arXiv:1509.02017v1 [math.PR] 7 Sep 2015

An Estimation Procedure for the Hawkes Process

Matthias Kirchner†

†RiskLab, Department of Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zurich,
Switzerland

(*This version: September 8, 2015*)

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. 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 an appropriate scaling, these time series estimates yield estimates for the underlying multivariate Hawkes process as well as formulas for their asymptotic distribution. 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. Finally, 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.

Keywords: Hawkes process; estimation; integer-valued autoregressive time series; contagion model; intraday financial econometrics

JEL Classification: C13, C14, C22, C51

1. Introduction

In this paper, we introduce a nonparametric estimation procedure for the multivariate Hawkes point process; see Definition 3.6 for the formal definition and Figure B1 for an illustrative summary of the main results. The Hawkes process is a model for event streams. Its alternative name, "selfexciting 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 datasets 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

---

Email: matthias.kirchner@math.ethz.ch

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 (1971a,b) as a model for event data from contagious processes. Theoretical cornerstones of the model are Hawkes (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 (2003). The main theoretical reference for the following presentation is our own contribution Kirchner (2015), 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 Embrechts *et al.* (2011), 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.* (2011), 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 and discusses its properties. 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 section concludes with a discussion on the implications of the presented results. Appendix A contains proofs and Appendix B presents illustrating figures. 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 Hawkes process specifies a distribution of points on one or more lines. Typically, the lines are interpreted as "time" and the points as "events". Selfexciting 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 ("selfexcites") 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 $\mathcal{P} = \{\ldots, T_{-1}, T_0, T_1, \ldots\}$ with $T_k \leq T_{k+1}$, $k \in \mathbb{Z}$,

having almost surely no limit points. Furthermore, we assume that the σ-algebras

$$\mathcal{H}_t^{\mathcal{P}} := \sigma\left(\left\{\omega \in \Omega : \#\left(\mathcal{P}(\omega) \cap (a,b]\right) = n : n \in \mathbb{N}_0, \ a < b \leq t\right\}\right), \quad t \in \mathbb{R},$$

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 $\left(\mathcal{H}_t^N\right) := \left(\mathcal{H}_t^{\mathcal{P}}\right)$ *history* of the point process. The *conditional intensity* of a point process N is

$$\Lambda_N(t) := \lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[N\left(t, t+\delta\right) | \mathcal{H}_t^N\right]}{\delta}, \quad t \in \mathbb{R}. \tag{1}$$

A *Hawkes process* is a stationary point process N with conditional intensity

$$\Lambda_N(t) = \eta + \int_{-\infty}^{t} h(t-s)N\left(\mathrm{d}s\right), \quad t \in \mathbb{R}. \tag{2}$$

The constant $\eta \geq 0$ is called *baseline intensity*, and the function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, measurable, is called *excitement function*. Necessary 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., $\mathbf{N} = \left(N^{(1)}, \ldots, N^{(d)}\right)^{\top}$. Each component process counts points from random point-sets $\mathcal{P}_1 \subset \mathbb{R}, \ldots, \mathcal{P}_d \subset \mathbb{R}$. In this multivariate setup, the counting processes $N^{(k)}$, $k = 1, \ldots, d$, do not only selfexcite but in general also interact with each other ("crossexcite"). The baseline intensity η is a d-variate vector in $\mathbb{R}_{\geq 0}^d$ and the excitement function is a measurable $d \times d$ matrix-valued function $H = (h_{i,j})_{1 \leq i,j \leq d} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}^{d \times d}$. The *conditional intensity of a d-variate Hawkes process* is $\mathbb{R}_{\geq 0}^d$-valued with

$$\boldsymbol{\Lambda}_{\mathbf{N}}(t) := \lim_{\delta \downarrow 0} \frac{\mathbb{E}\left[\mathbf{N}(t, t+\delta) | \mathcal{H}_t^{\mathbf{N}}\right]}{\delta} = \eta + \int_{-\infty}^{t} H(t-s)\mathbf{N}\left(\mathrm{d}s\right), \quad t \in \mathbb{R}, \tag{3}$$

where, for $i = 1, \ldots, d$,

$$\left(\int_{-\infty}^{t} H(t-s)\mathbf{N}\left(\mathrm{d}s\right)\right)_i := \left(\sum_{j=1}^{d} \int_{-\infty}^{t} h_{i,j}(t-s)N^{(j)}\left(\mathrm{d}s\right)\right)_i \tag{4}$$

and $\mathcal{H}_t^{\mathbf{N}} := \sigma\left(\left\{\omega \in \Omega : \mathbf{N}\left((a,b]\right) = \mathbf{n}\right\}, \mathbf{n} \in \mathbb{N}_0^d, \ a < b \leq t\right)$. 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 B2 for an example of a bivariate Hawkes process. In

Hawkes (1971b), we find the following sufficient condition for existence: if

$$\text{spr}(K) := \max\left\{|k| : k \text{ eigenvalue of matrix } K\right\} < 1, \tag{5}$$

where $K := \left(\int\limits_0^\infty h_{i,j}(t)\mathrm{d}t\right)_{1 \le i,j \le d}$, then a process with conditional intensity as in (3) exists. The matrix K in (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 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 $\text{Pois}\left(K_{ij}\right)$ 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 Ogata (1988). 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 B2. 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.

Note that in Bacry *et al.* (2012) another nonparametric method for the estimation of the multivariate Hawkes process is developed; it can be interpreted in our approximation framework. We will touch this alternative approach in Section 3.2.

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 $\tilde{X}_n^{(\Delta)} := N\left((n-1)\Delta, n\Delta\right)$, $n \in \mathbb{Z}$. We want to argue that for small $\Delta > 0$

and large $p \in \mathbb{N}$, we have that

$$\mathbb{E}\left[\tilde{X}_n^{(\Delta)}\middle|\sigma\left(\tilde{X}_{n-1}^{(\Delta)}, \tilde{X}_{n-2}^{(\Delta)}, \dots\right)\right] \approx \Delta\eta + \sum_{k=1}^{p}\Delta h(\Delta k)\tilde{X}_{n-k}^{(\Delta)}, \quad n \in \mathbb{Z}. \tag{6}$$

We divide the approximation above in three separate approximation-steps:

$$\mathbb{E}\left[\tilde{X}_n^{(\Delta)}|\mathcal{H}_{(n-1)\Delta}^N\right] \left(\overset{(1)}{=} \int_{(n-1)\Delta}^{n\Delta} \mathbb{E}\left[\Lambda(t)|\mathcal{H}_{(n-1)\Delta}^N\right] \mathrm{d}t\right)$$

$$\overset{(2)}{\approx} \Delta\eta + \Delta \int_{-\infty}^{(n-1)\Delta} h(n\Delta - u)N(\mathrm{d}u) \tag{7}$$

$$\approx \Delta\eta + \Delta \int_{(n-p-1)\Delta}^{(n-1)\Delta} h(n\Delta - u)N(\mathrm{d}u) \tag{8}$$

$$\approx \Delta\eta + \sum_{k=1}^{p}\Delta h(\Delta k)\tilde{X}_{n-k}^{(\Delta)}, \quad n \in \mathbb{Z}. \tag{9}$$

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 (7), a cut-off error (8) and a discretization error (8). There is an integer-valued time series that solves the approximative bin-count equation (6) to the point: the integer-valued autoregressive model of order $p \in \mathbb{N}$, the INAR(p) model.

2.4 The INAR(p) model

The INAR(p) process was first proposed by Li and Yuan (1991) as a time series model for count data. For the history and an exhaustive collection of properties of the model; see 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 thinning-operations. The following notation from Steutel and van Harn (1979) makes the analogy particularly obvious.

Definition 2.1: For an $\mathbb{N}_0$-valued random variable Y and a constant $\alpha \geq 0$ define the *thinning operator* $\circ$ by

$$\alpha \circ Y := \sum_{k=1}^{Y} \xi_k^{(\alpha)},$$

where $\xi_1^{(\alpha)}, \xi_2^{(\alpha)}, \dots$ 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 thinning operator and the multivariate version of the INAR(p):

Definition 2.2: For a $d \times d$ matrix $A = (\alpha_{i,j})_{1 \leq i,j \leq d} \in \mathbb{R}_{\geq 0}^{d \times d}$ and an $\mathbb{N}^d$-valued random variable $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$, define the *multivariate thinning 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 thinnings $(\alpha_{i,j} \circ \cdot)$ operate independently over $1 \leq i, j \leq d$.

Definition 2.3: Let $d, p \in \mathbb{N}$, $A_k \in \mathbb{R}_{\geq 0}^{d \times d}$, $k = 1, \ldots, p$, $\mathbf{a}_0 \in \mathbb{R}_{\geq 0}^d$ 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 *innovation-parameter vector* and to A_k, $k = 1, 2, \ldots, p$, as *thinning-coefficient matrices*.

This model has first been considered in Latour (1997). In the same paper we find that if all zeros of

$$z \mapsto \det \left(z 1_{d \times d} - \sum_{k=1}^p A_k \right), \quad z \in \mathbb{C}, \tag{10}$$

lie inside the unit circle, then a multivariate INAR(p) process as in Definition 2.3 exists.

Consider a univariate INAR(p) sequence (X_n) with innovation parameter α_0 and thinning 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 $X_n | X_{n-1}, X_{n-2}, \ldots \sim \text{Pois}\left(\alpha_0 + \sum_{k=1}^p \alpha_k X_{n-k}\right)$. In particular, $\mathbb{E}[X_n | \sigma(X_{n-1}, X_{n-2}, \ldots)] = \alpha_0 + \sum_{k=1}^p \alpha_k X_{n-k}$—which is the exact version of (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. $\text{Pois}(\alpha_0)$ variables. Each of these immigrants produces $\text{Pois}(\alpha_k)$ 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 2.1 would be Bernoulli. Note, however, that for small thinning 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 thinning 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 (2015). For a particularly obvious parallel, the reader may consider the analogy of the existence criteria (5) and (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 2.3 with $d = 1$ and $p = \infty$, we prove

Theorem 2.4 :   *Let N be a univariate Hawkes process with baseline intensity $\eta > 0$ and piecewise-continuous excitement function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ 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 innovation parameter $\alpha_0^{(\Delta)} := \Delta\eta$ and thinning 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 $\big(N^{(\Delta)}\big)$ converges weakly to the Hawkes process N.

