritical Hawkes process* is a simple stationary point process N admitting the $(\mathcal{H}_x^N)$-intensity

$$\lambda(x) = \eta + m \int_{(-\infty,t)} f(x-y)N(\mathrm{d}y), \quad x \in \mathbb{R}, \tag{1.2}$$

where $\eta \in (0, \infty)$ denotes the *immigration intensity*, $m \in [0, 1)$ the *branching coefficient*, and f the *displacement density* supported by $[0, \infty)$. It is well known that (1.2) specifies a unique distribution on $(M_p, \mathcal{M}_p)$; see Hawkes (1971a). The construction of a Hawkes process involves a branching mechanism: the first step of the construction is a Poisson random measure on $\mathbb{R}$ with intensity η. Each of these *immigrants* is the ancestor of a Hawkes family and has $\mathrm{Pois}(m)$ *children*. Each of these children again has $\mathrm{Pois}(m)$ children etc. The distance between parent and child is modeled by the displacement density f. All these offspring and displacement operations are independent. Each immigrant with all its *descendants* forms a *Hawkes family*. The Hawkes process is the random measure that counts the superposition of these Hawkes families in a given set. In other words, the Hawkes process is a sum of occupation measures of infinitely many particles performing (one-sided) branching random walks. Taking expectations on either side of (1.2) and applying stationarity, we note that the intensity of the process is $\lambda = \eta/(1-m)$. We obtain the same result by multiplying the intensity of the immigrants, that is, η, with the expected number of points in a Hawkes family, that is, $(1-m)^{-1}$. In Brémaud and Massoulié (2001), the authors consider a set of Hawkes processes $\{N^{(m)}\}_{m\in(0,1)}$, where the immigration intensity depends on the branching coefficient m by setting $\eta_m := (1-m)\lambda$, $m \in [0, 1)$, for some fixed $\lambda > 0$. Obviously, the intensity for all of these processes equals λ. The cited paper considers weak limits $N^{(m)}$ as $m \uparrow 1$. Note that increasing m corresponds to thinning immigrants and—at the same time—enlarging the remaining Hawkes families. Theorem 1 in the same reference states that if

$$\sup_{x\geqslant 0} f(x)x^{1+\alpha} \leqslant R \quad \text{and} \quad \lim_{x\to\infty} f(x)x^{1+\alpha} = r \tag{1.3}$$

for some constants $r, R \in (0, \infty)$ and $\alpha \in (0, 0.5)$, then $N^{(m)}$ converges weakly to a point process $N^{(1)}$ with average intensity λ and $(\mathcal{H}_x^{N^{(1)}})$-intensity

$$\lambda(x) = \int_{(-\infty,x)} f(x-y) N^{(1)}(\mathrm{d}y). \tag{1.4}$$

In view of the branching construction, the existence of such a critical Hawkes process $N^{(1)}$ is somewhat paradoxical: for an arbitrary point of a Hawkes process, consider the number of its children, of its grandchildren, of its great-grandchildren, etc. This sequence forms a Galton–Watson process with offspring distribution $\mathrm{Pois}(m)$. However, it is well known that not only subcritical ($m < 1$) but also critical ($m = 1$) Galton–Watson processes die out almost surely; for example, see Theorem 6.1 in Harris (1963). Consequently, in a critical Hawkes process, any realized point will almost surely have only a finite number of descendants. For illustration, consider $F(0-) = 0$ and $F(0) = 1$ for the displacement distribution (for the sake of example, we disregard simplicity of the process). In this case, the position of each point coincides with the position of its descendants. Therefore, if the immigrants are thinned, then all descendants are bound to vanish—and consequently the whole process with them. In fact, Proposition 1 in Brémaud and Massoulié (2001) shows that $\int x \mathrm{d}F(x) < \infty$ already has the same effect: if the expectation of the displacements is finite, no nontrivial solution to (2.3) can exist. In contrast, the conditions in (1.3) guarantee that the displacements are balanced in such a way that the Hawkes families grow larger and larger to fill the larger and larger gaps between the immigrants. We found that the resulting object, the critical Hawkes process, is related to many well-studied topics:

a) *Cluster invariance*: we observe that any critical Hawkes process must have this distributional property. This property opens the door to a complete theory presented in Chapter 12 of Matthes et al. (1978);

b) *Renewal theory*: we observe that critical Hawkes processes can be interpreted as Hawkes processes with renewal immigration, where the interarrival times have infinite mean. Thus, in the long run, we can no longer observe immigrants. This construction can be studied with standard tools of renewal theory as presented, e.g., in Chapter 8 of Bingham et al. (1987);

c) *Backward trees*: given a critical Hawkes process, we may pick any point and walk backwards the genealogical structure. In such a manner, we can (re-)construct the process and circumvent the lack of a root event (like immigrants). This method is related to the 'method of backwards trees' introduced in Kallenberg (1977);

d) *Kesten trees* or *size-biased trees*: the lack of immigrants—that is, root nodes—and the non-extinction property make branching representations of critical Hawkes processes not

straightforward. We find that the 'Kesten tree' (see, e.g., Lyons et al. (1995))—a special Galton-Watson tree with size-biased offspring distribution—may be used to encode the genealogical structure of a critical Hawkes process.

We hope that these connections will help to complete the understanding of critical Hawkes processes. In particular, the methods presented offer possible directions for the open discussion of multitype critical Hawkes processes as well as of integer-valued autoregressive time series. Finally note that the present paper gives rise to the following very general conjecture that connects the existence of critical Hawkes processes with the recurrence/transience dichotomy of random walks. We say that a distribution is *transient* (*recurrent*) if the random walk induced by this distribution is transient (recurrent). Furthermore, for any distribution F, let $\tilde{F}$ denote the distribution of $X_1 - X_2$, $X_i \sim F$, $i = 1, 2$, iid. We call $\tilde{F}$ the *symmetrized version of* F. Note that $\tilde{F} : \mathbb{R} \to [0, 1]$, $x \mapsto \int F(x + y)\mathrm{d}F(y)$.

