Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Mutually Exciting Point Processes with Latency

Yoann Potiron & Vladimir Volkov

To cite this article: Yoann Potiron & Vladimir Volkov (17 Jul 2025): Mutually Exciting Point Processes with Latency, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2516209

To link to this article: https://doi.org/10.1080/01621459.2025.2516209

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2025.2516209

Taylor & Francis
Taylor & Francis Group

Mutually Exciting Point Processes with Latency

Yoann Potiron[a] and Vladimir Volkov[b,c]

[a]Faculty of Business and Commerce, Keio University, Tokyo, Japan; [b]School of Business and Economics, University of Tasmania, Hobart, Australia; [c]HSE University, Moscow, Russia

ABSTRACT
A novel statistical approach to estimating latency, defined as the time it takes to learn about an event and generate response to this event, is proposed. Our approach only requires a multidimensional point process describing event times, which circumvents the use of more detailed datasets which may not even be available. We consider the class of parametric Hawkes models capturing clustering effects and define latency as a known function of kernel parameters, typically the mode of kernel function. Since latency is not well-defined when the kernel is exponential, we consider maximum likelihood estimation in the mixture of generalized gamma kernels case and derive the feasible central limit theory with in-fill asymptotics. As a byproduct, central limit theory for a latency estimator and related tests are provided. Our numerical study corroborates the theory. An empirical application on high frequency data transactions from the New York Stock Exchange and Toronto Stock Exchange shows that latency estimates for the United States and Canadian stock exchanges vary between 1 and 6 milliseconds from 2020 to 2021. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received May 2023
Accepted May 2025

KEYWORDS
In-fill asymptotics; Latency; Maximum likelihood estimation; Mixture of generalized gamma kernels; Parametric Hawkes processes

1. Introduction

Over the last decade, financial markets have undergone revolutionary institutional and technological changes. Massive increases in computer power have led to algorithmic trading, explosions in sub-second orders, and large increases in trading volume. These changes transformed the financial markets and made latency an inherent part of investment process. Low latency, or simply latency, can be broadly defined following Hasbrouck and Saar (2013) as the time it takes to learn about an event and generate response to this event.

Although the term latency is widely used in finance, it is also related to delay, a more common term in statistics. In fact, delay is present in datasets related to seismology, insurance, criminology, sociology and medicine as in for example, Harris (1990). Indeed, just like a time lag before a trading event is revealed to market participants, there is also a time lag before a tweet post becomes available on X, or a registration of medical incidents. Despite the fact that we mainly use the term latency and our empirical application focuses on finance, a statistician should keep this parallel in mind when reading the article.

In the finance literature an approach to obtaining latency is heavily dependent on datasets that are not available in many cases. Hasbrouck and Saar (2013) propose a widely used low latency measure based on strategic runs representing series of submissions, cancellations, and executions that are linked by direction, size, and timing, and which are likely to arise from a single algorithm. However, as the proposed approach requires detailed information about the timing and other characteristics of cancellations—which is not available for many markets—a unified statistical framework for accurate estimation of latency is a missing link in the literature.

An alternative method, proposed in this article, is to make use of a statistical model relying solely on multidimensional event times. These time series are normally available to a statistician, making our approach widely suitable for applications. Building upon the stylized fact that arrival times are not deterministic, an obvious choice is to use tractable Poisson processes, in which inter-arrival times are IID. Yet, Poisson processes are not well suited for modeling the arrival times as the empirical literature on inter-arrival durations points out that trades tend to cluster together over time. Accordingly, the class of autoregressive conditional duration (ACD) models is introduced in Engle and Russell (1998). ACD models are closely aligned with GARCH models and there are many multidimensional extensions of GARCH models. Most of these extensions are also applicable to ACD models, and copulas can be also used to provide appropriate solutions to such extensions, see Heinen and Rengifo (2007), Koopman et al. (2018), Barra, Borowska, and Koopman (2018) and the references therein. However, these models are hard to generalize to a set-up with asynchronous times.

This motivates using a more suitable class of multidimensional models, the so-called mutually exciting processes such that the occurrence of any event fuels the probability of the next events occurring. The d-dimensional intensity, which can be interpreted as the instantaneous expected number of events, is defined as

CONTACT Yoann Potiron potiron@fbc.keio.ac.jp Faculty of Business and Commerce, Keio University, 2-15-45 Mita, Minato-ku, 108-8345, Tokyo, Japan.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

$$\lambda(t) = \nu + \int_0^t h(t-s)\, dN_s, \qquad (1)$$

where ν is a d-dimensional Poisson baseline, $h(t)$ is a $d \times d$-dimensional kernel matrix whose diagonal components $h^{(i,i)}$ are self-exciting terms for the related ith process and non-diagonal components $h^{(i,j)}$ are cross-exciting terms made by event from the jth to the ith process.

Considering a classical parametric specification of model (1), the novelty in our model is that we define latency as a known function of parameters, typically the time corresponding to the peak of the kernel function, that is the mode. We assume that the latency is a $d \times d$-dimensional matrix. We insist on the fact that latency is not well-defined when the kernel is exponential as the mode is always equal to 0 in that case. Thus, we introduce a mixture of generalized gamma kernels in which latency is well-defined.

In this article, we focus on in-fill asymptotics, that is when T is fixed, since it is well-known that latency changes empirically across different days, and this is also corroborated by the findings in our empirical application. In the absence of latency, there already exists successful attempts to accommodate for in-fill asymptotics in Hawkes processes. Chen and Hall (2013) allow for a nonrandom parametric time-varying baseline. Their in-fill asymptotic results are based on random observation times of order n within the fixed time interval $[0, T]$. A single boosting of the baseline, that is $\lambda(t) = \alpha_n \nu_t + \int_0^t h(t-s)\, dN_s$, is considered where $\alpha_n \to \infty$ is a scaling sequence when $n \to \infty$. Chen and Hall (2013) derive a central limit theorem (CLT) for MLE parameters related to the baseline and kernel. Clinet and Potiron (2018) consider stochastic time-varying baseline and kernel parameters in the exponential kernel case, and introduce a joint boosting of the baseline and the kernel, that is $\lambda(t) = n\nu_t + \int_0^t n a_s \exp(-n b_s(t-s))\, dN_s$ to derive CLTs on integrated baseline and parameters with local MLE. Kwan, Chen, and Dunsmuir (2023) revisit Chen and Hall (2013) in the exponential kernel case and with the same in-fill asymptotics as in Clinet and Potiron (2018), that is $\lambda(t) = n\nu_t + \int_0^t na \exp(-nb(t-s))\, dN_s$. Kwan (2023) considers the non-exponential kernel case and advocates the use of in-fill asymptotics for statistical inference to better match high-frequency data. An example of Aït-Sahalia and Jacod (2014) for financial applications also confirms the feasibility of in-fill asymptotics for financial data.

In the absence of latency, the parametric Hawkes literature provides the results of large-T asymptotics, that is when the horizon time $T \to \infty$. Maximum likelihood estimation (MLE) is employed in the seminal paper of Ogata (1978), which shows the asymptotic normality of the MLE for an ergodic stationary point process. However, the definition of ergodicity is vague in that paper and most of the papers on parametric Hawkes models (e.g., Bowsher (2007), Large (2007), and Cavaliere et al. (2023), Assumption 1(b) and Remark 2.1) make this ergodicity assumption and point out this assumption is satisfied for Hawkes processes, whereas in fact this is hard to establish. As far as we know, there are only two results in the literature showing rigorously the ergodicity of Hawkes processes. Clinet and Yoshida (2017) provide a general point process framework where they obtain MLE based CLTs in Theorem 3.11 (p. 1809) when assuming ergodicity of the couplet of intensity process and an intensity process derivative. Their general machinery is verified in the case of a Hawkes process with exponential kernel in Theorem 4.6 (p. 1821) by proving first that the couplet of intensity process and the intensity process derivative is mixing and stable, and then ergodicity is implied. With in-fill asymptotics, that is, when T is fixed based on random observation times of order n and by exploiting a joint boosting of the baseline and the kernel, Kwan (2023) considers the non-exponential kernel case but the author mentions that such a setup is challenging since the resulting intensity process is non-Markovian, thus, rendering standard techniques for asymptotic inference of Markov processes futile. Consequently, the author can only show the ergodicity for the intensity procecss itself but not for the couplet of intensity process and intensity process derivative, and only the consistency of the MLE in Theorem 3.4.3 (p. 73) of this article is shown.

These two results are useful, but Clinet and Yoshida (2017) (Theorem 4.6) is restricted to the exponential kernel case and Kwan (2023) (Theorem 3.4.3) only obtains consistency of the MLE. Thus, no feasible CLTs are available when the kernel is not exponential, and tests on latency cannot be directly inferred from these two results. In our Theorem 1, we consider MLE in the mixture of generalized gamma kernels case and derive the feasible CLT with in-fill asymptotics. Our proof strategy builds on the general machinery of Clinet and Yoshida (2017) by proving first that the couplet of intensity process and intensity process derivative is mixing and stable, and then ergodicity is implied. The novelty in the proofs is in establishing the couplet is mixing and ergodic (Propoosition C1 in Supplement C). Consistent estimators of the asymptotic variance and feasible normalized CLTs are also provided.

Our latency estimator is defined as the known function of MLE kernel parameters. As a byproduct of Theorem 4.6, we obtain a feasible CLT for our latency estimator (Proposition 2 and Corollary 3). We also construct three Wald tests for latency. We first develop a test for equality between a latency matrix value for one particular index and a fixed value which can in particular be used to test for the inexistence of latency. Our second test focuses on equality between two latency matrix values. Our third test considers multidimensional linear hypotheses on the latency vector. The limit theory of the three tests is established (Corollaries 4, 5 and 6). A complementary nonparametric approach with applications to Covid-19 pandemic in France is given in Gámiz et al. (2022) and Gámiz et al. (2023) for the time-varying case.