Proof: This is Theorem 3 in Kirchner (2015). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Note that weak convergence of point processes is equivalent to convergence of the corresponding finite-dimensional distributions; see Daley and Vere-Jones (2003), Theorem 11.1.VII. The other result from Kirchner (2015) that is important for our estimation purpose is the fact that INAR(∞) processes can be approximated by INAR(p) processes, $p < \infty$:

Proposition 2.5:   *Let (X_n) be an INAR(∞) sequence with innovation parameter $\alpha_0 > 0$ and thinning coefficients $\alpha_k \geq 0$, $k \in \mathbb{N}$. Furthermore, let $\big(X_n^{(p)}\big)$ be a corresponding INAR(p) sequence, where the thinning coefficients are truncated after the p-th lag. That is, $\big(X_n^{(p)}\big)$ has innovation parameter $\alpha_0^{(p)} := \alpha_0$ and thinning coefficients $\alpha_k^{(p)} := 1_{\{k \leq p\}}\alpha_k$, $k \in \mathbb{N}$. 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 is Proposition 3 in Kirchner (2015). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

We have not worked out the multivariate versions of Theorem 2.4 and Proposition 2.5 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:

<div style="border:1px solid">

Basic Approximation

Let $\mathbf{N}$ be a d-variate Hawkes with baseline-intensity vector η and excitement function H as in (3). Let $\big(\mathbf{X}_n^{(\Delta)}\big)$ be a d-variate INAR(∞) sequence with innovation-parameter vector $\mathbf{a}_0^{(\Delta)} := \Delta\,\eta$ and thinning-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 thinning-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

$$\Big(N(0,\Delta), N(\Delta,2\Delta), \ldots, N\big((m-1)\Delta, m\Delta)\big)\Big)$$

$$\stackrel{d}{\approx} \Big(X_1^{(\Delta)}, X_2^{(\Delta)}, \ldots, X_m^{(\Delta)}\Big)$$

$$\stackrel{d}{\approx} \Big(X_1^{(\Delta,p)}, X_2^{(\Delta,p)}, \ldots, X_m^{(\Delta,p)}\Big), \quad m \in \mathbb{N}.$$

</div>

> If $\mathrm{supp}(H) \subset [0,s]$ for some finite $s > 0$, then the second approximation becomes an equality for all $p \geq \lceil s/\Delta \rceil$.

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 innovation-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 thinning-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 B1.

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. 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 rather than in terms of YW for three reasons: (i) For small sample-sizes, CLS is known to have a lower variance than YW. (ii) The CLS-estimator allows to present the estimation of excitement function and baseline intensity in the same formula. (iii) The asymptotic properties of the YW- and the CLS-estimator are the same. Their derivation, however, is typically done in the CLS-setting. In any case, even for medium sample-sizes, we note only very small differences between MLE-, YW- and CLS-estimates in simulation studies (not illustrated). Inference on CLS-estimation in the univariate INAR(p) context has been discussed, e.g., in Li and Yuan (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 nonlinear. However, as already noticed in Latour (1997), INAR(p) sequences can be represented as standard AR(p) models with white noise innovation terms. This yields ways for inference that are more direct.

Proposition 3.1: *Let* $(\mathbf{X}_n)$ *be a d-dimensional INAR(p) sequence as in Definition 2.3 with innovation-parameter vector* $\mathbf{a}_0 \in \mathbb{R}_{\geq 0}^d \setminus \{0_d\}$ *and thinning-coefficient matrices* $A_k \in \mathbb{R}_{\geq 0}^{d \times d}$, $k = 1, 2, \ldots, p$, *such that* (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*

$$\mathbb{E}\left[\mathbf{u}_n\mathbf{u}_{n'}^\top\right] = \begin{cases} \mathrm{diag}\left(\left(1_{d\times d} - \sum\limits_{k=1}^{p} A_k\right)^{-1}\right), & n = n', \\[2ex] 0_{d\times d}, & n \neq n'. \end{cases}$$

Proof: This can be shown by straightforward (if lengthy) calculations; see Appendix A.1. $\quad\square$

As a consequence of Proposition 3.1, a d-variate INAR(p) process can be represented as a standard d-variate autoregressive time series with (dependent) white-noise errors:

Corollary 3.2: *Let* $(\mathbf{X}_n)$ *be the multivariate INAR(p) sequence and* $(\mathbf{u}_n)$ *the white-noise sequence from Proposition 3.1. 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 3.3: 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

$$\hat{\theta}_{CLS}^{(p,n)} : \mathbb{R}^{d\times n} \longrightarrow \mathbb{R}^{d\times(dp+1)}$$

$$(\mathbf{x}_1,\ldots,\mathbf{x}_n) \longmapsto \hat{\theta}_{CLS}^{(p,n)}(\mathbf{x}_1,\ldots,\mathbf{x}_n) := \mathbf{Y}\,\mathbf{Z}^\top\left(\mathbf{Z}\,\mathbf{Z}^\top\right)^{-1},$$

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 3.4: 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 $a_{i,j}B,\ i = 1,\ldots,m,\ j = 1,\ldots,n$.

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 3.5: *Let* $(\mathbf{X}_n)$ *be a d-dimensional INAR(p) sequence as in Definition 2.3 with innovation-parameter (column) vector* $\mathbf{a}_0 \in \mathbb{R}^d_{\geq 0} \setminus \{0_d\}$ *and thinning-coefficient matrices* $A_k \in \mathbb{R}^{d \times d}_{\geq 0}$, $k \in \{1, 2, \ldots, p\}$, *such that* $\mathrm{spr}\left(\sum_{k=1}^{p} A_k\right) < 1$. *Let*

$$\mathbf{B} := \left(A_1, A_2, \ldots, A_p, \mathbf{a}_0\right) \in \mathbb{R}^{d \times (dp+1)} \quad and$$

$$\hat{\mathbf{B}}^{(n)} := \hat{\theta}_{CLS}^{(p,n)}\left((\mathbf{X}_k)_{k=1,\ldots,n}\right) \in \mathbb{R}^{d \times (dp+1)}$$

the CLS-estimator 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 from Definition 3.3 with respect to* $(\mathbf{X}_k)_{k=1,\ldots,n}$. *Assume that the limit*

$$\frac{1}{n-p} \mathbf{Z}\mathbf{Z}^\top \xrightarrow{\ \mathrm{p}\ } \colon \Gamma \in \mathbb{R}^{(dp+1) \times (dp+1)}, \quad n \longrightarrow \infty, \tag{11}$$

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* $\mathrm{vec}\left(\hat{\mathbf{B}}^{(n)}\right) \in \mathbb{R}^{d^2 p + d}$, *one has, for* $n \to \infty$,

$$\sqrt{n-p}\left(\mathrm{vec}(\hat{\mathbf{B}}^{(n)}) - \mathrm{vec}(\mathbf{B})\right)$$

$$\xrightarrow{\ \mathrm{d}\ } \mathcal{N}_{d^2 p + d}\left(0_{d^2 p + d}, \left(\Gamma^{-1} \otimes 1_{d \times d}\right) W \left(\Gamma^{-1} \otimes 1_{d \times d}\right)\right),$$

where

$$W := \mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d \times d}\right)\mathbf{u}_0\left(\left(\mathbf{Z}_0 \otimes 1_{d \times d}\right)\mathbf{u}_0\right)^\top\right] \in \mathbb{R}^{(d^2 p + d) \times (d^2 p + d)} \tag{12}$$

with

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

Proof: In view of Corollary 3.2, it suffices to prove Theorem 3.5 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 3.5 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 3.5 with the basic approximation from Section 2.5 yields the following estimator for multivariate Hawkes processes:

Definition 3.6: Let $\mathbf{N} = \left(N^{(1)}, N^{(2)}, \ldots, N^{(d)}\right)$ be a d-variate Hawkes process with baseline-intensity vector $\eta \in \mathbb{R}^d_{\geq 0} \setminus \{0_d\}$ and excitement function $H = (h_{i,j}) : \mathbb{R}_{\geq 0} \to \mathbb{R}^{d \times d}_{\geq 0}$ such that (5) holds. Let $T > 0$ and consider a sample of the process on the time interval $(0, T]$. For some

$\Delta > 0$, construct the $\mathbb{N}_0^d$-valued *bin-count sequence* from this sample:

$$\mathbf{X}_k^{(\Delta)} := \left(N^{(j)}\left(\big((k-1)\Delta, k\Delta \big] \right) \right)_{j=1,\ldots,d}, \quad k = 1, 2, \ldots, n := \lfloor T/\Delta \rfloor. \tag{13}$$

Define the *multivariate Hawkes estimator* with respect to some support s, $\Delta < s < T$, by applying the CLS-operator from Definition 3.3 with maximal lag $p := \lceil s/\Delta \rceil$ on these bin-counts:

$$\hat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta} \hat{\theta}_{CLS}^{(p,n)} \left(\left(\mathbf{X}_k^{(\Delta)} \right)_{k=1,\ldots,n} \right). \tag{14}$$

We collect the main properties of the estimator in the following remark.