Conjecture 1. *Let F be an absolutely continuous distribution function with $F(0) = 0$ and density f. Let $\tilde{F}$ be the symmetrized version of F and let $\lambda > 0$. A critical Hawkes process N with displacement density f and average intensity λ exists if and only if the symmetrized version of the displacement distribution $\tilde{F}$ is transient. In this case,* (1.4) *specifies a unique, stationary, and infinitely divisible distribution.*

Note that Conjecture 1 includes Proposition 1 of Brémaud and Massoulié (2001) stating the necessary condition $\int x\mathrm{d}F(x) = \infty$ for existence of a critical Hawkes process. Indeed: if $\int x\mathrm{d}F(x) < \infty$, we also have $\int x\mathrm{d}\tilde{F} = 0$. Consequently, by the Chung–Fuchs theorem (see Feller (1971), Section XVIII.6, Lemma 1) $\tilde{F}$ is recurrent.

2 Perspectives on critical Hawkes processes

For any distribution F, $m \in (0, \infty)$, and any point process $N \sim L$ with points $\{T_n\}_{n\in I}$, $I \subset \mathbb{Z}$, we define the *(Poisson) clustering operation*

$$[F, m] \star \{T_n\}_{n\in I} := \bigcup_{n\in I} \bigcup_{k=1}^{K_n} \{T_n + X_{n,k}\}, \tag{2.1}$$

where $\{K_n, X_{n,k} : k \in \mathbb{N}, n \in \mathbb{Z}\}$ are independent random *cluster variables* (also independent of N) with $K_n \sim \mathrm{Pois}(m)$ and $X_{n,k} \sim F$. We denote the point process resulting from (2.1) by $N_{[F,m]}$ and its distribution by $L_{[F,m]}$. We call $[F, m]$ a *cluster field induced by* F. We may construct a univariate subcritical Hawkes process N with immigration intensity $\eta \in (0, \infty)$, branching

coefficient $m \in (0,1)$, and displacement distribution F by applying this clustering operation:

$$N := \sum_{g \geqslant 0} N^{(g)} \text{ with } N^{(0)} \sim \mathrm{PRM}(\eta) \quad \text{and} \quad N^{(g)} := N^{(g-1)}_{[F,m]_g}, \quad g \in \mathbb{N}, \tag{2.2}$$

where the clustering operations '$[F,m]_g \star \cdot$' are applied independently over $g \in \mathbb{N}$. For an absolutely continuous distribution F with $F(0) = 0$ and density f, one can show that the point process N as in (2.2) solves (1.2). If $m = 1$, any construction as in (2.2) (with $\eta > 0$) would yield an infinite average intensity of the limit process N. However, one can argue from (1.4) that the case $m = 1$ can also be represented in terms of the cluster operation:

Definition 2. *Let F be a distribution on $\mathbb{R}$ with $F(0) = 0$ and $[F,1]$ the induced cluster field. Assume that N is an ergodic solution to*

$$N = N_{[F,1]} \tag{2.3}$$

with finite and constant average intensity $\lambda > 0$. Then we call N a critical (F,λ)-Hawkes process.

If F is absolutely continuous with density f, the critical cluster operation can be interpreted as attaching an inhomogeneous Poisson process with intensitiy $f(\cdot - T_n)$ to each point T_n. Thus, one can show that the critical (F,λ)-Hawkes process solves (1.4). Vice versa, (2.3) is more general than (1.4) in that the displacement distribution is not necessarily absolutely continuous. Also note that (2.3) specifies a unique parent point for every point T_n of N: indeed, for each $n \in \mathbb{Z}$, there exist $n' \in \mathbb{Z}$ and $k \in \mathbb{N}$ such that $k \leqslant K_{n'}$ and $T_{n'} + X_{n',k} = T_n$, where the random variables $K_{n'}$ and $X_{n',k}$ stem from the clustering operation; see (2.1). That is, the critical Hawkes process is 'eating its own tail'. The trivial—yet crucial—observation is that (2.3) also holds in distribution.

2.1 Critical cluster fields

For any distribution F on $\mathbb{R}$, denote by $[F] := [F,1]$ the *critical (Poisson) cluster field* induced by F. Chapter 12 in Matthes et al. (1978) (MKM) discusses (distributions of) nontrivial stationary point processes $N \sim L$ with the property

$$L = L_{[F]} \tag{2.4}$$

(in even higher generality). If (2.4) holds, then N and its distribution L are called *cluster invariant with respect to* $[F]$. Furthermore, F is called *stable* if such a distribution L exists. From (2.3), we get that critical Hawkes processes are obviously cluster invariant with respect to

the cluster field induced by their displacement distribution. Thus, we obtain several necessary conditions for the existence of critical Hawkes processes as corollaries from standard results:

Theorem 3. *Let F be a distribution on $[0,\infty)$ and $\lambda > 0$. Assume that a critical (F,λ)-Hawkes process N as in Definition 2 exists. Then the following holds:*

a) Definition 2 specifies a unique, infinitely divisible, and stationary distribution H on $(M_p, \mathcal{M}_p)$.

b) Let L be the distribution of a Poisson random measure with finite average intensity λ. For $g \in \mathbb{N}$, denote g independent clustering operations by '$[F^{[g]}] \star \cdot$'. Then, as $g \to \infty$, $L_{[F^{[g]}]}$ converges weakly to H.

c) The symmetrized displacement distribution $\tilde{F}$ is transient.

Proof. We note that, by definition, any possible distribution H of a critical (F,λ)-Hawkes process is cluster invariant with respect to $[F]$, has bounded average intensity λ, and is ergodic. In particular, F is stable. The statements of the theorem then follow from results in MKM: from Theorem 12.1.4. in MKM, we get that $L_{[F^{[g]}]}$ as in b) converges weakly to an infinitely divisible limit distribution. Stationarity of any possible H follows from Proposition 12.4.7. in MKM. From Theorem 12.4.1. in MKM, we get that H coincides with the limit distribution of $L_{[F^{[g]}]}$. Consequently, H is unique and infinite divisible. We have established a) and b). We obtain c) from Theorem 12.6.6. in MKM if the variance of the offspring distribution is neither zero nor infinity. This assumption obviously holds for our $\mathrm{Pois}(1)$ offspring. □

Next to the reference MKM the reader is referred to Kallenberg (1977) and Chapter 13.5 in Daley and Vere-Jones (2009), in particular, Proposition 13.5.II. therein. Theorem 3 shows that H can be seen as a steady distributional state that is reached by iterating clustering operations. Note that existence of a critical Hawkes process together with the assumptions on the average intensity *imply* stationarity. The observation of cluster invariance only yields *necessary* conditions for the existence of a critical Hawkes process. We now turn towards possible constructions to explore *sufficient* existence conditions.

