Hawkes Processes and Their Applications to High-Frequency Data Modeling

Chapter - April 2016
DOI: 10.1002/9781118593486.ch7

2 authors:
- Baron Law
- Frederi Viens

Some of the authors of this publication are also working on these related projects:
- STATISTICAL ESTIMATION AND GAUSSIAN FIELDS USING THE MALLIAVIN CALCULUS
- Stochastic calculus via regularization
Introduction

This short paper introduces and surveys an emerging class of stochastic point processes used in modeling the evolution of high-frequency data on stock markets at a high level of quantitative detail.

The information contained in a stock market’s Limit Order Books (LOB) is a multivariate time series which records the order arrival times and volumes at each price level of thousands of stocks trading on the exchange. An LOB exhibits a number of distinctive characteristics [1–3] including

1. irregular time interval between arrivals
2. discrete state space of price ticks and volume lot sizes
3. intraday seasonality (more activities around market open and close)
4. arrival clustering
5. self excitation from its own history
6. cross (mutual) excitation from the history of other assets
7. long memory of excitation effect

Consequently, classical time series models with fixed time intervals such as ARIMA and GARCH are not suitable to model High Frequency (HF) financial data. A standard approach commonly used in practice is to re-sample the data in 5-minutes intervals [4, 5], thereby avoiding the time scale for liquid stocks where many of the characteristics listed above can be observed, but this may amount to discarding more than 99% of the data for such stocks. On the other hand, Poisson processes, which are widely used in the market microstructure literature [6, 7], fail to depict the above prevalent features of HF data.

This survey paper, on the current research in HF financial data modeling, concentrates on the use of so-called Hawkes processes, a family of point processes designed to model self and cross excitation. In section 2, we offer an informal introduction to point processes while a more technical review can be found in Appendix A. All the presented material about point processes is thoroughly covered in major textbooks including our reference [8–12]. Sections 3 and 4 introduce Hawkes processes and their statistical inference; a brief history thereof is provided in Appendix B. Section 5 presents the applications of Hawkes processes to HF data modeling.

Point Processes

This section provides an informal introduction to point processes. A more rigorous treatment using random measures and martingale theory can be found in Appendix A.

A point process is a random countable set of points \( \{x_1, x_2, \ldots \} \) on a set \( X \). The \( \{x_n\} \)'s are \( X \)-valued random variables; they can have a highly non-trivial dependence structure. For example, \( \{x_n\} \)'s can be locations of earthquakes on \( X = \mathbb{R}^2 \), and if self excitation (property 5 above) is built into the model, which is the case with Hawkes processes as we will discuss, then \( \{x_n\} \)'s can exhibit random patterns of aftershocks surrounding major earthquakes.

Let \( N(A) \) be the number of points inside a region \( A \subseteq X \). If \( N(A) \) is known for all subsets \( A \) of \( X \), this essentially determines the locations of all the points and this representation is the principle tool to describe a point process. A point process \( N \) is called simple if each location has at most one point.
When $X$ is the positive half-line $\mathbb{R}_+$, this typically represents the time axis and the points are regarded as the times of event occurrences. In this case, $N(t) = N((0,t])$ denotes the number of occurrences at or before time $t$. For our purpose of modeling HF data, we will mainly deal with point processes on $\mathbb{R}_+$.

When an event happens at $t_n \in \mathbb{R}_+$, it may carry an additional information $y_n$ (its mark). For instance in finance, each order arrival is associated with an order quantity (volume); in seismology, each earthquake is reported with a magnitude. A point process with marks is called a marked point process (MPP). Let $Y$ be the mark space (i.e. $y_n \in Y$) and $A \subseteq Y$; then $N(t; A) = N((0,t] \times A)$ denotes the number of events that happened at or before time $t$ such that the marks fall within the set $A$. The ground process $N_e(t) = N((0,t] \times Y)$ is the count of all events in $(0,t]$ regardless of the marks. The marks of a MPP are called unpredictable if $y_n$ is independent of $\{(t_i, y_i)\}_{i \in n}$ and they are called independent if $y_n$ is independent of $\{(t_i, y_i)\}_{i \in n}^2$. A MPP $N$ is called simple if $N_e$ is simple.