Our newly developed model contributes to the literature where a compensator of intensity is interpreted as business time or economic time, see Engle (2000). These so-called models of time deformation deal with the relevant time scale as "economic time" rather than "calendar time". Intuitively, economic time measures the arrival rate of new information that influences trading intensity. The joint analysis of transaction times and latency facilitates analyzing an impact of trading characteristics observed at ultra-high frequency on the complex interaction between financial markets.

In our empirical analysis we focus on two major stock exchanges, the NYSE (US) and the TSX (Canada). These exchanges have simultaneous trading sessions and are directly comparable due to similarity in trading environment, which creates opportunities for investors to exploit price inefficiencies across venues. Gagnon and Karolyi (2010) show that deviations from price parity are economically small but volatile and can reach large extremes. They report that price parity deviations relate positively to proxies for holding costs that can limit arbitrage. Moreover, the cross-exchange interactions may happen due to this excessive level of volatility.

In both markets latency varies between 1 and 6 milliseconds in 2020 and 2021 with the traders in Canada being overall faster. The presence of interaction between trades and quotes implies that information in trading (quoting) events can be absorbed with delay in response to quoting (trading) events, which is associated with a phenomenon of co-latency. Our findings indicate the existence of co-latency channel working in both directions between trades and quotes in the NYSE and the TSX. We observe a faster reaction of trading co-latency on the response to quotes in both exchanges. This corresponds to Hoffmann (2014) where an ability of fast traders to revise their quotes quickly after news arrivals helps reducing market risks.

The rest of this article is organized as follows. The model is introduced in Section 2. Estimation and tests are given in Section 3. The theory is developed in Section 4. Our empirical application is provided in Section 5. Our numerical study is carried over in Supplement A. Examples of kernels, that meet the assumptions of the proposed framework, are given in Supplement B. All proofs of the theory are shown in Supplement C. Additional empirical results belong to Supplement D.

2. A Parametric Hawkes Model Accommodating for Latency

We start this section from a literature review. Then, we recall definitions of a point process and a classic parametric Hawkes model, see Ogata (1978), Bowsher (2007), Large (2007), Embrechts, Liniger, and Lin (2011), Clinet and Yoshida (2017), Cavaliere et al. (2023) and Kwan (2023). See also Potiron (2025). Finally, we introduce a definition of latency, which is novel in the point processes literature.

2.1. Literature Review

Hawkes (1971b) and Hawkes (1971a) introduce a family of models for point processes with stochastic intensity called "self-exciting and mutually exciting point processes" such that the occurrence of any event fuels the probability of the next occurring events. Importantly, these papers provide the Bartlett spectrum and the corresponding covariance density function, useful tools for analyzing point process models. Details about these models are discussed in Liniger (2009) and applications in finance are shown in Hawkes (2018) with the references therein.

Over the last few decades Hawkes processes have been widely used in the context of seismology. The classical MLE for point processes is originally described in Rubin (1972), and applied to Hawkes processes in Vere-Jones (1978) and Ozaki (1979). Vere-Jones and Ozaki (1982) rely on the MLE and provide applications to earthquake data. Ogata (1978) shows the asymptotic normality of the MLE for an ergodic stationary point process with large-T asymptotics.

Applications of Hawkes processes in finance have been evolving over the last two decades. Bowsher (2007) considers a two-dimensional Hawkes process model of the timing of trades and mid-quote price changes. Chavez-Demoulin, Davison, and McNeil (2005) introduce a marked Hawkes process to model extreme returns. A ten-dimensional Hawkes process model is used by Large (2007). Embrechts, Liniger, and Lin (2011) consider the application of Hawkes processes with marks using MLE. Bacry et al. (2013) provide a CLT for the multidimensional Hawkes point process with large-T asymptotics. Aït-Sahalia, Laeven, and Pelizzon (2014) model self- and cross-excitation shocks in CDS markets for several European countries using a standard multidimensional Hawkes process with exponential kernels. Aït-Sahalia, Cacho-Diaz, and Laeven (2015) study Hawkes jump-diffusion processes in different stock markets and use a parametric moment-based estimation.

The most recent use of point processes is also widespread. Corradi, Distaso, and Fernandes (2020) develop a test for conditional independence in quadratic variation jump components. Ikefuji et al. (2022) analyze the impact of earthquake risk on real estate prices with the use of ETAS Hawkes-based model. A bootstrap approach for Hawkes and more general point processes is developed in Cavaliere et al. (2023). In Karim, Laeven, and Mandjes (2021), the authors provide an analysis of the probabilistic behavior of the couplet of point process and intensity process. Kernel-based estimation of intensity with in-fill asymptotics is presented in van Lieshout (2021). Bennedsen et al. (2023) develop likelihood-based methods for estimation of continuous-time integer-valued trawl processes. Clements et al. (2023) consider nonparametric estimation.

None of these strands of literature provide a formal definition of latency using a point process framework.

2.2. Parametric Hawkes Model

Let T stand for the horizon time. A d-dimensional point process

$$(N_t)_{0\leq t\leq T} := (N_t^{(1)},\dots,N_t^{(d)})_{0\leq t\leq T},$$

corresponds to the accumulated number of market events at time t. In other words, the ith component, which corresponds to the ith event type of the point process, is formally defined as

$$\begin{aligned} dN_t^{(i)} := N_t^{(i)} - N_{t-}^{(i)} &= 1 \text{ if there is an event at time t,}\\ &= 0 \text{ otherwise.}\end{aligned}$$

We will refer to $(T_1^{(i)},\dots,T_{N^{(i)}}^{(i)})$ for the event times. A point process is driven by its d-dimensional intensity $\lambda(t)$, which can be interpreted as the instantaneous expected number of events since

$$\lambda(t) = \lim_{u\to 0} \mathbb{E}\left[\frac{N_{t+u}-N_t}{u}|\mathcal{F}_t^N\right],$$

where $\mathcal{F}_t^N = \sigma\{N_s, 0\leq s\leq t\}$ is defined as the canonical filtration generated by N_t. For formal definitions related to the theory of point processes, see Daley and Vere-Jones (2003), Daley and Vere-Jones (2008), and more generally Jacod and Shiryaev (2013).

The parametric mutually exciting processes have a d-dimensional intensity defined as

$$\lambda(t) = \nu^* + \int_0^t h(t-s,\theta_{ker}^*)\,dN_s, \tag{2}$$

where ν^* is a d-dimensional Poisson baseline, $h(t, \theta^*_{ker})$ is a $d \times d$-dimensional kernel matrix whose diagonal components $h^{(i,i)}$ are self-exciting terms for the related ith process and non-diagonal components $h^{(i,j)}$ are cross-exciting terms made by event from the jth process to the ith process.

2.3. Latency

The latency is defined as a $d \times d$-dimensional matrix which is a known function of the kernel parameter θ^*_{ker}, that is we assume that

$$L = F(\theta^*_{ker}). \tag{3}$$

When $L^{(i,j)} > 0$, a latency between an event in process j and its impact on process i is introduced. Typically, we set F such that the latency $L^{(i,j)}$ is specified as the time it takes before reaching the pick, that is the mode, of the kernel $h^{(i,j)}(t, \theta^*_{ker})$. This definition of latency is in agreement with the finance literature, for example Hasbrouck and Saar (2013), defining it as the time it takes to learn and generate response to a trading event. An advantage of this definition is that latency can be characterized by parameters $\theta^{(i,j)}_{ker}$ associated with factors affecting latency. Such a structural approach permits identifying different aspects of latency. As we show in Section 3.2, sub-parameters of θ, that is D, are interpreted as delay measures. This component of latency D identifies the time of learning about a trading event, which is critical for financial applications. When $L^{(i,j)} \leq 0$, there is no latency between an event in process j and its impact on process i.

3. Estimation and Tests

In this section, we first introduce MLE for the parametric Hawkes model and discuss the in-fill asymptotics used for theoretical analysis. Then, we introduce a mixture of generalized gamma kernels in which latency is well-defined highlighting that latency is not well-defined when the kernel is exponential. Finally, we introduce latency estimation and tests related to latency.

3.1. MLE

We assume that a stochastic basis $\mathcal{B}_n = (\Omega, \mathcal{F}, \boldsymbol{F}_n, \mathbb{P})$ is given, where the filtration is defined as $\boldsymbol{F}_n = (\mathcal{F}_t)_{t \in [0,T]}$, where T is the horizon time, that is 1 trading interval. The filtration contains all the necessary information to the statistician. We implicitly assume that the defined quantities depend on n, but we do not write explicitly such a dependence when it is clear from the context. Furthermore, we also assume that all the stochastic processes defined in the following are $\boldsymbol{F}_n$-adapted processes. In particular, this implies that $\mathcal{F}^N_t \subset \mathcal{F}_t$ for any $t \in [0, T]$.

For any space S such that $0 \in S$, we define the space without zero as S^*. For inference purposes, we consider in-fill asymptotics with joint boosting of the baseline and the kernel. Relying on a parametric approach we assume the existence of an unknown true value $\theta^* = (\nu^*, \theta^*_{ker})$ such that for $i = 1, \dots, d$ we have that the ith component of the $\boldsymbol{F}$-intensity is equal to

$$\lambda^{(i)}(t, \theta^*) = n\nu^{*,(i)} + \sum_{j=1}^{d} \int_0^t nh^{(i,j)}(n(t-s), \theta^{*,(i,j)}_{ker}) dN^{(j)}_s. \tag{4}$$

Here, we assume the existence of the parameter space Θ, consisting of m parameters, and the true parameter $\theta^* \in \Theta$. We assume that $m - d \geq d^2$, since at least one parameter should be used in each component of the kernel matrix. We denote the set of baseline parameters as Θ_ν, and the set of kernel parameters as Θ_h. By definition, we have $\Theta = (\Theta_\nu, \Theta_h)$. In (4), in-fill asymptotics are based on random observation times of order n within the time interval $[0, T]$ for a finite horizon time T. Kwan (2023) extends the asymptotic analysis of Clinet and Potiron (2018) and Kwan, Chen, and Dunsmuir (2023), also based on joint boosting, by not imposing an exponential kernel. Our case is different from in-fill asymptotics of Chen and Hall (2013) who consider no boosting of the kernel. See also Potiron et al. (2025b) and Potiron et al. (2025a). We rely on the log likelihood process (see Ogata 1978 and Daley and Vere-Jones 2003)

$$l_T(\theta) = \sum_{i=1}^{d} \int_0^T \log(\lambda^{(i)}(t, \theta)) dN^{(i)}_t - \sum_{i=1}^{d} \int_0^T \lambda^{(i)}(t, \theta) dt,$$

that is the MLE is defined as $\widehat{\theta}_T \in \operatorname{argmax}_{\theta \in \Theta} l_T(\theta)$.