2.2 Poisson embedding

From Theorem 3 b), a straightforward construction of a solution to (2.3) and thus of a critical (F,λ)-Hawkes process would be to start with a Poisson random field with intensity λ, and then iterating the clustering operations '$[F] \star \cdot$'. However, we cannot hope that the resulting sequences $N^{(g)}(B)$, $B \in \mathcal{B}_b(\mathbb{R})$ will converge *almost surely* to a nontrivial result when we apply independent (and in particular new) cluster variables at each new step. That is, the clustering operations of possible construction steps have to depend on each other. When the displacement distribution has a density f, we propose the following construction based on 'Poisson embedding'

similar to the (subcritical) Hawkes construction in Brémaud and Massoulié (1996) or Chapter 6.3 in Liniger (2009): let $\mathcal{N} : \Omega \to (\mathbb{R} \times \mathbb{R}_{\geqslant 0})$ be a Poisson random measure with intensity 1 on $\mathbb{R}^2$. We call $\mathcal{N}$ the *driving process*. For $\lambda > 0$, set $N^{(0)}(B) := \int_B \mathcal{N}\big(\mathrm{d}x \times (0, \lambda]\big)$, $B \in \mathcal{B}(\mathbb{R})$, and, for $g \in \mathbb{N}$, recursively define point processes $N^{(g)}$ by

$$\lambda^{(g)}(x) := \int_{(-\infty,x)} f(x-y) N^{(g-1)}(\mathrm{d}y), \quad N^{(g)}(B) := \int_B \mathcal{N}\big(\mathrm{d}x \times (0, \lambda^{(g)}(x)]\big), \quad B \in \mathcal{B}(\mathbb{R}). \tag{2.5}$$

Obviously, the average intensity equals λ for all $N^{(g)}$, $g \in \mathbb{N}_0$. The construction steps are very similar to the clustering operations '$[F] \star \cdot$'. Note however, that the clustering operations are not independent over $g \in \mathbb{N}$. In addition, note that the displacements for $N^{(g)}$ also depend on the positions of $N^{(g-1)}$. Thus, the marginal distributions of the point process sequence $(N^{(g)})$ are similar but not equal to $(L_{[F^{[g]}]})$, when L denotes the starting Poisson random field $N^{(0)}$. The symmetrized distribution $\tilde{F}$ comes into play when calculating second moment measures of $N^{(g)}$ recursively. We think that the transience condition in Conjecture 1 guarantees $\sup_{g\in\mathbb{N}} \mathrm{Var}(N^{(g)}(B)) < \infty$, $B \in \mathcal{B}_b(\mathbb{R})$, and thus non-triviality of the potential limit $N^{(\infty)}$. We summarize the above:

Conjecture 4. *For any absolutely continuous displacement distribution F, $\lambda > 0$, and $B \in \mathcal{B}_b(\mathbb{R})$, $N^{(g)}(B)$ as in* (2.5) *converges almost surely to a nonnegative integer as $g \to \infty$. These limits define a point process $N^{(\infty)}$. If $\tilde{F}$ is transient, then the average intensity of $N^{(\infty)}$ equals λ, otherwise it equals 0. Furthermore, the processes $\lambda^{(g)}$ converge almost surely pointwise to a limit $\lambda^{(\infty)}$ such that $\lambda^{(\infty)}(\cdot)$ is an $\mathcal{F}^{N^{(\infty)}}$-intensity.*

2.3 Renewal immigration

The average intensity of a subcritical Hawkes process equals the intensity of the immigrants times the expected number of points in a Hawkes family. Consequently, if $m = 1$, we need zero immigration intensity to obtain a locally finite mean measure. This can be thought of as immigrants stemming from a 'stationary renewal process on $\mathbb{R}$ with infinite interarrival expectation'. We know from the Renewal Theorem (for example, see (1.9) in Chapter XI of Feller (1971)), that the probability for observing such an immigrant in a finite interval will be zero.

Example 5. Let F be an absolutely continuous distribution on $[0, \infty)$ with infinite mean. For any $c \geqslant 0$, consider the truncated distribution $F_c(x) := \mathbb{P}\left[1_{X \leqslant c} X \leqslant x\right]$, $t \in \mathbb{R}$, where $X \sim F$, as well as the truncated mean

$$\mu : [0, \infty) \ni c \mapsto \int x \mathrm{d}F_c(x).$$

The function μ is non-decreasing and continuous with $\lim_{c\to\infty} \mu(c) = \infty$. Thus, we may define

$\mu^{\leftarrow} : [0,\infty) \ni y \mapsto \inf\{c \geqslant 0 : \ \mu(c) = y\}$, $y \geqslant 0$. Set $c : [0,1) \ni m \mapsto \mu^{\leftarrow}((1-m)^{-1})$. The function c is increasing and $\lim_{m\uparrow 1} c(m) = \infty$. For $m \in [0,1)$, let $N^{(m)}$ be a (subcritical) Hawkes process with branching coefficient $m \in [0,1)$, displacement distribution F, and stationary renewal immigration, where $F_{c(m)}$ is the interarrival distribution of the immigrants. This construction yields $\mu(c(m))^{-1} = (1-m)$ for the average intensity of the immigrants. Consequently, we obtain an average intensity 1 for the resulting Hawkes processes $\{N^{(m)}\}_{m\in[0,1)}$. In this construction, we may obtain arbitrary average intensities $\lambda > 0$ by scaling the interarrivals of the immigrants by λ^{-1}.