Remark 1: The following additional notation clarifies what the entries of the $\hat{\mathbf{H}}^{(\Delta,s)}$ matrix actually estimate:

$$\left(\hat{H}_1^{(\Delta,s)}, \ldots, \hat{H}_p^{(\Delta,s)}, \hat{\eta}^{(\Delta,s)} \right) := \hat{\mathbf{H}}^{(\Delta,s)}$$

From Theorem 3.5 on estimation of INAR(p) sequences together with the basic approximation in Section 2.5, we see that, for $0 < t < s$,

$$\left(\hat{H}_{\lfloor t/\Delta \rfloor}^{(\Delta,s)} \right)_{ij}, \quad i, j = 1, \ldots, d, \quad \text{respectively,} \quad \left(\hat{\eta}^{(\Delta,s)} \right)_i, \quad i = 1, \ldots, d,$$

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 3.5 that

$$\text{vec}\left(\hat{\mathbf{H}}^{(\Delta,s)} \right) \overset{\text{approx.}}{\sim} \mathcal{N}_{d^2 p + d}\left(\text{vec}\left(\mathbf{H} \right), \frac{1}{\Delta^2(n-p)} \left(\Gamma^{-1} \otimes 1_{d \times d} \right) W \left(\Gamma^{-1} \otimes 1_{d \times d} \right) \right), \tag{15}$$

where Γ and W are defined as in (11) and (12) with respect to the bin-count sequences. Substituting Γ and W with their empirical versions yields the covariance estimate

$$\hat{S^2} := \frac{1}{\Delta^2} \left(\left(\mathbf{Z}\mathbf{Z}^\top \right)^{-1} \otimes 1_{d \times d} \right) \sum_{k=p+1}^{n} \mathbf{w}_k \mathbf{w}_k^\top \left(\left(\mathbf{Z}\mathbf{Z}^\top \right)^{-1} \otimes 1_{d \times d} \right), \tag{16}$$

where $\mathbf{Z}$ is the design matrix from Definition 3.3 with respect to the bin-count sequence and, for $k = p+1, p+2, \ldots, n$,

$$\mathbf{w}_k := \left(\left(\left(\mathbf{X}_{k-1}^{(\Delta)} \right)^\top, \left(\mathbf{X}_{k-2}^{(\Delta)} \right)^\top, \ldots, \left(\mathbf{X}_{k-p}^{(\Delta)} \right)^\top, 1 \right)^\top \otimes 1_{d \times d} \right)$$

$$\cdot \left(\mathbf{X}_k - \Delta \hat{\eta}^{(\Delta,s)} - \sum_{l=1}^{p} \Delta \hat{H}_l^{(\Delta,s)} \mathbf{X}_{k-l} \right).$$

Following formulas are useful for implementation of confidence intervals:

$$\text{Cov}\left(\left(\hat{H}_{k_1}^{(\Delta,s)} \right)_{i_1,j_1}, \left(\hat{H}_{k_2}^{(\Delta,s)} \right)_{i_2,j_2} \right) = \hat{S}_{(k_1-1)d^2+(j_1-1)d+i_1,\, (k_2-1)d^2+(j_2-1)d+i_2}^2,$$

11

for $i_1, i_2, j_1, j_2 \in \{1, \ldots, d\}$ and $k_1, k_2 \in \{1, \ldots, p\}$.

$$\text{Cov}\left(\hat{\eta}_{i_1}^{(\Delta,s)}, \hat{\eta}_{i_2}^{(\Delta,s)}\right) = \hat{S}^2_{pd^2+i_1,\, pd^2p+i_2}, \quad \text{for } i_1, i_2 \in \{1, ..., d\}.$$

Applying Remark **1** above together with Definitions 3.3 and 3.6, our Hawkes estimation procedure may be implemented in a straightforward manner. However, we emphasize that the resulting matrix $\hat{\mathbf{H}}^{(\Delta,s)}$ in (14) 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 $\mathbb{R}_{\geq 0}$; 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 3.6 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 datasets; see Section 5.3. Finally, one is often only interested in the integral of the excitement; see comments after (5). In this case, it makes more sense to directly add up the estimates rather than to take the detour over some smoothing method.

Finally, we refer to the alternative nonparametric Hawkes estimation approach from Bacry *et al.* (2012). Here, an implicit equation for the autocovariance density of a Hawkes process is solved for the excitement function by Fourier analysis. This approach corresponds to directly applying YW-estimation to the bin-count sequences—if somewhat in disguise. Our procedure highlights the underlying approximation principle. This explicit connection with powerful time series theory seems more fertile than the manipulations in Fourier space: it is more intuitive, simpler to implement and yields much simpler ways for inference.

3.3 *Simulation studies*

We check the distributional properties of the Hawkes estimator collected in Remark **1** in a simulation study. The results are summarized in Figure B3. 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

$$H(t) = \begin{pmatrix} h_{1,1}(t) & h_{1,2}(t) \\ h_{2,1}(t) & h_{2,2}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1_{1<t\leq 3}0.25 \\ 0.5(1+t)^{-2} & 1_{t\leq\pi}0.2\sin(t) \end{pmatrix}; \tag{17}$$

see Figure B2 for this parametrization and Figure B1 for an estimation of a single realization. In each simulation, about $5\,000$ events in each component are generated and our Hawkes estimator (14) 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 crossexcitement 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 (16). 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 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.

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 B4. For various Δ, s and T, we calculate (16), the estimated covariance of the estimator matrix with respect to a single very large sample. 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.

4. Refinements

Our Hawkes estimator $\hat{\mathbf{H}}^{(\Delta, s)}$ from Definition 3.6 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 *The choice of support*

Estimating the support of the excitement function of a Hawkes process corresponds to estimating the largest lag of a nonzero thinning-coefficient (matrix) of the approximating INAR sequence. In view of the VAR(p) representation of INAR(p) sequences from Corollary 3.2, 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 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 such that there is about one event in average per bin and component is a good choice. With respect to this Δ_0, we calculate the bin-count sequence(s) as in (13). Now let $n_0 := \lfloor T/\Delta_0 \rfloor$ and $s_0 > 0$ be some very large support, e.g., $s_0 = T/10$. Then, for $p \in \{1, 2, \ldots, \lceil s_0/\Delta_0 \rceil\}$, we calculate Akaike's information criterion

$$\mathrm{AIC}(p) := \log\left(\det \hat{\Sigma}(p)\right) + \frac{2pd^2}{n_0 - p}, \tag{18}$$

where $\hat{\Sigma}(p) := 1/(n_0 - p) \sum_{k=p+1}^{n_0} \hat{\mathbf{u}}_k^{(p)} \hat{\mathbf{u}}_k^{(p)^\top}$. Here, with the notation from Remark **1**,

$$\hat{\mathbf{u}}_k^{(p)} := \mathbf{X}_k - \Delta_0 \hat{\eta}^{(\Delta_0, s)} - \sum_{l=1}^{p} \Delta_0 \hat{H}_l^{(\Delta_0, s)} \mathbf{X}_{k-l}, \quad k = p+1, p+2, \ldots, n_0,$$

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 (18). Finally, we choose $\hat{p}^{(\mathrm{AIC})} := \mathrm{argmin}_{p \leq \lceil s_0/\Delta_0 \rceil} \mathrm{AIC}(p)$ as estimated maximal lag for the approximating INAR model, respectively, we choose $\hat{s}^{(\mathrm{AIC})} := \hat{p}^{(\mathrm{AIC})} \Delta_0$ as support parameter in the calculation of the Hawkes estimator (14).

In parametric Hawkes setups, the support of the Hawkes excitement function is typically chosen infinite. Our estimation procedure, however, assumes finite excitement. This is not a big issue:

from (5), we get that the excitement-function components vanish for large times. In other words, the influence of the tail of the excitement on the model is negligible; see (8) and Proposition 2.5. 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. For this choice also, the AIC-approach from above can help. Applying AIC-selection naively on data generated from a Hawkes model with infinite support yields support values so that the truncated part of the true excitwent functions vanishes; see the simulation study below. As a side remark note that also the standard parametric approach suffers from cut-off errors as we only observe data in finite time-windows.

We perform a simulation study for the support-choice methods described above. These two univariate cases are summarized in Figure B6. We consider different Hawkes models: univariate and bivariate processes, both with finite as well as with infinite excitement-function support. For the univariate case, we first simulate from a Hawkes model with excitement function $h(t) := \exp(-t)1_{t \leq 3}$. and calculate the AIC-minimizing support with respect to different preliminary Δ_0. All results catch the true support very well. Next, we consider the case of infinite support $t \mapsto \exp(-\alpha t)$ with respect to three different decay parameters $\alpha \in \{1.1, 1.5, 2\}$. Again, we simulate large samples 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, $\int_{\hat{s}^{(\mathrm{AIC})}(\alpha)}^{\infty} \exp(-\alpha t)\mathrm{d}t$, are very small (less than 10^{-3}).