3.2. Mixture of Generalized Gamma Kernels

For any $i = 1, \dots, d$ and $j = 1, \dots, d$ the mixture of generalized gamma kernels is defined as

$$\begin{aligned} &h^{(i,j)}(t, \theta^{(i,j)}_{ker}) \\ &\quad = \sum_{k=1}^{K^{(i,j)}} \alpha^{(i,j)}_k \frac{p^{(i,j)}_k t^{(D^{(i,j)}_k - 1)} \exp(-(t/\beta^{(i,j)}_k)^{p^{(i,j)}_k})}{(\beta^{(i,j)}_k)^{D^{(i,j)}_k} \gamma(D^{(i,j)}_k / p^{(i,j)}_k)}, \end{aligned} \tag{5}$$

in which $\gamma(\cdot)$ is the gamma function, $\alpha^{(i,j)}_k \in \mathbb{R}^*_+$ is the size of the jump, $\beta^{(i,j)}_k \in \mathbb{R}^*_+$ is the scale parameter, $D^{(i,j)}_k \in \mathbb{R}^*_+$ and $p^{(i,j)}_k \in \mathbb{R}^*_+$ are shape parameters. In (5) the number of terms in the sum corresponding to the cross excitation between the ith and the jth market $K^{(i,j)}$ is fixed by the statistician, so they are not parameters to be estimated. We assume that the parameter related to the kernel is of the form

$$\begin{aligned} \theta_{ker} &= (\theta^{(i,j)}_{ker})_{1 \leq i,j \leq d} = (\theta^{(1,1)}_{ker}, \theta^{(1,2)}_{ker}, \dots, \theta^{(d,d-1)}_{ker}, \theta^{(d,d)}_{ker}) \\ \theta^{(i,j)}_{ker} &= (\alpha^{(i,j)}, \beta^{(i,j)}, D^{(i,j)}, p^{(i,j)}) \in (\mathbb{R}^*_+)^{K^{(i,j)}} \times (\mathbb{R}^*_+)^{K^{(i,j)}} \\ &\quad \times (\mathbb{R}^*_+)^{K^{(i,j)}} \times (\mathbb{R}^*_+)^{K^{(i,j)}}. \end{aligned} \tag{6}$$

For simplicity of exposition, we assume that each term in the sum of (5) is generalized gamma kernel. However, all the theory of this article also holds when some of parameters $\theta^{(i,j)}_{ker}$ are fixed to a value or equal to each other. In particular, the kernel can be exponential, gamma, or Weibull. Several examples covered by this framework are discussed in Supplement B.

3.3. Latency Estimation

We define the MLE restricted to the kernel parameter θ_{ker} as $\widehat{\theta}_{T,ker}$. The latency estimator is naturally defined as

$$\widehat{L}_T = F(\widehat{\theta}_{T,ker}). \tag{7}$$

3.4. Tests Related to Latency

We consider three Wald tests associated with latency. We first provide a test for equality between a latency value $L^{(i,j)}$ for an index $(i,j) \in \{1,\dots,d\}^2$ and a latency value $\widetilde{L} \in \mathbb{R}$, that is we define the null hypothesis as $H_0(\widetilde{L}) : \{L^{(i,j)} = \widetilde{L}\}$ and the alternative hypothesis as $H_1(\widetilde{L}) : \{L^{(i,j)} \neq \widetilde{L}\}$. We let our first test statistic be

$$W(\widetilde{L}) = nT \frac{(\widehat{L}_T^{(i,j)} - \widetilde{L})^2}{\widehat{\operatorname{Var}\left[\eta^{(i,j)}\right]}}, \tag{8}$$

where the variance estimator used in the denominator will be defined in (22). In the particular case when $\widetilde{L} = 0$, it can be interpreted as a test for the absence against the presence of latency, although this is not completely the case since latency can also be negative.

We second propose a test for equality between two latency values $L^{(i,j)}$ and $L^{(k,u)}$ for two indices $(i,j) \in \{1,\dots,d\}^2$ and $(k,u) \in \{1,\dots,d\}^2$, that is we define the null hypothesis as $H_0' : \{L^{(i,j)} = L^{(k,u)}\}$ and the alternative hypothesis as $H_1' : \{L^{(i,j)} \neq L^{(k,u)}\}$. We let our second test statistic be

$$W' = nT \frac{(\widehat{L}_T^{(i,j)} - \widehat{L}_T^{(k,u)})^2}{\widehat{\operatorname{var}\left[\eta^{(i,j)}\right]} + \widehat{\operatorname{var}\left[\eta^{(k,u)}\right]} - 2\widehat{\operatorname{cov}\left[\eta^{(i,j)}, \eta^{(k,u)}\right]}}. \tag{9}$$

where the variance and covariance estimators used in the denominator will be defined in (22).

For convenience we rewrite the $d \times d$-dimensional matrix of latencies $(L^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d}$ as a d^2-dimensional vector of latencies

$$\overline{L} = (L^{(i,j)})_{i,j=1,\dots,d} = (L^{(1,1)}, L^{(1,2)}, \dots, L^{(d,d)})^T.$$

We third introduce a test of q linear hypotheses on the d^2 latency vector which is expressed with the $q \times d^2$-dimensional matrix R, that is we define the null hypothesis as $\overline{H}_0(r) : \{R\overline{L} = r\}$ and the alternative hypothesis as $\overline{H}_1(r) : \{R\overline{L} \neq r\}$ for $r \in \mathbb{R}$. We let our third test statistic be

$$\overline{W}(r) = nT\left(R\widehat{\overline{L}}_T - r\right)^T \left(R\widehat{\overline{\Gamma}}_T R^T\right)^{-1} \left(R\widehat{\overline{L}}_T - r\right), \tag{10}$$

where the $d^2 \times d^2$-dimensional covariance matrix estimator used in the denominator will be defined in (33).

4. Theory

In this section, we first derive the feasible CLT of the MLE in the mixture of generalized gamma kernels case with in-fill asymptotics. This extends Clinet and Yoshida (2017) (Theorem 4.6) which is restricted to the exponential kernel case and Kwan (2023) (Theorem 3.4.3) which only obtains consistency of the MLE. Then, we derive the feasible CLT for the latency estimator. Finally, we obtain the limit theory for the tests related to latency.

4.1. CLT with Mixture of Generalized Gamma Kernels

We define $\overline{\Theta}$ as the closure space of Θ. We make the following conditions for the CLT.

$[A]$ (i) There exists $\nu_- \in \mathbb{R}_+^*$ such that for any $\nu \in \Theta_\nu$ we have that

$$\nu^{(i)} > \nu_-, \tag{11}$$

for any $i = 1,\dots,d$.

(ii) For any $\theta_{ker} \in \Theta_h$, we have that the kernel parameter θ_{ker} is of the form (6) and the kernel $h(t,\theta_{ker})$ is of the form (5).

(iii) There exists $p_- \in \mathbb{R}_+^*$ such that for any $i = 1,\dots,d$, any $j = 1,\dots,d$ and any $k = 1,\dots,K^{(i,j)}$ we have that

$$p_k^{(i,j)} > p_-. \tag{12}$$

(iv) There exists $D_- \in \mathbb{R}_+^*$ such that for any $i = 1,\dots,d$, any $j = 1,\dots,d$ and any $k = 1,\dots,K^{(i,j)}$ we have that

$$D_k^{(i,j)} > D_-. \tag{13}$$

(v) Let us define the matrix $\phi(\theta_{ker}) = (\phi^{(i,j)}(\theta_{ker}^{(i,j)}))_{i=1,\dots,d}^{j=1,\dots,d}$ where

$$\phi^{(i,j)}(\theta_{ker}^{(i,j)}) = \int_0^\infty h^{(i,j)}(s, \theta_{ker}^{(i,j)})ds,$$

and write $\rho(\phi(\theta_{ker}))$ being its spectral radius. There exists $0 < h_+ < 1$ such that for any $\theta \in \Theta$ we have that

$$\rho(\phi(\theta_{ker})) \leq h_+. \tag{14}$$

(vi) We have that $\Theta \subset (\mathbb{R}_+^*)^d \times \mathbb{R}^{m-d}$ is such that its closure $\overline{\Theta}$ is a compact space which satisfies the assumptions from the Sobolev embedding theorem (see Theorem 4.12 (p. 85) in Adams and Fournier 2003).

(vii) For any $i = 1,\dots,d$ and any $j = 1,\dots,d$ we have that $K^{(i,j)} = 1$.

Condition [A] (i), that is the positivity of the baseline, is well-known for Hawkes processes being well-defined. Condition [A] (ii) restricts to Hawkes processes with mixture of generalized gamma kernels. Condition [A] (iii) and (iv) put restriction on the parameter space. Condition [A] (v) is already required in the large-T asymptotics $T \to \infty$ to establish the existence of a stationary version of N_t on the same probability space and that N_t tends in distribution to this stationary process for a certain topology (see Theorem 7 in Brémaud and Massoulié (1996) and Proposition 4.4 in Clinet and Yoshida 2017). Condition [A] (vi) is necessary to apply the Sobolev embedding theorem. Condition [A] (vii) is required to obtain the classical nondegeneracy condition (Condition [A4] in Clinet and Yoshida 2017)). It is possible to weaken it, but its statement would be more cumbersome.

