Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference (2022) 1–19
JSPI

Alternative asymptotic inference theory for a nonstationary Hawkes process

Tsz-Kit Jeffrey Kwan[1,2]
j.t.kwan@unsw.edu.au
Feng Chen[3]
feng.chen@unsw.edu.au
William T.M. Dunsmuir [4]
w.dunsmuir@unsw.edu.au

Abstract

The Hawkes process is a popular point process model for events that exhibit a local clustering behaviour. Chen and Hall studied the asymptotic inference theory for a nonstationary Hawkes process where the baseline intensity is proportional to some time-varying function with the proportionality constant n tending to infinity. However, they assumed the excitation kernel to be independent of n, and therefore, as n increases, the waiting times from a baseline event to its excited events are of order $O_P(1)$ while the waiting times between baseline events is $O_P(1/n)$, suggesting the excitation effect is not local anymore, which defeats the purpose of choosing the exponential excitation kernel in the first place. To avoid this issue, we study the model in a more realistic setting where the excitation kernel also depends on the limit index n, so that the waiting times to excited events are of the same order of magnitude as those between baseline events. We establish consistency and asymptotic normality of the Maximum Likelihood Estimator (MLE). We also propose a score test to assess the constancy of the baseline intensity function and derive the asymptotic properties of the test. We apply the MLE and the score test to ultra-high frequency financial data as well as evaluating their finite sample performance via simulation experiments.

Keywords: Asymptotic power, Heavy-traffic asymptotics, Infill asymptotics, Locally stationary Hawkes process, Self-exciting.

1. Introduction

The self-exciting point process (SEPP), or Hawkes process [21], is a counting process N_t with stochastic intensity

$$\lambda_t = \nu + \int_0^{t-} g(t-s)\,\mathrm{d}N_s,\ t \geq 0, \tag{1}$$

[1]Designated corresponding author
[2]School of Mathematics and Statistics, The University of New South Wales, NSW 2052, Australia
[3]School of Mathematics and Statistics, The University of New South Wales, NSW 2052, Australia
[4]School of Mathematics and Statistics, The University of New South Wales, NSW 2052, Australia

where $\nu > 0$ denotes the baseline intensity, and the function g denotes the excitation kernel which measures the influence of past events on the intensity. Initial works by Hawkes [22, 21, 20] and Hawkes and Oakes [23] examined the probabilistic properties of the SEPP. Ogata [26] and Ozaki [28] studied the maximum likelihood estimator (MLE) of the SEPP. A comprehensive study of the mathematical development on the Hawkes process and point processes in general can be found in the standard reference on point processes by Daley and Vere-Jones [13, 14].

The SEPP is often used to model phenomena that exhibit a self-exciting nature and has seen a wide range of applications in areas such as neurosciences [30], limit order book modelling and trade arrival times for stocks [12], earthquake modelling and prediction [27, 36], social network interaction modelling [11] and infectious disease transmission modelling in epidemiology [17]. The Hawkes process has also seen application in conjunction with modern machine learning techniques [19, 25, 37].

While the aforementioned contributions mostly assume that the baseline intensity ν is constant with respect to time, the assumption of a constant baseline intensity is restrictive and unrealistic for modelling purposes in many situations. For instance, in the study of epidemiology, it is well known that baseline infection rates vary due to human activity, seasonality and other factors [24]. In modelling intra-day stock trading, Engle and Russell [16] observed that trading intensity is often much higher closer to market open and market close. In such situations, a SEPP with a constant baseline intensity fails to account for any time-varying trend in the event rate that cannot be explained by the excitation effect events.

This motivated Chen and Hall [6] to generalise the SEPP in (1) by allowing ν to vary with time, which leads to a point process N_t with stochastic intensity

$$\lambda_t = \nu(t) + \int_0^{t-} g(t-s)\,\mathrm{d}N_s. \tag{2}$$

Chen and Hall [6] referred to the generalized model as the SEPPVB (Self-Exciting Point Process with Varying Baseline intensity) and studied the SEPPVB under a parametric framework where the intensity function $\nu(\cdot)$ and the excitation kernel $g(\cdot)$ are known up to a finite dimensional parameter. They showed that the MLE of the parameters are consistent and asymptotically normally distributed.

Roueff et al. [31], Roueff and von Sachs [32] proceeded further by allowing both the baseline intensity and the excitation kernel to be time dependent, resulting in a process known as the locally stationary Hawkes process. They then considered its time-frequency analysis, which focused on the first two moments of the process. However, estimation of the first and second order structure of locally stationary Hawkes processes on the real line does not necessarily allow us to establish consistency and asymptotic normality of the MLE. Furthermore, without a full identification of the model, it is not possible to assess model performance. Clinet and Potiron [8] considered a varying coefficient SEPP with an exponential kernel where the vector of parameters θ_t is time-varying and stochastic. Their work focused on the estimation of the integrated parameter $\theta = \frac{1}{T}\int_0^T \theta_t\,\mathrm{d}t$ and a local central limit theorem for θ. Note that in these works, the models were not fully estimated as the underlying Hawkes process cannot be identified simply from the estimated features of the model.

The asymptotic theory developed in [6] considers a sequence of SEPPVB models N^n on the interval $[0, 1]$ indexed by n with the intensity process tending to infinity as n increases. Specifically, the stochastic intensity process of N_t^n is

$$\lambda_t^n = a_n\nu(t) + \int_0^{t-} g(t-s)\,\mathrm{d}N_s^n,\ t \in [0,1],$$

for some functions $\nu(\cdot) > 0$ and $g(\cdot) \geq 0$, and a scaling constant $a_n \to \infty$. Note that while the baseline intensity $a_n\nu(t)$ tends to ∞ with n, the excitation kernel remains fixed. This scenario for asymptotics has two undesirable features. The first is that the semiparametric SEPPVB model, in which the baseline intensity function was left unspecified while the excitation kernel is known up to a finite dimensional parameter, is asymptotically unidentifiable, as noted by Chen and Hall [7]. The second reason is that it does not agree well with some empirical evidence. For example, in the financial applications reported by Chen and Hall [6], it was noted that when the SEPPVB model was fitted to intraday stock transaction data, the estimated baseline intensity function and the initial value of the excitation kernel were often of similar magnitudes, suggesting that the excitation kernel should also be allowed to depend on the limiting index n, rather than be held fixed. Therefore, in this work we adopt a scenario for asymptotics where both the baseline intensity function and the excitation kernel depend on the limiting index n. As n tends to infinity, while the baseline

intensity function grows to infinity in the same fashion as in [6], the excitation kernel becomes thinner and taller with its integral held fixed, and converges to an improper Dirac δ-function in the limit. Specifically, we let the baseline event intensity function be $n\nu(\cdot)$ for a function $\nu(\cdot) > 0$ and the excitation kernel be $ng(n\cdot)$ for a function $g(\cdot) \geq 0$. Since clustering behaviour tends to be short lived, empirically speaking, to maintain the self-exciting nature of the Hawkes process, we require a setup which permits such fundamental characteristic.

We will establish the consistency and asymptotic normality of the MLE of a parametric SEPP with time-varying baseline intensity and an exponential kernel as the scaling constant n tends to infinity under certain regularity conditions. To establish consistency, we show that the normalised log-likelihood function converges to a deterministic function that has a unique maximum at the true parameter θ^0. We will do so by an approximation procedure. Let a_n be a sequence of positive constants that tends to infinity at a rate slower than n. Partition the time domain [0,1] into a_n blocks. As the partition gets finer, the original time-varying baseline intensity SEPP on each block is locally approximated by a stationary SEPP with constant baseline intensity. Although the original process itself is nonstationary due to the time-varying baseline intensity, the ergodicity of the stationary local approximations can be used to show the original SEPP, when suitably normalised, approaches some limiting process asymptotically. By analysing the limiting process, we can establish consistency and asymptotic normality of the MLE. This forms one of the main theoretical contributions of this work and distinguishes it from works of [8] and [32] since the goal of inference is different here as mentioned previously. Furthermore, this enables us to assess the adequacy of fit of a SEPP with constant baseline intensity by investigating the asymptotic distribution of the score test statistic under the null hypothesis that the baseline intensity is constant.

The remainder of this article is organised as follows. Section 2 formally introduces the abovementioned SEPP model and the setting of the asymptotics, and some relevant notation. In Section 3, we present the asymptotic properties of the MLE of the model parameters. Section 4 presents the asymptotic null distribution of the score-test statistic to test the constancy of the baseline intensity of the SEPP. In this section, we also study the local power of the score test. In Section 5, we report the results of numerical studies based on simulated data and real data from finance. Section 6 concludes with a discussion. Some further technical arguments are given in the online supplement. The Appendix contains the proofs for the results in Sections 3 and 4. The computations in this work were implemented in `R` [29] with the aid of the packages `IHSEP` [5] and `ibs` [4].

2. The Setup of the Asymptotic Theory

Consider a sequence of time-varying baseline intensity SEPPs $\tilde{N}^n$ on the interval [0,1], $n = 1, 2, \ldots$ For each n, let $\tilde{\mathcal{F}}^n = (\tilde{\mathcal{F}}^n_t)_{t\in[0,1]}$, with $\tilde{\mathcal{F}}^n_t = \sigma\{\tilde{N}^n_s : 0 \leq s \leq t\}$, denote the filtration induced by $\tilde{N}^n$. The $\tilde{\mathcal{F}}^n$-intensity of $\tilde{N}^n_t$ is

$$\tilde{\lambda}^n_t \equiv n\nu(t) + \int_0^{t-} ng(n(t-s))\,\mathrm{d}\tilde{N}^n_s, \quad t \in [0,1], \tag{3}$$

where $\nu(\cdot) > 0$ is the (normalised) baseline intensity function, and $g(\cdot) \geq 0$ is the excitation kernel such that $\int_0^\infty g(t)\,\mathrm{d}t < 1$. Assume the functions $\nu(\cdot) = \nu(\cdot;\theta)$ and $g(\cdot) = g(\cdot;\theta)$ are known up to a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$. The likelihood of θ relative to the sample path of N^n on $[0,1]$, up to a multiplicative constant, takes the form

$$\tilde{l}^n(\theta) = \exp\Big\{\int_0^1 \log\left(\tilde{\lambda}^n_{t,\theta}\right)\mathrm{d}\tilde{N}^n_t - \int_0^1 \tilde{\lambda}^n_{t,\theta}\,\mathrm{d}t\Big\},$$

where $\tilde{\lambda}^n_{t,\theta}$ is as in (3) with $\nu(\cdot)$ and $g(\cdot)$ replaced by $\nu(\cdot;\theta)$ and $g(\cdot;\theta)$ respectively. The normalised log-likelihood, score function and the negative Hessian matrix are respectively

