Noname manuscript No.
(will be inserted by the editor)

Accelerating the estimation of renewal Hawkes self-exciting point processes

Tom Stindl · Feng Chen

Received: date / Accepted: date

Abstract The renewal Hawkes process is a nascent point process model that generalizes the Hawkes process. Although it has shown strong application potential, fitting the renewal Hawkes process to data remains a challenging task, especially on larger datasets. This article tackles this challenge by providing two approaches that significantly reduce the time required to fit renewal Hawkes processes. Since derivative-based methods for optimization, in general, converge faster than derivative-free methods, our first approach is to derive algorithms for evaluating the gradient and Hessian of the log-likelihood function and then use a derivative- based method, such as the Newton-Raphson method, in maximizing the likelihood, instead of the derivative-free method currently being used. Our second approach is to seek linear time algorithms that produce accurate approximations to the likelihood function, and then directly optimize the approximation to the log-likelihood function. Our simulation experiments show that the Newton-Raphson method reduces the computational time by about 30 percent. Furthermore, the approximate likelihood methods produce equally accurate estimates compared to the methods based on the exact likelihood and are about 20-40 times faster on datasets with about 10000 events. We conclude with an analysis of price changes of several currencies relative to the US Dollar.

This research includes computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW Sydney, and the National Computational Infrastructure (NCI) supported by the Australian Government. Data sourced from Thomson Reuters Tick History (TRTH version 2) supplied by Refinitiv (formerly Thomson Reuters). Chen was partly supported by a UNSW SFRGP grant.

T. Stindl
Department of Statistics
E-mail: t.stindl@unsw.edu.au

F. Chen
Department of Statistics
E-mail: feng.chen@unsw.edu.au

Keywords Renewal Hawkes process · Likelihood approximation · Newton-Raphson · Derivative-based Optimization

1 Introduction

The Hawkes process (Hawkes, 1971a,b) is a point process model where a self-exciting effect exists among events of the process so that the occurrence of an event makes future events more likely to happen. The Hawkes process and all of its modifications have had profound impacts on many important application domains. Important financial applications, which motivate the need to conduct inferences on large data sets, include the estimation of risk measures (Chavez-Demoulin et al., 2005; Stindl and Chen, 2019), the arrival of insurance claims following a catastrophe (Klüppelberg and Mikosch, 1995) and high-frequency price fluctuations in financial markets (Embrechts et al., 2011). Other application domains include the clustering effects of earthquakes in seismology (Ogata, 1988), the spread of violence (Lewis et al., 2012) and crime (Lewis and Mohler, 2011), and electrical grid reliability, in which, Ertekin et al. (2015) predicted short term electrical grid failures such as outages, fires and explosions using reactive point processes.

Hawkes and Oakes (1974) showed that the Hawkes process is statistically equivalent to a cluster or branching process, in which immigrants (background events) arrive according to a homogeneous Poisson process, and any event (immigrant or offspring) can generate additional offspring of its own according to a common inhomogeneous Poisson process. As a generalization to the Hawkes process, the renewal Hawkes (RHawkes) process (Wheatley et al., 2016) allows the background event arrival rate to depend on the arrival time of the last immigrant event. More specifically, the background arrival rate resets upon the arrival of an immigrant, that is, immigrants arrive according to a general renewal process rather than the Poisson process as in the Hawkes process.

The additional flexibility afforded by the renewal background event process makes the RHawkes process substantially more flexible than the original Hawkes process. For instance, the point patterns of the RHawkes process can be either over- or under-dispersed compared to a Poisson process. This has shown to be very useful in financial applications. For instance, Chen and Stindl (2018) analyzed ultra-high frequency mid-price changes for the AUD/USD currency exchange rate and found that the shape parameter of the Weibull immigrant inter-renewals were generally estimated to be less than one for the majority of trading days considered. Note that the RHawkes process with Weibull inter-renewals can recover the classical Hawkes process by setting its shape parameter to be unity. This smaller than one shape parameter indicates a heavier clustering of exogenously driven mid-price changes than suggested by the competing Hawkes process. Furthermore, Stindl and Chen (2019) analyzed daily returns and, in particular, the occurrence of extreme negative returns, for several stocks traded on the ASX. They showed that the estimated shape parameter was generally larger than one, indicating a more regular occurrence

pattern for exogenously driven negative extreme returns. These case studies highlight the relevance and importance of the RHawkes process for financial data applications.

However, despite its strong application potential, fitting the RHawkes model to data is challenging, as the likelihood of the RHawkes process is challenging to calculate in general. This is because its intensity process relative to the natural filtration is not easily available, making it cumbersome to apply the general formula for the point process likelihood directly (see, e.g., Daley and Vere-Jones, 2003, Proposition 7.2.III.). To address this difficulty, Wheatley et al. (2016) proposed to obtain the maximum likelihood estimator (MLE) of the model parameters using the celebrated Expectation-Maximization (E-M) algorithm. However, when calculating the conditional expectation of the complete data log-likelihood in the E-step of their E-M algorithms, they used the conditional distribution of the missing data given part of the observed data instead of the whole observed data. Consequently, the resulting estimators are not consistent and, therefore, not the MLE desired. Chen and Stindl (2018) solved the problem by treating the likelihood function as the joint density of the observed data and calculating it using a recursive algorithm. In their simulation experiments, they found that the MLE obtained by directly maximizing the log-likelihood had satisfactory performances.

Although the likelihood evaluation algorithm of Chen and Stindl (2018) makes it viable to fit the RHawkes process model in practice, the algorithm requires the calculation, at each event time, the probabilities of the last immigrant being equal to every past event in the history of the process, and therefore its time-complexity is quadratic. This means it can be intolerably slow when applied to large data sets. However, in financial applications, the Hawkes process and its extensions have been applied extensively to model data that occur in very high frequencies and have an abundance of observations (Cont, 2011; Filimonov and Sornette, 2012, 2015; Large, 2007). Financial data of this type motivate the need to conduct inferences on large datasets expeditiously. This led us to propose faster methods for fitting RHawkes processes, which enables the process to be applicable in real-world trading applications or other applied domains. Our developments are based on two distinct frameworks. The first approach implements the Newton-Raphson method to optimize the log-likelihood function, which typically takes a much smaller number of iterations to converge than derivative-free methods such as the Nelder-Mead downhill simplex method. The second approach employs an approximation to the likelihood function by truncating the distribution of the last immigrant, which is similar in spirit to the approach used by Halpin (2013) to speed up the E-M algorithm for the classical Hawkes process with non-exponential excitation function.

The choice of optimization procedure profoundly impacts the computational time needed for estimation. In Chen and Stindl (2018), the derivative-free Nelder-Mead simplex method was used, which requires a large number of evaluations of the likelihood until convergence. To overcome this, the Newton-Raphson method which requires fewer iterations until convergence can be im-

plemented. To this end, we derive an algorithm for computing the gradient and Hessian of the log-likelihood for the RHawkes process. Once these have been found, the implementation of the Newton-Raphson algorithm is rather simple. The small number of iterations and exact Hessian computation reduces the computational time required for estimation significantly on larger datasets. As a by-product, the exact Hessian matrix calculated at the last cycle of Newton-Raphson iterations is automatically available and used to compute estimators of the variance of the MLE, thereby avoiding the need to compute a numerical approximation to the Hessian in a separate step after finding the maximizer of the log-likelihood function.

Our second approach to fast fitting of the RHawkes model is to optimize approximations to the likelihood rather than the exact value because it is often possible to compute good approximations to the RHawkes process likelihood in a much shorter time than the exact likelihood. We propose some modifications to the iterative algorithms of Chen and Stindl (2018) that achieves significant gains in computational efficiency and memory requirements at the small cost of a minor loss in accuracy of the estimation, which enables fast fitting of the RHawkes process using the maximum likelihood method on much larger data sets. The approximate likelihood approach has previously been successfully used in the estimation of multivariate renewal Hawkes processes (Stindl and Chen, 2018). The likelihood approximation works by truncating the possible candidates for the last immigrant event to events in the recent past. This truncation is justifiable in quite general conditions, since most of the past events, particularly those in the distant past have negligibly small probabilities of being the last immigrant at the current event time. Computing and storing these probabilities unnecessarily slows down the likelihood evaluation algorithm. In Stindl and Chen (2018), the potential candidates for the last immigrant at the i-th event time was restricted to the most recent B events. This speeds up the computational time for the MLE of the multivariate RHawkes processes substantially, with the estimation for some of the simulated data showing a reduction in computational time by a factor of forty while still producing comparable finite sample performance.

The tuning type parameter B, or B_i, if it is allowed to depend on the event index i, embodies a trade-off between estimation accuracy and both the computational complexity and memory requirements. However, as we will show, a well-chosen sequence B_i reduces the computational time significantly without degrading the accuracy of the estimates. We explore two different methods to determine the tuning parameters B_i. The first method is similar to Stindl and Chen (2018), in which $B_i = B$ is fixed before likelihood evaluation and we only compute the last immigrant probabilities for the most recent B events at each step of the algorithm. In a more dynamic approach, B_i can vary at each iteration of the likelihood evaluation algorithm according to the computed last immigrant probabilities at the previous iteration. The B_i at each iteration will depend on the waiting time distribution between successive immigrant events and the waiting time distribution between an event and its direct offspring event.

To explore the computational efficiency and statistical efficiency of the proposed estimation methods for theRHawkes process, we carry out a simulation study and compare the proposed estimation methods with the derivative-free Nelder-Mead algorithm used in Chen and Stindl (2018). The simulations show that the newly proposed estimation methods achieve comparable accuracy and require much shorter running times to obtain estimates and their standard errors. Furthermore, to highlight the applicability of our methodologies, we analyse ultra-high frequency price movement data on eight currency pairs for the American forex trading hours during the four trading weeks in July 2015.