The weak limit as $m \uparrow 1$ of the processes $\{N^{(m)}\}$ from Example 5 can be studied in a similar manner as the limit of the construction in Brémaud and Massoulié (2001). The interpretation, however, is different. We provide another example starting at time 0, where we specify the tails of the distributions as regularly varying:

Example 6. For $i = 1,2$, let F_i be a distribution on $[0,\infty)$ such that

$$1 - F_i(x) \sim \frac{l_i(x)}{x^{\alpha_i}\Gamma(1+\alpha_i)}, \quad x \to \infty,$$

where $\alpha_i \in (0,1]$ and l_i is slowly varying at infinity. We denote the renewal functions induced from F_i by U_i. Consider a renewal process on $[0,\infty)$ with interarrival distribution F_1. From each renewal epoch, we start a Hawkes family process with displacement distribution F_2. Note that a generic Hawkes family (starting with an ancestor in 0) has mean measure $U([0,x]) := \sum_{g\in\mathbb{N}_0} F^{g*}(x)$. (This can be shown by considering the generations separately.) For $x \in [0,\infty)$, denote the expected number of points in $[0,x]$ of the resulting point process by $\overline{U}(x)$. For any non-negative function G of bounded variation, we denote its Laplace–Stieltjes transform by $\hat{G} : [0,\infty) \ni s \mapsto \int_0^\infty e^{-sx}\mathrm{d}G(x)$ $(\in [0,\infty])$. We have that $\overline{U} = U_1 * U_2$ and, consequently,

$$\hat{\overline{U}}(s) = \frac{1}{1-\hat{F}_1(s)}\frac{1}{1-\hat{F}_2(s)} \quad s \in [0,\infty). \tag{2.6}$$

From Bingham et al. (1987) (BGT), Corollary 8.1.7, we get for $i = 1,2$ that $1 - \hat{F}_i(s) \sim s^{\alpha_i} l_i(1/s)$ as $s \downarrow 0$, . Thus, we obtain

$$\hat{\overline{U}}(s) \sim \frac{1}{s^{\alpha_1+\alpha_2} l_1(1/s) l_2(1/s)}, \quad s \downarrow 0.$$

As $(l_1(\cdot)l_2(\cdot))^{-1}$ is again slowly varying, we may apply Karamata's Tauberian Theorem (Theorem 1.7.1 in BGT) to find

$$\overline{U}(x) \sim \frac{x^{\alpha_1+\alpha_2}}{l_1(x)l_2(x)\Gamma(1+\alpha_1+\alpha_2)}, \quad x \to \infty. \tag{2.7}$$

Thus, setting $\alpha_1 := \alpha \in [0,1)$ and $\alpha_2 := 1-\alpha$ and choosing l_1 and l_2 such that $\lim_{x\to\infty}(l_1(x)l_2(x))^{-1} = \lambda > 0$, we get an 'elementary renewal behaviour' for $\overline{U}$:

$$\lim_{x\to\infty}\frac{\overline{U}(x)}{x} = \lim_{x\to\infty}\frac{x^{\alpha+1-\alpha}}{xl_1(x)l_2(x)\Gamma(1+\alpha+1-\alpha)} = \lim_{x\to\infty}\frac{1}{l_1(x)l_2(x)} = \lambda. \tag{2.8}$$

That is, for all $\alpha \in [0,1)$, the averaged expectation of such critical Hawkes processes with renewal immigration on $[0,\infty)$ converges. Naturally, this does not imply distributional convergence. Furthermore, the symmetry in $i = 1,2$ is only valid for the mean measure. Also note that $\alpha_1 = \alpha$ controls the interarrivals of the immigrant points and $\alpha_2 = 1-\alpha$ controls the displacement of offspring points. Intuitively, we want the distance between the immigration points to be 'larger' than between offspring points—otherwise, the offspring processes thin out faster than the immigrant process and thus vanish. In other words, for survival of the limit process, we will need $\alpha_1 < \alpha_2$, equivalently, $\alpha < 1-\alpha$ and, thus, $\alpha \in (0, 0.5)$—as in (1.3).

2.4 Backward tree and Palm process

We build a 'Palm version' of a critical (F,λ)-Hawkes process by reconstructing the process starting from some fixed point. Here, the role of the symmetrized distribution $\tilde{F}$ is most obvious. We use an approach similar to the 'method of backward trees'; see Kallenberg (1977) or Daley and Vere-Jones (2009, page 336). We pick an arbitrary point of a critical Hawkes process and shift the whole process in such a way that this arbitrary point has position 0. Obviously, from this point, we may walk back to its parent, then to its grandparent, etc. In other words, there is an infinite spine in the underlying (backwards) tree. In this way, we construct a Palm version of a critical (F,λ)-Hawkes process with (not necessarily absolutely continuous) displacement distribution F:

a) Fix a single special point at position 0 and attach a Hawkes family to it.

b) Generate its parent at position $-X_1$, with $X_1 \sim F$ and attach a Hawkes family to this parent.

c) Generate its grandparent at position $-X_1 - X_2$ with $X_2 \sim F$ and attach Hawkes family to this grandparent.

d) Continue in this way.

All random variables applied in this construction are chosen independent. As the construction is non-decreasing (when counting points in a fixed bounded Borel set), it will yield a random (possibly infinite) point measure N_0 on $\mathbb{R}$ with (possibly infinite) mean measure U_0. The limit process can be described as follows: the infinite spine of ancestors of the starting point forms a backward renewal process, with interarrival distribution F_-, where F_- denotes the distribution

function of the random variable $-X$, where $X \sim F$. This ancestor process has locally finite mean measure $U_- := \sum_{g\in\mathbb{N}_0} F_-^{g*}$. (For any distribution G, we write $G((a,b]) := G(b) - G(a)$, $a < b$, and use the notion 'G' as well for the distribution function as for the measure it defines.) Each renewal point marks the start of a new (forward) Hawkes family. Thus, we obtain for the mean measure of the limiting process N_0

$$U_0 = U_- * U = \sum_{g\in\mathbb{N}_0}\sum_{g'\in\mathbb{N}_0} F_-^{g*} * F^{g'*} = \sum_{g\in\mathbb{N}_0}\sum_{g'\geqslant g} F_-^{g*} * F^{g'*} + \sum_{g\in\mathbb{N}_0}\sum_{g'<g} F_-^{g*} * F^{g'*}. \tag{2.9}$$

Note that a random walk starting in 0 with step-size distribution $\tilde{F}$, the symmetrized version of F, has mean measure

$$\tilde{U} := \sum_{g\in\mathbb{N}_0} \tilde{F}^{g*} = \sum_{g\in\mathbb{N}_0} (F_- * F)^{g*} = \sum_{g\in\mathbb{N}_0} F_-^{g*} * F^{g*}. \tag{2.10}$$

Consequently, for the first summand of the right-hand side in (2.9), we obtain

$$\sum_{g\in\mathbb{N}_0}\sum_{g'\geqslant g} F_-^{g*} * F^{g'*} = \sum_{g\in\mathbb{N}_0}\sum_{g'\geqslant g} F_-^{g*} * F^{g*} * F^{*(g'-g)} = \sum_{g\in\mathbb{N}_0} F_-^{g*} * F^{g*} * \sum_{g'\geqslant g} F^{*(g'-g)} = \tilde{U} * U.$$