If a point process has multiple occurrences in the same location, we can treat it as a simple marked point process. Many point processes can be modeled in terms of their stochastic intensities. For a nonlinear stochastic volatility in diffusions. One important use of stochastic intensity is to allow it to change according to how time. When one generalizes the intensities to become stochastic processes on their own, they play a role similar to the marginal process for the points of type $i$. The ground process only specifies the intensity without any restriction on the mark distribution.

Many point processes can be modeled in term of their stochastic intensities. A Hawkes process 

\[ \lambda(t) = \lim_{h \to 0^+} \frac{\mathbb{E}(N((t-h,t])|\mathcal{F}_t-h)}{h} \quad (1) \]

where $\mathcal{F}_t$, which represents the information contained in all the events happened up to and including time $t$, denotes the filtration of $N$. For a multivariate point process, $\lambda_i(t)$ is the intensity of the marginal process $N_i(t)$.

The most well-known point processes are Poisson processes, where the intensities are deterministic functions of time. When one generalizes the intensities to become stochastic processes on their own, they play a role similar to stochastic volatility in diffusions. One important use of stochastic intensity is to allow it to change according to how events unfold over time. The Hawkes processes are directly exploring this feature for the purpose of modeling self excitation, as we are about to see.

### 3 Hawkes Processes

A Hawkes process [13] is a point process where its stochastic intensity has an autoregressive form. For a nonlinear multivariate marked Hawkes process, the intensity $\lambda(t) = (\lambda_1(t), \ldots, \lambda_d(t))$ is given by

\[ \lambda_i(t) = \Phi_i \left( \sum_{j=1}^{d} \int_{(\infty,t)\times Y} \gamma_{ij}(t-s,y)N_j(ds \times dy), t \right) = \Phi_i \left( \sum_{t_n \leq t} \gamma_{iy_n}(t-t_n,y_n), t \right) \quad (2) \]

where $w_n \in \{1, \ldots, d\}$ denotes the type of $t_n$ and $\Phi_i$ is known as rate function. Consider the special case

\[ \lambda_i(t) = \mu_i(t) + \sum_{j=1}^{d} \int_{(\infty,t)\times Y} \gamma_{ij}(t-s,y)N_j(ds \times dy) = \mu_i(t) + \sum_{t_n \leq t} \gamma_{iy_n}(t-t_n,y_n) \quad (3) \]

\[ \mu_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+, \quad \gamma_{ij} : \mathbb{R}_+ \times Y \rightarrow \mathbb{R}_+, \quad N_j : \mathcal{B}(\mathbb{R}_+ \times Y) \rightarrow \mathbb{N} \]

1. We may refer to the ground process of a MPP without the subscript “$g$” if the meaning is clear from the input parameter.
2. Notice that if marks are independent, future location $t_{n+1}$ cannot depend on previous mark $y_n$.
3. See Appendix A for formal definition and existence condition of stochastic intensity.
4. Notice that some authors use $\gamma_{yi}$, so that the first index is the source type and the second index is the destination type.
5. Hawkes process only specifies the intensity without any restriction on the mark distribution.
i.e. \( \Phi_i(x,t) = \mu_i(t) + x \). Such a Hawkes process determined by (3) is called linear, and \( \mu_i(t) \) is called the base or background rate. The function \( \gamma_j \) is called (marked) decay/exciting/fertility kernel and often \( \gamma_j(t,x) \) takes the separable form \( \gamma_j(t) g_{ij}(x) \) where \( g_{ij} \) is called mark impact kernel. Popular choices of decay kernel \( \gamma_j(t) \) include exponential \( \alpha_i e^{-\beta_i t} \) [13], power law \( \alpha_i (c_i + t)^{-\beta_i} \) [14] or Laguerre-type polynomial \( \sum_{k=0}^{K} \alpha_{ijk} t^k e^{-\beta_{ij} t} \) [15].