The rest of the article is as follows. In the next section, we provide a brief overview of the RHawkes process and recapitulate the likelihood evaluation algorithm of Chen and Stindl (2018). Section 3 proposes the methods as mentioned earlier for faster estimation of RHawkes processes and outlines their algorithms. Next, in Section 4, we provide numerical evidence of the improvements in computational efficiency and comparable accuracy of the estimates using a simulation study. Section 5 analyses the mid-price changes of several currencies relative to the US Dollar.

2 Model and Notation

Let $0 < \tau_1 < \tau_2 < \dots$ denote an increasing sequence of event times and $N(t)$ be the associated counting process that counts the number of events that have occurred prior to time t. Let $I(t) = \max\{i | \tau_i < t, M_i = 0\}$ indicate the index of the last immigrant prior to time t, where M_i is an (unobservable) indicator that indicates an immigrant event with $M_i = 0$ or an offspring event with $M_i = 1$. The intensity of the renewal Hawkes process with respect to the (non-natural) filtration $\tilde{\mathcal{F}}_t = \sigma\{N(s), I(s); s \le t\}$ takes the form,

$$\tilde{\lambda}(t) = \frac{\mathbb{E}\left[\mathrm{d}N(t)|\tilde{\mathcal{F}}_{t-}\right]}{\mathrm{d}t} = \mu(t - \tau_{I(t)}) + \sum_{j=1}^{N(t-)} g(t - \tau_j), \tag{1}$$

where $\mu : \mathbb{R}^+ \to \mathbb{R}^+$ is a hazard function of the waiting time between successive immigrant events, and $g : \mathbb{R}^+ \to \mathbb{R}^+$ is an offspring excitation function. Let $\phi(t) := \sum_{j=1}^{N(t-)} g(t - \tau_j)$ denote the cumulative sum of the contribution from past events to the event rate. For stability, we require that

$$\int_0^\infty \exp\left(-\int_0^t \mu(s)\mathrm{d}s\right)\mathrm{d}t < \infty \qquad \text{and} \qquad \int_0^\infty g(t)\mathrm{d}t < 1.$$

The former requirement implies that the waiting time between successive immigrants has a finite mean, and the latter requirement implies that the total number of offspring due to any event also has a finite mean.

The objective of this article is to provide faster methods of estimation for the parameter vector $\theta = (\theta_\mu, \theta_g)^\top$ of the RHawkes process, which contains

the parameters of the inter-renewal distribution θ_μ and the offspring excitation function θ_g. To this end, we present the likelihood function based on observations of the RHawkes process over the interval $[0, T]$. Suppose there are exactly n events in the interval, at times $\tau_{1:n} = (\tau_1, \tau_2, \ldots, \tau_n)$. Generally, the likelihood function is computed as a product of the marginal densities for each event time τ_i. However, to have an analytically tractable form for the hazard function between successive events, we must also condition on the index of the last immigrant event. Therefore, the likelihood function as stated in Chen and Stindl (2018) takes the following form,

$$L(\theta) = \mu(\tau_1)e^{-U(\tau_1)} \left\{ \prod_{i=2}^{n} \sum_{j=1}^{i-1} p_{ij}\, d_{ij} \right\} \left\{ \sum_{j=1}^{n} p_{n+1,j}\, S_{n+1,j} \right\}, \tag{2}$$

where $U(t) = \int_0^t \mu(s)\mathrm{d}s$ denotes the cumulative hazard function for immigrants, $p_{ij} = \mathbb{P}\left(I(\tau_i) = j | \tau_{1:i-1}\right)$ denotes the conditional last immigrant probabilities, $d_{ij} = \mathbb{P}\left(\tau_i | \tau_{1:i-1}, I(\tau_i) = j\right)$ denotes the density of τ_i conditional on $\tau_{1:i-1}$ and the index of the last immigrant $I(\tau_i) = j$ and the conditional probability that no events occurred in the interval $(\tau_n, T]$ is denoted by $S_{n+1,j} = \mathbb{P}\left(\tau_{n+1} > T | \tau_{1:n}, I(\tau_{n+1}) = j\right)$. Furthermore, the log-likelihood for which numerical optimization is generally more convenient, takes the form,

$$\ell(\theta) = \log \mu(\tau_1) - U(\tau_1) + \sum_{i=2}^{n} \log \left(\sum_{j=1}^{i-1} p_{ij}\, d_{ij} \right) + \log \left(\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j} \right). \tag{3}$$

The relationship that exists between hazard, density and survival functions (see e.g. Daley and Vere-Jones (2003) Eq. (1.1.1)-(1.1.3)) and from the hazard function of the waiting time between successive events of the RHawkes process conditional on the index of the last immigrant means that simple analytical expressions for the terms appearing in (2) and (3) are available. By conditioning on the previous event times $\tau_{1:i-1}$ and index of the last immigrant event $I(\tau_i) = j$, the hazard function of the inter-event waiting time $\tau_i - \tau_{i-1}$ is $\mu(\cdot + \tau_{i-1} - \tau_j) + \phi(\cdot + \tau_{i-1})$. From here, it can be shown that the conditional densities and survival probabilities are as follows,

$$d_{ij} = (\mu(\tau_i - \tau_j) + \phi(\tau_i)) e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\} - \{\Phi(\tau_i) - \Phi(\tau_{i-1})\}}, \tag{4}$$

$$S_{n+1,j} = e^{-\{U(T - \tau_j) - U(\tau_n - \tau_j)\} - \{\Phi(T) - \Phi(\tau_n)\}}, \tag{5}$$

where $\Phi(t) = \int_0^t \phi(s)\mathrm{d}s = \sum_{j=1}^{N(t-)} G(t - \tau_j)$ and $G(t) = \int_0^t g(s)\mathrm{d}s$. Furthermore, Chen and Stindl (2018) develop a recursive algorithm to compute the conditional last immigrant probabilities p_{ij}. Using the initial condition $p_{21} = 1$, then for each consecutive $i \in \{3, \ldots, n+1\}$ the conditional probabilities are computed as,

$$p_{ij} = \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1} - \tau_j) + \phi(\tau_{i-1})} \frac{d_{i-1,j} p_{i-1,j}}{\sum_{k=1}^{i-2} p_{i-1,k} d_{i-1,k}}, & j = 1, \ldots, i-2 \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1. \end{cases} \tag{6}$$

The proof for the log-likelihood formula in (3) and the last immigrant probabilities in (6) can be found in the Appendix to Chen and Stindl (2018).

3 Estimation methods

Direct likelihood evaluation is now feasible, and MLE can be performed using standard numerical optimization procedures, such as the derivative-free Nelder-Mead simplex method used in Chen and Stindl (2018). The Nelder-Mead algorithm only takes into account the value of the log-likelihood at each iteration. It does not explicitly account for the shape information, such as the slope and curvature, of the log-likelihood surface, and often requires many more iterations to converge than derivative-based optimization methods. Therefore we will consider the well-known derivative-based Newton-Raphson method here.

3.1 Likelihood optimization using the Newton-Raphson method

We start this section by briefly introducing the Newton-Raphson method and then follow with a procedure to compute the gradient vector and Hessian matrix. Let $\nabla(\theta)$ and $H(\theta)$ denote the gradient vector and Hessian matrix of the log-likelihood for the RHawkes process evaluated at the parameter vector θ, respectively. The Newton-Raphson method is an iterative procedure where at each iteration t, the current parameter estimate $\theta^{[t]}$ is updated using the following computation,

$$\theta^{[t+1]} = \theta^{[t]} - H(\theta^{[t]})^{-1}\nabla(\theta^{[t]}), \tag{7}$$

and then we set $t \leftarrow t + 1$ and repeat until some convergence criterion is satisfied. The optimization procedure is initialized at $\theta^{[0]}$. It would be beneficial to obtain an initial estimate using the approximate likelihood approach (see Section 3.2), which provides an initial starting point close to the true MLE. This approach can potentially reduce the computational time required for the implementation of the Newton-Raphson algorithm significantly, as generally only a small number of iterations will be needed until convergence when the initial parameter is close to the MLE.

Implementing the Newton Raphson algorithm above necessitates evaluation of the gradient vector and Hessian matrix for the RHawkes process. Recall that the log-likelihood for the RHawkes process is calculated by means of a recursion because the last immigrant probabilities are computed recursively. As such, the gradient vector and Hessian matrix also require a recursive procedure for their calculation. In the sequel, we outline the recursive algorithm to be used when computing the gradient vector and the Hessian matrix. As mentioned earlier, an additional advantage of this method is that the Hessian matrix needed for variance estimation is automatically obtained as a by-product at the end of the iterations.