We define the space E as $E = \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}^m$. We also denote by $C_\uparrow(E,\mathbb{R})$ the set of continuous functions $\psi : (u,v,w) \to \psi(u,v,w)$ from E to $\mathbb{R}$ that satisfy ψ is of polynomial growth in u, v, w, $\frac{1}{u}$, and $\frac{1}{v}$. For any $i = 1,\dots,d$ and any $\theta \in \Theta$, we also define the rescaled time-transformed intensity process as $\overline{\lambda}^{(i)}(t,\theta) = \frac{\lambda^{(i)}(\frac{t}{n},\theta)}{n}$, and the triplet as $X_t^{(i)} = (\overline{\lambda}^{(i)}(t,\theta^*), \overline{\lambda}^{(i)}(t,\theta), \partial_\theta \overline{\lambda}^{(i)}(t,\theta))$. Propositions C1 and

C2 from Supplement C state that $X_t^{(i)}$ is stable for any $\theta \in \Theta$, that is for any $i = 1, \dots, d$ there exists an $\mathbb{R}_+^*$-valued random variable $\overline{\lambda}_{lim}^{(i)}(\theta)$ such that $X_{nT}^{(i)} \to^{\mathcal{D}} (\overline{\lambda}_{lim}^{(i)}(\theta^*), \overline{\lambda}_{lim}^{(i)}(\theta), \partial_\theta \overline{\lambda}_{lim}^{(i)}(\theta))$. They also state that the triplet is ergodic, that is there exists a mapping $\pi^{(i)} : C_\uparrow(E, \mathbb{R}) \times \Theta \to \mathbb{R}$ such that for any $(\psi, \theta) \in C_\uparrow(E, \mathbb{R}) \times \Theta$ we have $\frac{1}{nT} \int_0^{nT} \psi(X_{s,n}^{(i)}) ds \to^{\mathbb{P}} \pi^{(i)}(\psi, \theta)$, where $\pi^{(i)}(\psi, \theta) = \mathbb{E}[\psi(\overline{\lambda}_{lim}^{(i)}(\theta^*), \overline{\lambda}_{lim}^{(i)}(\theta), \partial_\theta \overline{\lambda}_{lim}^{(i)}(\theta))]$. Finally, they state that there exists a probability measure $\Pi_\theta^{(i)}$ on $(E, \boldsymbol{B}(E))$ such that for any $\psi \in C_\uparrow(E, \mathbb{R})$ and any $\theta \in \Theta$, we have $\pi^{(i)}(\psi, \theta) = \int_E \psi(u, v, w) \Pi_\theta^{(i)}(du, dv, dw)$. If we consider a vector $z \in \mathbb{R}^m$, we define the tensor product as $z^{\otimes 2} = z \times z^T \in \mathbb{R}^{m \times m}$. We define the $m \times m$-dimensional Fisher information matrix Γ as

$$\Gamma = \sum_{i=1}^{d} \int_E w^{\otimes 2} \frac{1}{u} \Pi_{\theta^*}^{(i)}(du, dv, dw). \tag{15}$$

The Fisher information matrix measures the amount of information that the intensity $\lambda(t, .)$ carries about the parameter θ^*. Formally, it is the expected value of the observed information. The Fisher information matrix is used to calculate the covariance matrices associated with MLE. In other words, Γ^{-1} is the asymptotic covariance matrix. The asymptotic Fisher information matrix is estimated from

$$\widehat{\Gamma}_T = -\partial_\theta^2 \bar{l}_T(\widehat{\theta}_T), \tag{16}$$

where $\partial_\theta^2 \bar{l}_T(\theta)$ is the $m \times m$-dimensional Hessian matrix of the time-transformed likelihood defined as $\bar{l}_T(\theta) = \sum_{i=1}^d \int_0^{Tn} \log(\overline{\lambda}^{(i)}(t, \theta)) d\overline{N}_t^{(i)} - \sum_{i=1}^d \int_0^{Tn} \overline{\lambda}^{(i)}(t, \theta) dt$ with $\overline{N}_t^{(i)} = N_{\frac{t}{n}}^{(i)}$. This is a natural estimator since we can reexpress the asymptotic Fisher information matrix as $\Gamma = -\lim_{n \to \infty} \frac{1}{Tn} \mathbb{E}\big[\partial_\xi^2 \bar{l}(\theta^*)\big]$. Finally, ξ is defined as an m-dimensional standard normal vector.

We now provide the feasible CLT of the MLE in the mixture of generalized gamma kernels case with in-fill asymptotics. This extends Clinet and Yoshida (2017) (Theorem 4.6) which is restricted to the exponential kernel case and Kwan (2023) (Theorem 3.4.3) which only shows consistency of the MLE. See also Cavaliere et al. (2023) (Theorem 2, p. 138), who require stronger conditions.

Theorem 1. We assume that Condition [A] holds. As $n \to +\infty$, we have the consistency

$$\widehat{\theta}_T \to^{\mathbb{P}} \theta^* \tag{17}$$

and the CLT

$$\sqrt{n}(\widehat{\theta}_T - \theta^*) \to^{\mathcal{D}} \sqrt{T} \Gamma^{-1/2} \xi. \tag{18}$$

We show the consistency of the Fisher information matrix estimator

$$\widehat{\Gamma}_T \to^{\mathbb{P}} \Gamma. \tag{19}$$

Moreover, we show the feasible normalized CLT

$$\widehat{\Gamma}_T^{1/2} \sqrt{nT}(\widehat{\theta}_T - \theta^*) \to^{\mathcal{D}} \xi. \tag{20}$$

4.2. CLT for Latency

We introduce the parameters of the kernel used in the definition of latency as θ_l of dimension l. As latency is equal to a function of kernel parameters, we have by definition that $\theta_l \subset \theta_{ker}$ and $l \leq m - d$. Moreover, we can rewrite the latency function as $L = F(\theta_l^*)$. For any $i = 1, \dots, d$ and $j = 1, \dots, d$ we define the l-dimensional differential vector corresponding to the (i, j)-index of F at point θ_l as

$$dF^{(i,j)}(\theta_l) := \Big(\frac{\partial F^{(i,j)}}{\partial \theta_l^{(1)}}(\theta_l), \dots, \frac{\partial F^{(i,j)}}{\partial \theta_l^{(l)}}(\theta_l) \Big).$$

We denote the Fisher information matrix and its estimator, restricted to the latency parameter θ_l, by Γ_l and $\widehat{\Gamma}_{T,l}$. Also ξ_l is defined as an l-dimensional standard normal vector. We then define the limit random matrix in the CLT as

$$(\eta^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d} = (dF^{(i,j)}(\theta_l^*) \Gamma_l^{-1/2} \xi_l)_{i=1,\dots,d}^{j=1,\dots,d}. \tag{21}$$

We denote the space of latency parameters by Θ_l. We make the following condition for deriving the CLT of latency estimation.

[B] We assume that $F : \Theta_l \to \mathbb{R}^{d \times d}$ is twice continuously differentiable and $dF^{(i,j)}(\theta_l^*) \Gamma_l^{-1/2}$ is not null for any $i = 1, \dots, d$ and $j = 1, \dots, d$.

Condition [B] puts some regular smoothness restrictions on F that are required to use Taylor expansion, while the non nullity of the vectors $dF^{(i,j)}(\theta_l^*) \Gamma_l^{-1/2}$ is required for the non nullity in the limit random matrix of the CLT (21).

We estimate the covariance between $\eta^{(i,j)}$ and $\eta^{(k,u)}$ as

$$\begin{aligned} \mathrm{cov}\left[\widehat{\eta^{(i,j)}, \eta^{(k,u)}}\right] = \sum_{q=1}^{l} \Big(\sum_{r=1}^{l} dF^{(i,j,r)}(\widehat{\theta}_{T,ker}) \big(\widehat{\Gamma}_{T,ker}^{-1/2}\big)^{(r,q)} \Big) \\ \times \Big(\sum_{r=1}^{l} dF^{(k,u,r)}(\widehat{\theta}_{T,l}) \big(\widehat{\Gamma}_{T,l}^{-1/2}\big)^{(r,q)} \Big). \end{aligned} \tag{22}$$

We also estimate the correlation between the normalized asymptotic covariance matrix (i, j)-index and (k, u)-index as

$$\mathrm{cor}\left[\widehat{\widetilde{\xi}^{(i,j)}, \widetilde{\xi}^{(k,u)}}\right] = \frac{\mathrm{cov}[\widehat{\eta^{(i,j)}, \eta^{(k,u)}}]}{\sqrt{\mathrm{var}\left[\widehat{\eta^{(i,j)}}\right] \mathrm{Var}\left[\widehat{\eta^{(k,u)}}\right]}}. \tag{23}$$

Now we provide the feasible CLT for the latency estimator with in-fill asymptotics. This complements the related approach on nonparametric Hawkes processes and applications to Covid-19 pandemic in France given in Gámiz et al. (2022) and Gámiz et al. (2023) for the time-varying case.