$$\begin{aligned}
\tilde{\mathcal{L}}^n(\theta) &= \frac{1}{n}\log(\tilde{l}^n(\theta)) = \frac{1}{n}\int_0^1 \log\left(\tilde{\lambda}^n_{t,\theta}\right) \mathrm{d}\tilde{N}^n_t - \frac{1}{n}\int_0^1 \tilde{\lambda}^n_{t,\theta}\,\mathrm{d}t,\\
\tilde{\mathcal{S}}^n(\theta) &= \partial_\theta \tilde{\mathcal{L}}^n(\theta) = \frac{1}{n}\int_0^1 \frac{\partial_\theta \tilde{\lambda}^n_{t,\theta}}{\tilde{\lambda}^n_{t,\theta}} \mathrm{d}\tilde{N}^n_t - \frac{1}{n}\int_0^1 \partial_\theta \tilde{\lambda}^n_{t,\theta}\,\mathrm{d}t,\\
\tilde{\mathcal{I}}^n(\theta) &= -\partial^{\otimes 2}_\theta \tilde{\mathcal{L}}^n(\theta) = -\frac{1}{n}\int_0^1 \left(\frac{\partial^{\otimes 2}_\theta \tilde{\lambda}^n_{t,\theta}}{\tilde{\lambda}^n_{t,\theta}} - \left(\frac{\partial_\theta \tilde{\lambda}^n_{t,\theta}}{\tilde{\lambda}^n_{t,\theta}}\right)^{\otimes 2}\right) \mathrm{d}\tilde{N}^n_t + \frac{1}{n}\int_0^1 \partial^{\otimes 2}_\theta \tilde{\lambda}^n_{t,\theta}\,\mathrm{d}t,
\end{aligned}$$

where $\partial_\theta = \frac{\partial}{\partial\theta}$, and $\partial^{\otimes 2}_\theta = \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}^\top$. The MLE $\hat{\theta}^n$ maximizes the log-likelihood $\tilde{\mathcal{L}}^n(\theta)$ over the parameter space Θ. Applying a time change, we can stretch the time domain from $[0,1]$ to $[0,n]$. Let $N^n_t = \tilde{N}^n_{t/n}$, and $\mathcal{F}^n_t = \tilde{\mathcal{F}}^n_{t/n}$, $t \in [0,n]$. Then the intensity of N^n relative to the filtration $\mathcal{F}^n = (\mathcal{F}^n_t)_{t\in[0,n]}$ is

$$\lambda^n_{t,\theta} = \frac{1}{n}\tilde{\lambda}^n_{t/n,\theta} = \nu\left(\frac{t}{n};\theta\right) + \int_0^{t-} g(t-s;\theta)\,\mathrm{d}N^n_s, \quad t \in [0,n]. \tag{4}$$

The associated normalised log-likelihood, score function and the negative Hessian are now given respectively by

$$\mathcal{L}^n(\theta) = \frac{1}{n}\int_0^n \log(\lambda^n_{t,\theta})\,\mathrm{d}N^n_t - \frac{1}{n}\int_0^n \lambda^n_{t,\theta}\,\mathrm{d}t, \tag{5}$$

$$\mathcal{S}^n(\theta) = \partial_\theta \mathcal{L}^n(\theta) = \frac{1}{n}\int_0^n \frac{\partial_\theta \lambda^n_{t,\theta}}{\lambda^n_{t,\theta}}\,\mathrm{d}N^n_t - \frac{1}{n}\int_0^n \partial_\theta \lambda^n_{t,\theta}\,\mathrm{d}t, \tag{6}$$

$$\mathcal{I}^n(\theta) = -\partial^{\otimes 2}_\theta \mathcal{L}^n(\theta) = -\frac{1}{n}\int_0^n \left(\frac{\partial^{\otimes 2}_\theta \lambda^n_{t,\theta}}{\lambda^n_{t,\theta}} - \left(\frac{\partial_\theta \lambda^n_{t,\theta}}{\lambda^n_{t,\theta}}\right)^{\otimes 2}\right) \mathrm{d}N^n_t + \frac{1}{n}\int_0^n \partial^{\otimes 2}_\theta \lambda^n_{t,\theta}\,\mathrm{d}t. \tag{7}$$

It is clear that $\mathcal{L}^n(\theta) = \tilde{\mathcal{L}}^n(\theta) - \log(n)\tilde{N}^n_1/n$, $\mathcal{S}^n(\theta) = \tilde{\mathcal{S}}^n(\theta)$, $\mathcal{I}^n(\theta) = \tilde{\mathcal{I}}^n(\theta)$, and therefore the MLE is not affected by the time change.

3. Consistency and Asymptotic Normality of the MLE

In this section, we outline the procedure to establish consistency and asymptotic normality of the MLE $\hat{\theta}^n$. The following is a list of regularity conditions under which we will establish consistency and asymptotic normality of $\hat{\theta}^n$.

[C1] The parameters of $\nu(\cdot)$ and $g(\cdot)$ are separable; that is $\theta = (\theta_1^\top, \theta_2^\top)^\top \in \Theta_1 \times \Theta_2 = \Theta \subset \mathbb{R}^{d_1+d_2}$, with $\nu(t;\theta) = \nu(t;\theta_1)$ depending only on θ_1 and $g(t;\theta) = g(t;\theta_2)$ depending only on θ_2. Furthermore, we assume that Θ_1 and Θ_2 are both compact and contain a nonempty open ball in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, which in turn contains θ_1^0 and θ_2^0 respectively, with $(\theta_1^{0\top}, \theta_2^{0\top})^\top = \theta^0$ being the true parameter.

[C2] All components of $\nu(t;\theta_1)$, $\partial_{\theta_1}\nu(t;\theta_1)$, $\partial^{\otimes 2}_{\theta_1}\nu(t;\theta_1)$ and $\partial^{\otimes 3}_{\theta_1}\nu(t;\theta_1)$ are Lipschitz continuous in t and θ_1, where $\partial^{\otimes 3}_{\theta_1} = \frac{\partial}{\partial\theta_1}\partial^{\otimes 2}_{\theta_1}$.

[C3] There exists a $t \in [0,1]$ for each $i \in \{1,\cdots,d_1\}$ such that $\partial_{\theta_1^i}\nu(t;\theta_1)$ is non-zero for any $\theta_1 \in \Theta_1$ with $(\theta_1^1,\ldots,\theta_1^{d_1})^\top = \theta_1$.

[C4] The excitation kernel $g(t;\theta_2)$ is exponential; that is, it depends on the parameter $\theta_2 = (\eta,\beta)^\top$ in this form,

$$g(t;\theta_2) = \eta\beta e^{-\beta t}$$

where η is known as the branching ratio and β is known as the decay parameter. Furthermore, for any $(t, \theta_1) \in [0,1] \times \Theta_1$, there exist $\underline{\nu}, \bar{\nu}, \underline{\eta}, \bar{\eta}, \underline{\beta}$ and $\bar{\beta} \in \mathbb{R}$ such that

$$\begin{aligned} 0 < \underline{\nu} \leq \nu(t;\theta_1) \leq \bar{\nu} < \infty; \\ 0 < \underline{\eta} \leq \eta \leq \bar{\eta} < 1; \\ 0 < \underline{\beta} \leq \beta \leq \bar{\beta} < \infty. \end{aligned}$$

For the rest of the article, $\xrightarrow{\mathbb{P}}$ denotes convergence in probability and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. We now present the theorems for consistency and asymptotic normality of the MLE $\hat{\theta}^n$.

Theorem 3.1. *Under [C1] - [C4], with probability tending to 1, $\hat{\theta}^n$ exists as a maximizer to the normalised log-likelihood function $\mathcal{L}^n(\theta)$ and $\hat{\theta}^n \xrightarrow{\mathbb{P}} \theta^0$.*

Theorem 3.2. *Under [C1] - [C4], there exists an invertible matrix $\Gamma(\theta^0) \in \mathbb{R}^{d\times d}$ such that*

$$\sqrt{n}(\hat{\theta}^n - \theta^0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Gamma(\theta^0)^{-1}\right) \tag{8}$$

where $\Gamma(\theta^0)$ denotes the deterministic limit of the empirical information matrix $\mathcal{I}^n(\theta)$ evaluated at θ^0 as in (7). Moreover, $\mathcal{I}^n(\hat{\theta}^n) \xrightarrow{\mathbb{P}} \Gamma(\theta^0)$.

The challenge in proving Theorem 3.1 arises from the time-varying baseline intensity $\nu(\cdot;\theta_1)$ since stationarity and ergodicity of point processes often play an important role in establishing consistency of MLEs [26, 15, 9]. Therefore, with a time-varying baseline intensity, standard techniques would not suffice in our setting. Our procedure to establish consistency is as follows. We first show that

$$\sup_{\theta\in\Theta} \left|\mathcal{L}^n(\theta) - \mathcal{L}^n(\theta^0) - \tilde{\pi}(\theta)\right| \xrightarrow{\mathbb{P}} 0 \tag{9}$$

for some mapping $\tilde{\pi}$ with a unique maximum at θ^0. Then, by the M-estimator master theorem [see e.g. 35, Theorem 5.7], $\hat{\theta}^n$ converges to θ^0 in probability. We will show (9) through an approximation procedure. Let a_n be a monotone sequence of positive integers $\to \infty$ such that $a_n = o(n)$. Split the integral in $\mathcal{L}^n(\theta) - \mathcal{L}^n(\theta^0)$ into a_n equal-lengthed blocks with length $b_n = n/a_n$, and denote by $\mathcal{L}^{i,n}(\theta)$ the normalised log-likelihood of N^n on the ith block $((i-1)b_n, ib_n]$, $i = 1, \ldots, a_n$. Then, $\mathcal{L}^n(\theta) - \mathcal{L}^n(\theta^0)$ can be expressed as

$$\begin{aligned} &\frac{1}{a_n}\sum_{i=1}^{a_n} \mathcal{L}^{i,n}(\theta) - \mathcal{L}^{i,n}(\theta^0) \\ &= \frac{1}{a_n}\sum_{i=1}^{a_n} \frac{1}{b_n}\left\{\int_{(i-1)b_n}^{ib_n} \log\frac{\lambda^n_{t,\theta}}{\lambda^n_{t,\theta^0}}\,\mathrm{d}N^n_t - \int_{(i-1)b_n}^{ib_n} (\lambda^n_{t,\theta} - \lambda^n_{t,\theta^0})\,\mathrm{d}t\right\}. \end{aligned}$$

On each block, we approximate $\mathcal{L}^{i,n}(\theta) - \mathcal{L}^{i,n}(\theta^0)$ by $\mathcal{L}^{i,n,c}(\theta) - \mathcal{L}^{i,n,c}(\theta^0)$, where $\mathcal{L}^{i,n,c}(\theta)$ denotes the normalised log-likelihood of an auxiliary SEPP $N^{i,n,c}$ on $[0, b_n]$, which has a constant baseline intensity $\nu((i-1)/a_n)$ and excitation kernel g.

For such an approximation procedure to work, we make use of the Poisson embedding technique adopted in [8] where via a fixed point argument, both N^n and $N^{i,n,c}$ can be represented as an integral with respect to a unit rate Poisson process on $\mathbb{R}^2$ such that they are defined on the same probability space and they are generated by the same realisation of the unobserved process N^P. For more details on Poisson embedding techniques, see [14, Section 14.7], [2] or [3]. It should be noted that the model under $N^{i,n,c}$ is by definition misspecified, thus $\mathcal{L}^{i,n,c}(\theta)$ is not the normalised log-likelihood function of the original model under N^n on the i-th block.