Further to the notation introduced in Section 2, we introduce some additional notation to provide convenient expressions for the gradient vector and the Hessian matrix. Let $\partial_\theta = \partial/\partial\theta$ and $\partial^2_{\theta\theta^\top} = \partial^2/\partial\theta\partial\theta^\top$ denote the first and second order partial derivatives with respect to the model parameters. By taking the partial derivatives of the log-likelihood with respect to the parameter vector θ, the gradient vector $\nabla(\theta) = \partial_\theta \ell(\theta)$ takes the form,

$$\begin{aligned}\nabla(\theta) = \frac{\partial_\theta \mu(\tau_1)}{\mu(\tau_1)} - \partial_\theta U(\tau_1) + \sum_{i=2}^{n} \left\{ \frac{\sum_{j=1}^{i-1} d_{ij}\partial_\theta p_{ij} + p_{ij}\partial_\theta d_{ij}}{\sum_{j=1}^{i-1} p_{ij} d_{ij}} \right\} \\ + \frac{\sum_{j=1}^{n} S_{n+1,j}\partial_\theta p_{n+1,j} + p_{n+1,j}\partial_\theta S_{n+1,j}}{\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j}},\end{aligned} \tag{8}$$

where

$$\begin{aligned}\partial_\theta d_{ij} &= \Psi_{ij}\Big[\partial_\theta \mu(\tau_i - \tau_j) + \partial_\theta \phi(\tau_i) + (\mu(\tau_i - \tau_j) + \phi(\tau_i))\,\partial_\theta \psi_{ij}\Big], \\ \partial_\theta S_{n+1,j} &= \Psi_{n+1,j}\partial_\theta \psi_{n+1,j}, \\ \Psi_{ij} &= e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\} - \{\Phi(\tau_i) - \Phi(\tau_{i-1})\}}, \\ \psi_{ij} &= -\left\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\right\} - \left\{\Phi(\tau_i) - \Phi(\tau_{i-1})\right\},\end{aligned}$$

and henceforth we use the convention $\tau_{n+1} := T$. We have provided explicit equations to compute most of the terms in (8). However, we still lack a procedure to compute the derivative of the last immigrant probabilities $\partial_\theta p_{ij}$ with respect to each parameter of the model. A recursive procedure to compute the derivatives of p_{ij} is derived in Appendix A.

Next, we calculate the Hessian matrix $H(\theta) = \partial^2_{\theta\theta^\top}\ell(\theta)$ by computing the partial derivatives of the (transposed) gradient vector $\nabla(\theta)^\top$ in (8) with respect to θ, which gives the following,

$$\begin{aligned}H(\theta) = \frac{\partial^2_{\theta\theta^\top}\mu(\tau_1)}{\mu(\tau_1)} - \frac{\{\partial_\theta \mu(\tau_1)\}^{\otimes 2}}{\mu(\tau_1)^2} - \partial^2_{\theta\theta^\top} U(\tau_1) \\ + \sum_{i=2}^{n} \left\{ \frac{\sum_{j=1}^{i-1} d_{ij}\partial^2_{\theta\theta^\top} p_{ij} + 2\partial_\theta d_{ij} \odot \partial_\theta p_{ij} + p_{ij}\partial^2_{\theta\theta^\top} d_{ij}}{\sum_{j=1}^{i-1} p_{ij} d_{ij}} \right. \\ \left. + \frac{\left(\sum_{j=1}^{i-1} d_{ij}\partial_\theta p_{ij} + p_{ij}\partial_\theta d_{ij}\right)^{\otimes 2}}{\left(\sum_{j=1}^{i-1} p_{ij} d_{ij}\right)^2} \right\} \\ + \left\{ \frac{\sum_{j=1}^{n} S_{n+1,j}\partial^2_{\theta\theta^\top} p_{n+1,j} + 2\partial_\theta S_{n+1,j} \odot \partial_\theta p_{n+1,j} + p_{n+1,j}\partial^2_{\theta\theta^\top} S_{n+1,j}}{\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j}} \right. \\ \left. + \frac{\left(\sum_{j=1}^{i-1} S_{n+1,j}\partial_\theta p_{n+1,j} + p_{n+1,j}\partial_\theta S_{n+1,j}\right)^{\otimes 2}}{\left(\sum_{j=1}^{i-1} p_{n+1,j} S_{n+1,j}\right)^2} \right\},\end{aligned} \tag{9}$$

where and henceforth $x^{\otimes 2} := xx^\top$ denotes the outer product of a vector x with itself, and $x \odot y := \frac{1}{2}(xy^\top + yx^\top)$ denotes the symmetrized outer product of two vectors x and y of the same dimension, the first derivatives are as in (8), and

$$\begin{aligned}\partial^2_{\theta\theta^\top} d_{ij} = \Psi_{ij} \Bigg[& \{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \left\{ (\partial_\theta \psi_{ij})^{\otimes 2} + \partial^2_{\theta\theta^\top} \psi_{ij} \right\} \\ & + 2\partial_\theta \psi_{ij} \odot \partial_\theta \{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \\ & + \partial^2_{\theta\theta^\top} \{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \Bigg], \\ \partial^2_{\theta\theta^\top} S_{n+1,j} = \Psi_{n+1,j} \Bigg[& (\partial_\theta \psi_{n+1,j})^{\otimes 2} + \partial^2_{\theta\theta^\top} \psi_{n+1,j} \Bigg].\end{aligned}$$

Similar to the gradient vector computation above, a recursion to compute the second derivatives of the last immigrant probabilities $\partial^2_{\theta\theta^\top} p_{ij}$ are required. These can be found using the recursive procedure as presented in Appendix B.

Remark 1 *The recursive algorithms to compute the gradient vector and Hessian matrix both have quadratic computational time. The storage requirements are linear, in that a model with m parameters requires m vectors of dimension n to store the first derivatives and a total of $m(m+1)/2$ vectors of dimension n to store the second-derivatives. There is no need to store m^2 vectors due to the symmetry of the second order partial derivatives.*

Therefore, by using the expressions in (8) and (9) above, the implementation of the Newton-Raphson method is reasonably straightforward. Using the recursion defined in (7), the parameter estimates $\theta^{[t]}$ converge to the MLE $\hat{\theta}$ quite rapidly, particularly when the initial starting point $\theta^{[0]}$ is not far from $\hat{\theta}$.

3.2 Estimation using approximate likelihood functions

In Section 3.1, the estimation procedure was enhanced by increasing the computational efficiency of the optimization procedure using the Newton-Raphson method. However, there are situations where a good start value for the Newton-Raphson iterations is hard to come by, and the evaluation of the gradient and Hessian of the log-likelihood still demands quadratic time. Therefore, we consider an alternative approach of speeding up the fitting of the RHawkes process by maximizing an approximate rather than the exact log-likelihood of the RHawkes process. We use two methods to approximate the log-likelihood. In both methods, we truncate the distribution of the last immigrant and the range of the excitation effect. We restrict the potential candidates for the last immigrant at an event time to a relatively small number of past events, instead of considering all the past events, and we also restrict the excitation effect of an event to a small number of events in the future rather than let it last forever. The two methods differ in that the first method uses fixed

ranges in the truncations at different event times, while the second uses varying ranges of truncation at different event times. The first approach has been used by Stindl and Chen (2018) in the maximum likelihood estimation of the multivariate RHawkes process model, and the second is somewhat similar to the approach used by Halpin (2013) to approximate the complete data log-likelihood of the classical Hawkes process. It might also be worth mentioning that the first method is similar to the use of a fixed number of particles in the sequential Monte Carlo approximation to the likelihood of hidden Markov models.

We begin by discussing the truncation to the distribution of the last immigrant. Specifically, at each event time τ_i, we only consider, at most, the most B_i recent events as the possible candidates for the last immigrant. In other words, we effectively assume that the events which occur before the B_i-th most recent event have negligible probabilities of being the last immigrant event. Two choices of B_i are considered in this article. The first assumes that $B_i \equiv B$ for a large integer B, and the second allows the B_i to depend on the event time τ_i. In the case of dynamically chosen B_i's, at each step of the likelihood evaluation algorithm, the B_i is the value such that the sum of the last immigrant probabilities sums up to q or larger, where $q = 1 - \epsilon$ for a small $\epsilon > 0$, is the threshold probability. If we let $c_{i,j}$ denote the cumulative last immigrant probability of the most recent j events at time τ_i, so that $c_{i,j} = \sum_{k=1}^{j} p_{i,i-k}$, then the dynamically chosen B_i is given by $B_i = \min\{j \geq 1 : c_{ij} \geq q\}$. Next the probabilities p_{ij}, $j = i - B_i, \ldots, i-1$, are renormalized to sum to 1, and the p_{ij} for $j = 1, \ldots, i - B_i - 1$ are all set to 0. By a slight misuse of notation, these slightly modified probabilities are still denoted by p_{ij}. The inner summation terms in the log-likelihood function (3) are then approximated as follows at each iteration,

$$\sum_{j=1}^{i-1} p_{ij} d_{ij} \approx \sum_{j: i-B_i \leq j \leq i-1} p_{ij} d_{ij}, \tag{10}$$

$$\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j} \approx \sum_{j: n-B_n \leq j \leq n} p_{n+1,j} S_{n+1,j}, \tag{11}$$

with the last immigrant probabilities $p_{i+1,j}$ at the beginning of the next iteration still calculated using the recursion in (6), before they are subsequently truncated.

At this point, it might appear that we can already approximate the RHawkes process log-likelihood in linear time even without the truncation to the excitation effect, at least when $B_i \equiv B$, since at each iteration we only need at most B computations to calculate the last immigrant probabilities p_{ij}, $j = i - B, \ldots, i - 1$, at most B computations to calculate the conditional densities d_{ij} or the conditional survival probabilities $S_{n+1,j}$, $j = i - B, \ldots, i - 1$, given in (4) and (5), and a final summation of at most B terms in (10) or (11). However, this is not true in general, because the computation of p_{ij}, d_{ij} and S_{ij} involves $\phi(\tau_i) = \sum_{j=1}^{i-1} g(\tau_i - \tau_j)$ and $\Phi(\tau_i) = \sum_{j=1}^{i-1} G(\tau_i - \tau_j)$, both of which

require linearly growing time to compute in general, due to their dependence on all past event times τ_j, $j = 1, \ldots, i-1$.