Secondly, we consider a bivariate Hawkes model with the excitement function H from (17). We realize a single large sample from this model. Then we simulate another sample from a truncated version of the model, with

$$H^{(\mathrm{tr})}(t) = \begin{pmatrix} h_{1,1}(t) & h_{1,2}(t) \\ h_{2,1}^{(\mathrm{tr})}(t) & h_{2,2}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1_{1<t\leq 3}0.25 \\ 1_{t\leq 4}0.5(1+t)^{-2} & 1_{t\leq\pi}0.2\sin(t) \end{pmatrix}.$$

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*

In the following, we discuss the choice of the bin size $\Delta > 0$ for the Hawkes estimator $\hat{\mathbf{H}}^{(\Delta,s)}$ from Definition 3.6. We suppose a reasonable support $s > 0$ has already been chosen by a procedure as described in Section 4.1. 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 (7) and (9). At the same time, due to the $1/\Delta$ factor in the calculation of the estimator matrix $\hat{\mathbf{H}}^{(\Delta,s)}$ from (14), its (componentwise) variance increases when Δ decreases. In a simulation study, we simulate 100 times from a Hawkes model with excitement function $h(t) = \exp(-1.1t)$. For each sample, we calculate the Hawkes estimator with respect to three different bin sizes $\Delta \in \{0.1, 0.5, 1\}$. Figure B5 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. Concerning the large variance, we should keep in mind that the final goal may be an estimation for the excitement-function components h_{ij}—and not only 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 balances the increase in pointwise variance. So if the goal of the estimation procedure is a completely specified Hawkes model, then the smallest Δ that is numerically convenient may be chosen; see the discussion after Remark **1**. The following toy-example clarifies things:

We consider a smoothing method for which we can approximately calculate the variance. Namely, we define a box moving-average with window size $\tau > 0$. For some bin size $\Delta > 0$,

we consider a univariate Hawkes selfexcitement estimate $\hat{\mathbf{H}}^{(\Delta,p\Delta)} = (\hat{h}_1^{(\Delta)}, \ldots, \hat{h}_p^{(\Delta)}, \hat{\eta}^{(\Delta)})$. As a smoothed function-estimate, we set

$$\hat{h}^{(\Delta)} : t \mapsto \frac{1}{\#\{k : k\Delta \in [t \pm \tau/2]\}} \sum_{k:\, k\Delta \in [t \pm \tau/2]} \hat{h}_k^{(\Delta)}, \quad t \geq 0.$$

Then, using that $\mathrm{Var}\left(\hat{h}_k^{(\Delta)}\right) \approx c/\Delta$, see Figure B4, $\mathrm{Cov}(\hat{h}_k^{(\Delta)}, \hat{h}_l^{(\Delta)}) \approx 0$, $k \neq l$, and that $\#\{k : k\Delta \in [t \pm \tau/2]\} \approx \tau/\Delta$, we obtain

$$\mathrm{Var}\left(\hat{h}^{(\Delta)}(t)\right) \approx \frac{1}{(\tau/\Delta)^2} \sum_{k:\, k\Delta \in [t \pm \tau/2]} \mathrm{Var}\left(\hat{h}_k^{(\Delta)}\right) \approx \frac{c}{\tau}.$$

In other words, the variance of the smoothed estimate is approximately constant in Δ. The same effect can be observed empirically with more sophisticated smoothing methods: for a single large simulated sample from a Hawkes process, we calculate the pointwise Hawkes estimator from Definition 3.6 with respect to three different bin-sizes Δ. Applying a cubic smoothing-spline procedure on the three results one findes that the smoothed functions are hardly affected by the choice of Δ.

If we choose a very small bin-size Δ, computation time becomes an issue. The calculations in (14) require the construction of the design matrix $\mathbf{Z}$ from Definition 3.3 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).

We now understand that the trade-off related to the bin-size choice is not so much a bias/variance trade-off but more a bias/numerical-issues trade-off! To check, if we have chosen Δ small enough, we propose to calculate the (biased) estimate of the baseline intensity vector $\eta := (\eta_i)_{1 \leq i \leq d}$ for a decreasing sequence of bin sizes $\Delta_0 > \Delta_1 > \Delta_2 > \ldots$ The variance of the estimates $\hat{\eta}_i^{(s,\Delta_n)}$, $n = 0, 1, \ldots$ is approximately constant over the different bin-sizes; see Figure B4. This makes the estimates comparable. For $i = 1, 2, \ldots, d$, we plot the values $\left(\hat{\eta}_i^{(s,\Delta_n)}\right)_{n=0,1,\ldots}$ against $(\Delta_n)_{n=0,1,\ldots}$. 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 3.6. Reasonable graphical results as in Figure B1 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 (2003), 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 $\left(T_k^{(i)}\right)_{k \in \mathbb{Z}}$, $i = 1, \ldots, d$, from a d-variate point process with conditional intensity $\Lambda = \left(\Lambda^{(i)}\right)_{i=1,\ldots,d}$, one has that $\int_{T_k^{(i)}}^{T_{k+1}^{(i)}} \Lambda^{(i)}(t) \, \mathrm{d}t \sim \mathrm{Exp}(1)$ independently over $i = 1, \ldots, d$ 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 $\mathrm{Exp}(1)$-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-datasets 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 3.6 allows to analyze the order flow in a Hawkes setup. 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 the context and the preparation of the data quite detailed.

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 *et al.* (2013). Our particular data sample was provided by TickData inc. It stems from September 2013. We have a separate dataset 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) Cancelation of some limit orders

In the trade data set, we see the traded price and the number of contracts traded.

In both datasets, 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. As data is so relatively sparse, the multiple events cannot be accidental. This leaves two possibilities: either the multiples stem from a single order that has been split (for some technical reason) or the multiples are almost instantaneous responses to each other that are reported at the same millisecond due to rounding. We may safely rule out this second possibility: as yet, it is impossible to "react" (i.e.: observe an order, send an order, let the electronic book record or match the new order) in less than a millisecond. In addition, 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 datasets only once. After the thinning procedure, we derive two one-dimensional event datasets 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 busy trading hours, i.e., between 8:30am and 3:15pm (Chicago 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 of 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 3.6, we fit a bivariate Hawkes process $(N^{(\mathcal{T})}, N^{(\mathcal{L})})$ on a single 30min-sample of the data $(\mathcal{T}, \mathcal{L})$, 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:

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. 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 2.8 sec. 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 B6. 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 only take not more than the last three seconds of the LOB-history into account.

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 $s = 3$ sec fixed and, for different bin-size candidates Δ, calculate the baseline-intensity estimate $\hat{\eta}_i^{(\Delta)}$, $i = 1, 2$, together with the corresponding confidence intervals; see (14) and (15) for the necessary calculations. We observe a monotone relation between the bin-size candidates and the corresponding baseline-estimates. However, for $\Delta \leq 0.01$ sec, the differences of the estimates are of a lower order than their (estimated) confidence intervals. So it is sensible to assume that, for this particular sample, the bias of our estimation method becomes negligible for bin-size choices of $\Delta \leq 0.01$ sec. From the bivariate event dataset, we finally calculate the Hawkes estimator from

Definition 3.6 with respect to support $s = 3\,\text{sec}$ and bin size $\Delta = 0.01\,\text{sec}$. Figure B8 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 selfexcitement. The crossexcitement, however, is obviously directed: we observe a very strong crossexcitement from $\mathcal{T}$ on $\mathcal{L}$, but hardly an effect from $\mathcal{L}$ on $\mathcal{T}$. The estimated interactions can be summarized in the branching-matrix estimate

$$\begin{pmatrix} 0.62(\pm 0.04) & 0.03(\pm 0.01) \\ 0.55(\pm 0.06) & 0.54(\pm 0.03) \end{pmatrix}, \text{i.e.,} \quad {}^{``}\begin{pmatrix} \mathcal{T} \stackrel{0.62}{\rightsquigarrow} \mathcal{T} & \mathcal{L} \stackrel{0.03}{\rightsquigarrow} \mathcal{T} \\ \mathcal{T} \stackrel{0.55}{\rightsquigarrow} \mathcal{L} & \mathcal{L} \stackrel{0.54}{\rightsquigarrow} \mathcal{L} \end{pmatrix}{}^{"}. \tag{19}$$

See Remark **1** for the calculation of the point estimates as well as the 95%-confidence bounds of the branching-matrix components. Also see the explanations after (5) for the interpretation of the branching-matrix that is indicated in the right matrix. The largest eigenvalue of matrix (19), i.e., the stability-criterion estimate, is 0.72. The strong asymmetry in (19) may be interpreted such that the trades cause the limit orders (and cancelations) and not vice versa. In further analysis, we found that the estimated branching-matrix, and in particular the asymmetric crossexcitement, is quite stable over all thirty minute windows of the busy trading hours (not illustrated). In the crossexcitement 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. This time however, we ignore the best support choice and set it naively to $s = 0.1\,\text{sec}$ only. In addition, we apply an extremely small bin size of $\Delta = 0.002\,\text{sec}$. In the first milliseconds after each event, the results indicate an inhibitory effect; the Hawkes model does not allow for negative excitement. Still, this result is not surprising. It reflects the fact that it takes at least 3 or 4 milliseconds for any market participant to observe and react to a change in the LOB. In the smoothed function-estimate of the selfexcitement of the first component (the trades process), we detect local maxima at 0.1 sec multiples. As above, this may be is a sign, that 0.1 sec is the "resolution" of some of the algorithms that drive the market. 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 $\hat{H} : \mathbb{R}_{\geq 0} \to \mathbb{R}^{2 \times 2}$. 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.

5.3 *Interpretation of the estimation results*

The interpretation of the estimation results from Section 5.2 is not straightforward: observing the income 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 our modeling approach however, any fluctuation of exogenous processes that influence the observed event-process will also be detected as selfexcitement. Candidates for such 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 will yield new problems. So the best way to get rid of artificial selfexcitement 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 crossexcitement in data. However, we ought to be careful with interpretation of these terms.

The Hawkes fit is meaningful and fertile despite of the criticism above and despite of the vanishing p-values in our application: plots of excitement estimates as in Figure B8 are visualizations of huge event-datasets 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, an excitement-function estimate $\hat{h}_{ij}(\cdot)$ is also theoretically meaningful. It is an estimate for the best linear filter of $\mathbb{E}\left[N_i(\{t\})/\mathrm{d}t|\sigma\left(N_j(\{s\}), s < t\right)\right]$ 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 the discretization involved. 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.

Due to space constraints, the presentation leaves out obvious subsequent topics. *Confidence bounds for the rate of endogeneity* (the branching parameter of a univariate Hawkes process), estimation of the excitement function on a *nonequidistant estimation-grid*, derivation of *power-law decay-parameter estimates* in the parametric case via a linear regression on the log/log values of the pointwise estimates, and *estimation of marked Hawkes processes*: using the concept of our paper as a starting point, all these aims can be achieved in a straightforward manner.