It turns out that the intensity process of $N^{i,n,c}$ is asymptotically ergodic in the sense that there exists some mapping π such that $\mathcal{L}^{i,n,c}(\theta) - \mathcal{L}^{i,n,c}(\theta^0) - \pi(\nu((i-1)/a_n;\theta_1), \theta_2, \theta^0)$ converges to 0 in probability, uniformly in i and θ. From the assumed regularity conditions on ν, the mapping $\pi(\nu(x;\theta_1), \theta_2, \theta^0)$ is Riemann integrable in x. Therefore, (9) holds

with $\tilde{\pi}(\theta) = \int_0^1 \pi(\nu(x;\theta_1),\theta_2,\theta^0)\,\mathrm{d}x$. Here, the integrability of π is key in transferring the ergodicity result from the blocks to the whole time domain, in spite of the global nonstationarity of the process N^n.

For the proof of Theorem 3.2, we first utilise the martingale central limit theorem to establish the asymptotic normality of the score function in the neighbourhood of θ^0, and then use the Mean Value Theorem (MVT) to transfer the normality to $\hat{\theta}$. Furthermore, we will use a similar approximation procedure mentioned above to identify the variance-covariance matrix from the observed information matrix. The proofs for Theorem 3.1 and 3.2 are detailed in the Appendix.

4. The Score Test for the Constancy of the Baseline Intensity

To justify the use of a time-varying baseline intensity SEPP, a test of the constancy of the baseline intensity is needed. For this purpose, standard inferential techniques such as the Wald or likelihood ratio test might be considered. However, we shall use the score test instead due to its computational convenience, since it only requires the computation of the MLE under the null hypothesis. The score test has been used to test the presence of mark effect in the marked Hawkes process [10]. We will study the asymptotic distribution of the score test statistic under the null hypothesis that the SEPP has a constant baseline intensity. Suppose the baseline intensity function can be parameterised in the form $\nu(\cdot;\theta_1) = \theta_{1,1} + \psi(\cdot;\theta_{1,2})$ for $\theta_1 = (\theta_{1,1},\theta_{1,2}) \in \Theta_{1,1}\times\Theta_{1,2}$, $\Theta_{1,1}\subset\mathbb{R}_+$, $\Theta_{1,2}\subset\mathbb{R}^{d_1-1}$, $\Theta_1 = \Theta_{1,1}\times\Theta_{1,2}$ and $\psi : (x,\theta_{1,2}) \mapsto \psi(x;\theta_{1,2})$ satisfying condition [S1] below. Note that the regularity condition [C2] on $\nu(\cdot)$ implies the condition [S2] below for $\psi(\cdot)$.

[S1] There exists a unique $\theta_{1,2}^{0,*} \in \Theta_{1,2}$ such that $\psi(x;\theta_{1,2}^{0,*}) = 0$ for all $x \in [0,1]$;

[S2] All components of $\partial^{\otimes i}_{\theta_{1,2}}\psi(x;\theta_{1,2})$ for $i \in \{0,1,2,3\}$ are Lipschitz continuous in x.

The null hypothesis to be tested can now be formulated as

$$H_0 : \theta_{1,2} = \theta_{1,2}^{0,*}.$$

The score test statistic is defined as

$$\mathcal{T}^n = \sqrt{n}\partial_\theta \mathcal{L}^n(\theta)^\top \mathbb{E}\left[-\partial^{\otimes 2}_\theta \mathcal{L}^n(\theta)\right]^{-1} \sqrt{n}\partial_\theta \mathcal{L}^n(\theta)\Big|_{\theta=\hat{\theta}^{0,n}}, \tag{10}$$

where $\hat{\theta}^{0,n}$ is the MLE of the parameter under H_0. Our goal is to identify the asymptotic distribution of $\mathcal{T}^n$. Notice that if $\nu(t;\theta_1) = \theta_{1,1} + \psi(t;\theta_{1,2})$ and g is again an exponential kernel as denoted in [C4], then $\lambda^n_{t;\theta}$ satisfies [C1]-[C4]. Under H_0, the score function evaluated at $\hat{\theta}^{0,n}$ is given by

$$\partial_\theta \mathcal{L}^n(\hat{\theta}^{0,n}) = \left[0, \partial_{\theta_{1,2}}\mathcal{L}^n(\hat{\theta}^{0,n}), 0\right]^\top.$$

Hence, we only require the $(2:d_1)\times(2:d_1)$ block of the matrix in (10) evaluated at $\hat{\theta}^{0,n}$, denoted by

$$\mathbb{E}\left[-\partial^{\otimes 2}_\theta \mathcal{L}^n(\hat{\theta}^{0,n})\right]^{-1}_{[2:d_1,2:d_1]},$$

in order to compute the score test statistics. Note that the above expectation is typically not computable in closed form. An empirical evaluation by the time average over event times using the estimated parameter $\hat{\theta}^{0,n}$ could be used to replace the expectation. Therefore, the score test statistic can be implemented in practice as

$$\hat{\mathcal{T}}^n = \sqrt{n}\partial_{\theta_{1,2}}\mathcal{L}^n(\theta)^\top \left[\frac{1}{n}\int_0^n \left(\frac{\partial_\theta \lambda^n_{t,\theta}}{\lambda^n_{t,\theta}}\right)^{\otimes 2} \mathrm{d}N^n_t\right]^{-1}_{[2:d_1,2:d_1]} \sqrt{n}\partial_{\theta_{1,2}}\mathcal{L}^n(\theta)\Big|_{\theta=\hat{\theta}^{0,n}}. \tag{11}$$

We now state the asymptotic null distribution of the score statistic.

Theorem 4.1. *In addition to [C1] - [C4], [S1] and [S2], suppose H_0 is true. Then $\hat{\mathcal{T}}^n \xrightarrow{\mathcal{D}} \chi^2_{d_1-1}$, where χ^2_p denotes a chi-square distribution with p degrees of freedom.*

Next, we investigate the local power of the score test. To this end, we need to identify the asymptotic distribution of $\hat{\mathcal{T}}^n$ under a sequence of local alternatives $H_a^n : \theta_{1,2} = \theta_{1,2}^{a,n}$ where $\left|\theta_{1,2}^{a,n} - \theta_{1,2}^{0,*}\right| = O(\frac{1}{\sqrt{n}})$, which is given in the following result.

Theorem 4.2. *In addition to [C1] - [C4], [S1] and [S2], suppose H_a^n is true. Then, there exist some vector $k \in \mathbb{R}^{d_1-1}$ and some positive-definite matrix valued function $\tilde{\Gamma} : \Theta \mapsto \mathbb{R}^{d_1-1 \times d_1-1}$ such that*

$$\hat{\mathcal{T}}^n \xrightarrow{\mathcal{D}} \chi^2_{d_1-1, k^\top \tilde{\Gamma}(\theta^0_{1,1}, \theta^{0,*}_{1,2}, \theta^0_2)k}.$$

Here $\chi^2_{p,c}$ denotes a noncentral chi-square distribution with p degrees of freedom and c as the noncentrality parameter.

The proofs for Theorem 4.1 and 4.2 are presented in the Appendix.

5. Empirical Study and Simulations

5.1. Empirical study: modelling bid trade-throughs using a time-varying baseline SEPP

The limit order book records the arrival time, size, side of trade and other features of limit and market orders. The primary function of the limit order book is to match buyers and sellers in the market. Market participants often attempt to minimise the impact of their order by splitting their orders based on the liquidity of the stocks they are trading. This is to ensure the prices do not change unfavourably against them due to the size of their order. However, when the volume of a market order submitted is greater than what is available at the best quote price, the order will move on to consume the liquidity available at the second best price and and continue to do so until the order is fulfilled. A trade-through is defined as a transaction that occurs at the second or worse price levels in an order book and hence provide insights for market dynamics and price movements [34]. Stindl and Chen [33] analysed trade-throughs for both the bid and ask side of the Euronext-traded limit order book for the stock BNP Paribas through a bivariate renewal Hawkes process. The data consists of 109 trading days from 1st June to 29th October, 2010.

While the renewal Hawkes process is stationary in nature, trade-through events are empirically nonstationary since trading activity is observed to be higher during market open and market close than in the middle of the day. Thus, to fit the renewal Hawkes process to the trade-through data, the authors of [33] had to transform the trade-through processes to make them stationary before model fitting. However, the transformation adjusts the waiting times between successive trade-through events via a function that depend on future as well as past events. Hence, the inference conducted in [33] does not translate easily onto the original data. For this reason, we want to avoid any ad hoc transformations and model the nonstationary nature of the original trade-through data directly by fitting a nonstationary Hawkes process with a time-varying baseline intensity and an exponential excitation kernel.

Since the previous findings by Stindl and Chen [33] and Toke and Pomponio [34] suggest that the cross-excitation effects between ask and bid trade-throughs are not significant, we model the ask and bid trade-through sequences as two separate processes. For illustration, we report here only the results of fitting the nonstationary Hawkes process on the bid trade-through sequence. We use an exponential excitation kernel, $g(t;\theta_2) \equiv \eta\beta e^{-\beta t}$. For the baseline intensity function, we use a B-spline function of order $p = 3$ with k equally spaced internal knots; that is, the complete knot sequence $\xi_i, i = 1, 2, \ldots, k + 2p$ of the B-spline is $\xi_1 = \cdots = \xi_p = 0$, $\xi_{p+j} = jT/(k+1)$, $j = 1, \ldots, k$, $\xi_{k+p+1} = \cdots = \xi_{k+2p} = T$, with T denoting the end of the trading day, which equals 30,600 seconds here, since the Euronext Paris is open for 8.5 hours from 9am to 5:30pm local time and we choose to measure time in seconds. Denote by basis functions. Then, the baseline intensity function takes the form $\nu(x;\theta_1) \equiv \sum_{i=1}^{p+k} \theta_{1,i} B_i^{p,k}(x)$. To determine the appropriate value of k for each trading day, we fit the model with increasing values of k starting from 0 and calculate the Akaike Information Criterion (AIC; [1]) value. The optimal value of k is chosen as the last value of k such that the AIC value stops decreasing.

Out of the 109 trading days, 45 days have optimal $k = 0$, 47 have $k = 1$, 10 have $k = 2$, 5 have $k = 3$ and the remaining 2 have $k = 4$. Figure 2 presents the plot of estimated branching ratio $\hat{\eta}$ and decay parameter $\hat{\beta}$ with a shaded 95% pointwise confidence interval on different trading days, from which we note that the estimated branching ratios are fluctuating around 0.13, consistent with the findings by Stindl and Chen [33]. Figure 2 also shows that the estimated mean waiting time to an excited trade-through event is around 60 seconds, which is also consistent with those reported by Stindl and Chen [33], although slightly smaller.

We also tested the constancy of the baseline intensity function using the score test discussed in Section 4 for each trading day. At the conventional significance level of 0.05, 30 out of the 45 days with 0 internal knots rejected H_0, 45 out of the 47 days with 1 internal knot rejected H_0 and all days with 2, 3 and 4 internal knots rejected H_0 (cf. Table 1). This results in a 84% overall rejection of the null hypothesis out of the 109 trading days, which suggests that a SEPP with a time-varying B-spline baseline intensity is more suitable in modelling the arrival of bid trade-throughs compared to a SEPP with a constant B-spline baseline intensity. The estimated baseline intensity functions are shown in Figure 1, from which we see the phenomenon that trading intensity is almost always much higher during market open and market close [16].

Table 1: Summary of the results of the score test for constancy of the baseline trade-through event intensity of the stock BNP Paribas, for the 109 trading days.