For a genuine linear-time approximation algorithm, we shall use another approximation to make sure that the computation time required for the d_{ij}'s and S_{ij}'s for each i stays bounded. To this end, we first observe that their dependence on $\Phi(\tau_i) - \Phi(\tau_{i-1})$ can be avoided. Specifically, we note from the definition of d_{ij} and S_{ij} in (4) and (5) and the expression of the log-likelihood (3) that the likelihood can be expressed in this alternative form,

$$\ell(\theta) = \log \mu(\tau_1) - U(\tau_1) + \sum_{i=2}^{n} \log \left(\sum_{j=1}^{i-1} p_{ij} \tilde{d}_{ij} \right) + \log \left(\sum_{j=1}^{n} p_{n+1,j} \tilde{S}_{n+1,j} \right) + \Phi(T), \tag{12}$$

where $\tilde{d}_{ij}$ and $\tilde{S}_{ij}$ are free from $\Phi(\tau_i)$ or $\Phi(\tau_{i-1})$ and are as follows,

$$\begin{aligned} \tilde{d}_{ij} &= (\mu(\tau_i - \tau_j) + \phi(\tau_i)) e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\}}, \\ \tilde{S}_{ij} &= e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\}}, \end{aligned}$$

where we recall $\tau_{n+1} = T$ by convention. Moreover, the recursion on the last immigrant probabilities p_{ij} in (6) can also be expressed in terms of $\tilde{d}_{ij}$'s and $\tilde{S}_{ij}$'s as follows,

$$\begin{aligned} p_{ij} &= \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1} - \tau_j) + \phi(\tau_{i-1})} \frac{p_{i-1,j} \tilde{d}_{i-1,j}}{\sum_{k=1}^{i-2} p_{i-1,k} \tilde{d}_{i-1,k}}, & j = 1, \ldots, i-2, \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1, \end{cases} \\ &= \begin{cases} \frac{p_{i-1,j} \phi(\tau_{i-1}) \tilde{S}_{i-1,j}}{\sum_{k=1}^{i-2} p_{i-1,k} \tilde{d}_{i-1,k}}, & j = 1, \ldots, i-2, \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1. \end{cases} \end{aligned} \tag{13}$$

Note also that, with the truncation on the last immigrant distribution applied, the p_{ij}'s at the start of the ith iteration, before the truncation and renormalization at the current iteration happens, take the form,

$$p_{ij} = \begin{cases} \frac{p_{i-1,j} \phi(\tau_{i-1}) \tilde{S}_{i-1,j}}{\sum_{k=i-1-B_{i-1}}^{i-2} p_{i-1,k} \tilde{d}_{i-1,k}}, & j = i-1-B_{i-1}, \ldots, i-2 \\ 1 - \sum_{k=i-1-B_{i-1}}^{i-2} p_{ik}, & j = i-1. \end{cases}$$

The computation of $\Phi(T)$ in (12) still takes linear time. In considering the approximate likelihood inference for the classical Hawkes process, some authors (Lewis and Mohler, 2011; Veen and Schoenberg, 2008) have suggested approximating $\Phi(T)$ by $nG(\infty)$, which is a close approximation when the parameters of the excitation kernel are such that the kernel decays rapidly. However, this approximation might be too crude for some parameter values and cause efficiency loss on the parameters of the excitation kernel. Therefore, we will calculate its exact value without approximation, which does not cause concern here, since the computation needs only to happen once, and we are after a linear-time algorithm after all.

To speed up the evaluation of $\phi(\tau_i)$, we assume that the excitation effect due to events in the distant past is negligible. This is justifiable since the integrability condition on the kernel g implies $g(\tau_i - \tau_j) \approx 0$ when τ_i and τ_j are far apart. Specifically, we use the approximation

$$\phi(\tau_i) \approx \sum_{j=i-F_i}^{i-1} g(\tau_i - \tau_j),$$

for large values of F_i. Again we consider two choices of F_i. The first assumes $F_i = \min(F, i-1)$ for a fixed (large) positive integer F. With the second approach we choose F_i to be the smallest integer $j \leq i-1$ such that $G(\tau_i - \tau_{i-j}) \geq (1-\delta)G(\infty)$, for a small $\delta > 0$. That is,

$$\begin{aligned} F_i &= \min\left\{j \leq i-1; G(\tau_i - \tau_{i-j}) \geq (1-\delta)G(\infty)\right\} \\ &= \min\left\{j \leq i-1; \tau_i - \tau_{i-j} \geq \tilde{G}^{-1}(1-\delta)\right\}, \end{aligned}$$

where $\tilde{G}^{-1}(1-\delta)$ denotes the $(1-\delta)$-quantile of the normalised excitation kernel $\tilde{g}(\cdot) = g(\cdot)/G(\infty)$, and we define $F_i := i-1$, when the set is empty.

Remark 2 *Our dynamic approximation to $\phi(\tau_i)$ above is similar to, but different than, that used by Halpin (2013), which approximates $\phi(\tau_i)$ by discarding the terms that are smaller than or equal to a small value δ from the summation:*

$$\phi(\tau_i) \approx \sum_{j \in W_i} g(\tau_i - \tau_j), \text{ with } W_i = \left\{j \leq i-1;\, g(\tau_i - \tau_j) > \delta\right\}.$$

Compared to Halpin's approach, our approach seems easier to implement, as we can determine F_i from $\tilde{G}^{-1}(1-\delta)$ easily using a binary search, while Halpin's approach requires the determination of the set W_i, which is not straightforward when the excitation kernel is not monotonically decreasing.

Remark 3 *When $B_i \equiv B$ and $F_i \equiv F$, it is clear that our likelihood approximation takes linear time to compute. On the other hand, when the B_i's and F_i's are dynamically determined using the afore-discussed approach, it is not instantly clear that the likelihood approximation algorithm is still a linear time one. However, there is strong numerical evidence that the sequences $B_1, \ldots, B_n$ and $F_1, \ldots, F_n$ are both asymptotically stationary, suggesting that the algorithm is a linear time one in a stochastic sense.*

Remark 4 *When the excitation kernel is an exponential function, or a finite linear combination of exponential functions, then the approximation to the $\phi(\tau_i)$'s are not necessary, since they can be evaluated exactly in linear time. Specifically, if $g(t) = ae^{-bt}$ for some a, b, then the $\phi(\tau_i)$'s can be calculated using this recursion, starting with $\phi(\tau_i) = 0$:*

$$\phi(\tau_i) = \left\{\phi(\tau_{i-1}) + a\right\} e^{-b(\tau_i - \tau_{i-1})}, \quad i = 2, \ldots, n.$$

Now if $g(t) = \sum_{k=1}^{K} a_k e^{-b_k t}$ for some K and a_k, b_k, $k = 1, \ldots, K$, then from $\phi(\tau_i) = \sum_{k=1}^{K} \phi_k(\tau_i)$ with $\phi_k(\tau_i) := \sum_{j=1}^{i-1} a_k e^{-b_k(\tau_i - \tau_j)}$, and the fact that each of the K sequences $\left\{\phi_k(\tau_i), i = 1, \ldots, n\right\}$, $k = 1, \ldots, K$, can be evaluated in linear time, we see that $\phi(\tau_i)$, $i = 1, \ldots, n$ can also be evaluated in linear time.

4 Simulations

In this section we report the results of the simulation experiments to assess the performance of the Nelder-Mead simplex, the Newton-Raphson and approximate likelihood methods for parameter estimation, the accuracy and speed of the two likelihood approximation methods, and the influence of the tuning parameters on the two likelihood approximation methods.

4.1 Estimation results

We investigate the numerical performance of the Newton-Raphson and Nelder-Meld optimization routines and the exact likelihood compared to the approximate likelihood estimation procedures by analysing the estimation results for the following four methods:

- Exact likelihood optimization using the Nelder-Mead simplex method (NM);
- Exact likelihood optimization using the Newton-Raphson algorithm (NR);
- Approximate likelihood optimization with fixed B and F values using the Nelder-Mead simplex method (AL1), where $B = F$ are both set to 10, or the number of events divided by 20 and then rounded up, whichever is smaller;
- Approximate likelihood optimization consisting of a dynamically chosen B_i and F_i with $\epsilon = 10^{-6}$ and $\delta = 10^{-6}$, using the Nelder-Mead simplex method (AL2).

The simulations consists of $N = 1000$ realizations from the RHawkes process with Weibull distributed inter-renewal waiting times with hazard function given by, $\mu(t) = \kappa t^{\kappa-1}/\beta^{\kappa}$, with shape parameter $\kappa = 3$ or $1/3$ and scale $\beta = 1.2,$ or 0.2 which implies that the mean waiting time between successive immigrants is close to one (1.07 and 1.2 respectively). The excitation function takes an exponential form $g(t) = \eta \exp(-t/\gamma)/\gamma$ with mean waiting time between offspring events $\gamma = 1$ and branching ratio $\eta = 0.5$. This means that approximately half of the events are immigrants and the other half are offspring events. The censoring time T is 550 and 700 to ensure that the mean number of events is close to 1000 in both simulation models.

For both the Newton-Raphson and Nelder-Mead optimization routines the procedures had the same initialization at the true parameter values. The optimization routines were considered converged when they were unable to decrease the value of the log-likelihood by $10^{-8}(|\ell| + 10^{-8})$ at an iteration, that is, `reltol` was set at `1e-8` when the `optim` function in R (R Core Team, 2018) was called. Table 1 contains the bias, empirical standard error (SE), average standard error estimate ($\hat{\text{SE}}$) of the estimated parameters for each of the estimation methods described above. The average running time of the R code [on Intel Xeon Gold 6130 (22M Cache, 2.1GHz) Skylake processors] to perform the optimization procedure and compute the approximate Hessian matrix (exact Hessian matrix in the case of NR) and the total number of iterations until convergence are also presented.