If the decay function is exponential with \( \beta_j = \beta_i \), the intensity \( \lambda(t) \) and the vector \( (N(t), \lambda(t)) \) are both Markov processes [6, 16, 17]. Moreover, provided that \( \mu_i(t) = \mu_i \), then \((\lambda_1(t),...\lambda_d(t))\) satisfies the system of stochastic differential equations (SDE)

\[
d\lambda_i(t) = \beta_i (\mu_i - \lambda_i(t))dt + \sum_{j=1}^{d} \alpha_{ij} dN_j(t) \tag{4}
\]

This specification has the simple interpretation that the events of \( N_j \) which happened just before time \( t \) increase the intensity \( \lambda_i(t) \) by \( \alpha_{ij} \geq 0 \) and thus trigger further events. Yet if the intensity \( \lambda_i(t) \) is higher than \( \mu_i \), the first term becomes negative (\( \beta_i > 0 \)) and draws the intensity back to the equilibrium level \( \mu_i \). In other words, the intensity \( \lambda_i(t) \) is a mean-reverting process driven by its own point process. The Markov property and this intuitive interpretation may explain why the exponential decay kernels are so widely used.

For linear Hawkes processes, \( \mu_i(t), \gamma_j(t) \) and \( g_{ij}(t) \) must be non-negative for all \( t \), in order to ensure the positivity of \( \lambda_i(t) \). As a result, unlike nonlinear Hawkes processes, linear Hawkes processes cannot model inhibitory effect (negative excitation). Nonetheless, the linear Hawkes processes are easier to handle, their properties are better understood and most importantly, they have a branching structure representation, which is extremely useful in simulation, estimation and interpretation of the models.

### 3.1 Branching Structure Representation

Linear Hawkes processes have a very elegant branching structure representation [18]. We describe here the version for the multivariate Hawkes processes with unpredictable marks [19].

There are \( d \) types of immigrants arriving according to Poisson processes with rates \( \mu_1,...,\mu_d \). Each individual (descendant or immigrant) will carry an unpredictable mark when born or arrived. An individual of type \( j \) born at time \( t_n \) with mark \( y_n \) will give birth to an individual of type \( i \) according to a non-homogeneous Poisson process with rate \( \gamma_i(t-t_n,y_n) \). All the non-homogeneous Poisson processes are independent of one another.

Let \( N_i(t) \) be the total number of individuals of type \( i \) born/arrived at or before time \( t \) under the above scenario, then \( N(t) = (N_1(t),...N_d(t)) \) will follow the linear marked Hawkes process (3). This representation forms the basis of the Expectation Maximization (EM) algorithm in Section 4.2.2 and we will also see how it is used to measure the endogeneity of a point process in Section 5.4.

### 3.2 Stationarity

Considering a Hawkes process \( N \) with intensity (2) such that \( \Phi_i(x,t) = \Phi_i(x), N \) has an unique stationary version\(^7\) if either of the following conditions are satisfied [18, 20]:
1. \( \Phi_i(x) \) is \( k \)-Lipschitz\(^6\) and the spectral radius\(^8\) \( \rho(A) < 1 \) for the \( d \times d \) matrix \( A = [\gamma_j(t)]_{i,j} \)
2. \( \Phi_i(x) \) is Lipschitz, \( \Phi_i(x) \leq M, \int_0^\infty |\gamma_j(t)|dt < \infty \) and \( \int_0^\infty t |\gamma_j(t)|dt < \infty \)

Technically speaking, \( N \) may have other non-stationary versions together with the stationary one; however, the non-stationary version will converges weakly to the stationary version when \( t \to \infty \) (see [21] for exact meaning). Since the Hawkes process starts at \( -\infty \), \( N((0,t]) \) will have the stationary distribution for all \( t > 0 \).

For the case of an exponential decay kernel \( \alpha e^{-\beta t} \), we have a simpler result. Let \( A = [\int_0^\infty \alpha_{ij} e^{-\beta_{ij} t}dt]_{i,j} = [\alpha_{ij}/\beta_{ij}]_{i,j} \), then \( N \) has an unique stationary version under either of the following conditions [22]:
1. \( \Phi_i(x) = \mu_i + x, \alpha_{ij} \geq 0, \beta_{ij}, \mu_i > 0, \rho(A) < 1 \) (linear Hawkes process)
2. \( \Phi_i(x) = \max(\mu_i + x, \varepsilon_i), \alpha_{ij} \in \mathbb{R}, \beta_{ij}, \mu_i > 0, \varepsilon_i > 0, \rho(A) < 1 \) (T-Hawkes process)
3. \( \Phi_i(x) = \min(\mu_i + \exp(x), M_i), \alpha_{ij} \in \mathbb{R}, \beta_{ij} > 0, M_i > \mu_i > 0 \) (E-Hawkes process)

\(^6\)\( N \) itself is not a Markov process as its intensity at time \( t \) depends on its full history before time \( t \).