Number of internal knots	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Number of trading days	45	47	10	5	2
Number of trading days rejected H_0	30	45	10	5	2

Figure 1: Spaghetti plots of the estimated baseline intensities of bid trade-throughs of the stock BNP Paribas for the 109 trading days from 01/06/2010 - 29/10/2010.

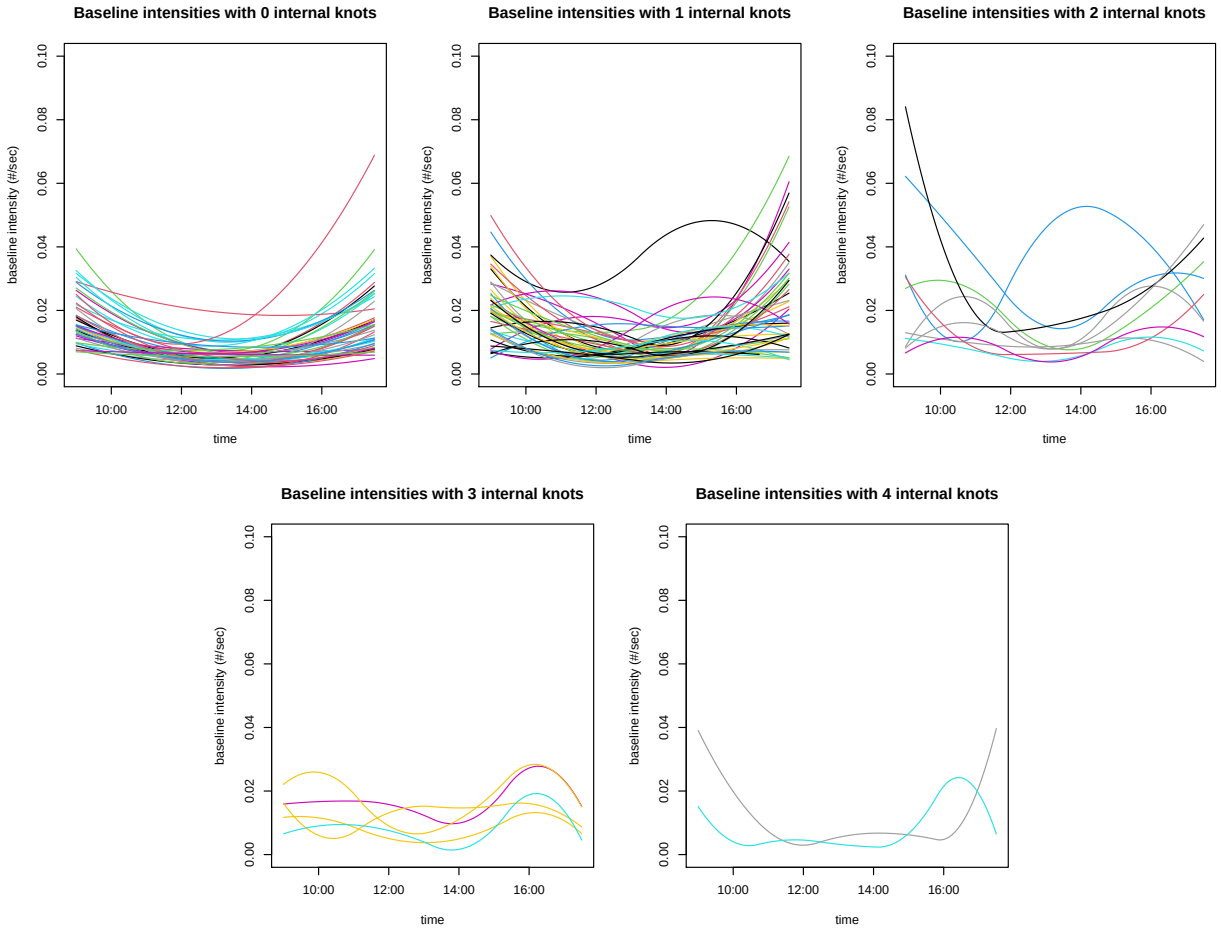
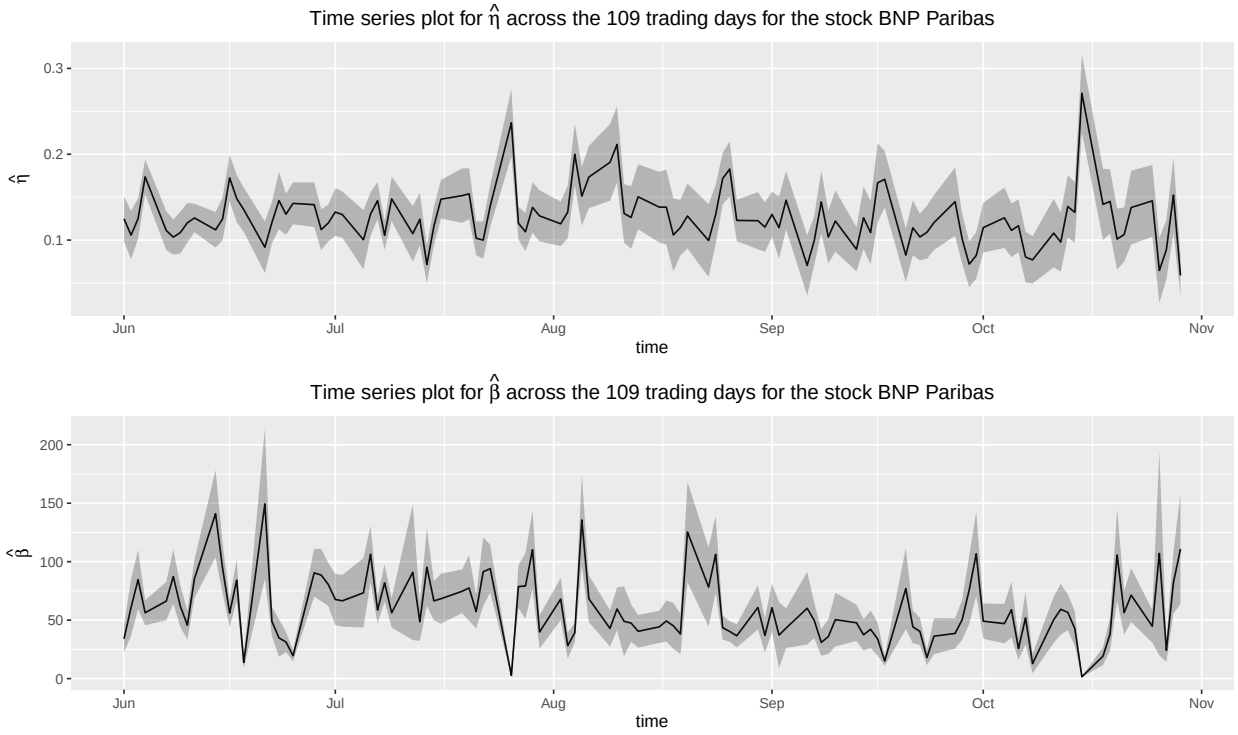


Figure 2: Time series plot of the MLE's for the branching ratio and decay parameter of the excitation kernel for the stock BNP Paribas over the 109 trading days from 01/06/2010 - 29/10/2010 with pointwise 95% confidence interval.



To assess how well our model fits to the trade-through data, we inspect the point process residuals of the fitted model, $\hat{U}_i = \hat{\Lambda}(\tau_i)$ where $\hat{\Lambda}(t) = \int_0^t \lambda(s;\hat{\theta})\,\mathrm{d}s$ is the fitted cumulative intensity process. When the model specification is correct, $\hat{\Lambda}(\tau_i)$ should approximately correspond to the jump times of a unit rate Poisson process by the random time-change theorem [13, Theorem 7.4.I] and therefore they should be roughly uniformly distributed in the interval $[0, \hat{\Lambda}(T)]$. Therefore, as a check of the goodness-of-fit of the point process model, we can test the uniformity of the residuals $\hat{U}_i$ on the interval $[0, \hat{\Lambda}(T)]$, using e.g. the Kolmogorov-Smirnov (KS) test. Out of the 109 trading days, 82 (75%) days have a p-value > 0.05 for the KS test of uniformity of the residuals. If we use the Bonferroni corrected significance level $0.05/109 = 0.00046$, then the number increases to 99 (91%). This suggests that the model we have considered can provide acceptable fit to the majority of the trade-through data. On the days where the model is inadequate, better fit might be achieved by considering more flexible excitation kernels, given the substantial flexibility of the fitted baseline intensity function. Further discussion on enhancing the model is given in the concluding remarks in Section 6.

5.2. *Simulation study*

This section is dedicated to studying the finite sample performance of the MLE and the score test. Motivated by our study on the bid trade-throughs in Section 5.1, we simulated a SEPP model with an exponential excitation kernel and a baseline intensity function that is a B-spline of order 3 with 1 internal knot on the interval $[0, T] = [0, 30600]$. The parameter vector of the B-spline function and the excitation kernel is the same as that of the fitted model on trading day 47 in the empirical study, as this day is "typical" in the sense that it has the median number 428 of bid trade-through events among the 45 days with $k = 1$ where the constancy of the baseline intensity was rejected by the

score test. To be specific, the baseline intensity function and the excitation kernel are, respectively,

$$\begin{aligned} \nu(t;\theta_1) &= \theta_{1,1}B_1^{3,1}(t) + \theta_{1,2}B_2^{3,1}(t) + \theta_{1,3}B_3^{3,1}(t) + \theta_{1,4}B_4^{3,1}(t), \\ g(t;\theta_2) &= \eta\beta\exp(-\beta t), \end{aligned} \tag{12}$$

where

$$(\theta_{1,1},\theta_{1,2},\theta_{1,3},\theta_{1,4}) = (0.0213, 0.0043, 0.0123, 0.0126),\ \eta = 0.1999 \text{ and } \beta = 39.25. \tag{13}$$

We simulated this SEPP model on the interval $[0, 30600]$ a total of 1000 times. For the simulated data set, the parameter θ was estimated by maximizing the log-likelihood function, and the variance-covariance matrix of the estimator was estimated by inverting the negative of the Hessian matrix.

The results are shown in Table 2. In this table, the first column gives the parameter, the second column the true parameter value and the third column the mean of the 1000 estimates. The fourth column gives the mean of the estimated standard errors, that is, the square-root of the negative diagonal values of the inverted standardised observed information matrix. The fifth column shows the empirical standard error, that is, the standard deviation of the 1000 estimates. The last column gives the p-value of a two-sided KS test of the normality of the 1000 estimates for each parameter, where the mean and variance of the theorized normal distribution are chosen as the empirical mean and variance of the 1000 estimates respectively. Figure 3 presents the normal QQ-plots of the parameter estimates.

From Table 2, we observe that the mean estimates and their respective true parameter values are close, relative to their standard errors. In addition, the mean estimated standard errors are close to the respective empirical standard errors. The QQ-plots in Figure 3 indicate that $\hat{\beta}$ is slightly positively skewed, while the other parameters are relatively normal looking. The p-values of the KS tests against normality are 0.05 or higher for all the estimators. This gives a similar conclusion as the QQ-plots, that the distribution of the estimates are not significantly different from normality. These findings are in line with Theorems 3.1 and 3.2.

Table 2: The results of estimating the parameters using the ML method.