The four estimation methods NM, NR, AL1, and AL2 all produce very comparable estimation results. The biases for each parameter using the two exact likelihood with the NM and NR methods are very close to zero, suggesting that these estimates are approximately unbiased. The average standard errors and empirical standard error estimates are close for all methods, and in particular, the results from NM and AL1 and AL2 are nearly identical, suggesting the log-likelihood approximations are highly accurate. An important observation that is beneficial and of practical importance is the similarly small biases for the approximate likelihood methods AL1 and AL2.

In terms of the computation time required, the NR method is about 30% faster than the NM method, despite the number of iterations required to achieve convergence being substantially (over 30 times) smaller than that of the NM method (with exact or approximate log-likelihood). The approximate likelihood methods AL1 and AL2 are the fastest, both about five times as fast as the NM method, and about three times as fast as the NR method. On larger data sets, the speed gains by using the two approximate likelihood methods should be more impressive due to their linear time complexity in contrast to the quadratic time complexity of the NM and the NR methods.

4.2 Accuracy and speed of the log-likelihood approximation methods

We next examine the accuracy and speed of the two log-likelihood approximation methods. To this end, we generated 100 realizations of the sample path of the second simulation model in the previous section, where $\kappa = 1/3$, $\beta = 0.2$, $\gamma = 1$, and $\eta = 0.5$, up to the censoring time of 8000. We then evaluated the exact and the two approximate log-likelihoods of four parameter vectors relative to each of the 100 sample paths up to four censoring times, to see the influence of the parameter and the amount of data on the accuracy and speed of the approximation. The four parameter vectors are $\theta_1 = (0.1, 0.1, 0.1, 0.1)$, $\theta_2 = (0.3, 0.2, 0.5, 0.5)$, $\theta_3 = (1, 1, 1, 0.8)$, and $\theta_4 = (3, 2, 10, 0.9)$, which are chosen from different regions of the parameter space. The four censoring times are $T_1 = 1000$, $T_2 = 2000$, $T_3 = 4000$, and $T_4 = 8000$. For the tuning parameters, we used $B = F = 100$ in AL1, and in AL2 the B_i's and F_i's were dynamically determined using the threshold probabilities $1 - \epsilon = 1 - \delta = 1 - 0.001$. The average running time in seconds and the median absolute relative error (MARE) of the log-likelihood approximation are shown in Table 2 together with the average number of events, and the mean of the average value of B_i and F_i in AL2. The average running time and MARE of the two log-likelihood approximation method against the average number of events are graphed in Figure 1.

From Table 2 and Figure 1 observe that the running times of the two log-likelihood approximation methods increase roughly linearly with the number of events, which is to be expected since in AL1 the B and F values are preset, and the average B_i and F_i values are also very stable over time, as shown in the table. From the table, also notice that the time required to evaluate the exact

		$\kappa = 3$	$\beta = 1.2$	$\gamma = 1$	$\eta = 0.5$	Iter	Time (s)
	Bias	0.0180	-0.0034	0.0147	-0.0033		
NM	SE	0.2622	0.0505	0.2132	0.0302	169.96	95.4
	$\widehat{\text{SE}}$	0.2538	0.0489	0.2021	0.0304		
	Bias	0.0242	-0.0022	0.0256	-0.0025		
NR	SE	0.2590	0.0496	0.2056	0.0296	4.26	69.2
	$\widehat{\text{SE}}$	0.2545	0.0489	0.2041	0.0303		
	Bias	0.0179	-0.0034	0.0145	-0.0033		
AL1	SE	0.2622	0.0505	0.2128	0.0302	170.3	23.1
	$\widehat{\text{SE}}$	0.2537	0.0489	0.2019	0.0304		
	Bias	0.0180	-0.0034	0.0147	-0.0033		
AL2	SE	0.2622	0.0505	0.2132	0.0302	169.62	19.3
	$\widehat{\text{SE}}$	0.2526	0.0488	0.2011	0.0303		
		$\kappa = 1/3$	$\beta = 0.2$	$\gamma = 1$	$\eta = 0.5$	Iter	Time (s)
	Bias	-0.0073	0.0155	0.0088	0.0043		
NM	SE	0.0178	0.0418	0.1171	0.0391	143.36	85.9
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1115	0.0353		
	Bias	-0.0073	0.0155	0.0092	0.0042		
NR	SE	0.0178	0.0419	0.1167	0.0391	3.89	68.0
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1116	0.0353		
	Bias	-0.0072	0.0170	-0.0123	0.0045		
AL1	SE	0.0177	0.0440	0.1138	0.0391	144.60	19.8
	$\widehat{\text{SE}}$	0.0142	0.0355	0.1054	0.0353		
	Bias	-0.0073	0.0155	0.0088	0.0043		
AL2	SE	0.0178	0.0418	0.1171	0.0391	143.39	19.8
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1114	0.0353		

Table 1 Estimation results for $N = 1000$ realizations from two simulation models using the Nelder-Mead (NM) and Newton-Rapshon (NR) methods based on the exact likelihood and the Nelder-Mead method based on the two likelihood approximation likelihoods (AL1, AL2) methods. The NR method requires about four iterations to converge, while the Nelder-Mead method with exact or approximate log-likelihoods needs substantially more (about 30 times as many) iterations to converge. The estimates using different methods are mostly identical. In terms of speed, AL1 and AL2 are roughly five times as fast as the NM method, while the Newton-Raphson is about about 30% faster than the NM method.

log-likelihood increases roughly quadratically with the amount of data. The speed of the method AL1 does not seem to depend on the parameter vector, while the speed of AL2 seems to depend on the parameter vector because in AL2 the B_i and F_i values need to adapt to the parameter vectors under consideration to make sure the preset requirements on the truncations to the last immigrant distribution and the excitation kernel are met. For example, for parameter vector θ_4, where the mean of the offspring birth time distribution γ is 10, the dynamically selected F_i values are around 130 on average to make sure the truncation to the excitation kernel accounts for 99.9% ($= 1 - 0.001$) of the excitation effect.

Both approximation methods are reasonably accurate in that the median absolute errors in all cases are well below 1% of the exact log-likelihood value.

			T	1000	2000	4000	8000
			$\mathbb{E}[N(T)]$	1676.1	3309.0	6478.0	11385.1
θ_1 : $\kappa = 0.1$, $\beta = 0.1$, $\gamma = 0.1$, $\eta = 0.1$		Time (s.)	Exact	1.1	3.86	14.76	45.22
			AL1	0.22	0.46	0.91	1.38
			AL2	0.3	0.35	0.87	1.73
		MARE	AL1	0.000e+00	0.000e+00	0.000e+00	0.000e+00
			AL2	1.557e-04	1.577e-04	1.528e-04	1.484e-04
		$E[B]$ in AL2		5.12	5.10	5.11	5.08
		$E[F]$ in AL2		4.99	4.95	4.96	4.93
θ_2 : $\kappa = 0.3$, $\beta = 0.2$, $\gamma = 0.5$, $\eta = 0.5$		Time (s.)	Exact	1.1	3.99	15.33	46.1
			AL1	0.22	0.45	0.9	1.36
			AL2	0.17	0.39	0.94	1.84
		MARE	AL1	1.73e-14	4.13e-13	3.98e-12	9.13e-12
			AL2	3.85e-04	3.95e-04	3.93e-04	4.04e-04
		$E[B]$ in AL2		11.50	11.46	11.55	11.49
		$E[F]$ in AL2		14.74	14.61	14.66	14.55
θ_3 : $\kappa = 1$, $\beta = 1$, $\gamma = 1$, $\eta = 0.8$		Time (s.)	Exact	1.05	3.9	14.82	42.31
			AL1	0.22	0.45	0.88	1.33
			AL2	0.18	0.41	1	1.91
		MARE	AL1	7.65e-08	4.56e-07	1.45e-06	7.13e-07
			AL2	3.99e-03	5.17e-03	6.33e-03	4.52e-03
		$E[B]$ in AL2		18.39	18.23	18.30	18.18
		$E[F]$ in AL2		23.20	22.96	23.10	22.95
θ_4 : $\kappa = 3$, $\beta = 2$, $\gamma = 10$, $\eta = 0.9$		Time (s.)	Exact	1.08	4.02	15.11	41.54
			AL1	0.23	0.47	0.92	1.38
			AL2	0.18	0.42	1.02	1.94
		MARE	AL1	2.43e-04	1.88e-04	1.67e-04	5.89e-05
			AL2	2.26e-05	1.60e-05	1.37e-05	4.85e-06
		$E[B]$ in AL2		14.33	14.21	14.26	14.17
		$E[F]$ in AL2		127.22	128.74	131.00	130.40

Table 2 The average running time (Time) and the median absolute relative error (MARE) of the two log-likelihood approximation methods AL1 and AL2 for four different parameter vectors at four increasingly large censoring times. The average time required to calculate the exact log-likelihood is also reported. In AL1, $B = F = 100$; In AL2, the B_i's and the F_i's are dynamically determined from cut-off probabilities $1 - \epsilon$ and $1 - \delta$ respectively, with $\epsilon = \delta = 0.001$.

The accuracy of both approximation methods seems to be affected by the parameter vector under consideration. However, it seems more so for AL1 than for AL2, as in AL1, the fixed B and F values can not adapt to the parameter vector under consideration and therefore, might be too small for specific parameter values to produce accurate approximations. Given these observations, it seems reasonable to use AL2 in order to achieve a more consistent relative approximation error in practical applications.