Proposition 2. We assume that Condition [A] and Condition [B] hold. As $n \to +\infty$, we have the consistency

$$\widehat{L}_T \to^{\mathbb{P}} L \tag{24}$$

and the CLT

$$\sqrt{n}(\widehat{L}_T^{(i,j)} - L^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d} \to^{\mathcal{D}} \sqrt{T}(\eta^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d}. \tag{25}$$

Moreover, $\eta^{(i,j)}$ can be re-expressed for any $i = 1, \dots, d$ and any $j = 1, \dots, d$ as

$$\eta^{(i,j)} = \sum_{q=1}^{l} \Big(\sum_{r=1}^{l} dF^{(i,j,r)}(\theta_l^*)\big(\Gamma_l^{-1/2}\big)^{(r,q)} \Big) \xi_l^{(q)}. \tag{26}$$

We can deduce that

$$\begin{aligned}\operatorname{cov}[\eta^{(i,j)}, \eta^{(k,u)}] = {} & \sum_{q=1}^{l} \Big(\sum_{r=1}^{l} dF^{(i,j,r)}(\theta_l^*)\big(\Gamma_l^{-1/2}\big)^{(r,q)} \Big) \\ & \times \Big(\sum_{r=1}^{l} dF^{(k,u,r)}(\theta_l^*)\big(\Gamma_l^{-1/2}\big)^{(r,q)} \Big).\end{aligned} \tag{27}$$

We obtain the consistency of the covariance estimator

$$\widehat{\operatorname{cov}\left[\eta^{(i,j)}, \eta^{(k,u)}\right]} \to^{\mathbb{P}} \operatorname{cov}[\eta^{(i,j)}, \eta^{(k,u)}]. \tag{28}$$

We show the feasible normalized CLT

$$\sqrt{nT}\left(\frac{\widehat{L}_T^{(i,j)} - L^{(i,j)}}{\sqrt{\widehat{\operatorname{var}\left[\eta^{(i,j)}\right]}}} \right)_{i=1,\dots,d}^{j=1,\dots,d} \to^{\mathcal{D}} (\widetilde{\xi}^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d}, \tag{29}$$

where each component $\widetilde{\xi}^{(i,j)}$ is a standard normal variable with correlation that satisfies

$$\operatorname{cor}\left[\widetilde{\xi}^{(i,j)}, \widetilde{\xi}^{(k,u)}\right] = \frac{\operatorname{cov}[\eta^{(i,j)}, \eta^{(k,u)}]}{\sqrt{\operatorname{var}\left[\eta^{(i,j)}\right] \operatorname{Var}\left[\eta^{(k,u)}\right]}}. \tag{30}$$

The consistency of the correlation estimator is obtained as

$$\widehat{\operatorname{cor}\left[\widetilde{\xi}^{(i,j)}, \widetilde{\xi}^{(k,u)}\right]} \to^{\mathbb{P}} \operatorname{cor}\left[\widetilde{\xi}^{(i,j)}, \widetilde{\xi}^{(k,u)}\right]. \tag{31}$$

For any $i = 1, \dots, d$ and $j = 1, \dots, d$ each component $\widetilde{\xi}^{(i,j)}$ is a standard normal variable, but the limit matrix $(\widetilde{\xi}^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d}$ is not a standard normal vector in (29). In what follows, we give a feasible CLT with a standard normal vector in the limit since this is useful for providing the asymptotic theory of the multidimensional Wald test in Corollary 6. To obtain a standard normal vector in the limit, we rewrite the $d \times d$-dimensional matrix of latency estimators $(\widehat{L}_T^{(i,j)})_{i=1,\dots,d}^{j=1,\dots,d}$ as a d^2-dimensional vector of latencies

$$\widehat{\overline{L}}_T = (\widehat{L}_T^{(i,j)})_{i,j=1,\dots,d} = (\widehat{L}_T^{(1,1)}, \widehat{L}_T^{(1,2)}, \dots, \widehat{L}_T^{(d,d)})^T$$

and we introduce the $d^2 \times d^2$-dimensional asymptotic covariance matrix $\overline{\Gamma}^{-1}$ satisfying

$$(\overline{\Gamma}^{-1})_{i,j=1,\dots,d}^{k,l=1,\dots,d} = \operatorname{cov}[\eta^{(i,j)}, \eta^{(k,l)}]. \tag{32}$$

We estimate the asymptotic covariance matrix with

$$(\widehat{\overline{\Gamma}}_T^{-1})_{i,j=1,\dots,d}^{k,l=1,\dots,d} = \widehat{\operatorname{cov}\left[\eta^{(i,j)}, \eta^{(k,l)}\right]}. \tag{33}$$

We make the following condition to derive the CLT of latency estimation with a standard normal vector in the limit.

[*C*] We assume that the $d^2 \times l$-dimensional matrix

$$\Big(\sum_{r=1}^{l} dF^{(i,j,r)}(\theta_l^*)\big(\Gamma_l^{-1/2}\big)^{(r,q)} \Big)_{i,j=1,\cdots,d}^{q=1,\cdots,l}$$

has rank d^2.

Condition [C] ensures existence of a standard normal vector in the limit of the CLT. Thus, Condition [C] is slightly stronger than Condition [B]. In practice, this implies that $d^2 \leq l$. However, this condition is automatically satisfied since we use at least one parameter for each index of the latency matrix. Finally, we introduce the d^2-dimensional standard normal vector $\overline{\xi}$.

Corollary 3. We assume that Condition [A], Condition [B], and Condition [C] hold. As $n \to +\infty$, we have the CLT

$$\sqrt{n}(\widehat{\overline{L}}_T - \overline{L}) \to^{\mathcal{D}} \sqrt{T}\overline{\Gamma}^{-1/2}\overline{\xi}. \tag{34}$$

We obtain the consistency for the asymptotic covariance matrix

$$\widehat{\overline{\Gamma}}_T \to^{\mathbb{P}} \overline{\Gamma}. \tag{35}$$

Moreover, we provide the feasible normalized CLT

$$\widehat{\overline{\Gamma}}_T^{1/2} \sqrt{nT}(\widehat{L}_T^{(i,j)} - L^{(i,j)})_{i,j=1,\dots,d} \to^{\mathcal{D}} \overline{\xi}. \tag{36}$$

4.3. Tests Related to Latency

The following corollary shows that the first Wald test statistic converges in distribution to a Chi-squared distribution with one degree of freedom under the null hypothesis and is consistent under the alternative hypothesis. The proof is based on Proposition 2. We denote Q_u as the quantile function of the Chi-squared distribution with one degree of freedom.

Corollary 4. We assume that Condition [A] and Condition [B] hold. As $n \to +\infty$ and for any latency value $\widetilde{L} \in \mathbb{R}$, the first Wald test statistic $W(\widetilde{L})$ converges in distribution to a Chi-squared distribution with one degree of freedom under the null hypothesis $H_0(\widetilde{L}) : \{L^{(i,j)} = \widetilde{L}\}$ and is consistent under the alternative hypothesis $H_1(\widetilde{L}) : \{L^{(i,j)} \neq \widetilde{L}\}$, that is for any $0 < \alpha < 1$:

$$\operatorname{size}(\alpha) = \mathbb{P}\{W(\widetilde{L}) > Q_{1-\alpha} \mid H_0(\widetilde{L})\} \to \alpha, \tag{37}$$

$$\operatorname{power}(\alpha) = \mathbb{P}\{W(\widetilde{L}) > Q_{1-\alpha} \mid H_1(\widetilde{L})\} \to 1. \tag{38}$$

The second Wald test statistic converges in distribution to a Chi-squared distribution with one degree of freedom under the null hypothesis and is consistent. The proof is based on Proposition 2.

Corollary 5. We assume that Condition [A] and Condition [B] hold. We also assume that $(\eta^{(i,j)}, \eta^{(k,l)})$ is a two-dimensional random vector. As $n \to +\infty$, the second Wald test statistic W' converges in distribution to a Chi-squared distribution with one degree of freedom under the null hypothesis $H_0' : \{L^{(i,j)} = L^{(k,l)}\}$ and is consistent under the alternative hypothesis $H_1' : \{L^{(i,j)} \neq L^{(k,l)}\}$, that is for any $0 < \alpha < 1$ we have

$$\operatorname{size}'(\alpha) = \mathbb{P}\{W' > Q_{1-\alpha} \mid H_0'\} \to \alpha, \tag{39}$$

$$\operatorname{power}'(\alpha) = \mathbb{P}\{W' > Q_{1-\alpha} \mid H_1'\} \to 1. \tag{40}$$

Finally, the following corollary shows the third Wald test statistic converging to a Chi-squared distribution with q degrees of freedom under the null hypothesis and being consistent. The proof is based on Corollary 3. We denote $Q_u^{(q)}$ as the quantile function of the Chi-squared distribution with q degrees of freedom.

Corollary 6. We assume that Condition [A], Condition [B] and Condition [C] hold. As $n \to +\infty$ and for any $r \in \mathbb{R}$, the third Wald test statistic $\overline{W}(r)$ converges in distribution to a Chi-squared distribution with q degrees of freedom under the null hypothesis $\overline{H}_0(r) : \{R\overline{L} = r\}$ and is consistent under the alternative hypothesis $\overline{H}_1(r) : \{R\overline{L} \neq r\}$, that is for any $0 < \alpha < 1$:

$$\overline{\text{size}}(\alpha) = \mathbb{P}\{\overline{W}(r) > Q_{1-\alpha}^{(q)} \mid \overline{H}_0(r)\} \to \alpha, \quad (41)$$

$$\overline{\text{power}}(\alpha) = \mathbb{P}\{\overline{W}(r) > Q_{1-\alpha}^{(q)} \mid \overline{H}_1(r)\} \to 1. \quad (42)$$

5. Empirical Application

In this section we analyze the performance of the proposed model using the transaction data for the New York Stock Exchange (NYSE) and the Toronto Stock Exchange (TSX).

5.1. Data

Our sample period runs from January 2, 2020 to December 31, 2021. Each day, we construct a sample of stocks included in the S&P 500 index and the TSX composite index and that are traded in the NYSE and the TSX. We obtain trade and mid-quote price, that is the average price between best bid and ask prices, and time stamps from the consolidated trade history in the transaction Datascope database. Following Hasbrouck (2018) all the stock trades and quotes between 9.45 a.m. and 3.45 p.m. (EST) are considered. Our selection of characteristics for each stock and each day includes millisecond time stamps. These stock characteristics are used to obtain the estimates of latency and to test the hypotheses formulated in the following section.

We apply additional filters in the following order. First, we exclude trades and quotes with zero volumes and prices. Second, we drop a stock-day observation if it takes extreme values falling in the top or bottom 1% of the monthly cross-sections. Finally, each daily sample comprises the 798 stocks traded in the NYSE and the TSX.