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

References

Aït-Sahalia, Y., Cacho-Diaz, J. and Laeven, R., Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 2015, p. in print.

Akaike, H., Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, Budapest*, 1973, pp. 276–281.

Bacry, E., Dayri, K. and Muzy, J., Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency data. *The European Physical Journal B*, 2012, **85**, 157.

Bacry, E., Delattre, S., Hofmann, M. and Muzy, J., Scaling limits for Hawkes processes and financial data modeling. *The European Physical Journal*, 2011 `http://arxiv.org/abs/****`.

Bates, D. and Maechler, M., Matrix: Sparse and Dense Matrix Classes and Methods. *R package version 1.1-5*, 2015 `http://CRAN.R-project.org/package=Matrix`.

Bowsher, C., Modelling security market events in continuous time: intensity based, multivariate point process models. *Nuffield College Economics Discussion Papers*, 2002, pp. 1–55.

Brémaud, P. and Massoulié, L., Stability of nonlinear Hawkes processes. *The Annals of Probability*, 1996, **24**, 1563–1588.

Brémaud, P. and Massoulié, L., Hawkes branching processes without ancestors. *Journal of Applied Probability*, 2001, **38**, 122–135.

Brown, T. and Nair, M., A simple proof of the multivariate random time change theorem for point processes. *Journal of Applied Probability*, 1988, **25**, 210–214.

Chavez-Demoulin, V., Davison, A. and McNeil, A., Estimating value-at-risk: a point process approach. *Quantitative Finance*, 2005, **5**, 227–234.

Chavez-Demoulin, V. and McGill, J., High-frequency financial data modeling using Hawkes processes. *Journal of Banking and Finance*, 2012, **36**, 3415–3426.

Crane, R. and Sornette, D., Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 2008, **105**, 15649–15653.

da Silva, I.M., Contributions to the Analysis of Discrete-Valued Time Series. PhD thesis, Departamento de Matematica Aplicada Faculdade de Ciencias da Universidade do Porto, 2005.

Daley, D. and Vere-Jones, D., *An Introduction to the Theory of Point Processes*, Second Edition , Vol. I and II, , 2003, Springer.

Durrett, R., *Probability: Theory and Examples*, Second , 1995, Duxbury Press.

Embrechts, P., Liniger, T. and Lu, L., Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 2011, **48(A)**, 367–378.

Errais, E., Gieseke, K. and Goldberg, L., Affine point processes and portfolio credit risk. *Society for Industrial and Applied Mathematics: Journal on Financial Mathematics*, 2010, **1**, 642–665.

Filiminov, V. and Sornette, D., Scaling limits for Hawkes processes and financial data modeling. *preprint*, 2012.

Fokianos, K. and Kedem, B., *Regression Models for Time Series Analysis*, 2012, Wiley.

Gould, M., Porter, M., Williams, S., McDonald, M., Fenn, D. and Howison, S., Limit order books. , 2013 `http://arxiv.org/pdf/1012.0349.pdf`.

Hamilton, J., *Time Series Analysis*, 1994, Princeton University Press.

Hannan, E., *Multiple Time Series*, 1970, Wiley.

Hardiman, S., Bercot, N. and Bouchaud, J.P., Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B*, 2013, **86**, 442.

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

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

Hawkes, A., A cluster representation of a self-exciting point process. *Journal of Applied Probability*, 1974, **11**, 493–503.

Kim, H., Spatio-Temporal Point Process Models for the Spread of Avian Influenza Virus (H5N1). PhD thesis, University of California, Berkeley, 2011.

Kirchner, M., Hawkes and INAR(∞) processes. Working Paper, ETH Zurich, 2015.

Klimko, L. and Nelson, P., On conditional least squares estimation for stochastic processes. *The Annals of Statistics*, 1978, **6**, 629–642.

Latour, A., The multivariate GINAR(p) process. *Advances in Applied Probability*, 1997, **29**, 228–248.

Li, J.G. and Yuan, Y., The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*, 1991, **12**, 129–142.

Liniger, T., Multivariate Hawkes Processes,. PhD thesis, ETH Zurich, 2009.

Lütkepohl, H., *New Introduction to Multiple Time Series Analysis*, 2005, Springer.

McNeil, A., Frey, R. and Embrechts, P., *Quantitative Risk Management*, 2005, Duxbury press.

Meyer, P., Démonstration simplifiée d'un théorème de Knight.. *Lecture Notes in Mathematics*, 1971, **191**, 191–195.

Mikosch, T. and Stărică, C., Is it really long memory that we see in financial returns?. *Extremes and integrated risk management*, 2000, pp. 149–168.

Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P. and Tita, G.E., Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2011, **106**, 100–108.

Muzy, J. and Bacry, E., Hawkes model for price and trades high frequency dynamics. *Quantitative Finance*, 2014, pp. 1–10.

Ogata, Y., Statistical models for earthquake occurences and residual analysis for point processes. *Journal of the American Statistical Association*, 1988, **83**, 9–27.

Reynaud-Bouret, P. and Schbath, S., Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 2010, **38**, 2781–2822.

Steutel, F. and van Harn, K., Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 1979, **7**, 893–899.

Zhang, H., Wang, D. and Zhu, F., Inference for INAR(p) processes with signed generalized power series thinning operator. *Journal of Statistical Planning and Inference*, 2010, **140**, 676–683.

Zhao, H., A Dynamic Contagion Process for Modelling Contagion Risk in Finance and Insurance. PhD thesis, The London School of Economics and Political Science, 2012.

Appendix A: Proofs

A.1 *Proof of Proposition 3.1*

First, we establish that

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

defines a white noise sequence. Stationarity of $(\mathbf{u}_n)$ follows from the stationarity of $(\mathbf{X}_n)$. For the sequel of the proof, fix any $n \in \mathbb{Z}$. From the property $\mathbb{E}\left[A \circledast \mathbf{X}_n\right] = A \, \mathbb{E}\, \mathbf{X}_n$, $A \in \mathbb{R}_{\geq 0}^{d \times d}$, of the thinning operation from Definition 2.2 we get

$$
\begin{aligned}
\mathbb{E}\,\mathbf{u}_n &= \mathbb{E}\left[\mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^{p} A_k \mathbf{X}_{n-k}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{p} A_k \circledast \mathbf{X}_{n-k} + \varepsilon_n - \mathbf{a}_0 - \sum_{k=1}^{p} A_k \mathbf{X}_{n-k}\right] \\
&= \sum_{k=1}^{p} A_k \, \mathbb{E}\left[\mathbf{X}_{n-k}\right] + \mathbf{a}_0 - \mathbf{a}_0 - \sum_{k=1}^{p} A_k \, \mathbb{E}\left[\mathbf{X}_{n-k}\right] \\
&= 0_d.
\end{aligned}
$$

For the autocovariances of the sequence $(\mathbf{u}_n)$, first note that the error $\mathbf{u}_n$ is uncorrelated with any previous value $\mathbf{X}_{n'}, n' < n$, of the original sequence. Indeed:

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{u}_n \mathbf{X}_{n'}^\top\right] &= \mathbb{E}\left[\left(\mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^p A_k \mathbf{X}_{n-k}\right)\mathbf{X}_{n'}^\top\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}_n - \mathbb{E}\left[\mathbf{X}_n \,|\, \sigma\left(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots\right)\right]\right)\mathbf{X}_{n'}^\top\right] \\
&= \mathbb{E}\left[\mathbf{X}_n \mathbf{X}_{n'}^\top\right] - \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_n \mathbf{X}_{n'}^\top \,|\, \sigma\left(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots\right)\right]\right] \\
&= 0_{d\times d}.
\end{aligned}
\tag{A1}
$$

So that we get, for $n' < n$ (and then, by symmetry, for $n' \neq n$),

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{u}_n \mathbf{u}_{n'}^\top\right] &= \mathbb{E}\left[\mathbf{u}_n \left(\mathbf{X}_{n'} - \mathbf{a}_0 - \sum_{k=1}^p A_k \mathbf{X}_{n'-k}\right)^\top\right] \\
&= \mathbb{E}\left[\mathbf{u}_n \mathbf{X}_{n'}^\top\right] - \mathbb{E}\left[\mathbf{u}_n \mathbf{a}_0^\top\right] - \sum_{k=1}^p A_k \mathbb{E}\left[\mathbf{u}_n \mathbf{X}_{n'-k}^\top\right] \\
&\overset{(A1)}{=} -\mathbb{E}\left[\mathbf{u}_n\right]\mathbf{a}_0^\top \\
&= 0_{d\times d}.
\end{aligned}
$$

We have established that $(\mathbf{u}_n)$ is a white noise sequence. In a second step, we derive its marginal covariance matrix. Since $\mathbb{E}\,\mathbf{u}_n = 0_d$ and $\mathbb{E}\left[\mathbf{u}_n X_{n-k}^\top\right] = 0_{d\times d}$, $k = 1, \dots, p$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{u}_n \mathbf{u}_n^\top\right] \\
&= \mathbb{E}\left[\mathbf{u}_n \left(\mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^p A_k \mathbf{X}_{n-k}\right)^\top\right] \\
&\overset{(A1)}{=} \mathbb{E}\left[\mathbf{u}_n \mathbf{X}_n^\top\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}_n - \mathbf{a}_0 - \sum_{k=1}^p A_k X_{n-k}\right)\left(\varepsilon_n + \sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top\right] \\
&= \mathbb{E}\left[\left(\varepsilon_n + \sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k} - \mathbf{a}_0 - \sum_{k=1}^p A_k X_{n-k}\right)\left(\varepsilon_n + \sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top\right] \\
&= \mathbb{E}\left[\varepsilon_n \varepsilon^\top + \varepsilon_n \left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top + \sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\, \varepsilon^\top + \sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top\right. \\
&\qquad \left. -\mathbf{a}_0\, \varepsilon_n^\top - \mathbf{a}_0 \left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top - \sum_{k=1}^p A_k X_{n-k} \varepsilon_n^\top - \sum_{k=1}^p A_k X_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top\right].
\end{aligned}
$$