Par.	True	Mean Est.	Mean $\widehat{\text{S.E.}}$	Emp. S.E.	p-value (KS test)
$\theta_{1,1}$	2.130e-2	2.134e-2	2.711e-3	2.581e-3	0.971
$\theta_{1,2}$	4.304e-3	4.294e-3	2.134e-3	2.079e-3	0.444
$\theta_{1,3}$	1.233e-2	1.238e-2	2.231e-3	2.131e-3	0.741
$\theta_{1,4}$	1.257e-2	1.261e-2	2.344e-3	2.389e-3	0.793
η	1.999e-1	1.987e-1	2.161e-2	2.170e-2	0.992
β	39.25	41.90	4.943	5.099	0.473

To assess the performance of the score test, we extend the simulation study considered above as follows. We simulate the model in (12) with $\theta_2 = (\eta,\beta)^\top$ the same as in (13) but with θ_1 varying from $\theta_1^s = (0.0112, 0.0112, 0.0112, 0.0112)^\top$ to $\theta_1^t = (0.0213, 0.0043, 0.0123, 0.0126)^\top$ along the sequence $\theta_1^i = \theta_1^s + (\theta_1^t - \theta_1^s)\times\frac{i}{70}$, $i = 0, 1, \ldots, 70$. Note that $\theta_1^{70} = \theta_1^t$ is the same as the θ_1 in (13). When $\theta_1 = \theta_1^0 = \theta_1^s$, the baseline intensity function is a constant, $\nu(\cdot;\theta_1) \equiv 0.0112$, corresponding to the null hypothesis. Note also that the value of θ_1^s here has been chosen such that for all θ_1 values along the sequence θ_1^i, $i = 0, 1, \ldots, 70$, the integral of the baseline intensity function $\nu(\cdot;\theta_1)$ on the interval $[0, T]$ always equals 342.21, so that the expected total number of events over the interval remains 428 approximately while the baseline intensity gradually moves away from a constant towards the curve $\sum_{j=1}^4 \theta_{1,j}^t B^{3,1}(\cdot)$. For each $i = 0, 1, \ldots, 70$, the model with baseline intensity $\nu(\cdot;\theta_1^i)$ and excitation kernel $g(\cdot;\theta_2)$ was simulated 1000 times. The score statistic (11) was calculated based on each simulated dataset.

Table 3 shows the empirical type-I error of the size 0.05 score test against $H_0: \theta_1 = \theta_1^0$, that is, the percentage of score test statistic values greater than the 0.95 quantile of the χ^2 distribution with 3 degrees of freedom $\chi^2_{0.95;3} = 7.81$, and the empirical 0.95 quantile of the score test statistic. We observe that the empirical type-I error is close to the nominal significance level of 0.05, and the empirical level 0.05 critical value is also close to the theoretical value of 7.81, as expected by Theorem 4.1. The left panel of Figure 4 shows the QQ-plot of the values of the probability integral transformed score test statistic $F_{\chi^2_3}(\hat{\mathcal{T}})$ when the data was simulated with $\theta = \theta_1^0$ together with the p-value of the KS test of the hypothesis that the distribution of $F_{\chi^2_3}(\hat{\mathcal{T}})$ equals uniform (0,1), with $F_{\chi^2_3}(\cdot)$ denoting the cumulative

Figure 3: Normal QQ-plots of the estimated parameters in the simulation study. The columns in the first row correspond to the parameters $\hat{\theta}_{1,1}$, $\hat{\theta}_{1,2}$ and $\hat{\theta}_{1,3}$ from left to right, and the columns in the second row correspond to the parameters $\hat{\theta}_{1,4}$, $\hat{\eta}$ and $\hat{\beta}$ from left to right.

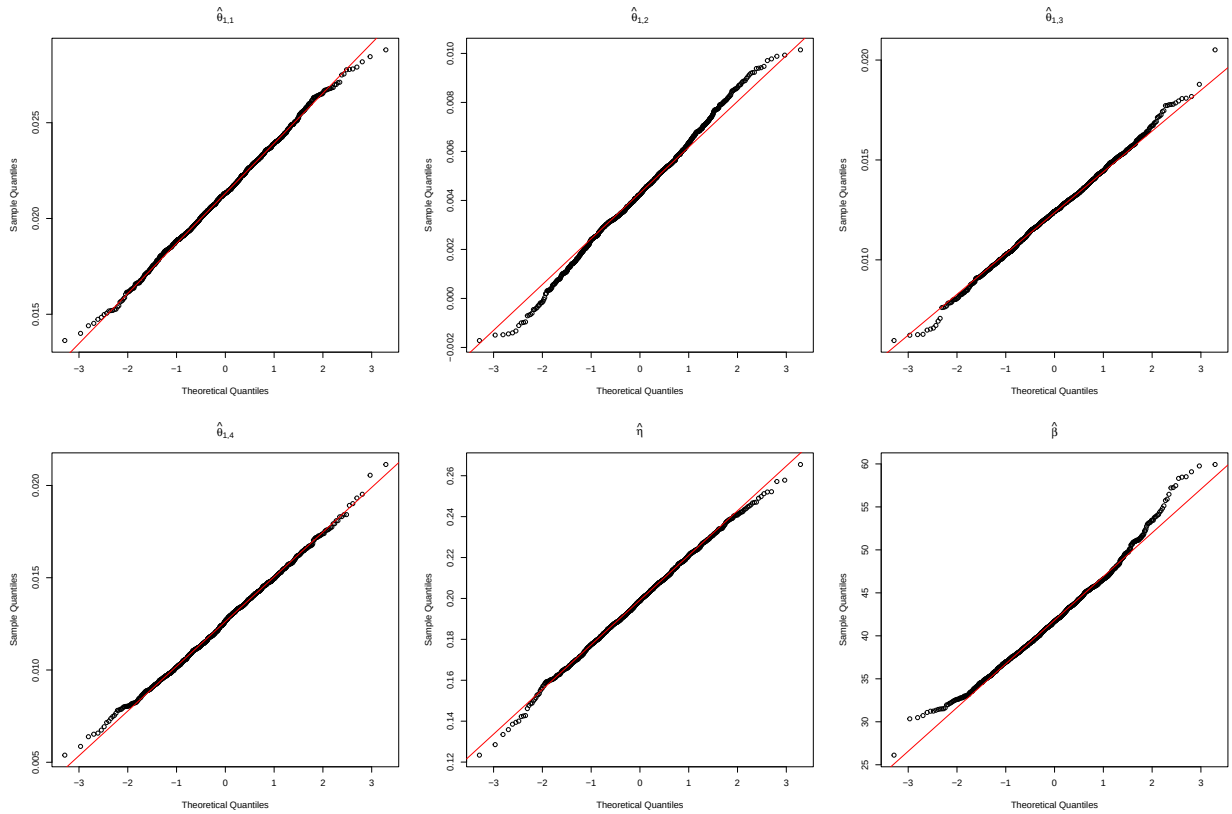


Table 3: The empirical type-I error, and the 95th percentile of the empirical distribution of the score test when $\theta_1 = \theta_1^0$.

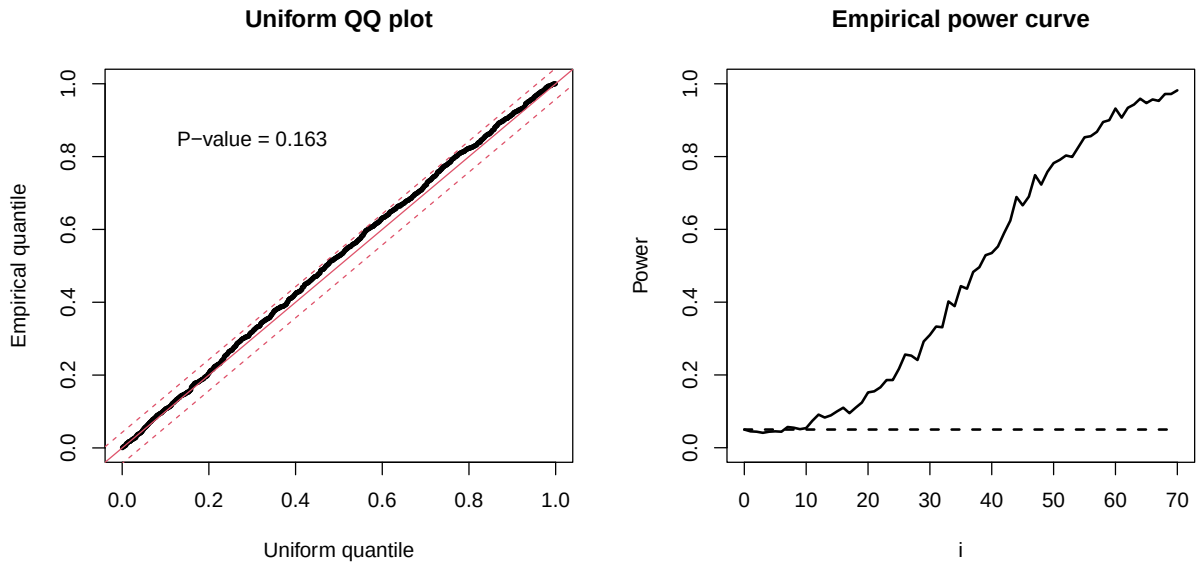
Empirical type-I Error	Empirical 5% Critical Value
0.064	8.304

distribution function of χ^2_3. From the QQ-plot, it is evident that empirical distribution of the transformed score test statistic resembles the uniform (0,1) distribution, which is also confirmed by the p-value 0.16 (> 0.05) of the KS-test. This suggests that the empirical distribution of $\hat{\mathcal{T}}$ can be well approximated by the χ^2 distribution, again in line with Theorem 4.1.

The right panel of Figure 4 shows the empirical power of the score test of $H_0 : \theta_1 = \theta_1^0$ vs $H_a : \theta_1 = \theta_1^i$, where the empirical power is calculated as the proportion of score test statistic values $> \chi^2_{0.95;3}$ when the data was simualted with $\theta_1 = \theta_1^i, i = 0, 1, \ldots, 70$. The power curve in Figure 4 indicates that the power of the score test steadily increases as θ_1 moves away from $\theta_1^s = \theta_1^0$ and approaches 1 when θ_1 approaches $\theta_1^t = \theta_1^{70}$, suggesting the score test is powerful enough to detect the non-constancy of the baseline intensity function when the magnitude of the deviation from constancy and the amount of data are typical of those seen in our emipircal studies.

Remark 5.1. *Recall that the QQ-plot for $\hat{\beta}$ in Figure 3 indicates that the empirical distribution of $\hat{\beta}$ is slightly positively skewed. In our unreported simulation experiments, we replicated the above simulations over the time domain* $[0, 61200]$ *and* $[0, 122400]$*, which has the effect of doubling and quadrupling the amount of data respectively, and observed that the resemblance to normality of the empirical distribution of $\hat{\beta}$ improves as the amount of data increases. This is also confirmed by the large (*> 0.05*) and increasing p-value of the KS-test against normality.*

Figure 4: QQ-plot of the probability integral transformed score test statistic against uniform (0,1) and the empirical power curve for the score test.



Similar trends are also observed for the other parameters.

6. Concluding Remarks

We established consistency and asymptotic normality of the MLEs for a time-varying baseline intensity Hawkes process with an exponential excitation kernel under an alternative scenario of asymptotics alluded to in the discussion of [6]. We also proposed a score test of the constancy of the baseline intensity function, and established the asymptotic distribution of the score test statistic. Our simualtions confirmed the good numerical performance of the MLE and the score test. We alos illustrated the proposed methodology using stock trade-through data.