4.3 Influence of the tuning parameters on the speed and accuracy of log-likelihood approximation

Finally we look at the influence of the tuning parameters on the speed and accuracy of the likelihood approximations. With large B_i and F_i values, the likelihood approximations will be more accurate, but the running times will be longer, and similarly, smaller B_i and F_i values mean faster but less accurate

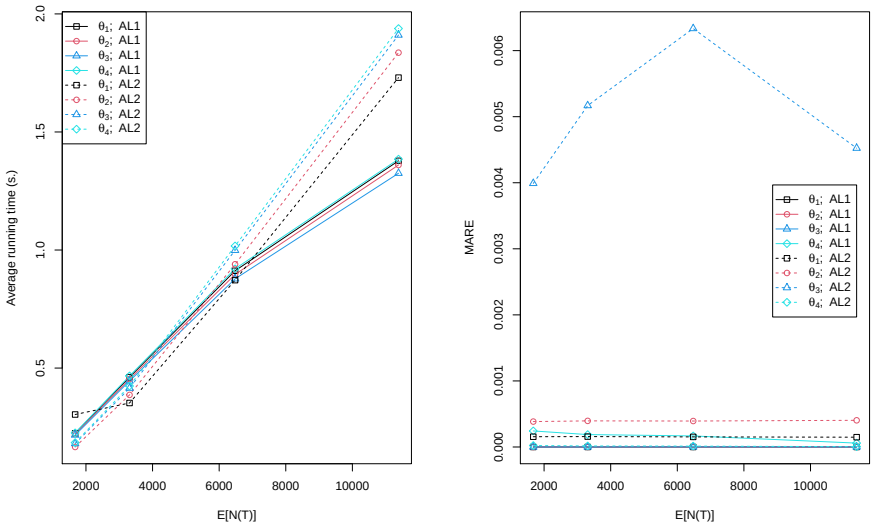

Fig. 1 The average running times and median absolute relative errors (MAREs) of the two log-likelihood likelihood methods (AL1 and AL2) on different parameters at different censoring times. Left: running times; right: MAREs. Solid lines for AL1; dashed lines for AL2. Different points indicate different parameter vectors.

approximations. In this section, we explore to what extent the choice of the tuning parameters influences the speed and accuracy of the log-likelihood. To this end, we applied the log-likelihood approximation methods AL1 and AL2 with varying tuning parameters to find the log-likelihoods of the parameter vectors θ_1, θ_2, θ_3 and θ_4 relative to the 100 simulated sample paths up to censoring time $T = 8000$ of the RHawkes model. The tuning parameter values used in AL1 are $B = F = 10, 20, \ldots, 160$; and the tuning parameter values used in AL2 are $\epsilon = \delta = 10^{-1}, 10^{-2}, \ldots, 10^{-16}$. The average running times of our R code on the 100 data sets, the MARE of the log-likelihood approximations, for each of the four parameter vectors in Table 2 with different values of the tuning parameters are shown in Figure 2.

The average of the mean values of the $\mathbb{E}[B]$ and $\mathbb{E}[F]$ in AL2 on the 100 data sets are also shown. From Figure 2, notice that the average running time of AL1 increases roughly linearly with the value of the tuning parameter B and F. The running times of AL2 also tend to increase as ϵ and δ shrinks, or equivalently the approximations to the last immigrant distribution and the excitation kernel function become more and more accurate. It is interesting to note that although AL2 seems slower than AL1 for the tuning parameters considered, the running time of AL2 increases much slower than AL1 and plateaus when ϵ and δ are smaller enough, despite the average back-looking lag $\mathbb{E}[B]$ for the last immigrant and the average of forward-looking lag $\mathbb{E}[F]$ for the range of the excitation effect both increasing linearly (cf. lower panels of Figure 2) as

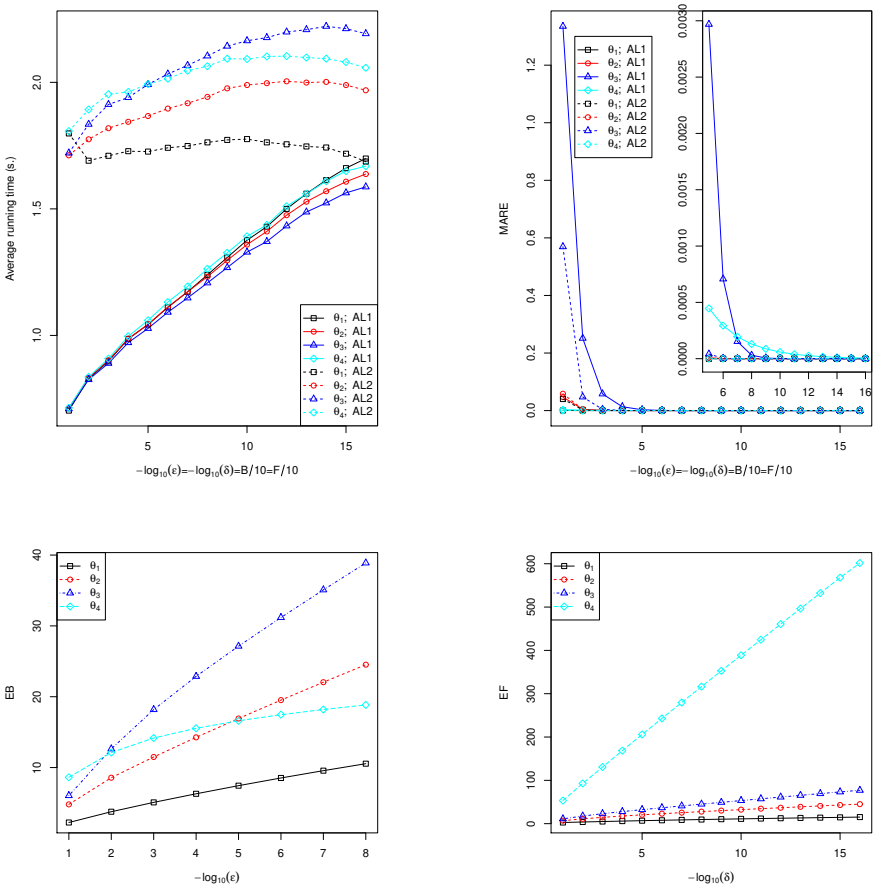

Fig. 2 The average running time (top left), and the median absolute relative error (MARE, top right) of AL1 and AL2 at four different parameter vectors, and the average of the means of B_i in AL2 ($\mathbb{E}[B]$, bottom left), and the average of the means of the F_i in AL2 ($\mathbb{E}[F]$, bottom right) against the (transformed) tuning parameter. In the top right panel, the smaller inset graph is a zoom-in of the right part of the larger graph.

their deterministic counterparts in AL1. This might be because, in AL2, the computational overhead to determine the values of B_i and F_i dominate the computational cost in each iteration when the values of B_i and F_i are small relative to the total sample size. We also note that AL2 seems to be more accurate than AL1, with the relative approximation error of AL2 on all the four parameters considered being practically zero when ϵ and δ are 10^{-6} or smaller. This is also the reason why in our simulation experiments to compute the MLE of the parameters (cf. Table 1), the average number of iterations of the Nelder-Mead optimization routine applied with the exact log-likelihood and is nearly identical to that with the AL2 approximate log-likelihood. The implication for

fitting RHawkes processes (therefore including classical Hawkes processes) in practice is that we can simply apply any derivative-free optimization routines on the approximate log-likelihood calculated using AL2 with very small values of ϵ and δ, e.g., 10^{-6} or smaller, to achieve significant gains in computational speed without having to worry about loss in statistical efficiency. Another practically useful strategy of fitting the RHawkes process on big datasets is to obtain an initial estimate of the parameters quickly using AL1 with smaller values of B and F, and then use the initial estimate for the starting value in a subsequential optimization using the Newton-Raphson method, or using a derivative-free method with a more accurate log-likelihood approximation.

5 Mid-price changes of foreign currency exchange rates

This analysis intends to quantify the level of endogeneity (self-exciting effects) in the foreign exchange market (forex) using RHawkes processes. For this purpose, we study the mid-price changes of the following currency pairs: AUD/USD (Australian Dollar against US Dollar), USD/CAD (USD against Canadian Dollar), USD/CHF(USD against Swiss Franc), EUR/USD (euro against USD), GBP/USD (British Pound against USD), USD/HKD (USD against Hong Kong Dollar), USD/JPY (USD against Japanese Yen) and USD/SEK (USD against Swedish Krona). These currency pairs represent some of the most traded currency pairs in terms of value traded. This analysis examines the mid-price changes for these pairs, for the four trading weeks (excluding weekends) from 1st July 2019 until 26th July 2019, and we consider the hours between 12:00:00 GMT until 21:00:00 GMT, the official operating hours for the New York forex market. We compare the different currency pairs by analyzing the level of endogenous and exogenous trading activities and their impact on the mid-price changes.