Using independency of ε_n from the past of the process and plugging in $\mathbb{E}\,\varepsilon_n = \mathbf{a}_0$ yields

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{u}_n \mathbf{u}_n^\top\right] \\
&= \mathrm{Cov}(\varepsilon_n) + \mathbb{E}\left[\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top - \sum_{k=1}^p A_k \mathbf{X}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top\right]
\end{aligned}
$$
(A2)

As the components of ε_n are Poisson distributed and mutually independent, we have $\mathrm{Cov}\left(\varepsilon_n\right) = \mathrm{diag}(\mathbf{a}_0)$. For the second term in (A2), we condition the difference in the expectation on the past of the process. Then the thinnings become the only source of randomness. Furthermore, the counting series of the thinnings are independent of $\sigma\left(X_{n-1}, X_{n-2}, \dots\right)$. So:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top - \sum_{k=1}^p A_k \mathbf{X}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{X}_{n-k}\right)^\top \middle| \sigma\left(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots\right)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\right)^\top - \sum_{k=1}^p A_k \mathbf{x}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\right)^\top\right]_{\mathbf{x}_{n-k}=\mathbf{X}_{n-k}, k=1,\dots,p} \\
&= \mathbb{E}\left[\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\left(\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\right)^\top - \sum_{k=1}^p A_k \mathbf{x}_{n-k}\left(\sum_{k=1}^p A_k \mathbf{X}_{n-k}\right)^\top\right]_{\mathbf{x}_{n-k}=\mathbf{X}_{n-k}, k=1,\dots,p} \\
&= \mathrm{Cov}\left(\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\right)_{\mathbf{x}_{n-k}=\mathbf{X}_{n-k}, k=1,\dots,p}
\end{aligned}
$$

For fixed $\mathbf{x} \in \mathbb{N}_0^d$ and $(\alpha_{i,j}) := A \in \mathbb{R}_{\geq 0}^{d\times d}$ we have

$$
A \circledast \mathbf{x} = \begin{pmatrix} \sum_{j=1}^d \alpha_{1,j} \circ \mathbf{x}_j \\ \dots \\ \sum_{j=1}^d \alpha_{d,j} \circ \mathbf{x}_j \end{pmatrix}.
$$

The components $\sum_{j=1}^d \alpha_{i,j} \circ \mathbf{x}_j$, $i = 1, \dots, d$, of this vector are Poisson distributed with parameters $\sum_{j=1}^d \alpha_{i,j}\,\mathbf{x}_j$; see Definition 2.2. As the thinnings involved are all independent by definition, the components are uncorrelated. Therefore, the covariance matrix of the vector is $\mathrm{Cov}\left(A \circledast \mathbf{x}\right) = \mathrm{diag}\left(A\,\mathbf{x}\right)$ and

$$
\mathrm{Cov}\left(\sum_{k=1}^p A_k \circledast \mathbf{x}_{n-k}\right) = \sum_{k=1}^p \mathrm{diag}\left(A_k\,\mathbf{x}_{n-k}\right) = \mathrm{diag}\left(\sum_{k=1}^p A_k\,\mathbf{x}_{n-k}\right).
$$

Plugging in the random variables $\mathbf{X}_{n-k}$ and taking the expectation, we continue from (A2) and find

$$\mathbb{E}\left[\mathbf{u}_n\mathbf{u}_n^\top\right] = \text{diag}\left(\mathbf{a}_0\right) + \mathbb{E}\,\text{diag}\left(\sum_{k=1}^{p} A_k\,\mathbf{X}_{n-k}\right)$$

$$= \text{diag}\left(\mathbf{a}_0 + \sum_{k=1}^{p} A_k\,\mathbb{E}\,\mathbf{X}_{n-k}\right)$$

$$= \text{diag}\left(\mathbf{a}_0 + \sum_{k=1}^{p} A_k\left(1_{d\times d} - \sum_{j=1}^{p} A_j\right)^{-1}\mathbf{a}_0\right)$$

$$= \text{diag}\left(\left(1_{d\times d} - \sum_{j=1}^{p} A_j\right)^{-1}\mathbf{a}_0\right).$$

$\square$

A.2 *Proof of Theorem 3.5*

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 3.2. 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 3.3 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

$$\frac{1}{n-p}\,\mathbf{Z}\,\mathbf{Z}^\top \xrightarrow{p} \colon \Gamma \in \mathbb{R}^{(dp+1)\times(dp+1)}, \quad n \longrightarrow \infty, \tag{A3}$$

exists and is invertible. In addition, we use that, for $n \longrightarrow \infty$,

$$\frac{1}{\sqrt{n-p}}\text{vec}\left(\mathbf{U}\,\mathbf{Z}^\top\right)$$

$$\xrightarrow{\text{d}} \mathcal{N}_{d^2p+d}\left(0_{d^2p+d}, \mathbb{E}\left[\left(\mathbf{Z}_0\otimes 1_{d\times d}\right)\mathbf{u}_0\left(\left(\mathbf{Z}_0\otimes 1_{d\times d}\right)\mathbf{u}_0\right)^\top\right]\right), \tag{A4}$$

where $\mathbf{Z}_0 := \left(\mathbf{X}_{-1}^\top, \mathbf{X}_{-2}^\top, \ldots, \mathbf{X}_{-p}^\top, 1\right)^\top \in \mathbb{N}_0^{(pd+1)\times 1}$ has the same distribution as any of the columns of the design matrix $\mathbf{Z}$. We postpone the reasoning for (A4) 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

$$\mathbf{Y} := (\mathbf{X}_{p+1}, \mathbf{X}_{p+1}, \ldots, \mathbf{X}_n) = \mathbf{B}\,\mathbf{Z} + \mathbf{U}\left(\in \mathbb{N}_0^{d\times(n-p)}\right); \tag{A5}$$

see Definition 3.3.

$$
\begin{aligned}
\hat{\mathbf{B}}^{(n)} - \mathbf{B} &= \mathbf{Y}\,\mathbf{Z}^{\top}\left(\mathbf{Z}\,\mathbf{Z}^{\top}\right)^{-1} - \mathbf{B} \\
&\overset{(A5)}{=} \left(\mathbf{B}\,\mathbf{Z} + \mathbf{U}\right)\mathbf{Z}^{\top}\left(\mathbf{Z}\,\mathbf{Z}^{\top}\right)^{-1} - \mathbf{B} \\
&= \mathbf{U}\,\mathbf{Z}^{\top}\left(\mathbf{Z}\,\mathbf{Z}^{\top}\right)^{-1} \\
&= \frac{\mathbf{U}\,\mathbf{Z}^{\top}}{n-p}\left(\frac{\mathbf{Z}\,\mathbf{Z}^{\top}}{n-p}\right)^{-1}.
\end{aligned}
$$

By (A3), the second factor converges in probability to the constant matrix Γ. By (A4), 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 $\tilde{W}/\sqrt{n-p} \overset{p}{\longrightarrow} 0_{d\times(dp+1)}$ and therefore $\hat{\mathbf{B}}^{(n)} - \mathbf{B} \overset{p}{\longrightarrow} 0_{d\times(dp+1)}$. For establishing the asymptotic distribution, we treat the difference of the estimated and true vectorized parameter-matrix in a similar way:

$$
\begin{aligned}
\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}\left(\mathbf{B}\right) &= \operatorname{vec}\left(\hat{\mathbf{B}}^{(n)} - \mathbf{B}\right) \\
&= \operatorname{vec}\left(\mathbf{U}\,\mathbf{Z}^{\top}\left(\mathbf{Z}\,\mathbf{Z}^{\top}\right)^{-1}\right) \\
&= \left(\left(\mathbf{Z}\,\mathbf{Z}^{\top}\right)^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\mathbf{U}\,\mathbf{Z}^{\top}\right) \\
&= \frac{1}{\sqrt{n-p}}\left(\left(\frac{\mathbf{Z}\,\mathbf{Z}^{\top}}{n-p}\right)^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\frac{\mathbf{U}\,\mathbf{Z}^{\top}}{\sqrt{n-p}}\right). \quad\quad (A6)
\end{aligned}
$$

In the third step of the calculation above we use that

$$
\operatorname{vec}\left(AB\right) = \left(B^{\top} \otimes I\right)\operatorname{vec}\left(A\right), \quad\quad (A7)
$$

for matrices A, B and identity matrix I such that the calculations are consistent dimensionwise; see A.12 in Lütkepohl (2005). It follows from (A6) together with (A3) that $\sqrt{n-p}\left(\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}\left(\mathbf{B}\right)\right)$ has the same asymptotic distribution as

$$
\left(\Gamma^{-1} \otimes 1_{d\times d}\right)\operatorname{vec}\left(\frac{\mathbf{U}\,\mathbf{Z}^{\top}}{\sqrt{n-p}}\right). \quad\quad (A8)
$$

With (A4), we then find that the asymptotic distribution of (A8)–and therefore of $\sqrt{n-p}\left(\operatorname{vec}\left(\hat{\mathbf{B}}^{(n)}\right) - \operatorname{vec}\left(\mathbf{B}\right)\right)$–is centered normal with covariance matrix

$$
\begin{aligned}
&\left(\Gamma^{-1} \otimes 1_{d\times d}\right)\operatorname{cov}\left(\operatorname{dlim}_{n\to\infty}\operatorname{vec}\left(\frac{\mathbf{U}\,\mathbf{Z}^{\top}}{\sqrt{n-p}}\right)\right)\left(\Gamma^{-1} \otimes 1_{d\times d}\right) \\
&\overset{(A4)}{=} \left(\Gamma^{-1} \otimes 1_{d\times d}\right)\mathbb{E}\left[\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_k\left(\left(\mathbf{Z}_0 \otimes 1_{d\times d}\right)\mathbf{u}_k\right)^{\top}\right]\left(\Gamma^{-1} \otimes 1_{d\times d}\right).
\end{aligned}
$$