Table 1 presents summary sample statistics over the final sample. The sample statistics are computed for the United States and Canadian stock exchanges separately. The U.S. market is characterized by a shorter duration, while the TSX experienced a lower standard deviation of trade durations.

Table 1. Summary statistics are reported for all S&P 500 and TSX composite index stocks traded in the NYSE and TSX.

	Obs.	Mean	Std dev	Min	Max
	NYSE				
Trade duration	273,845,398	33.304	149.052	1	63247.101
Mid-quote duration	3,361,459,591	2.987	4.983	1	72.370
	TSX				
Trade duration	120,801,281	77.245	127.918	1	4336.670
Mid-quote duration	1,409,644,068	6.798	10.286	1	222.901

NOTE: The daily average statistics are obtained over the sample from January 02, 2020 to December 31, 2021. Durations are expressed in milliseconds.

5.2. Hypotheses

To understand the evolution of latency in the NYSE and TSX we formulate testable hypotheses of interest. All hypotheses are tested for the whole sample including all trading days and p-values can be obtained with Corollary 6. To verify that our results of testing hypotheses are not distorted due to a multiple statistical inference problem, we implement a Bonferroni adjustment of Holm (1979) for all p-values. The adjusted p-values computed at the 1% level provide the identical conclusions about all hypotheses confirming the statistical robustness of our results. Another robustness check of our testing results is conducted following Bajgrowicz, Scaillet, and Treccani (2016) and the results are in agreement with the Bonferroni corrected tests.

First, we conjecture if latency exists in both exchanges and for mid-quote and trade events.

Latency exists in the NYSE and TSX stock exchanges, that is $\overline{H}_0(0) : L^{(i,i)} = 0$ *for all* $i = 1, \dots, 4$.

A p-value of 0.001 confirms the existence of latency at the 5% level. One may argue that latency is solely characterized by technological development of a stock exchange. In this case a technology arms race would not be defined by timing of orders and latency estimates are expected to be similar in both exchanges. If this is not the case, transforming competition on speed into competition on price is possible when firms strategically consider the timing of order submissions, see for example Budish, Cramton, and Shim (2015).

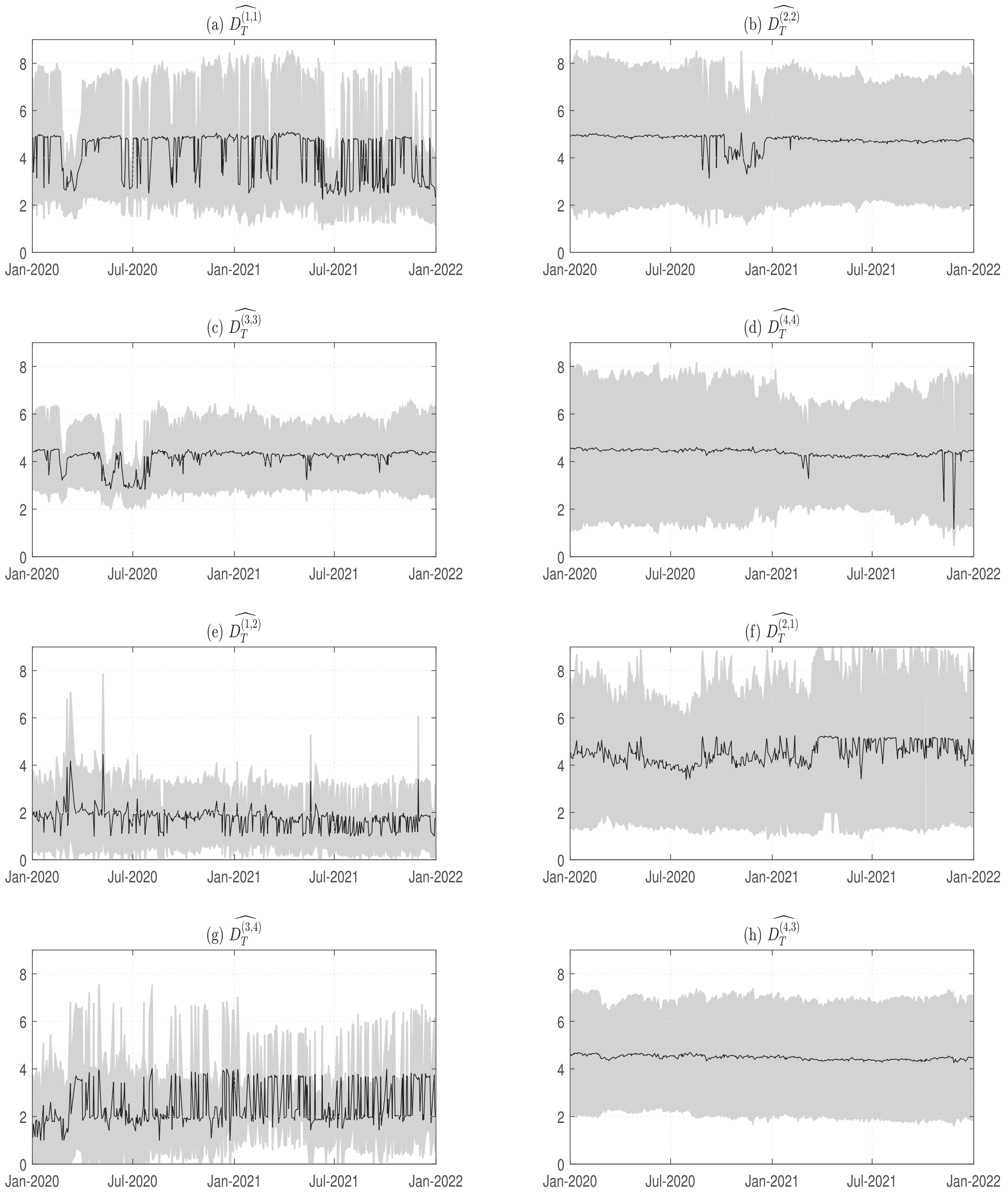
Latency varies across the stock exchanges, that is $\overline{H}_0(0) : L^{(1,1)} = L^{(3,3)}$ *and* $L^{(2,2)} = L^{(4,4)}$.

A p-value of 0.002 provides evidence of rejecting the null hypothesis at the 5% level. This suggests that there exist additional sources of market information that must be taken into consideration when a new measure of latency is designed. This conjecture extends an idea of Riordan and Storkenmaier (2012) about interrelation between latency and price discovery.

The presence of latency is a feature of modern financial markets, but it is unclear if changes in latencies across stock exchanges create market co-movements. This conjecture is supported by Baron et al. (2019) who find latency to be used as a channel for cross-exchange interactions. We call this phenomenon co-latency. Co-latency can be considered as a channel of emerging spillovers between the market events corroborating findings of Malceniece, Malcenieks, and Putniņš (2019) about the substantial impact of trading activity on co-movements in stock returns. Following Aït-Sahalia, Cacho-Diaz, and Laeven (2015) the presence of spillovers is associated with statistically significant cross-excitation effects.

Co-latency is observed for different events (trades, quotes) within the stock exchanges, that is (*) $\overline{H}_0(0) : L^{(1,2)} = L^{(2,1)} = 0$ *and* $L^{(3,4)} = L^{(4,3)} = 0$ *or (**)* $\overline{H}_0(0) : L^{(3,1)} = L^{(3,2)} = L^{(4,1)} = L^{(4,2)} = 0$ *and* $L^{(1,3)} = L^{(1,4)} = L^{(2,3)} = L^{(2,4)} = 0$.

Rejecting (*) with p-value=0.003 confirms the existence of co-latency in both exchanges. However, (**) is not rejected (p-value=0.203), justifying the co-location argument of Brogaard et al. (2015) and confirming that co-latency does not spread across exchanges.



(a) $\widehat{D_T^{(1,1)}}$

(b) $\widehat{D_T^{(2,2)}}$

(c) $\widehat{D_T^{(3,3)}}$

(d) $\widehat{D_T^{(4,4)}}$

(e) $\widehat{D_T^{(1,2)}}$

(f) $\widehat{D_T^{(2,1)}}$

(g) $\widehat{D_T^{(3,4)}}$

(h) $\widehat{D_T^{(4,3)}}$

Figure 1. Parameter estimates $\widehat{D_T^{(i,j)}}$ obtained from (4). Parameter estimates for each day in the NYSE and TSX are shown. 90% confidence intervals are also presented.

5.3. Latency in the United States and Canadian Stock Exchanges

We now discuss the estimates of latency obtained by the MLE procedure presented in Section 3. Trade and mid-quote time stamps of all the stocks are used to estimate Model (4) for each day and for both the NYSE and the TSX. In this case the kernel matrix $h^{(i,j)}$ is 4×4-dimensional and the market interaction between the United States and Canadian stock exchanges is captured by the cross excitation terms $h^{(i,j)}$ when $i \neq j$. The shape of the kernel matrix $h^{(i,j)}$ follows the gamma specification discussed in Supplement B.2 which is a special case of mixture of generalized gamma kernels presented in Section 3.2 with parameters α, β and D. Following the results from the previous section we discuss only (co)latency estimates within the NYSE and the TSX captured by $h^{(1,1)}, h^{(2,2)}, h^{(3,3)}, h^{(4,4)}$ and $h^{(1,2)}, h^{(2,1)}, h^{(3,4)}, h^{(4,3)}$. This is verified by the testing results in the previous section and in agreement with the colocation argument of Shkilko and Sokolov (2020) implying the region specific nature of trading activity.

Delay D is discussed now as an important factor contributing to latency. Parameter estimates $\widehat{D_T^{(i,j)}}$ are presented in Figure 1 for each trading day.[1] The delay parameters $D^{(1,1)}$, $D^{(2,2)}$ for the NYSE and $D^{(3,3)}$, $D^{(4,4)}$ for the TSX trade and mid-quote

[1] The estimates of parameters $\alpha^{(i,j)}$ and $\beta^{(i,j)}$ are not presented here, but available in Supplement D.