After a similar calculation for the second summand, we finally get that

$$U_0 = \tilde{U} * (U + U_- - \delta_0). \tag{2.11}$$

Note that U_0 is a symmetric measure (as a convolution of symmetric measures), and that $U_0(\{0\}) \geqslant 1$—with equality if F is absolutely continuous. These two facts were already visible in the first expression of (2.9). More importantly, we learn from (2.11) that U_0 coincides with the expectation of the occupation measures of infinitely many random walks, with each walk starting at renewal times of a two-sided renewal process (with a single renewal in 0). Local finiteness of U_0 means that the construction yields a point process with finite intensity whose law can be identified with a Palm distribution. We summarize:

Conjecture 7. *If F is transient, then N_0 defines a point process (that is, a random locally finite counting measure) with distribution L_0 and λL_0 is the Palm measure of a stationary point process with finite average intensity λ—the critical (λ, F)-Hawkes process.*

We were not able to show local finiteness of the measure U_0 from (2.11) in such full generality. Instead, we supply another example in the case of a displacement distribution F with regularly varying tails:

Example 8. Let $F(x) \sim x^{-\alpha} l(x)$, $x \to \infty$, with l slowly varying at infinity and $\alpha \in (0, 1]$. For

all $h > 0$, we have that

$$U_0([0,h]) = U_- * U([0,h]) = \int_{(-\infty,0]} U\big([-x, h-x]\big) U_-(\mathrm{d}x) = \int_{[0,\infty)} U\big([x-h, x]\big) U(\mathrm{d}x), \tag{2.12}$$

where we use in the last equality that $U_-(B) = U(-B)$, $B \in \mathcal{B}_b(\mathbb{R})$. For $x \to \infty$, the integrand is of the same order as $x^{\alpha-1}/l(x)$ (apart from a set of measure 0) by Theorem 8.6.6 in BGT. For the density u of U (assuming it exists and is ultimately monotone), we have that $x \sim \alpha l(x) x^{\alpha-1}$, $x \to \infty$, by the Monotone Density Theorem; see Theorem 1.7.2 in BGT. Thus, for large x, the integrand with respect to Lebesgue measure is of the same or lower order as $x^{2(\alpha-1)}$ for $x \to \infty$. Consequently, we find that U_0 defines a locally bounded measure if $\alpha \in (0, 0.5)$—as in (1.3)

2.5 Kesten tree

The constructions of the critical Hawkes process in Sections 2.3 and 2.4 are related to so-called *Kesten trees* or *size-biased trees*, a generalization of Galton–Watson trees, where the nodes are either of a *normal* or of a *special* type; see Lyons et al. (1995). The distribution of the (independent) offspring operations of the nodes depends on their type. If (p_k) is the offspring distribution of a normal node and $m \in (0, \infty)$ the corresponding expectation, then (kp_k/m) is the *size-biased offspring distribution* of a special node. In our case, where the normal offspring distribution is $\mathrm{Pois}(1)$, one can check that the size bias corresponds to adding $+1$ to a $\mathrm{Pois}(1)$ random variable. The root node $\varnothing$ of a Kesten tree is special. Normal nodes have only normal children whereas special nodes have exactly one special child and otherwise normal children. Obviously, such a Kesten tree never dies out. In fact, it is well known that a Kesten tree is distributed like the corresponding Galton–Watson tree conditional on non-extinction. Denote such a Kesten tree (with respect to $\mathrm{Pois}(1)$ normal offspring and '$\mathrm{Pois}(1)+1$' special offspring) by $\mathbf{T}$, its nodes by $\{\sigma\}_{\sigma\in\mathbf{T}}$, and its root node by $\varnothing$. We write σ^- for the unique parent node of $\sigma \in \mathbf{T}\backslash\{\varnothing\}$. Every node is supplied with a position in $\mathbb{R}$ in a recursive (random) way:

$$S_\varnothing := 0, \quad S_\sigma := \begin{cases} S_{\sigma^-} + Y_\sigma^{(1)}, & \text{if } \sigma \in \mathbf{T} \text{ is a normal node and} \\ S_{\sigma^-} + Y_\sigma^{(2)}, & \text{if } \sigma \in \mathbf{T} \text{ is a special node and } \sigma \neq \varnothing, \end{cases} \tag{2.13}$$

where $Y_\sigma^{(i)} \sim F_i$, $i = 1, 2$, $\sigma \in \mathbf{T}$, are independent. The distribution F_1 coincides with F, the displacement distribution; the distribution F_2, $F_2(0) = 0$, controls the desired limiting average intensity λ. The chain of nodes along the special nodes of $\mathbf{T}$ form an infinite spine. The position of these nodes along the infinite spine correspond to the immigrant renewal process from Section 2.3. Obviously, one could represent the specific constructions from Examples 5

and 6 in a similar manner as (2.13).

Similarly, the backward or Palm construction from Section 2.4 may be written in terms of Kesten trees. The underlying tree is exactly the same as for the case with renewal immigration in (2.13). However, the position labels change. Namely, we set

$$S_\varnothing := 0, \quad S_\sigma := \begin{cases} S_{\sigma^-} + Y_\sigma, & \text{if } \sigma \text{ is a normal node and} \\ S_{\sigma^-} - Y_\sigma, & \text{if } \sigma \text{ is a special node and } \sigma \neq \varnothing, \end{cases} \tag{2.14}$$

where $Y_\sigma \sim F$, iid, with F the displacement distribution. Thus, for studying the distribution of a critical Hawkes process, we may apply limit theorems for Kesten trees to its genealogical structure and then—with the tree given—study the positions separately by standard renewal or random walk theory.

3 Discussion

The presented methods open the door for discussions of multitype critical Hawkes processes as well as of critical integer-valued autoregressive time series:

3.1 Critical multitype Hawkes processes

The Poisson embedding, the connection to critical cluster fields, as well as the renewal immigration representation and the Palm construction allow for multitype generalizations. These generalizations will involve regime switching renewal processes and/or regime switching random walks. Algebraic properties of the branching matrix such as irreducibility will be important. Results on multitype critical cluster fields as in Ivanoff (1982) can be applied. A possible formalization for multiple size-biased trees is given in Kurtz et al. (1997).