In developing the inference theory, we relied on the exponential kernel assumption. While this assumption allowed us to exploit the ergodic theory for Hawkes processes developed by Clinet and Yoshida [9], such assumption is rather restrictive in practice. For example, in infectious disease transmission modelling, the exciting effect can be considered as the infectiousness of an infected individual, which may increase initially until it peaks at some point and then dedays towards zero. Non-monotonic excitation effect as such can not be modelled by an exponential kernel, but it can be easily accommodated by a gamma or Weibull density curve. In some other applications such as earthquake modelling, the power law decay or Pareto kernel is very popular. In our (unreported) numerical experiments, we have noted that the MLE for the SEPPVB model with such non-exponential kernels has similar asymptotic behaviors. However, the proof of asymptotic results similar to those reported in this work is not straightforward, as with such non-exponential kernels, the excitation process $\int_0^{t-} g(t-s;\theta_2^0)\,\mathrm{d}N^{i,n,c}$ becomes non-Markovian and we will not be able to apply Theorem 4.6 of [9] to establish (17). Ergodicity results for non-Markovian process are needed to establish the asymptotic theory of the MLE for SEPPs with general kernels.

One of the motivations of the alternative scenario of asymptotics considered in this work is its potential to overcome the semi-parametric unidentifiability issue pointed out by Chen and Hall [7]. In our simulation experiments (unreported) we have found that a sieve maximum likelihood estimator of the semiparametric Hawkes model with a nonparametric baseline intensity function and a parametric excitation kernel seems to converge to the true model parameter. The asymptotic properties of such semi-parametric estimators of the nonstationary Hawkes model need further investigation.

Another interesting extension of the SEPP model is to allow a time-varying excitation kernel, in addition to a time-varying baseline intensity. For example, when the kernel is exponential with time-vaying parameters, the intensity process of N^n may look like this,

$$\lambda^n_{t,\theta} = \nu\left(\frac{t}{n};\theta_\nu\right) + \int_0^{t-} \eta\left(\frac{s}{n};\theta_\eta\right)\beta\left(\frac{s}{n};\theta_\beta\right) e^{-(t-s)\beta\left(\frac{s}{n};\theta_\beta\right)}\,\mathrm{d}N^n_s.$$

Denote by t_i the time of occurrence of event i, the integral in the above expression can be written as

$$\sum_{i:t_i<t} \eta\left(\frac{t_i}{n};\theta_\eta\right)\beta\left(\frac{t_i}{n};\theta_\beta\right) e^{-(t-t_i)\beta\left(\frac{t_i}{n};\theta_\beta\right)}.$$

Under this formulation, the excitation effect of each event depends only on its time of occurrence. Adopting the approximation procedure outlined in this work, the intensity process of $N^{i,n}_t$ on each block is

$$\begin{aligned}\lambda^{i,n}_{t,\theta} = \nu\left(\frac{\tau^n_{i-1}+t}{n};\theta_\nu\right) &+ \int_0^{\tau^n_{i-1}-} \eta\left(\frac{s}{n};\theta_\eta\right)\beta\left(\frac{s}{n};\theta_\beta\right) e^{-(\tau^n_{i-1}+t-s)\beta\left(\frac{s}{n};\theta_\beta\right)}\,\mathrm{d}N^n_s \\ &+ \int_0^{t-} \eta\left(\frac{\tau^n_{i-1}+s}{n};\theta_\eta\right)\beta\left(\frac{\tau^n_{i-1}+s}{n};\theta_\beta\right) e^{-(t-s)\beta\left(\frac{\tau^n_{i-1}+s}{n};\theta_\beta\right)}\,\mathrm{d}N^{i,n}_s,\end{aligned}$$

and the intensity of the approximating process $N^{i,n,c}_t$, a SEPP with a constant baseline intensity and a non-time-varying exponential excitation kernel, is

$$\lambda^{i,n,c}_{t,\theta} = \nu\left(\frac{\tau^n_{i-1}}{n};\theta_\nu\right) + \int_0^{t-} \eta\left(\frac{\tau^n_{i-1}}{n};\theta_\eta\right)\beta\left(\frac{\tau^n_{i-1}}{n};\theta_\beta\right) e^{-(t-s)\beta\left(\frac{\tau^n_{i-1}}{n};\theta_\beta\right)}\,\mathrm{d}N^{i,n,c}_s.$$

Assuming there exist $\underline{\nu}$, $\bar{\nu}$, $\underline{\eta}$, $\bar{\eta}$, $\underline{\beta}$ and $\bar{\beta}$ such that $0 < \underline{\nu} \le \nu(t;\theta_\nu) \le \bar{\nu} < \infty$, $0 < \underline{\eta} \le \eta(t;\theta_\eta) \le \bar{\eta} < 1$ and $0 < \underline{\beta} \le \beta(t;\theta_\beta) \le \bar{\beta} < \infty$ for all $t \in [0,1]$, $\theta_\nu \in \Theta_\nu$, $\theta_\eta \in \Theta_\eta$ and $\theta_\beta \in \Theta_\beta$ (where Θ_ν, Θ_η and Θ_β are all compact), with moderate modifications in the technical arguments presented in this work and by imposing certain smoothness conditions on $\eta(\cdot;\cdot)$ and $\beta(\cdot;\cdot)$ (similar to [C2]), it should be possible to adopt the approximation procedure to derive similar asymptotic results for the time-varying excitation kernel case.

Supplementary Materials

In the online supplement, we introduce a few new notations, provide some preliminary results, all proofs of the Lemmas presented in the Appendix, and discuss the aforementioned approximation procedure in details.

Credit author statement

Tsz-Kit Jeffrey Kwan: Conceptualization, Software, Writing - Original Draft, Writing - Review & Editing
Feng Chen: Conceptualization, Writing - Review & Editing, Supervision
William T.M. Dunsmuir: Conceptualization, Writing - Review & Editing, Supervision

Acknowledgements

Kwan was supported by the Australian government research training program scholarship and the faculty top up scholarships from the School of Mathematics and Statistics, University of New South Wales Sydney. Chen was partly supported by a UNSW SFRG grant. Chen would like to acknowledge the helpful discussions with Nakahiro Yoshida, Simon Clinet and Yoann Potiron on inference for Hawkes processes. We would also like to thank the anonymous Executive Editors and an anonymous reviewer for helpful comments and advice which improved the content and presentation of this work.

Declarations of interest

None.

7. Appendix: Technical proofs

In the Appendix, we present the technical proofs of the Theorems in Sections 3 and 4, and we resume the notations therein.

Proof of Theorem 3.1. Recall that it suffices to establish that there exists some deterministic function $\tilde{\pi}$ such that

$$\sup_{\theta\in\Theta}\left|\mathcal{L}^n(\theta)-\mathcal{L}^n(\theta^0)-\tilde{\pi}(\theta)\right|\xrightarrow{\mathbb{P}}0, \tag{14}$$

and that θ^0 exists as a unique maximum for $\tilde{\pi}$. Recall also that a_n is a monotone sequence of positive integers $\to\infty$ such that $a_n=o(n)$. The time interval $[0,n]$ is split into a_n equal-lengthed blocks, each with length $b_n=n/a_n$. Denote by $\tau^n_{i-1}=(i-1)b_n$ the starting time of the i-th block. Then, for any $i\in\{1,\cdots,a_n\}$, consider the time change operator $\zeta^{i,n}:t\mapsto t+\tau^n_{i-1}$ and the restriction of N^n on $((i-1)b_n,ib_n]$ to define the SEPP $N^{i,n}$ associated with the ith block by

$$N^{i,n}_t=N^n_{\zeta^{i,n}_t}-N^n_{\tau^n_{i-1}},\quad t\in[0,b_n].$$

Let $\mathcal{F}^{i,n}=(\mathcal{F}^{i,n}_t)_{t\in[0,b_n]}$ with $\mathcal{F}^{i,n}_t=\mathcal{F}^n_{\zeta^{i,n}_t}$, the intensity of $N^{i,n}$ relative to $\mathcal{F}^{i,n}$ is

$$\lambda^{i,n}_{t,\theta}=\nu\left(\frac{\tau^n_{i-1}+t}{n};\theta_1\right)+\int_0^{\tau^n_{i-1}-}g(\tau^n_{i-1}+t-s;\theta_2)\,\mathrm{d}N^n_s+\int_0^{t-}g(t-s;\theta_2)\,\mathrm{d}N^{i,n}_s.$$

The second term of $\lambda^{i,n}_{t,\theta}$ describes the pre-excitation effects prior to ith block evaluated at $\tau^n_{i-1}+t$. We can also express the original SEPP as

$$N^n_t=\sum_{i=1}^{\lfloor t/b_n\rfloor}N^{i,n}_{b_n}+N^{\lfloor t/b_n\rfloor+1,n}_{t-\lfloor t/b_n\rfloor b_n}.$$

Denote by N^P a unit rate Poisson process on $\mathbb{R}^2$ and by $\mathcal{F}^{N^P}_t$ the σ-field generated by the random variables $N^P(C)$ for $C\in\mathcal{B}([0,t]\times\mathbb{R})$. Let $\mathcal{F}^{N^P}=(\mathcal{F}^{N^P}_t)_{t\in[0,n]}$ be a filtration induced by N^P on $[0,n]\times\mathbb{R}$ and set $\mathcal{F}^n_t=\mathcal{F}^{N^P}_t$. By Poisson embedding, N^n has the representation

$$N^n_t=\int_{[0,t]\times\mathbb{R}}\mathbb{1}_{\{0\le z\le\lambda^n_{s,\theta^0}\}}N^P(\mathrm{d}s\times\mathrm{d}z).$$

Similarly, $N^{i,n}$ can be represented as integrals with respect to a sequence of Poisson processes $N^{P,i,n}$ of unit intensity on $\mathbb{R}^2$ as

$$N^{i,n}_t=\int\int_{[0,t]\times\mathbb{R}}\mathbb{1}_{\{0\le z\le\lambda^{i,n}_{s,\theta^0}\}}N^{P,i,n}(\mathrm{d}s\times\mathrm{d}z)$$

where $N^{P,i,n}$ is the time changed version of the initial bivariate unit rate Poisson process N^P defined by $N^{P,i,n}(A\times B)=N^P(\zeta^{i,n}(A)\times B)$ for $A\in\mathcal{B}([0,b_n])$ and $B\in\mathcal{B}(\mathbb{R})$.