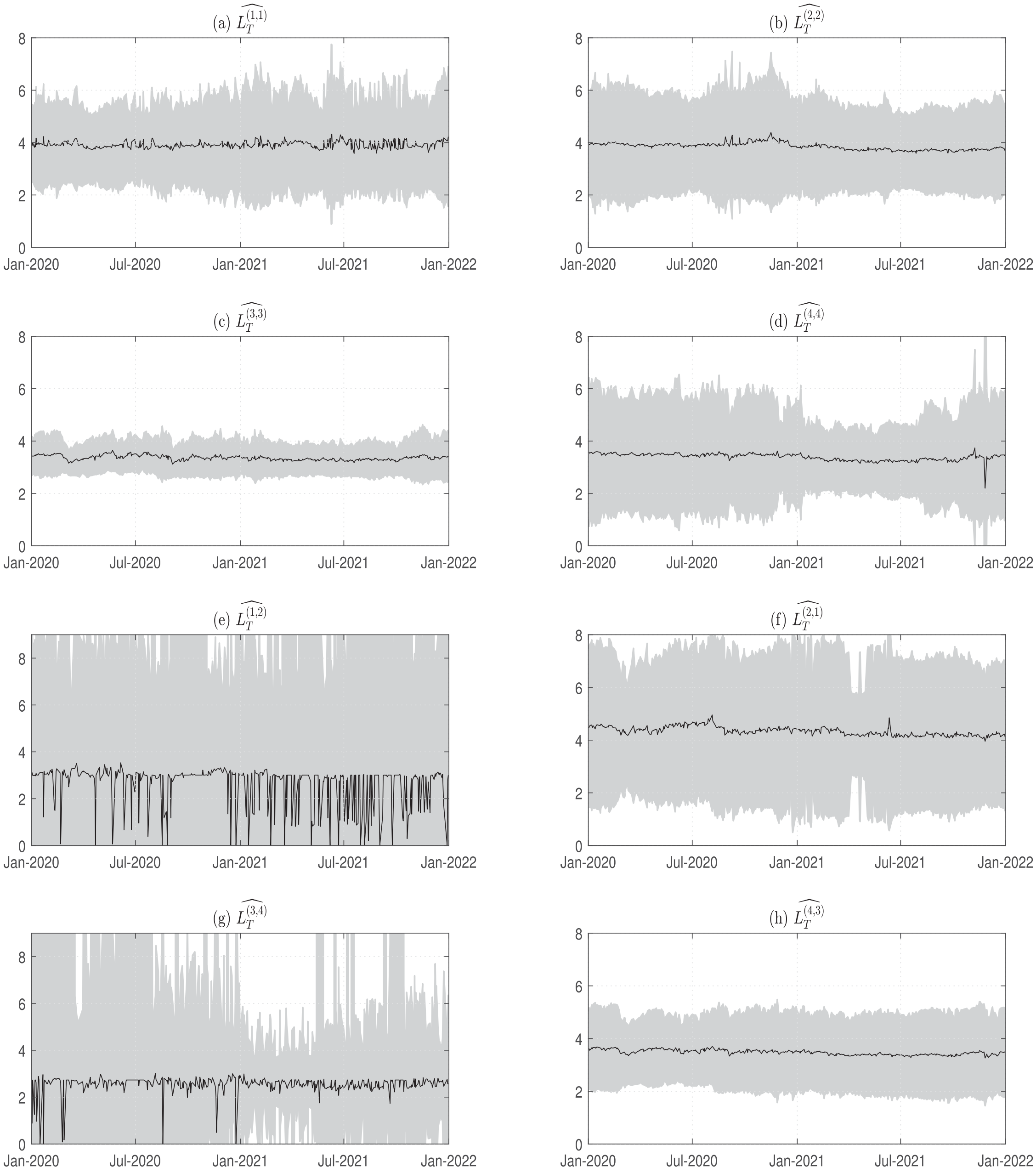


Figure 2. Latency and co-latency estimates $\widehat{L_T^{(i,j)}}$ obtained from Model (4). Each graph shows co-latency estimates for each day across all events in the NYSE and TSX. 90% confidence intervals are reported.

events are statistically different from zero changing between 1 and 5 milliseconds. A substantial drop in delay from almost 4 to 2 milliseconds observed in Figure 1(d) happened in November 26th, 2021 which was the worst day of year for North American stock markets. On this day S&P 500 index dropped more than 2% due to a new Covid variant found in South Africa triggering a shift from risk assets and accelerating the speed of trading. The cross-exciting parameter $D^{(1,2)}$ capturing delay in the NYSE trades due to mid-quote events in the same exchange fluctuates between 1.4 and 4 milliseconds. In February and March 2020, active trading during the start of COVID pandemic creates longer delays $D^{(1,2)}$ and $D^{(3,4)}$ showing that market participants respond more quickly to trades when information is flowing. Delays of responding quotes to trades $D^{(2,1)}$ and $D^{(4,3)}$ are longer for both the NYSE and the TSX suggesting that in both markets trades absorb market information faster.

The estimates of latency and co-latency $\widehat{L_T^{(i,j)}}$ for the NYSE and the TSX obtained for each day over the sample are presented in Figure 2. The latency estimates $\widehat{L_T^{(i,i)}}$ for the United States and Canada change between 2 and 5 milliseconds over the sample. Overall the TSX is faster with the latency just below 4 milliseconds for trades and mid-quotes. In February and March 2020, the start of COVID-19 pandemic, co-latency in the NYSE and TSX observed from Figure 2(e) and (g) jumped a few times between 1 and almost 4 milliseconds. This pattern is attributed to suspending floor trading due to COVID-19 pandemic in February and March 2020. Another common pattern for both exchanges observed from Figure 2(e)–(h) is a faster reaction of trading co-latency in response to quotes. Confidence intervals for $\widehat{L_T^{(1,2)}}$ and $\widehat{L_T^{(3,4)}}$ are wider comparing to $\widehat{L_T^{(2,1)}}$ and $\widehat{L_T^{(4,3)}}$ indicating uncertainty about the impact of quote events on trades which reduces in 2021 for $\widehat{L_T^{(4,3)}}$. Our findings indicate the existence of co-latency channel working in both directions between the United States and Canada. This corresponds to Hoffmann (2014) where an ability of fast traders to revise their quotes quickly after news arrivals helps reducing market risks in some markets, that is the TSX and the NYSE in our case.

6. Conclusion

A novel statistical approach to estimating latency, defined as the time it takes to learn about an event and generate response to this event, is described. Outside of finance this definition helps understanding and modeling delay in reactions to events for point processes. The problem is formulated to be solved by the comprehensive use of stochastic analysis techniques. More specifically, we have considered the class of parametric Hawkes models, which circumvents the use of more detailed datasets which may not even be available. We define latency as a known function of kernel parameters, typically the mode of kernel function. Since latency is not well-defined when the kernel is exponential, we consider maximum likelihood estimation in the mixture of generalized gamma kernels case and derive the feasible CLT with in-fill asymptotics. As a byproduct, CLT for a latency estimator, defined as the function of parameter estimates, and three tests were deduced. Latency estimates for the United States and Canadian stock exchanges are found to vary between 1 and 6 milliseconds from 2020 to 2021. The existence of co-latency channel working in both directions between the United States and Canada is also confirmed.

A more realistic framework with polynomial periodic kernel is studied in Erdemlioglu et al. (2025a). This is based on the theoretical results given in Potiron (2025), which extend the theoretical results from Clinet and Yoshida (2017). Finally, there is an extension to time-dependent latency in Erdemlioglu et al. (2025b). This time-dependent latency work builds on Clinet and Potiron (2018).

Supplementary Materials

Our numerical study is carried over in Supplement A. Examples of kernels covered by our framework are given in the Supplement B. All proofs of the theoretical results are shown in Supplement C. An additional empirical result belongs to Supplement D.

Acknowledgments

We thank Dylan Small (the co-Editor), the anonymous Associate Editor, four anonymous reviewers, Simon Clinet, Dobrislav Dobrev, participants of the Intrinsic Time Conference in Finance 2022 for discussions. Research assistantship from Fedor Fradkin is acknowledged.

Disclosure Statement

The authors report there are no competing interests to declare

Funding

Financial support from the Japanese Society for the Promotion of Science under grants 17K13718, 20K13470 and 23H00807 (Potiron) is gratefully acknowledged. Volkov gratefully acknowledges support from the Basic Research Program of the National Research University Higher School of Economics (HSE University).