3.2 Critical autoregressive time series

In the same way as this paper analyzes critical monotype Hawkes processes, one could discuss critical univariate integer-valued autoregressive time series of infinite order—that is, critical INAR(∞) processes; see Kirchner (2016) for the subcritical case and for the INAR–Hawkes relation. More explicitely, one could consider time series (X_n) solving

$$X_n = \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \xi_{n,k}^{(\alpha_k)}, \quad \mathbb{E}\, X_n \equiv \lambda > 0, \quad n \in \mathbb{Z}, \tag{3.1}$$

for some independent random variables $\{\xi_{n,k}^{(\alpha_k)}\}$ with $\xi_{n,k}^{(\alpha_k)} \sim \text{Pois}(\alpha_k)$, $\alpha_k \geqslant 0, k \in \mathbb{N}$, $n \in \mathbb{Z}$ and $\sum_{k\in\mathbb{N}} \alpha_k = 1$. Arguing in a similar manner as in Section 2 for the Hawkes process, we may rewrite (3.1) in terms of a (critical) cluster operation with offspring distribution Pois(1) and displacement distribution (α_k). The link between the clustering and the counting variables $(\xi_{n,k})$ in (3.1) is provided by the property

$$\xi_{n,k} := \#\{Y_{n,l} = k : l = 1, \ldots, K_n\} \sim \text{Pois}(\alpha_k), \quad \text{independently over } k \in \mathbb{N},\ n \in \mathbb{Z}, \tag{3.2}$$

when $\{K_n, Y_{n,l} : l \in \mathbb{N}, n \in \mathbb{Z}\}$ are independent random variables with $K_n \sim \text{Pois}(1)$ and $Y_{n,l} \sim (\alpha_k)$. (In (3.2), we use the convention that $K_n = 0 \Rightarrow \xi_{n,k} = 0$.) In analogy to Theorem 3, one can then show that solutions to (3.1) necessarily specify a unique stationary time series distribution. Thus, our conjecture is that such a critical INAR(∞) process exists if and only if the symmetric random walk with step size distribution $(\sum_{l=1}^{\infty} \alpha_l \alpha_{k+l})_{k\in\mathbb{Z}}$ ($\alpha_k := 0, k \leqslant 0$) is transient. Note that we may then argue as after Conjecture 1 that necessarily $\sum_{k=1}^{\infty} \alpha_k k = \infty$. So in particular, we need $\alpha_k > 0$, infinitely often, and a critical INAR(p) process with $p < \infty$ cannot exist. It is interesting to study the more familiar autoregressive representation of (3.1), namely

$$X_n = \sum_{k=1}^{\infty} \alpha_k X_{n-k} + u_n, \quad n \in \mathbb{Z}, \tag{3.3}$$

with $u_n := \sum_{k=1}^{\infty} \sum_{l=1}^{X_{n-k}} \xi_{n,k}^{(\alpha_k)} - \sum_{k=1}^{\infty} \mathbb{E} \sum_{l=1}^{X_{n-k}} \xi_{n,k}^{(\alpha_k)}$. The innovations (u_n) are stationary and have zero marginal means. In addition, one can show that $\text{Cov}(u_n, u_{n'}) = 0$, $n \neq n'$. In other words, the critical INAR(∞) time series is a nontrivial critical (='unit root') *and stationary* autoregressive process. This stands in putative contradiction to standard time series theory; see e.g. Exercise 4.28.* in Brockwell and Davis (1991). However, note that stationarity does not imply 'weak stationarity', respectively, 'covariance stationarity' (the meaning of 'stationarity' in the mentioned exercise) if $\text{Var}(X_n) = \infty$. We conclude that X_n as in (3.1) or in (3.3) has infinite variance. Note that $\text{Var}(u_n) < \infty$ might still be possible because the conclusion '$\text{Var}(Z_1) = \text{Var}(Z_2) = \infty \Rightarrow \text{Var}(Z_1 + Z_2) = \infty$' is in general wrong.

3.3 Conclusion

We have identified the distribution of a critical Hawkes process: it coincides with the distribution of a cluster-invariant point process. From Theorem 3b), we get that this distribution can be constructed by starting with a Poisson random field of ancestors, then applying iterated clustering—only considering children, then only grandchildren, etc. In other words, the points of a critical Hawkes process are related like 'cousins of a very, very high degree'.

Bibliography

Adamopoulos, L. (1975). Some counting and interval properties of the mutually-exciting processes. *Journal of Applied Probability*, 12:78–86. [12]

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117:585–606. [14, 108]

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, Budapest*, pages 276–281. [130]

Akaike, H. and Ogata, Y. (1982). On linear intensity models for mixed doubly stochastic poisson and self-exciting point processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 44(1):102–107. [15]

Al-Osh, M. and Alzaid, A. (1987). First order integer valued autoregressive process. *Journal of Time Series Analysis*, 8:261–275. [18, 44]

Alfonsi, A. and Blanc, P. (2015). Extension and calibration of a Hawkes-based optimal execution model. *arXiv:1506.08740*. [15, 113, 127]

Bacry, E., Dayri, K., and Muzy, J. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *European Physical Journal B*, 85(5):157. [14, 16, 108, 113, 128]

Bacry, E., Delattre, S., Hofmann, M., and Muzy, J. (2013). Scaling limits for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499. [14, 108]

Bacry, E., Gaïffas, S., and Muzy, J. (2015a). A generalization error bound for sparse and low-rank multivariate Hawkes processes. *arXiv:1501.00725*. [14, 160, 194]

Bacry, E., Jaisson, T., and Muzy, J. (2014). Estimation of slowly decreasing Hawkes kernels: Application to high frequency order book modelling. *arXiv:1412.7096*. [14, 16, 33, 113, 128, 146, 148, 149, 194, 213]

Bacry, E., Mastromatteo, I., and Muzy, J. (2015b). Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01). 1550005. [14]

Bacry, E. and Muzy, J. (2015). Second order statistics characterization of Hawkes processes and non-parametric estimation. *arXiv:1401.0903v2*. [13, 14, 113, 128, 129, 195]

Bartlett, M. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 25(2):264–296. [11]

Bates, D. and Maechler, M. (2015). Matrix: Sparse and dense matrix classes and methods. R package version 1.1-5. [136, 151, 189, 227]