We still have to establish (A4). To that aim, we rewrite the left-hand side of (A4) as

$$\frac{1}{\sqrt{n-p}}\mathrm{vec}\left(\mathbf{U}\,\mathbf{Z}^{\top}\right) = \frac{1}{\sqrt{n-p}}\mathrm{vec}\left(\left(\sum_{j=1}^{n-p}\mathbf{Z}_{j,1}^{\top}\,\mathbf{U}_{\cdot,j},\ldots,\sum_{j=1}^{n-p}\mathbf{Z}_{j,dp+1}^{\top}\,\mathbf{U}_{\cdot,j}\right)\right)$$

$$= \frac{1}{\sqrt{n-p}}\sum_{j=1}^{n-p}\mathrm{vec}\left((\mathbf{Z}_{1,j}\,\mathbf{U}_{\cdot,j},\ldots,\mathbf{Z}_{dp+1,j}\,\mathbf{U}_{\cdot,j})\right)$$

$$= \frac{1}{\sqrt{n-p}}\sum_{j=1}^{n-p}\mathrm{vec}\left((\mathbf{U}_{1,j},\ldots,\mathbf{U}_{d,j})^{\top}\,(\mathbf{Z}_{1,j},\mathbf{Z}_{2,j},\ldots,\mathbf{Z}_{dp+1,j})\right)$$

$$= \frac{1}{\sqrt{n-p}}\sum_{k=p+1}^{n}\mathrm{vec}\left(\mathbf{u}_k\cdot\mathbf{Z}_k^{\top}\right),$$

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 $n \longrightarrow \infty$,

$$\frac{1}{\sqrt{n-p}}\sum_{k=p+1}^{n}\mathbf{w}_k \xrightarrow{d} \mathcal{N}_{\tilde{d}}(0_{\tilde{d}}, S).$$

Proof of (a) Define the filtration $(\mathcal{H}_k)$ by setting

$$\mathcal{H}_k := \sigma\left(\left(\mathbf{u}_i,\mathbf{X}_{i-1},\mathbf{X}_{i-2},\ldots,\mathbf{X}_{i-p}\right): \ i \le k\right), k \in \mathbb{Z}.$$

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

$$\mathbb{E}\left[\mathbf{X}_k\,\Big|\mathcal{H}_{k-1}\right] = \mathbb{E}\left[\varepsilon_k + \sum_{m=1}^{p}A_m\circledast\mathbf{X}_{k-m}\,\Big|\mathcal{H}_{k-1}\right] = \mathbf{a}_0 + \sum_{m=1}^{p}A_m\,\mathbf{X}_{k-m},$$

we obtain the martingale difference property:

$$\mathbb{E}\left[\mathbf{u}_k \big| \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\mathbf{X}_k - \mathbf{a}_0 - \sum_{m=1}^{p} A_m \, \mathbf{X}_{k-m} \Big| \mathcal{H}_{k-1}\right]$$

$$= \mathbb{E}\left[\mathbf{X}_k \left| \mathcal{H}_{k-1}\right.\right] - \mathbf{a}_0 - \sum_{m=1}^{p} A_m \, \mathbf{X}_{k-m}$$

$$= 0_d.$$

Proof of (b) Independency of k follows from stationarity of $(\mathbf{w}_k)$. Choose $k = 0$. We need to show that, for $b \in \mathbb{R}^{d(pd+1)} \setminus \{0_{d(pd+1)}\}$,

$$b^{\top} \, \mathbb{E}\left[\mathbf{w}_k \, \mathbf{w}_k^{\top}\right] b = \mathbb{E}\left[b^{\top} \, \mathbf{w}_0 \, \mathbf{w}_0^{\top} \, b^{\top}\right] = \mathrm{Var}\left(b^{\top} \, \mathbf{w}_0\right) > 0. \tag{A9}$$

With (A7), we find

$$\mathbf{w}_0 = \mathrm{vec}\left(\mathbf{u}_0 \cdot \mathbf{Z}_0^{\top}\right) = (\mathbf{Z}_0 \otimes 1_{d \times d}) \, \mathrm{vec}\left(\mathbf{u}_0\right) = (\mathbf{Z}_0 \otimes 1_{d \times d}) \, \mathbf{u}_0 \tag{A10}$$

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 (A9), we define the σ-algebra

$$\mathcal{F} := \sigma\left(\mathbf{X}_{-1}, \dots, \mathbf{X}_{-p}, A_1 \circledast \mathbf{X}_{-1}, \dots, A_p \circledast \mathbf{X}_{-p}\right).$$

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

$$\begin{aligned}
\mathrm{Var}\left(b^{\top} \, \mathbf{w}_0\right) &= \mathbb{E}\left[\mathrm{Var}\left(b^{\top} \, \mathbf{w}_0 \,|\, \mathcal{F}\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[b^{\top} \, \mathbf{w}_0 \,|\, \mathcal{F}\right]\right) \\
&\geq \mathbb{E}\left[\mathrm{Var}\left(b^{\top} \, \mathbf{w}_0 \,|\, \mathcal{F}\right)\right] \\
&\overset{(A10)}{=} \mathbb{E}\left[\mathrm{Var}\left(b^{\top} \left(\mathbf{Z}_0 \otimes 1_{d \times d}\right) \mathbf{u}_0 |\, \mathcal{F}\right)\right] \\
&= \mathbb{E}\left[b^{\top} \left(\mathbf{Z}_0 \otimes 1_{d \times d}\right) \mathrm{Cov}\left(\mathbf{u}_0 |\, \mathcal{F}\right) \left(b^{\top} \left(\mathbf{Z}_0 \otimes 1_{d \times d}\right)\right)^{\top}\right]. \tag{A11}
\end{aligned}$$

Since

$$\mathbf{u}_0 = \mathbf{X}_0 - \mathbf{a}_0 - \sum_{i=1}^{p} A_i \, \mathbf{X}_{-i} = \varepsilon_0 + \sum_{i=1}^{p} A_i \circledast \mathbf{X}_{-i} - \mathbf{a}_0 - \sum_{i=1}^{p} A_i \, \mathbf{X}_{-i}, \tag{A12}$$

the summand ε_0 is the only term that contributes to the conditional covariance matrix in (A11)—the other summands in (A12) are constant with respect to $\mathcal{F}$ and ε_0 is independent of $\mathcal{F}$. So we

have $\mathrm{Cov}\left(\mathbf{u}_0 | \mathcal{F}\right) = \mathrm{Cov}\left(\varepsilon_0 | \mathcal{F}\right) = \mathrm{Cov}\left(\varepsilon_0\right) = \mathrm{diag}(\mathbf{a}_0)$ and continuing with (A11) we find

$$\mathrm{Var}\left(b^\top \mathbf{w}_0\right) \geq \mathbb{E}\left[b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right) \mathrm{diag}\left(\mathbf{a}_0\right) \left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right)\right)^\top\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{d} \mathbf{a}_{0,i} \left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right)\right)_{1,i}^2\right]$$

$$\geq \mathbf{a}_{0,i_0} \mathbb{E}\left[\left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right)\right)_{1,i_0}^2\right] > 0, \tag{A13}$$

where $i_0 \in \{1, 2, \ldots, d\}$ in (A13) 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 (A13) follows because, for $j_0 \in \{1, 2, \ldots, np + d\}$ such that $b_{j_0} \neq 0$, we have that

$$\mathbb{P}\left[\left(b^\top \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right)\right)_{1,i_0} \neq 0\right] = \mathbb{P}\left[b^\top \cdot \left(\mathbf{Z}_0^\top \otimes 1_{d \times d}\right)_{\cdot,i_0} \neq 0\right]$$

$$\geq \mathbb{P}\left[b_{j_0} \mathbf{X}_{k_0,l_0} \neq 0\right]$$

$$= \mathbb{P}\left[\mathbf{X}_{k_0,l_0} \neq 0\right] > 0,$$

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 (2015). 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-dimensional $INAR(1)$ sequence $(\tilde{\mathbf{X}}_k)$; see Latour (1997). It is easily checked that the latter is an irreducible, aperiodic Markov chain on $\mathbb{N}_0^{pd}$. So $(\tilde{\mathbf{X}}_k)$ is ergodic; see Durrett (1995), 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 $(\mathbf{w}_k)$.