ORCID

Yoann Potiron http://orcid.org/0000-0002-3621-3943
Vladimir Volkov http://orcid.org/0000-0001-9721-9700

References

Adams, R. A., and Fournier, J. J. F. (2003), *Sobolev Spaces*, Amsterdam: Elsevier. [5]

Aït-Sahalia, Y., and Jacod, J. (2014), *High-Frequency Financial Econometrics*, Princeton: Princeton University Press. [2]

Aït-Sahalia, Y., Laeven, R. J. A., and Pelizzon, L. (2014), "Mutual Excitation in Eurozone Sovereign CDS," *Journal of Econometrics*, 183, 151–167. [3]

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. J. A. (2015), "Modeling Financial Contagion Using Mutually Exciting Jump Processes," *Journal of Financial Economics*, 117, 585–606. [3,8]

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J. F. (2013), "Some Limit Theorems for Hawkes Processes and Application to Financial Statistics," *Stochastic Processes and their Applications*, 123, 2475–2499. [3]

Bajgrowicz, P., Scaillet, O., and Treccani, A. (2016), "Jumps in High-Frequency Data: Spurious Detections, Dynamics, and News," *Management Science*, 62, 2198–2217. [8]

Baron, M., Brogaard, J., Hagströmer, B., and Kirilenko, A. (2019), "Risk and Return in High-Frequency Trading," *Journal of Financial and Quantitative Analysis*, 54, 993–1024. [8]

Barra, I., Borowska, A., and Koopman, S. J. (2018), "Bayesian Dynamic Modeling of High-Frequency Integer Price Changes," *Journal of Financial Econometrics*, 16, 384–424. [1]

Bennedsen, M., Lunde, A., Shephard, N., and Veraart, A. E. D. (2023), "Inference and Forecasting for Continuous-Time Integer-Valued Trawl Processes," *Journal of Econometrics*, 236, 105476. [3]

Bowsher, C. G. (2007), "Modelling Security Market Events in Continuous Time: Intensity based, Multivariate Point Process Models," *Journal of Econometrics*, 141, 876–912. [2,3]

Brémaud, P., and Massoulié, L. (1996), "Stability of Nonlinear Hawkes Processes," *The Annals of Probability*, 24, 1563–1588. [5]

Brogaard, J., Hagströmer, B., Nordén, L., and Riordan, R. (2015), "Trading Fast and Slow: Colocation and Liquidity," *The Review of Financial Studies*, 28, 3407–3443. [8]

Budish, E., Cramton, P., and Shim, J. (2015), "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, 130, 1547–1621. [8]

Cavaliere, G., Lu, Y., Rahbek, A., and Stærk-Ostergaard, J. (2023), "Bootstrap Inference for Hawkes and General Point Processes," *Journal of Econometrics*, 235, 133–165. [2,3,6]

Chavez-Demoulin, V., Davison, A. C., and McNeil, A. J. (2005), "Estimating Value-at-Risk: A Point Process Approach," *Quantitative Finance*, 5, 227–234. [3]

Chen, F., and Hall, P. (2013), "Inference for a Nonstationary Self-Exciting Point Process with an Application in Ultra-High Frequency Financial Data Modeling," *Journal of Applied Probability*, 50, 1006–1024. [2,4]

Clements, A. E., Hurn, A. S., Lindsay, K. A., and Volkov, V. (2023), "Estimating a Non-parametric Memory Kernel for Mutually-Exciting Point Processes," *Journal of Financial Econometrics*, 21, 1759–1790. [3]

Clinet, S., and Potiron, Y. (2018), "Statistical Inference for the Doubly Stochastic Self-Exciting Process," *Bernoulli*, 24, 3469–3493. [2,4,11]

Clinet, S., and Yoshida, N. (2017), "Statistical Inference for Ergodic Point Processes and Application to Limit Order Book," *Stochastic Processes and their Applications*, 127, 1800–1839. [2,3,5,6,11]

Corradi, V., Distaso, W., and Fernandes, M. (2020), “Testing for Jump Spillovers Without Testing for Jumps,” *Journal of the American Statistical Association*, 115, 1214–1226. [3]

Daley, D. J., and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes*, (Vol. 1, 2nd ed.), New York: Springer-Verlag. [3,4]

——— (2008), *An Introduction to the Theory of Point Processes: General Theory and Structure* (Vol. 2, 2nd ed.), New York: Springer Verlag. [3]

Embrechts, P., Liniger, T. J., and Lin, L. (2011), “Multivariate Hawkes Processes: An Application to Financial Data,” *Journal of Applied Probability*, 48, 367–378. [3]

Engle, R. F. (2000), “The Econometrics of Ultra-High-Frequency Data,” *Econometrica*, 68, 1–22. [2]

Engle, R. F., and Russell, J. R. (1998), “Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data,” *Econometrica*, 66, 1127–1162. [1]

Erdemlioglu, D., Potiron, Y., Xu, T., and Volkov, V. (2025a), “Estimation of Latency for Hawkes Processes with a Polynomial Periodic Kernel,” working paper available at *https://www.fbc.keio.ac.jp/~potiron/Erdemlioglu2025workingpaperestimationlatency.pdf*. [11]

——— (2025b), “Estimation of Time-Dependent Latency with Locally Stationary Hawkes Processes,” working paper available at *https://www.fbc.keio.ac.jp/~potiron/Erdemlioglu2025workingpaper.pdf*. [11]

Gagnon, L., and Karolyi, G. (2010), “Multi-Market Trading and Arbitrage,” *Journal of Financial Economics*, 97, 53–80. [2]

Gámiz, M. L., Mammen, E., Martínez-Miranda, M. D., and Nielsen, J. P. (2022), “Missing Link Survival Analysis with Applications to Available Pandemic Data,” *Computational Statistics & Data Analysis*, 169, 107405. [2,6]

——— (2023), “Monitoring a Developing Pandemic with Available Data,” arXiv preprint arXiv:2308.09919. [2,6]

Harris, J. (1990), “Reporting Delays and the Incidence of AIDS,” *Journal of the American Statistical Association*, 85, 915–924. [1]

Hasbrouck, J. (2018), “High-Frequency Quoting: Short-Term Volatility in Bids and Offers,” *Journal of Financial and Quantitative Analysis*, 53, 613–641. [8]

Hasbrouck, J., and Saar, G. (2013), “Low-Latency Trading,” *Journal of Financial Markets*, 16, 646–679. [1,4]

Hawkes, A. G. (1971a), “Point Spectra of Some Mutually Exciting Point Processes,” *Journal of the Royal Statistical Society*, Series B, 33, 438–443. [3]

——— (1971b), “Spectra of Some Self-Exciting and Mutually Exciting Point Processes,” *Biometrika*, 58, 83–90. [3]

——— (2018), “Hawkes Processes and their Applications to Finance: A Review,” *Quantitative Finance*, 18, 193–198. [3]

Heinen, A., and Rengifo, E. (2007), “Multivariate Autoregressive Modeling of Time Series Count Data Using Copulas,” *Journal of Empirical Finance*, 14, 564–583. [1]

Hoffmann, P. (2014), “A Dynamic Limit Order Market with Fast and Slow Traders,” *Journal of Financial Economics*, 113, 156–169. [3,11]

Holm, S. (1979), “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65–70. [8]

Ikefuji, M., Laeven, R. J. A., Magnus, J. R., and Yue, Y. (2022), “Earthquake Risk Embedded in Property Prices: Evidence from Five Japanese Cities,” *Journal of the American Statistical Association*, 117, 82–93. [3]

Jacod, J., and Shiryaev, A. (2013), *Limit Theorems for Stochastic Processes* (Vol. 288), Berlin: Springer. [3]

Karim, R., Laeven, R. J. A., and Mandjes, M. (2021), “Exact and Asymptotic Analysis of General Multivariate Hawkes Processes and Induced Population Processes,” arXiv preprint arXiv:2106.03560. [3]

Koopman, S. J., Lit, R., Lucas, A., and Opschoor, A. (2018), “Dynamic Discrete Copula Models for High-Frequency Stock Price Changes,” *Journal of Applied Econometrics*, 33, 966–985. [1]

Kwan, T. K. J. (2023), “Asymptotic Analysis and Ergodicity of the Hawkes Process and its Extensions,” Ph.D. thesis, UNSW Sydney. [2,3,4,5,6]

Kwan, T. K. J., Chen, F., and Dunsmuir, W. T. M. (2023), “Alternative Asymptotic Inference Theory for a Nonstationary Hawkes Process,” *Journal of Statistical Planning and Inference*, 227, 75–90. [2,4]

Large, J. (2007), “Measuring the Resiliency of an Electronic Limit Order Book,” *Journal of Financial Markets*, 10, 1–25. [2,3]

Liniger, T. J. (2009), “Multivariate Hawkes Processes,” Ph.D. thesis, ETH Zurich. [3]

Malceniece, L., Malcenieks, K., and Putniņš, T. (2019), “High Frequency Trading and Comovement in Financial Markets,” *Journal of Financial Economics*, 134, 381–399. [8]

Ogata, Y. (1978), “The Asymptotic Behaviour of Maximum Likelihood Estimators for Stationary Point Processes,” *Annals of the Institute of Statistical Mathematics*, 30, 243–261. [2,3,4]

Ozaki, T. (1979), “Maximum Likelihood Estimation of Hawkes’ Self-Exciting Point Processes,” *Annals of the Institute of Statistical Mathematics*, 31, 145–155. [3]

Potiron, Y. (2025), “Inference for Hawkes Processes with a General Kernel,” working paper available at *https://www.fbc.keio.ac.jp/~potiron/Potiron2025inferenceworkingpaper.pdf*. [3,11]

Potiron, Y., Scaillet, O., Volkov, V., and Yu, S. (2025a), “Estimation of Branching Ratio for Hawkes Processes with Itô Semimartingale Baseline,” working paper available at *https://www.fbc.keio.ac.jp/~potiron/Potiron2025estimationworkingpaper.pdf*. [4]

——— (2025b), “High-Frequency Estimation of Itô Semimartingale Baseline for Hawkes Processes,” working paper available at *https://www.fbc.keio.ac.jp/~potiron/Potiron2025workingpaper.pdf*. [4]

Riordan, R., and Storkenmaier, A. S. (2012), “Latency, Liquidity and Price Discovery,” *Journal of Financial Markets*, 15, 416–437. [8]

Rubin, I. (1972), “Regular Point Processes and their Detection,” *IEEE Transactions on Information Theory*, 18, 547–557. [3]

Shkilko, A., and Sokolov, K. (2020), “Every Cloud has a Silver Lining: Fast Trading, Microwave Connectivity, and Trading Costs,” *The Journal of Finance*, 75, 2899–2927. [9]

van Lieshout, M. N. M. (2021), “Infill Asymptotics for Adaptive Kernel Estimators of Spatial Intensity,” *Australian & New Zealand Journal of Statistics*, 63, 159–181. [3]

Vere-Jones, D. (1978), “Earthquake Prediction-A Statistician’s View,” *Journal of Physics of the Earth*, 26, 129–146. [3]

Vere-Jones, D., and Ozaki, T. (1982), “Some Examples of Statistical Estimation Applied to Earthquake Data: I. Cyclic Poisson and Self-Exciting Models,” *Annals of the Institute of Statistical Mathematics*, 34, 189–207. [3]