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons, New York. [53, 72, 74]

Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular Variation*. Cambridge University Press, Cambridge. [241, 246]

Bowsher, C. (2002). Modelling security market events in continuous time: intensity based, multivariate point process models. *Nuffield College Economics Discussion Papers*, pages 1–55. [14, 108]

Box, G. and Jenkins, G. (1970). *Time Series Analysis - Forecasting and Control*. Holden Day, San Francisco. [43]

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24(3):1563–1588. [12, 14, 22, 36, 37, 44, 59, 102, 108, 245]

Brémaud, P. and Massoulié, L. (2001). Hawkes branching processes without ancestors. *Journal of Applied Probability*, 38:122–135. [13, 19, 34, 36, 44, 54, 56, 108, 240, 241, 242, 246]

Brémaud, P., Nappo, G., and Torrisi, G. L. (2002). Rate of convergence to equilibrium of marked Hawkes processes. *Journal of Applied Probability*, 39(1):123–136. [13]

Brillinger, D. (1975). The identification of point process systems. *The Annals of Probability*, 3(6):909–929. [13]

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, 2nd edition. [49, 70, 251]

Brown, T. and Nair, M. (1988). A simple proof of the multivariate random time change theorem for point processes. *Journal of Applied Probability*, 25:210–214. [138]

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208. [226]

Chavez-Demoulin, V., Davison, A., and McNeil, A. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5:227–234. [14, 108]

Chavez-Demoulin, V. and McGill, J. (2012). High-frequency financial data modeling using Hawkes processes. *Journal of Banking and Finance*, 36:3415–3426. [108]

Cont, R. and de Larrard, A. (2013). Price dynamics in a Markovian limit order market. *Journal of Financial Mathematics*, 4(1):1–25. [33, 194, 195, 205]

Cox, D. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 17(2):129–164. [12]

Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 105(41):15649–15653. [14, 108]

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. [189]

Daley, D. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York. [13]

Daley, D. and Vere-Jones, D. (2003/2009). *An Introduction to the Theory of Point Processes*, volume I and II. Springer, New York, 2nd edition. [6, 9, 13, 15, 44, 53, 72, 101, 108, 116, 138, 160, 223, 244, 247]

Delattre, S., Fournier, N., and Hoffmann, M. (2016). Hawkes processes on large networks. *Annals of Applied Probability*, 26(1):216 – 261. [160]

Du, Y. and Li, J.-G. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*, 12(2):129–142. [18, 44, 114, 118]

Durrett, R. (1996). *Probability: Theory and Examples*. Duxbury Press, Belmont, 2nd edition. [156]

Eichler, M., Dahlhaus, R., and Dueck, J. (2016). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*. [16, 102]

Embrechts, P. and Kirchner, M. (2017). Hawkes graphs. *Theory of Probability and Its Applications*, 62(1):163–193. [84, 101, 127, 128, 129, 136, 144, 150, 194, 195, 198, 206, 215, 224, 233]

Embrechts, P., Liniger, T., and Lu, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378. [14]

Errais, E., Gieseke, K., and Goldberg, L. (2010). Affine point processes and portfolio credit risk. *Society for Industrial and Applied Mathematics: Journal on Financial Mathematics*, 1:642–665. [13, 14, 108, 113]

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, New York, 2nd edition. [242, 245]

Filiminov, V. and Sornette, D. (2012). Quantifying critical reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108. [139]

Fokianos, K. and Kedem, B. (2012). *Regression Models for Time Series Analysis*. John Wiley and Sons, New York. [36, 44, 59, 114]

Fokianos, K. and Tjøstheim, D. (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics*, 64:1205–1225. [36, 59]

Gould, M., Porter, M., Williams, S., McDonald, M., Fenn, D., and Howison, S. (2013). Limit order books. *arXiv:1012.0349*. [31, 139, 193, 195]

Greenwood, M. (1946). The statistical study of infectious diseases. *Journal of the Royal Statistical Society, Series A*, 109(2):85–110. [1, 12]

Gunawardana, A., Meek, C., and Xu, P. (2014). A model for temporal dependencies in event streams. *Microsoft Research*. [165]

Haccou, P., Jagers, P., and Vatutin, V. (2005). *Branching Processes*. Cambridge University Press, Cambridge. [164]

Hall, E. and Willett, R. (2016). Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346. [160]

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press, Princeton. [154]

Hannan, E. (1970). *Multiple Time Series*. John Wiley and Sons, New York. [119]

Hansen, N., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143. [113, 128]

Hardiman, S.J., Bercot, N., and Bouchaud, J.-P. (2013). Critical reflexivity in financial markets: a Hawkes process analysis. *European Physical Journal B*, 86(10):442. [33, 130, 139, 146, 149, 213]

Harris, T. (1963). *The Theory of Branching Processes*. Die Grundlehren der mathematischen Wissenschaften. Springer, Berlin. [34, 241]

Hawkes, A. (1971a). Spectra of some self-exciting and mutually-exciting point processes. *Biometrika*, 58:83–90. [5, 11, 13, 44, 50, 57, 108, 112, 160, 165, 223, 240]

Hawkes, A. (1971b). Point spectra of some mutually-exciting point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 33:438–443. [11, 44, 50, 56, 108, 128, 160, 165, 195, 223]

Hawkes, A. and Adamopoulos, L. (1973). Cluster models for earthquakes – regional comparisons. *Bulletin International Statistical Institute*, 45(3):454–461. [14]

Hawkes, A. and Oakes, D. (1974). A cluster representation of a self-exciting point process. *Journal of Applied Probability*, 11(3):493–503. [9, 12, 44, 50, 51, 52, 84, 108, 112, 160, 162, 165, 223, 225]

Hawkes, T. (1968). On the class of the Sylow tower groups. *Mathematische Zeitschrift*, 105:393–398. [160]

Ivanoff, G. (1982). The multitype branching random walk, II. *Journal of Multivariate Analysis*, 12:526–548. [250]

Jaisson, T. and Rosenbaum, M. (2015). Limit theorems for nearly unstable Hawkes processes. *The Annals of Applied Probability*, 25(2):600–631. [13, 130]

Kallenberg, O. (1977). Stability of critical cluster fields. *Mathematische Nachrichten*, 77:7–43. [24, 241, 244, 247]