Note that $N^{i,n}_{b_n}$ counts the number of events in the i-th block of N^n. If we split the integral in $\mathcal{L}^n(\theta)-\mathcal{L}^n(\theta^0)$ into the a_n blocks of equal length, we can write

$$\mathcal{L}^n(\theta)-\mathcal{L}^n(\theta^0)=M^n(\theta,\theta^0)+\frac{1}{a_n}\sum_{i=1}^{a_n}\left(L^{i,n}(\theta,\theta^0)-L^{i,n}(\theta^0,\theta^0)\right),$$

where

$$M^n(\theta,\theta^0)=\frac{1}{n}\int_0^n\log\left(\frac{\lambda^n_{t,\theta}}{\lambda^n_{t,\theta^0}}\right)\mathrm{d}\bar{N}^n_t,$$

with $\bar{N}_t^n = N_t^n - \int_0^t \lambda_{s,\theta^0}^n \, \mathrm{d}s$ denoting the compensated counting process, and

$$L^{i,n}(\theta, \theta^0) = \frac{1}{b_n} \int_0^{b_n} \left(\log(\lambda_{t,\theta}^{i,n}) \lambda_{t,\theta^0}^{i,n} - \lambda_{t,\theta}^{i,n} \right) \mathrm{d}t.$$

From Lemma S.1 in the online supplement S1 and the proof of Lemma 3.10 of the work by [9], we deduce that $M^n(\theta, \theta^0)$ converges to 0 in $\mathbb{L}^p$ uniformly in θ. Therefore, it suffices to show the following instead of (14):

$$\sup_{\theta \in \Theta} \left| \frac{1}{a_n} \sum_{i=1}^{a_n} \left(L^{i,n}(\theta, \theta^0) - L^{i,n}(\theta^0, \theta^0) \right) - \tilde{\pi}(\theta) \right| \xrightarrow{\mathbb{P}} 0. \tag{15}$$

Next associate each $N^{i,n}$ with another SEPP $N^{i,n,c}$ with a constant baseline intensity equal to $\nu(\tau_{i-1}^n / n; \theta_1)$, the value of the baseline intensity at the start of the ith block. Let $\mathcal{F}^{i,n,c} = (\mathcal{F}_t^{i,n,c})_{t \in \mathbb{R}_+}$ with $\mathcal{F}_t^{i,n,c} = \mathcal{F}_{\zeta_t^{i,n}}^n$, the intensity of $N^{i,n,c}$ relative to $\mathcal{F}^{i,n,c}$ is

$$\lambda_{t,\theta}^{i,n,c} = \nu\left(\frac{\tau_{i-1}^n}{n}; \theta_1 \right) + \int_0^{t-} g(t - s; \theta_2) \, \mathrm{d}N_s^{i,n,c}$$

and by Poisson embedding again, we have

$$N_t^{i,n,c} = \int \int_{[0,t] \times \mathbb{R}} \mathbb{1}_{\{0 \le z \le \lambda_{s,\theta^0}^{i,n,c}\}} N^{P,i,n}(\mathrm{d}s \times \mathrm{d}z).$$

Define similarly

$$L^{i,n,c}(\theta, \theta^0) = \frac{1}{b_n} \int_0^{b_n} \left(\log(\lambda_{t,\theta}^{i,n,c}) \lambda_{t,\theta^0}^{i,n,c} - \lambda_{t,\theta}^{i,n,c} \right) \mathrm{d}t.$$

If

$$\sup_{\theta \in \Theta} \frac{1}{a_n} \sum_{i=1}^{a_n} \left| L^{i,n}(\theta, \theta^0) - L^{i,n}(\theta^0, \theta^0) - \left(L^{i,n,c}(\theta, \theta^0) - L^{i,n,c}(\theta^0, \theta^0) \right) \right| \xrightarrow{\mathbb{P}} 0, \tag{16}$$

$$\sup_{\theta \in \Theta} \frac{1}{a_n} \sum_{i=1}^{a_n} \left| L^{i,n,c}(\theta, \theta^0) - L^{i,n,c}(\theta^0, \theta^0) - \pi\left(\nu\left(\frac{\tau_{i-1}^n}{n}; \theta_1 \right), \theta_2, \theta^0 \right) \right| \xrightarrow{\mathbb{P}} 0, \tag{17}$$

$$\sup_{\theta \in \Theta} \left| \frac{1}{a_n} \sum_{i=1}^{a_n} \pi\left(\nu\left(\frac{\tau_{i-1}^n}{n}; \theta_1 \right), \theta_2, \theta^0 \right) - \tilde{\pi}(\theta) \right| \xrightarrow{\mathbb{P}} 0, \tag{18}$$

for some function $\pi : [\underline{\nu}, \bar{\nu}] \times \Theta_2 \times \Theta \mapsto \mathbb{R}$, then (15) holds. Note that (16) holds if

$$\sup_{\theta \in \Theta} \sup_{1 \le i \le a_n} \left| L^{i,n}(\theta, \theta^0) - L^{i,n,c}(\theta, \theta^0) \right| \xrightarrow{\mathbb{P}} 0.$$

Since a_n increases with n and the above supremum is taken with respect to i, we will define two auxiliary SEPPs, $N^{x,n}$ and N^x, to avoid the complications with the convergence of a triangular array. For any x and n, denote another time change operator $\zeta^{x,n} : t \mapsto nx + t$. Let $\mathcal{F}^{x,n} = (\mathcal{F}_t^{x,n})_{t \in [0, b_n]}$ and $\mathcal{F}^x = (\mathcal{F}_t^x)_{t \in \mathbb{R}_+}$ with $\mathcal{F}_t^{x,n} = \mathcal{F}_t^x = \mathcal{F}_{\zeta_t^{x,n}}^n$, the intensities of $N^{x,n}$ and N^x relative to $\mathcal{F}^{x,n}$ and $\mathcal{F}^x$ are respectively

$$\lambda_{t,\theta}^{x,n} = \nu\left(x + \frac{t}{n}; \theta_1 \right) + \int_0^{xn-} g(xn + t - s; \theta_2) \, \mathrm{d}N_s^n + \int_0^{t-} g(t - s; \theta_2) \, \mathrm{d}N_s^{x,n}$$

where $(x, t) \in [0, 1] \times [0, b_n]$ such that $n - b_n \ge nx$, and

$$\lambda_{t,\theta}^x = \nu(x; \theta_1) + \int_0^{t-} g(t - s; \theta_2) \, \mathrm{d}N_s^x$$

where $(x,t) \in [0,1]\times[0,\infty)$. Denote by $N^{P,x,n}$ the time changed version of N^P defined by $N^{P,x,n}(A\times B) = N^P(\zeta^{x,n}(A)\times B)$ for A and B any two Borel sets of $\mathbb{R}$. By Poisson embedding, $N^{x,n}$ and N^x have the following representations respectively:

$$N_t^{x,n} = \int_{[0,t]\times\mathbb{R}} \mathbb{1}_{\{0\le z\le \lambda^{x,n}_{s,\theta^0}\}} N^{P,x,n}(\,\mathrm{d}s\times\,\mathrm{d}z)$$
$$N_t^{x} = \int_{[0,t]\times\mathbb{R}} \mathbb{1}_{\{0\le z\le \lambda^{x}_{s,\theta^0}\}} N^{P,x,n}(\,\mathrm{d}s\times\,\mathrm{d}z).$$

Define similarly

$$L^{x,n}(\theta,\theta^0) = \frac{1}{b_n}\int_0^{b_n}\left(\log\left(\lambda^{x,n}_{t,\theta}\right)\lambda^{x,n}_{t,\theta^0} - \lambda^{x,n}_{t,\theta}\right)\mathrm{d}t$$

and

$$L^{x,n,c}(\theta,\theta^0) = \frac{1}{b_n}\int_0^{b_n}\left(\log\left(\lambda^{x}_{t,\theta}\right)\lambda^{x}_{t,\theta^0} - \lambda^{x}_{t,\theta}\right)\mathrm{d}t.$$

Then

$$\sup_{\theta\in\Theta}\sup_{x\in[0,1]}\left|L^{x,n}(\theta,\theta^0) - L^{x,n,c}(\theta,\theta^0)\right| \xrightarrow{\mathbb{P}} 0 \tag{19}$$

is sufficient to assert (16). Similarly, it suffices to show that

$$\sup_{\theta\in\Theta}\sup_{x\in[0,1]}\left|L^{x,n,c}(\theta,\theta^0) - L^{x,n,c}(\theta^0,\theta^0) - \pi\left(\nu\left(x;\theta_1\right),\theta_2,\theta^0\right)\right| \xrightarrow{\mathbb{P}} 0 \tag{20}$$

holds instead of (17). Note that (19) and (20) hold as a consequence of Lemmas 7.1 and 7.2 below. To show θ^0 exists as a unique maximizer of $\tilde{\pi}$, we will examine the derivatives of $\tilde{\pi}$ with respect to θ. Taking this into consideration, the ergodic property stated in Lemma 7.2 includes derivatives of $\lambda^x_{t,\theta}$ with respect to θ as well.

Lemma 7.1. *Under [C1], [C2] and [C4], (19) holds.*

Let us introduce some notations for the next lemma. Let $E_1 = \mathbb{R}_+\times\mathbb{R}^d\times\mathbb{R}^{d\times d}\times\mathbb{R}_+$, $E_2 = [\underline{\nu},\bar{\nu}]\times\mathbb{R}^{d_1}\times\mathbb{R}^{d_1\times d_1}\times\Theta_2\times\Theta$. Write $C_\uparrow(E_1,\mathbb{R})$ the set of functions that maps from E_1 to $\mathbb{R}$ that are continuous and of polynomial growth as in [9]. For any $x\in[0,1]$, N^x coincides with the definition of the exponential Hawkes model in [9]. Let $Y_t^x(\theta,\theta^0) = \left(\lambda^x_{t,\theta}, \partial_\theta\lambda^x_{t,\theta}, \partial^{\otimes 2}_\theta\lambda^x_{t,\theta}, \lambda^x_{t,\theta^0}\right)$ and $Z^x(\theta,\theta^0) = \left(\nu\left(x;\theta_1\right), \partial_{\theta_1}\nu(x;\theta_1), \partial^{\otimes 2}_{\theta_1}\nu(x;\theta_1), \theta_2, \theta^0\right)$.

Lemma 7.2. *Under [C1], [C2] and [C4], for any $\psi\in C_\uparrow(E_1,\mathbb{R})$, $n\ge 1$ and $x\in[0,1]$, there exists a mapping $\gamma: E_2\times C_\uparrow(E_1,\mathbb{R})\mapsto\mathbb{R}$ such that as $n\to\infty$,*

$$\sup_{x\in[0,1]}\sup_{\theta\in\Theta}\left|\frac{1}{b_n}\int_0^{b_n}\psi\left(Y_t^x(\theta,\theta^0)\right)\mathrm{d}t - \gamma\left(Z^x(\theta,\theta^0);\psi\right)\right| \xrightarrow{\mathbb{P}} 0.$$

Note that (18) ensures the ergodic limit from each block to be "stitched" together such that $\tilde{\pi}(\theta)$ exists as the asymptotic ergodic limit of $L^n(\theta,\theta^0) - L^n(\theta^0,\theta^0)$. The following lemma ensures (18) holds.

Lemma 7.3. *Under [C1], [C2] and [C4], (18) holds.*

The following lemma guarantees that θ^0 exists as a unique maximizer of $\tilde{\pi}$.

Lemma 7.4. *Under [C1] - [C4], θ^0 exists as a unique maximizer of $\tilde{\pi}$.*

To complete the proof, it remains to show Lemmas 7.1 - 7.4, the proofs of which are given in the online supplement S2. ■

Up until now, we applied a time-change argument to $\tilde{N}^n$ from $[0,1]$ to $[0,n]$ and established consistency of $\hat{\theta}^n$ under N^n. For the proof of Theorem 3.2, we will return to our original setup under $\tilde{N}^n$.