The mid-price is defined as the mean of the best bid and ask price, and a mid-price change happens when the value of this mid-price changes. This change occurs when either the bid, ask, or a combination thereof, changes price. It is well observed that trading activity does not happen in a stationary manner during the trading day. Extremely evident is the drastically different features of trades exhibited during the opening and closing times of international forex markets. To see this, in Figure 3, we plot the expected duration between mid-price changes conditional on the time of day that the mid-price change transpired. The expectation is estimated using a cubic regression spline, where the knots were chosen every two hours throughout the day. This procedure has previously been employed in the work of Engle and Russell (1998). The figure shows a clear diurnal pattern with the New York forex hours having a shorter duration during the opening hours than the closing hours. It is apparent that when the majority of the forex markets are in operation, the expected waiting time between mid-price changes is much shorter than when fewer markets are in operation. The non-stationary arrival time regime of the mid-price changes

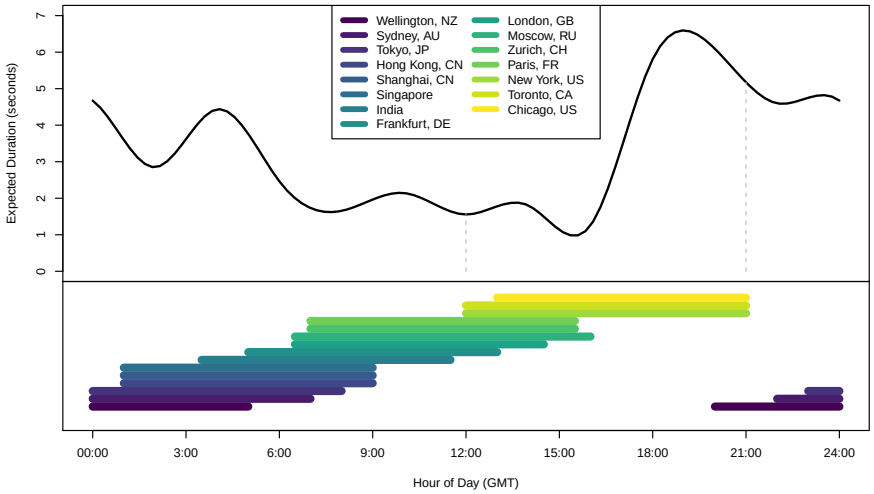

Fig. 3 A non-parametric estimate of the daily pattern of the duration between mid-price changes, and in the bottom panel, the operating hours of the major forex centers around the world.

over any particular trading day necessitates a transformation to remove the intra-day patterns and hence obtain stationarity.

To this end, we follow the work of Engle and Russell (1998) and discount the observed durations by a factor proportional to the corresponding expected duration subjected to the requirement that the sum of the modified durations in a trading day is equivalent to the original durations. The purpose is to supply less weight to mid-price changes occurring in the opening of the US forex markets, when high activity is to be expected, as the majority of the international forex markets are open; cf. Figure 3. Also note that, although the timestamps in the dataset are accurate down to microseconds (10^{-6} seconds), on several occasions, two or more price changes are recorded to have transpired at identical times. On these occasions, small random noises were added to these event times to break the ties while still preserving the time order of the price moves.

For each of the currency pairs, the mid-price changes for the twenty trading days (four trading weeks) are each fitted with a RHawkes processes with Weibull immigration. The motivation for the Weibull distribution follows from the work of Chen and Stindl (2018), in which the Weibull renewal distribution was found to provide a superior fit than the gamma and exponential distributions. However, the datasets described above are too big for the exact likelihood methods of Chen and Stindl (2018) to be practical. For instance, the currency pair GBP/USD has on average above 22,000 mid-price changes during the period of analysis, and an exact likelihood evaluation takes several minutes, and therefore application of the Nelder-Mead optimzation algorithm,

which typically takes thousands of function evaluations before convergence, would be intolerably slow. Therefore, the linear time algorithms introduced herein or, the Newton Raphson algorithm which requires much fewer calculations of the log-likelihood are required for practical fitting of the RHawkes process.

Therefore, to accomplish the model fitting tasks we use the AL2 approximate likelihood method with $\epsilon = \delta = 10^{-20}$. The optimization routine was initialized by first fitting a classical Hawkes process to the mid-price changes using its exact likelihood and then utilizing the fitted Hawkes model parameters as the starting values for the AL2 method for the RHawkes process. The parameter estimates for each of the currency pairs are plotted in Figure 4, whereby the time series of estimates evolve over the four trading weeks. Note that, for the scale parameter of the Weibull distribution β, the results are presented on the log-scale for better visualization. We also remark that this analysis only deals with a fixed window of time and does not take into account the mid-price changes before the start of the observation period. As such, edge effects may influence the parameter estimates, and the results presented in Figure 4, but the large sample size should guarantee that these effects are inconsequential.

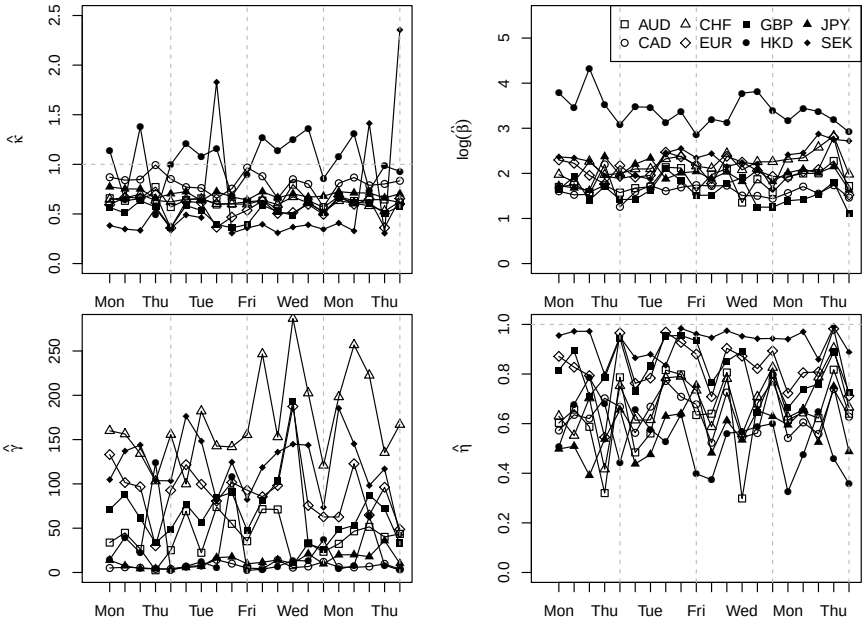

Fig. 4 Time series plot of the estimated model parameters of the RHawkes process as they evolve over the four trading weeks from 1st July 2019 until 26th July 2019 between the hours of 12:00:00 GMT until 21:00:00 GMT for all the pairs.

First, let us examine the exogenous features of the mid-price arrival regime before investigating the endogenous components of the process after that. From Figure 4, it can be observed that the arrival process of exogenously driven mid-price changes, has a Weibull shape parameter that remains below one for the majority of the currency pairs. This implies that exogenously driven mid-price changes occur in heavier bursts than would be suggested by a Poisson process. For instance, consider the pair GBP/USD, the estimated shape parameter $\hat{\kappa}$ has the following five-number summary for the four trading weeks: $(0.36, 0.50, 0.55, 0.59, 0.66)$ with mean 0.53 and standard deviation 0.09. This implies a heavy-tailed distribution for the arrival process of exogenous changes and therefore price changes due to external influences tend to occur in a more bursty fashion rather than uniformly through time. Such departure from Poisson immigration suggests that the RHawkes process is preferable to model this structure of price moves since the classical Hawkes process would be inadequate to accommodate this type of fit. This general departure from Poisson immigration suggests that exogenously driven mid-price changes occur in heavy bursts, or close in time, which might be appropriate due to the dependence of these prices moves on external factors such as the current state of the market (i.e., recession) and global trends or events. However, for the pair USD/HKD, this is not evident, and the arrival process of exogenous mid-price changes seem to exhibit complete randomness and does not deviate significantly from a Poisson process, as the estimated shape parameter $\hat{\kappa}$ is not very different from one.

It is incredibly evident that the scale parameter for the currency pair USD/HKD has a significantly larger value and deviates significantly from the other pairs. This is because the total number of mid-price changes for this pair is much smaller compared to the other pairs. For instance, the mean number of mid-price changes on a given day is $2,399$, and this is three times smaller than the next smallest, which is the currency pair USD/CHF with $8,299$ mid-price changes. For each trading week, there tends to be a trough that occurs on Fridays for the estimated parameters $\hat{\beta}$ by looking at Figure 3. Again, this can be attributed to the high level of trading that tends to occurs on Fridays in comparison to the rest of the trading week, and this phenomenon is present for the majority of the currency pairs, and therefore, the mean waiting time between exogenous mid-price changes are much smaller on these days.

Market participants generally take more time to react to mid-price movements than to external factors or exogenous mid-price changes when executing currency trades. For instance, in Figure 5, we plot the expected waiting time between exogenously driven mid-price changes in solid lines and endogenously driven mid-price changes in dashed lines. It can be observed that the waiting time between endogenous mid-price changes are much longer than exogenous ones, implying that market participants take more prolonged time to react to previous price movements than to external market news or event. Note that the mean waiting time between exogenous events is computed using $\hat{\beta}\Gamma(1+1/\hat{\kappa})$, and endogenous events using $\hat{\gamma}$. These two waiting times tend to move in the

same direction, although, on some occasions, the two times move in opposite directions.

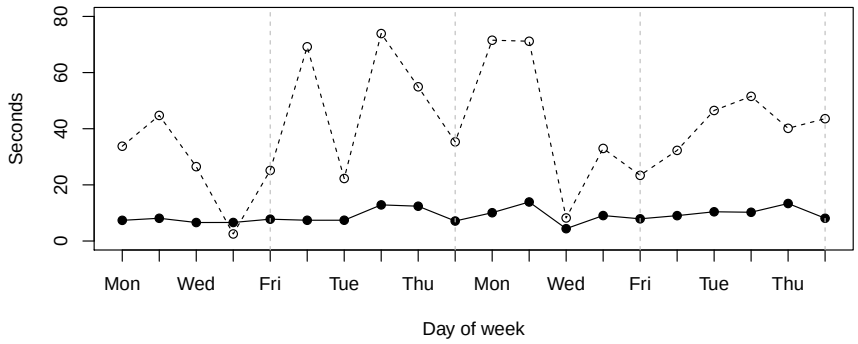

Fig. 5 Mean waiting time between exogenously driven mid-price changes in solid lines, and the mean waiting time between endogenously driven mid-price changes in dashed lines, for the currency pair AUD/USD.