\(^7\)See Appendix A for definition of stationarity of point processes.

\(^8\)\( f : \mathbb{R} \to \mathbb{R} \) is called \( k \)-Lipschitz \((k>0)\) if \( |f(x) - f(y)| \leq k|x - y| \forall x, y \in \mathbb{R} \).

\( \rho(A) = \max(|\pi_i|), \{\pi_i\} \) are eigenvalues of \( A \).
For the univariate linear case with $\mu = 0$, if there exists $r, R > 0, c \in (0, 1/2)$ such that $\int_0^\infty \gamma(t) dt = 1$, $\sup_{t \geq 0} t^{1+c} \gamma(t) \leq R$, $\lim_{t \to \infty} t^{1+c} \gamma(t) = r$, Brémaud and Massoulié [23] show that there exists a unique stationary non-trivial Hawkes process having such an intensity and he calls it the critical Hawkes process or Hawkes process without ancestors ($\mu = 0$).

3.3 Convergence

In this section, we state the various results about the convergence of Hawkes processes. A properly scaled linear Hawkes process will converge weakly to a Brownian diffusion when the spectral radius of decay functions' $L^1$-norm is less than one [24]. When the spectral radius is close to one in a certain sense, it converges to the integrated Cox-Ingersoll-Ross (CIR) process [25]. For the non-linear Hawkes processes, we only have the result for the univariate case and the sufficient conditions depends on the Lipschitz constant of $\Phi$ [26].

3.3.1 Law of Large Numbers for Multivariate Linear Hawkes processes

Assuming the model (3) without marks, if the spectral radius $\rho(A) < 1$ where $A = [\int_0^\infty \gamma_j(t) dt]_{i,j}$, then [24]

$$
\sup_{t \in [0,1]} \left\| \frac{N(mt)}{n} - t(I_d - A)^{-1} \mu \right\| \xrightarrow{a.s./L^2, n \to \infty} 0 \quad 10, 11
$$

where $\mu = (\mu_1, ..., \mu_d)$. When $d = 1$ and we take $t = 1$, it implies

$$
\frac{N(T)}{T} \xrightarrow{a.s./L^2, t \to \infty} \mu \int_0^\infty \gamma(t) dt \quad 6
$$

3.3.2 Functional Central Limit Theorem for Multivariate Linear Hawkes processes

Assuming the model (3) without marks, $N = (N_1, ..., N_d)$, if the spectral radius $\rho(A) < 1$ where $A = [\int_0^\infty \gamma_j(t) dt]_{i,j}$ and $\int_0^\infty \sqrt{\gamma_j(t)} dt < \infty \forall i, j$, then [24]

$$
\sqrt{n} (N(\bullet n)/n - \bullet (I_d - A)^{-1} \mu) \xrightarrow{weak, n \to \infty} (I_d - A)^{-1} \Sigma^{1/2} W(\bullet) \quad 12
$$

$$
\Sigma = \text{diag}((I_d - A)^{-1} \mu) \quad 13
$$

$W$ is standard $d-$ dimensional Brownian Motion

3.3.3 Functional Central Limit Theorem for Univariate Non-linear Hawkes processes

Assuming the model (2) without marks and $d = 1$, if $\gamma(t)$ is decreasing, $\int_0^\infty t \gamma(t) dt < \infty$, $\Phi(x, t) = \Phi(x)$ is increasing and $k$-Lipschitz, $\int_0^\infty k \gamma(t) dt < 1$ then [26]

$$
\sqrt{n} (N(\bullet n)/n - \bullet v) \xrightarrow{weak, n \to \infty} \sigma W(\bullet) \quad 9
$$

$$
\sigma^2 = \mathbb{E}((N(0, 1) - v)^2) + 2 \sum_{n=1}^\infty \mathbb{E}((N(n, 1) - v)(N(n+1, 1) - v)), v = \mathbb{E}(N(0, 1)) \quad 10
$$

3.3.4 Convergence of Nearly Unstable Univariate Linear Hawkes processes

Considering the linear model (3) without marks and $d = 1$, $N(T)/T$ converges to $\mu/(1 - \int_0^\infty \gamma(t) dt)$ when $\int_0^\infty \gamma(t) dt < 1$ by (6), while it explodes when $\int_0^\infty \gamma(t) dt = 1$. However, Jaisson and Rosenbaum [25] find that the properly scaled