Proof of Theorem 3.2. The proof is based on the Mean Value Theorem (MVT) and the martingale central limit theorem [see 18, Theorem 5.1.1]. By the MVT, we have

$$0 = \sqrt{n}\tilde{\mathcal{S}}^n(\hat{\theta}^n) = \sqrt{n}\tilde{\mathcal{S}}^n(\theta^0) - \tilde{\mathcal{I}}^n(\theta^*)\sqrt{n}(\hat{\theta}^n - \theta^0),$$

where θ^* lies on the line segment joining $\hat{\theta}^n$ and θ^0. Rearranging, we get

$$\sqrt{n}(\hat{\theta}^n - \theta^0) = \tilde{\mathcal{I}}^n(\theta^*)^{-1}\sqrt{n}\tilde{\mathcal{S}}^n(\theta^0).$$

If we can show that

$$\sqrt{n}\tilde{\mathcal{S}}^n(\theta^0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma(\theta^0)),$$
$$\tilde{\mathcal{I}}^n(\theta^*) \xrightarrow{\mathbb{P}} \Gamma(\theta^0)$$

where $\Gamma(\theta^0)$ is invertible, then, by Slutsky's theorem, the proof is complete. To complete the proof, it remains to show the following two lemmas, the proofs of which are relegated to the online supplement S2.

Lemma 7.5. *Under [C1], [C2] and [C4],* $\sqrt{n}\tilde{\mathcal{S}}^n(\theta^0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma(\theta^0))$.

Lemma 7.6. *Under [C1] - [C4],* $\tilde{\mathcal{I}}^n(\theta^*) \xrightarrow{\mathbb{P}} \Gamma(\theta^0)$ *and* $\Gamma(\theta^0)$ *is invertible.* ■

Proof of Theorem 4.1. For the proof of Theorem 4.1, we require Lemmas 7.7 and 7.8 below, the proofs of which are given in the online supplement S3. Denote by $\theta^{0,*} = (\theta^0_{1,1}, \theta^{0,*}_{1,2}, \theta^0_2)$ the true parameter value assuming H_0 is true, $\hat{\theta}^{0,n} = (\hat{\theta}^{0,n}_{1,1}, \theta^{0,*}_{1,2}, \hat{\theta}^{0,n}_2)$ the constrained MLE under H_0 and $\hat{\theta}^n = (\hat{\theta}^n_{1,1}, \hat{\theta}^n_{1,2}, \hat{\theta}^n_2)$ the unconstrained MLE.

Lemma 7.7. *In addition to [C1] - [C4], [S1] and [S2], suppose* H_0 *is true,*

$$\sqrt{n}\partial_{\hat{\theta}^{0,n}_{1,2}}\mathcal{L}^n(\hat{\theta}^{0,n}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (\Gamma(\theta^{0,*})^{-1}_{[2:d_1,2:d_1]})^{-1}).$$

Lemma 7.8. *In addition to [C1] - [C4], [S1] and [S2], suppose* H_0 *is true, there exists a positive-definite matrix valued function* $\Gamma(\theta^{0,*})$ *such that*

$$\frac{1}{n}\int_0^n \left(\frac{\partial_{\hat{\theta}^{0,n}}\lambda^n_{t,\hat{\theta}^{0,n}}}{\lambda^n_{t,\hat{\theta}^{0,n}}}\right)^{\otimes 2} \mathrm{d}N^n_t \xrightarrow{\mathbb{P}} \Gamma(\theta^{0,*}). \tag{21}$$

Furthermore $\Gamma(\theta^{0,*})_{[2:d_1,2:d_1]}$ *coincides with the variance covariance matrix of the asymptotic distribution given in Lemma 7.7.*

Lemma 7.7 shows that under H_0, the asymptotic distribution of $\sqrt{n}\partial_{\hat{\theta}^{0,n}_{1,2}}\mathcal{L}^n(\hat{\theta}^{0,n})$ is given by $\mathcal{N}(0, (\Gamma(\theta^{0,*})^{-1}_{[2:d_1,2:d_1]})^{-1})$. We can deduce from Lemma 7.8 that the middle term of (11) converges to $\Gamma(\theta^{0,*})^{-1}_{[2:d_1,2:d_1]}$, which coincides with the variance covariance matrix of the limiting distribution given in Lemma 7.7, and is positive-definite. Since the chi-square distribution is the distribution of the sum of squared standard normal random variables with the degrees of freedom equal to the number of terms being summed, we deduce from Lemmas 7.7 and 7.8 that $\hat{\mathcal{T}}^n$ converges in distribution to $\chi^2_{d_1-1}$. ■

Proof of Theorem 4.2. In this proof, we require Lemmas 7.9 - 7.11 from below, the proofs of which are detailed in the online supplement S3. Denote by $\theta^{a,n} = (\theta^0_{1,1}, \theta^{a,n}_{1,2}, \theta^0_2)$ the true parameter value assuming H^n_a is true.

Lemma 7.9. *In addition to [C1] - [C4], [S1] and [S2], suppose* H^n_a *is true,* $\hat{\theta}^{0,n} \xrightarrow{\mathbb{P}} \theta^{0,*}$ *and* $\hat{\theta}^n \xrightarrow{\mathbb{P}} \theta^{0,*}$.

Lemma 7.10. *In addition to [C1] - [C4], [S1] and [S2], suppose* H^n_a *is true, there exists some finite constant* $k \in \mathbb{R}^{d_1-1}$ *such that*

$$\sqrt{n}\partial_{\hat{\theta}^{0,n}_{1,2}}\mathcal{L}^n(\hat{\theta}^{0,n}) \xrightarrow{\mathcal{D}} \mathcal{N}\Big((\Gamma(\theta^{0,*})^{-1}_{[2:d_1,2:d_1]})^{-1}k, (\Gamma(\theta^{0,*})^{-1}_{[2:d_1,2:d_1]})^{-1}\Big).$$

Lemma 7.11. *In addition to [C1] - [C4], [S1] and [S2], suppose H_a^n is true, there exists a positive-definite matrix valued function $\Gamma(\theta^{0,*}) \in \mathbb{R}^{d\times d}$ such that*

$$\frac{1}{n}\int_0^n \left(\frac{\partial_{\hat{\theta}^{0,n}}\lambda^n_{t,\hat{\theta}^{0,n}}}{\lambda^n_{t,\hat{\theta}^{0,n}}}\right)^{\otimes 2} \mathrm{d}N_t^n \xrightarrow{\mathbb{P}} \Gamma(\theta^{0,*}). \tag{22}$$

Furthermore $\Gamma(\theta^{0,*})_{[2:d_1,2:d_1]}$ coincides with the variance covariance matrix of the asymptotic distribution given in Lemma 7.10.

Note that under H_a^n, the data generating parameter $\theta^{a,n}$ varies with n. Keeping in mind our goal to identify the asymptotic distribution of the score test statistic, similar to the proof for Theorem 3.1, we first show that under H_a^n, both $\hat{\theta}^{0,n}$ and $\hat{\theta}^n$ converge to $\theta^{0,*}$ in probability. Then we follow a similar structure as in the proof of Theorem 4.1 to establish the asymptotic distribution of $\hat{\mathcal{T}}^n$ via the above lemmas. ■

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

[2] Pierre Brémaud and Laurent Massoulié. Imbedded construction of stationary sequences and point processes with a random memory. *Queueing Systems*, 17(1):213–234, 1994. doi: 10.1007/BF01158695. URL https://doi.org/10.1007/BF01158695.

[3] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[4] Feng Chen. *ibs: Integral of B-Spline Functions*, 2018. URL https://CRAN.R-project.org/package=ibs. R package version 1.4.

[5] Feng Chen. *IHSEP: Inhomogeneous Self-Exciting Process*, 2021. URL https://cran.r-project.org/package=IHSEP. R package version 0.2.

[6] Feng Chen and Peter Hall. Inference for a Nonstationary Self-Exciting Point Process with an Application in Ultra-High Frequency Financial Data Modeling. *Journal of Applied Probability*, 50(4):1006–1024, 2013.

[7] Feng Chen and Peter Hall. Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach. *Journal of Computational and Graphical Statistics*, 25(1):209–224, 2016.

[8] Simon Clinet and Yoann Potiron. Statistical inference for the doubly stochastic self-exciting process. *Bernoulli*, 24(4B):3469–3493, 2018.

[9] Simon Clinet and Nakahiro Yoshida. Statistical inference for ergodic point processes and application to limit order book. *Stochastic Processes and their Applications*, 127(6):1800–1839, 2017.

[10] Simon Clinet, William T. M. Dunsmuir, Gareth W. Peters, and Kylie-Anne Richards. Asymptotic distribution of the score test for detecting marks in Hawkes processes. *Statistical Inference for Stochastic Processes*, 2021. Forthcoming. DOI:https://doi.org/10.1007/s11203-021-09245-5.

[11] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[12] José Da Fonseca and Riadh Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

[13] Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. Springer, New York, 2003.

[14] D.J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer, New York, 2008.

[15] Randal Douc, Paul Doukhan, and Eric Moulines. Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications*, 123(7):2620–2647, 2013.

[16] Robert F Engle and Jeffrey R Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162, 1998.

[17] Juan V Escobar. A hawkes process model for the propagation of covid-19: Simple analytical results. *EPL (Europhysics Letters)*, 131(6):68005, 2020.

[18] Thomas R Fleming and David P Harrington. *Counting Processes and Survival Analysis*, volume 169. John Wiley & Sons, 2011.

[19] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[20] AG Hawkes. Spectra of some mutually exciting point processes with associated variables. *Stochastic point processes*, pages 261–271, 1972.

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

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

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

[24] Minkyoung Kim, Dean Paini, and Raja Jurdak. Modeling stochastic processes in disease spread across a heterogeneous social system. *Proceedings of the National Academy of Sciences*, 116(2):401–406, 2019.

[25] Hongyuan Mei and Jason M Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

[26] Yoshiko Ogata. The Asymptotic Behaviour of Maximum Likelihood Estimators for Stationary Point Processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.
[27] Yosihiko Ogata. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.
[28] Tohru Ozaki. Maximum Likelihood Estimation of Hawkes' Self-Exciting Point Processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.
[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.
[30] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. Inference of functional connectivity in Neurosciences via Hawkes processes. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 317–320. IEEE, 2013.
[31] François Roueff, Rainer Von Sachs, and Laure Sansonnet. Locally stationary Hawkes processes. *Stochastic Processes and their Applications*, 126(6):1710–1743, 2016.
[32] François Roueff and Rainer von Sachs. Time-frequency analysis of locally stationary Hawkes processes. *Bernoulli*, 25(2):1355–1385, 05 2019. doi: 10.3150/18-BEJ1023. URL https://doi.org/10.3150/18-BEJ1023.
[33] Tom Stindl and Feng Chen. Likelihood based inference for the multivariate renewal Hawkes process. *Computational Statistics & Data Analysis*, 123:131–145, 2018.
[34] Ioane Muni Toke and Fabrizio Pomponio. Modelling Trades-Through in a Limit Order Book Using Hawkes Processes. *Economics*, 6 (1):20120022, 2012. doi: doi:10.5018/economics-ejournal.ja.2012-22. URL https://doi.org/10.5018/economics-ejournal.ja.2012-22.
[35] Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000.
[36] David Vere-Jones. Forecasting Earthquakes and Earthquake Risk. *International Journal of Forecasting*, 11(4):503–538, 1995.
[37] Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-Attentive Hawkes Process. In *International Conference on Machine Learning*, pages 11183–11193. PMLR, 2020.