Kallenberg, O. (1983). *Random Measures*. Akademie-Verlag, 3rd edition. [50]

Kim, H. (2011). *Spatio-Temporal Point Process Models for the Spread of Avian Influenza Virus (H5N1)*. PhD thesis, University of California, Berkeley. [14, 108]

Kirchner, M. (2016). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 162:2494–2525. [83, 95, 96, 100, 107, 108, 116, 117, 129, 156, 161, 170, 172, 189, 195, 198, 233, 250]

Kirchner, M. (2017a). An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4):571–595. [45, 58, 84, 101, 102, 161, 165, 170, 171, 172, 194, 195, 198, 223, 224, 226, 227, 231]

Kirchner, M. (2017b). Hawkes forests. *Submitted*. [198]

Kirchner, M. and Vetter, S. (2017). Hawkes model specification for the limit order book. *Submitted*. [129, 144, 150]

Klimko, L. and Nelson, P. (1978). On conditional least squares estimation for stochastic processes. *The Annals of Statistics*, 6(3):629–642. [118]

Kurtz, T., Lyons, R., Pemantle, R., and Peres, Y. (1997). A conceptual proof of the Kesten–Stigum Theorem for multi-type branching processes. In Athreya, K. and Jagers, P., editors, *Classical and Modern Branching Processes*, pages 181–185. Springer New York, New York. [250]

Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability*, 29:228–248. [44, 84, 95, 115, 118, 156]

Lemonnier, R. and Vayatis, N. (2014). Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In Daelemans, W. and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, Berlin. [15, 113, 127, 128]

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. `http://paleo.sscnet.ucla.edu/Lewis-Molher-EM_Preprint.pdf`. Accessed: 2016-05-12. [15, 113, 127]

Lewis, E., Mohler, G., Brantingham, J., and Bertozzi, A. (2012). Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25(3):244–264. [14]

Liniger, T. (2009). *Multivariate Hawkes Processes*. PhD thesis, ETH Zurich. [12, 13, 15, 36, 44, 59, 84, 108, 110, 113, 129, 160, 162, 195, 196, 223, 245]

Lomnitz, C. (1974). *Global Tectonics and Earthquake Risk*, volume 5 of *Developments in Geotectonics*. Elsevier. [14]

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. [119, 120, 121, 130, 132, 152, 153]

Lyons, R., Pemantle, R., and Peres, Y. (1995). Conceptual proofs of l log l criteria for mean behavior of branching processes. *The Annals of Probability*, 23(3):1125–1138. [23, 242, 249]

Marques da Silva, I. (2005). *Contributions to the Analysis of Discrete-Valued Time Series*. PhD thesis, Departamento de Matematica Aplicada Faculdade de Ciencias da Universidade do Porto. [44, 114, 130]

Matthes, K., Kerstan, J., and Mecke, J. (1978). *Infinitely Divisible Point Processes*. John Wiley and Sons, New York. [35, 241, 243]

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin*, 21(4):645–650. [44]

McNeil, A., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press, Princeton. [14, 108]

Meek, C. (2014). Toward learning graphical and causal process models. *Microsoft Research*. [165]

Meyer, P. (1971). Démonstration simplifiée d'un théorème de Knight. *Lecture Notes in Mathematics*, 191:191–195. [138]

Mikosch, T. and Stărică, C. (2000). Is it really long memory that we see in financial returns? In Embrechts, P., editor, *Extremes and integrated risk management*, pages 149–168. Risk Books, London. [140, 148]

Mohler, G., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108. [14, 108]

Muzy, J. and Bacry, E. (2014). Hawkes model for price and trades high frequency dynamics. *Quantitative Finance*, 14(7):1–10. [14, 108]

Nakayama, M., Shahabuddin, P., and Sigman, K. (2004). On finite exponential moments for branching processes and busy periods for queues. *Journal of Applied Probability*, 41:273–280. [66]

Neveu, J. (1986). Arbres et processus de Galton–Watson. *Annales d'Institute Henri Poincaré*, 22:199–207. [85]

Oakes, D. (1975). The Markovian self-exciting process. *Journal of Applied Probability*, 12(1):69–77. [12]

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–262. [15]

Ogata, Y. (1988). Statistical models for earthquake occurences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27. [14, 108, 160]

Ogata, Y. (1999). Seismicity analysis through point-process modeling: a review. *Pure and Applied Geophysics*, 155(2):471–507. [14]

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point process. *Annals of the Institute of Statistical Mathematics*, 31(Part B):145–155. [12, 13, 15, 44, 160]

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition. [29, 165]

Reinsel, G. (1997). *Elements of Multivariate Time Series Analysis*. Springer, New York, 2nd edition. [130]

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York. [50, 53]

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Mal, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1–41. [14, 16, 113, 128]

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822. [14, 16, 108]

Seneta, E. (1969). Functional equations and the Galton–Watson process. *Advances in Applied Probability*, 1:1–42. [46]

Shi, Z. (2015). *Branching Random Walks*, volume 2151 of *Lecture Notes in Mathematics*. Springer, Cham. [163, 225]

Stephenson, R. (2016). Local convergence of large critical multi-type Galton–Watson trees and applications to random maps. *Journal of Theoretical Probability*, pages 1–47. [85]

Steutel, F. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7(5):893–899. [18, 46, 114, 116]

Vere-Jones, D. (1975). Stochastic models for earthquake sequences. *Geophysical Journal International*, 42(2):811–826. [14]

Watson, C. (2015). The geometric series of a matrix. `http://www.math.uvic.ca/~dcwatson/work/geometric.pdf`. Accessed: 2016-07-01. [88, 164]

Wheatley, S. (2016). The Hawkes process with renewal immigration & its estimation with an EM algorithm. *Computational Statistics & Data Analysis*, 94(C):120–135. [15]

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. PhD thesis, University of Uppsala. [43]

Wiener, N. (1932). Tauberian theorems. *Annals of Mathematics*, 33(1):1–100. [70]

Zhang, H., Wang, D., and Zhu, F. (2010). Inference for INAR(p) processes with signed generalized power series thinning operator. *Journal of Statistical Planning and Inference*, 140:676–683. [118]

Zhao, H. (2012). *A Dynamic Contagion Process for Modelling Contagion Risk in Finance and Insurance*. PhD thesis, The London School of Economics and Political Science. [36, 59, 150]

Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31, pages 641–649. [160]