$\square$

Appendix B: Figures

Excitation from component 1 on component 1

Excitation from component 2 on component 1

Excitation from component 1 on component 2

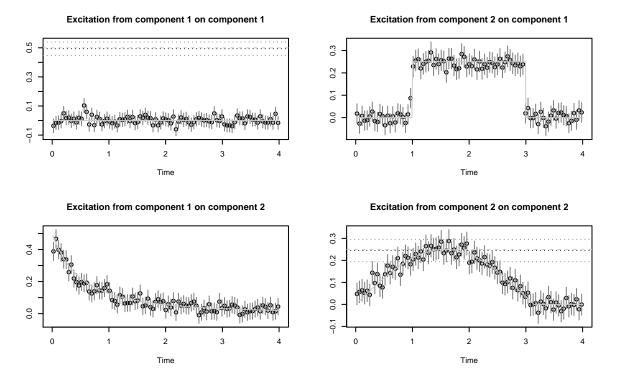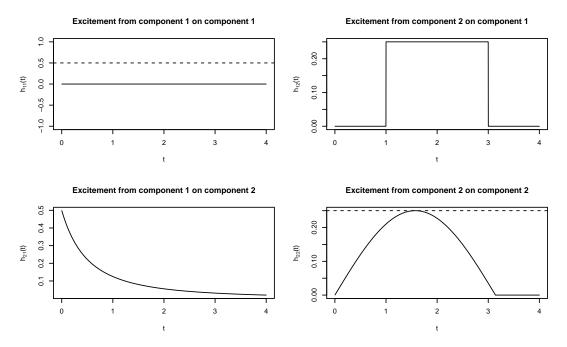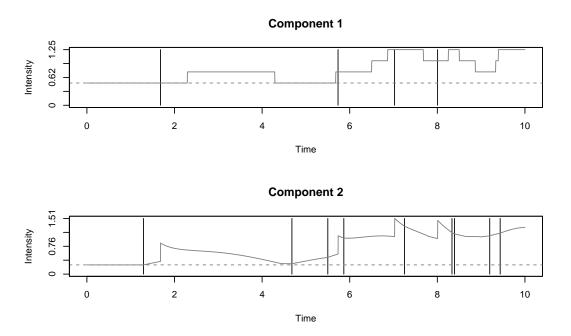**Excitation from component 2 on component 2**

Figure B1. Summary of the main result of the paper. From the bivariate Hawkes model presented in Figure B2, around 100 000 events in each component are simulated. From this single large sample, we calculate the estimator from Definition 3.6. 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 **1**. All solid and dashed lightish-grey lines refer to the true underlying parameters; compare with Figure B2. Eyeball examination shows that the estimation method approximates the form of the true excitement functions well. Also the non-monotonicities 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 B3.

Excitement from component 1 on component 1

Excitement from component 2 on component 1

Excitement from component 1 on component 2

Excitement from component 2 on component 2

(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 selfexcitement functions $h_{1,1}(t) \equiv 0$ and $h_{2,2}(t) = 1_{t \leq \pi} 0.25 \sin(t)$ as well as the two crossexcitement functions $h_{1,2}(t) \equiv 1_{1 < t \leq 3} 0.25$ 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 B1.

Component 1

Component 2

(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 crossexcitement from component 1 on component 2 and also the delayed rectangle impuls impact from component 2 on component 1 are particularly visible.

Figure B2. 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.
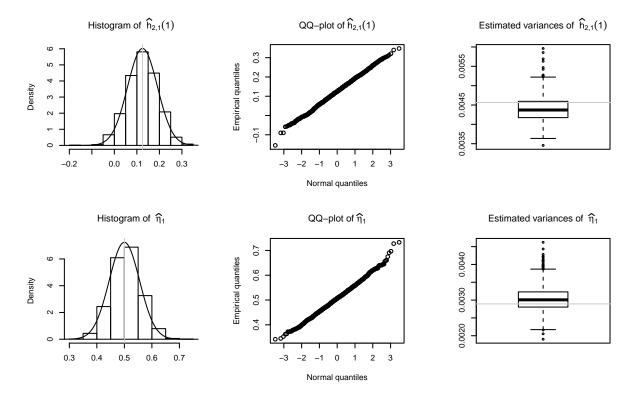
Figure B3. Illustration of the simulation study described in Section 3.3. The study confirms the distributional properties of our estimation procedure collected in Remark **1**: We simulate 2 000 times from the bivariate Hawkes process introduced in Figure B2. In each simulation, we realize about 5 000 events in each component. For all of these samples, we calculate our estimator from Definition 3.6 as well as the covariance estimator from (16). 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 (16). The horizontal grey lines refer to the empirical variance of the 2 000 estimates.
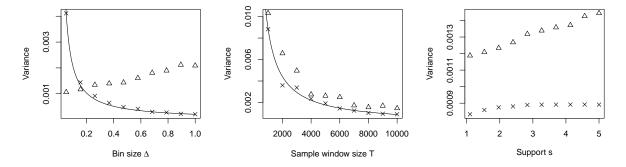
Figure B4. The Hawkes estimator from Definition 3.6 depends on the bin size Δ, on the size of the sample window T and on the support parameter s. We examine empirically how the variances of the estimates depend on these three parameters. We simulate a very large sample from a univariate Hawkes process with excitement function $h : t \mapsto 1_{t \leq 3}(1 + t)^{-2}$ 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 (16). 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.
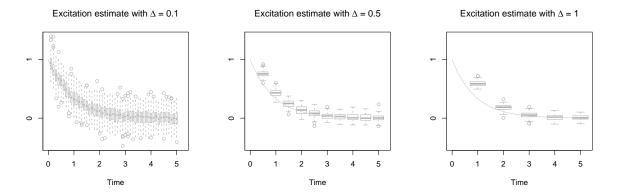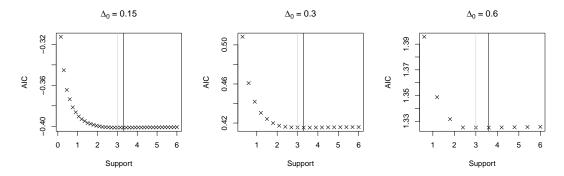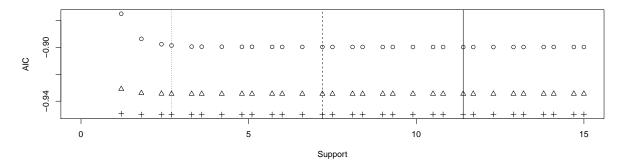

Figure B5. Illustration of the bias/variance trade-off in the choice of the bin size Δ; see Section 4.2. We simulate 100 realizations of a Hawkes process. For each of these 100 samples, we calculate the estimator from Definition 3.6 with respect to three different bin-sizes $\Delta \in \{0.1, 0.5, 1\}$. 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.
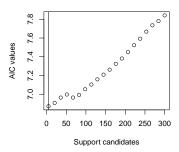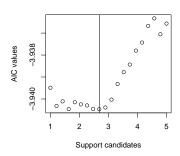
32

(a) We consider a univariate Hawkes process with excitement function $h(t) := 1_{t \leq 3} \exp(-t)$. 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".

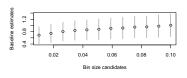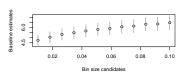

(b) We examine the infinite support case. To that aim, we consider simulated data from three univariate Hawkes process with excitement functions $h_\alpha(t) := \exp(-\alpha t)$, where $\alpha = 1.1$ (circles for AIC-values and solid vertical line for AIC-minimizing support), $\alpha = 1.5$ (triangles 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 (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 B6. Simulation study on the choice of the support parameter of our estimator from Definition 3.6; 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.
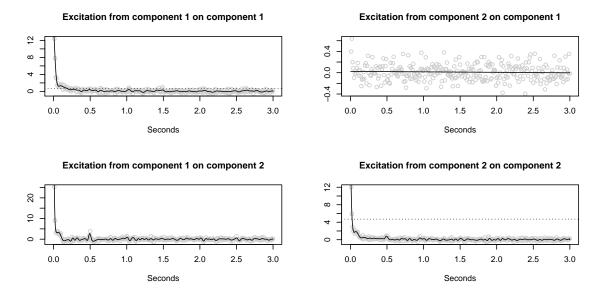
(a) Support analysis with respect to a very coarse preliminary bin-size $\Delta_0 = 1\,\mathrm{sec}$; see Section 4.1. The estimator from Definition 3.6 is calculated for different support candidates (in seconds). The corresponding AIC-values are calculated as in (18). We establish a quite short AIC-optimal support of the excitement function.
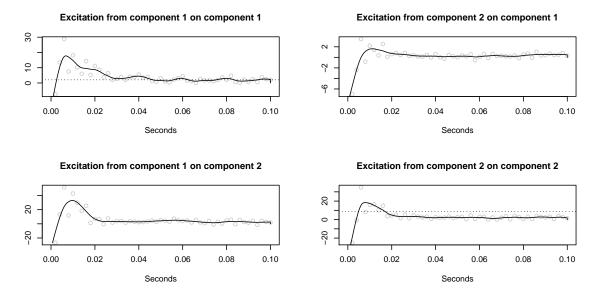
(b) After the rough analysis illustrated in Figure 7(a), we repeat the procedure for smaller support candidates and with respect to a much finer bin-size $\Delta = 0.01\,\mathrm{sec}$. We find an AIC-optimal support value of about 2.7 seconds. This value is stable over other choices of the bin-size. The attained minumum is remarkably clear-cut compared to the attained minimum in the simuation study illustrated in Figure B6.

(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 $0.01\,\mathrm{sec}$, the decrease is of a lower magnitude than the 95%-confidence intervals. We conclude that, for $\Delta \leq 0.01\,\mathrm{sec}$, the bias of our estimation method becomes negligible.

Figure B7. Preliminary analysis for the bivariate data example $(\mathcal{T}, \mathcal{L})$ (trades/limit orders); see Section 5.2. Our nonparametric Hawkes estimator from Definition 3.6 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 $s = 3\,\mathrm{sec}$ and $\Delta = 0.01\,\mathrm{sec}$ are reasonable choices.

Excitation from component 1 on component 1

Excitation from component 2 on component 1

Excitation from component 1 on component 2

Excitation from component 2 on component 2

(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 B7. Eyeball examination reveals local maxima in the lower panels at half seconds.

Excitation from component 1 on component 1

Excitation from component 2 on component 1

Excitation from component 1 on component 2

Excitation from component 2 on component 2

(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-estimate of the selfexcitement of the first component (the trades process), we detect local maxima at 0.1 sec multiples.

Figure B8. Exemplary Hawkes fits of the bivariate data example $(\mathcal{T}, \mathcal{L})$ described in Section 5.2 with respect to 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, respectively, cancelations $\mathcal{L}$. 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.