This analysis concludes by analysing the level of endogeneity in the forex markets. As seen in Figure 4, the estimated branching ratio $\hat{\eta}$ shows fluctuating levels of self-excitation with estimated branching ratios ranging between 0.54 for the pair USD/HKD and 0.93 for the pair USD/SEK. This suggests that a significant portion of price movements are endogenously generated rather than related to the arrival of news in the market place. The overall weaker endogeneity in the price movements of the HKD relative to the USD compared to the other currency pairs could be because the price of the HKD is pegged to the USD, and therefore traders of this currency pair might be less sensitive to its price movements. Another critical aspect to acknowledge is the waiting time between endogenous mid-price changes. The USD/CHF exchange rate exhibits highly dispersed values of $\hat{\gamma}$ compared to the other currency pairs. For instance, the average of the mean waiting times for the pair USD/CAD is 7.14 seconds, and for the pair USD/CHF, it is significantly longer with 171.03 seconds.

6 Discussion

Two methods for accelerating estimation for RHawkes processes have been proposed in this article. The first procedure makes use of the Newton-Raphson algorithm which includes shape information about the log-likelihood surface, such as slope and curvature, to make more informed movements in the parameter space to converge to the MLE more rapidly than derivative-free opti-

mization methods. The second procedure computes two approximations to the likelihood by excluding negligible contributions to the log-likelihood value, either in the form of truncating the possible candidates to be the last immigrant event or the influence of distant past events on the intensity rate.

In the simulation results reported in Section 4.1, we have used the true parameter values as the initial values for the Newton-Raphson method and the Nelder-Mead method with approximate likelihoods. Although not reported in the paper, we have also tried to initialize the optimization routines with different starting values such as those obtained by fitting a classical Hawkes process. The Newton-Raphson method was found to be very sensitive to the starting values, while the approximate likelihood method is robust to the choice of starting values with the final estimates nearly identical to those obtained with the initial values set at the true parameter values. Therefore, our recommendation for practical applications is to use the parameter estimates obtained by fitting a simpler model nested in the RHawkes model, such as the classical Hawkes process model.

As noted by a referee, an obvious alternative procedure is to combine likelihood approximation and Newton-Raphson method of optimization. However, given the numerical stability issue faced by the Newton-Raphson method with the exact likelihood, we anticipate a similar stability issue with this procedure, and therefore do not pursure it in this work.

A First derivative of the last immigrant probabilities

In this appendix, we outline the derivation to compute the derivatives of the last immigrant probabilities with respect to the parameter θ using a recursive algorithm. For any fixed $i \in \{3, \ldots, n\}$ and for $j \leq i-2$, the last immigrant probabilities $p_{ij} = \mathbb{P}\left(I(\tau_i) = j | \tau_{1:i-1}\right)$ given in Chen and Stindl (2018) take the form,

$$p_{ij} = \frac{\phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}}{\sum_{k=1}^{i-2}\left(\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right)\Psi_{i-1,k}p_{i-1,k}}. \tag{14}$$

where the notations are as in Section 2. The derivative of p_{ij} is obtained by application of the quotient rule. The derivative of the p_{ij} numerator in (14) with respect to parameter vector θ of the model is given by,

$$\Psi_{i-1,j}\Big[p_{i-1,j}\partial_\theta\phi(\tau_{i-1}) + \phi(\tau_{i-1})\partial_\theta p_{i-1,j} + \phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j}\Big], \tag{15}$$

and the derivative of the denominator of p_{ij} is

$$\sum_{k=1}^{i-2}\Psi_k(\tau_{i-1})\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k)+\partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k} + \left\{\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k}+\partial_\theta p_{i-1,k}\right)\Big]. \tag{16}$$

We now apply the quotient rule with the aid of (15) and (16). Then the following recursion for the derivative of the last immigrant probabilities $\partial_\theta p_{ij}$ holds for $i \in \{3, \dots, n+1\}$,

$$\partial_\theta p_{ij} = \frac{A}{C} - \frac{B}{C^2}, \qquad j = 1, \dots, i-2, \tag{17}$$

$$\partial_\theta p_{i,i-1} = -\sum_{j=1}^{i-2} \partial_\theta p_{ij}, \tag{18}$$

where

$$A := \Psi_{i-1,j}\Big[p_{i-1,j}\partial_\theta\phi(\tau_{i-1}) + \phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j} + \phi(\tau_{i-1})\partial_\theta p_{i-1,j}\Big],$$

$$B := \phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}\sum_{k=1}^{i-2}\Psi_{i-1,k}\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k) + \partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k} + \left\{\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k} + \partial_\theta p_{i-1,k}\right)\Big],$$

$$C := \sum_{k=1}^{i-2}\left(\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right)\Psi_k(\tau_{i-1})p_{i-1,k}.$$

The recursion is initialized with the initial condition that $\partial_\theta p_{21} = 0$.

B Second derivative of the last immigrant probabilities

Now for the second derivative of the last immigrant probabilities, the terms in (17) is differentiated, again using the quotient rule, and we obtain the following,

$$\partial^2_{\theta\theta^\top} p_{ij} = \frac{\partial_\theta A^\top}{C} - \frac{A\partial_{\theta^\top}C + \partial_\theta B^\top}{C^2} + 2\frac{B\partial_{\theta^\top}C}{C^3}, \quad j = 1, \dots, i-2, \tag{19}$$

where

$$\partial_\theta A^\top = \Psi_{i-1,j}\Big[\phi(\tau_{i-1})\partial^2_{\theta\theta^\top}p_{i-1,j} + \left(\partial^2_{\theta\theta^\top}\phi(\tau_{i-1})\right)p_{i-1,j} + 2\partial_\theta\phi(\tau_{i-1})\odot\partial_\theta p_{i-1,j} + 2\left\{\left(\partial_\theta\phi(\tau_{i-1})\right)p_{i-1,j} + \phi(\tau_{i-1})\partial_\theta p_{i-1,j}\right\}\odot\partial_\theta\psi_{i-1,j} + \phi(\tau_{i-1})p_{i-1,j}\left\{\left(\partial_\theta\psi_{i-1,j}\right)^{\otimes 2} + \partial^2_{\theta\theta^\top}\psi_{i-1,j}\right\}\Big]$$

$$A\partial_{\theta^\top}C + \partial_\theta B^\top = 2\left\{\Psi_{i-1,j}\Big[p_{i-1,j}\partial_\theta\phi(\tau_{i-1}) + \phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j} + \phi(\tau_{i-1})\partial_\theta p_{i-1,j}\Big]\right\} \odot \left\{\sum_{k=1}^{i-2}\Psi_{i-1,k}\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k) + \partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k} + \left\{\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k} + \partial_\theta p_{i-1,k}\right)\Big]\right\}$$

and

$$B\partial_{\theta^\top}C = \phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}\left\{\sum_{k=1}^{i-2}\Psi_{i-1,k}\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k) + \partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k} + \left\{\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k} + \partial_\theta p_{i-1,k}\right)\Big]\right\}^{\otimes 2}.$$

The recursion is initialized with $\partial^2_{\theta\theta^\top} p_{21} = 0$. Then for each consecutive $i \in \{3, \ldots, n+1\}$ compute (19) for $j \leq i-2$, and then when $j = i-1$ simply use

$$\partial^2_{\theta\theta^\top} p_{i,i-1} = -\sum_{k=1}^{i-2} \partial^2_{\theta\theta^\top} p_{ik}.$$

References

Chavez-Demoulin, V., Davison, A. C., and McNeil, A. J. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5(2):227–234.

Chen, F. and Stindl, T. (2018). Direct likelihood evaluation for the renewal Hawkes process. *Journal of Computational and Graphical Statistics*, 27:119–131.

Cont, R. (2011). Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine*, 28(5):16–25.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*. Springer-Verlag, New York, 2nd edition.

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

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66:1127–1162.

Ertekin, S., Rudin, C., and McCormick, T. H. (2015). Reactive point processes: A new approach to predicting power failures in underground electrical systems. *Ann. Appl. Stat.*, 9(1):122–144.

Filimonov, V. and Sornette, D. (2012). Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Phys. Rev. E*, 85:056108.

Filimonov, V. and Sornette, D. (2015). Apparent criticality and calibration issues in the hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15(8):1293–1314.

Halpin, P. F. (2013). A scalable EM algorithm for Hawkes processes. In Millsap, R. E., van der Ark, L. A., Bolt, D. M., and Woods, C. M., editors, *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting*, pages 403–414. Springer New York, New York, NY.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443.

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

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):pp. 493–503.

Klüppelberg, C. and Mikosch, T. (1995). Explosive poisson shot noise processes with applications to risk reserves. *Bernoulli*, 1(1/2):125–147.

Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1 – 25.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. In *Joint Statistical Meetings*, Miami, Florida.

Lewis, E., Mohler, G., Brantingham, P. J., and Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stindl, T. and Chen, F. (2018). Likelihood based inference for the multivariate renewal Hawkes process. *Computational Statistics & Data Analysis*, 123:131–145.

Stindl, T. and Chen, F. (2019). Modeling extreme negative returns using marked renewal Hawkes processes. *Extremes*, 22(4):705–728.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Wheatley, S., Filimonov, V., and Sornette, D. (2016). The Hawkes process with renewal immigration & its estimation with an EM algorithm. *Computational Statistics & Data Analysis*, 94:120 – 135.