10A sequence of random variables $X_n \xrightarrow{a.s./L^2, n \to \infty} X$ if $\mathbb{P}(|\lim_{n \to \infty} X_n - X| = 1) = 0$

11A sequence of probability measure $\mu_n \xrightarrow{L^2, n \to \infty} \mu$ if $\mathbb{E}(|X_n - X|^2) = 0$

12A sequence of probability measure $P_n$ converges weakly to $P$ if $\int_a f dP_n \to \int_a f dP$ for all bounded continuous function $f$. A sequence of stochastic process $X_n : \Omega \to D[0, 1]$ converges weakly (in distribution) to $X$ if the law of $X_n = (P_n \circ X^{-1})$ converges weakly to law of $X$ in the sense of probability measure, $D[0, 1]$ is the Skorokhod space of càdlàg (right continuous with left limits) functions (see [27, 28]).

$v \in \mathbb{R}^d, \text{diag}(v) = [a_{ij}]_{d \times d}, a_{ii} = v_i, a_{ij} = 0 \forall i \neq j.$
Hawkes process converges to the integrated Cox-Ingersoll-Ross (CIR) process when one has a sequence of decay kernel \( \psi^{(n)}(t) \) whose integral converges to one at the speed of \( n^{-1} \) (see (13)). More precisely, let

\[
\lambda^{(n)}(t) = \mu + \int_{(0,t]} \psi^{(n)}(t-s) dN^{(n)}(s), \quad \mu > 0, \quad \psi^{(n)}(t) = \alpha^{(n)} \gamma(t)
\]

(11)

\( \gamma: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \quad \int_0^\infty \gamma(t) dt = 1, \quad \int_0^\infty t \gamma(t) dt = m < \infty, \quad \int_0^\infty |\gamma(t)| dt < \infty, \quad \sup_{t \in [0,\infty)} |\gamma(t)| < \infty \)

(12)

\( \alpha^{(n)} \in (0,1), \quad \lim_{n \to \infty} \alpha^{(n)} = 1, \quad \lim_{n \to \infty} n(1-\alpha^{(n)}) = c > 0 \)

(13)

\[
\psi^{(n)}(t) = \sum_{k=1}^\infty \psi^{(n)\otimes k}(t), \quad R^{(n)}(t) = \frac{n\psi^{(n)}(nt)}{\int_0^\infty \psi^{(n)}(t) dt}, \quad |R^{(n)}(t)| \leq M \forall n \forall t
\]

(14)

where \( f^{\otimes k} \) denotes the k-fold self-convolution of \( f \). If the sequence of Hawkes process \( N^{(n)} \) has intensity \( \lambda^{(n)} \) satisfying (11-14), then the scaled intensity converges to the CIR process and the scaled Hawkes process converges to the integrated CIR process [25] as follows:

\[
(1-\alpha^{(n)})\lambda^{(n)}(u) \xrightarrow{\text{weak}} X(u) \quad n \to \infty
\]

(15)

\[
(1-\alpha^{(n)})\frac{N^{(n)}(u)}{n} \xrightarrow{\text{weak}} \int_0^u X(s) ds \quad n \to \infty
\]

(16)

\[
dX = \frac{c}{m} (\mu - X_t) dt + \sqrt{\frac{c}{m}} \sqrt{X_t} dW_t, \quad X_0 = 0
\]

(17)

### 4 Statistical Inference of Hawkes Processes

#### 4.1 Simulation

In this section, we will give an overview of the algorithms which simulate Hawkes processes. Assume we know all the parameters in the functional form of \( \mu(t) \) and \( \gamma(t,y) \), our goal is to simulate the points \((t_1, y_1), (t_2, y_2), \ldots\) on the interval \([0,T]\).

If the marks distribution depends only on \( t_n \), we can simply generate \( y_n \) conditioned on the generated \( t_n \). Next, \( t_{n+1} \) can be generated from the intensity \( \lambda(t) \) for \( t > t_n \) which depends on \( \{(t_1, y_1), \ldots, (t_n, y_n)\} \). If the distribution of \( y_n \) also depends on \( \{(t_{n-1}, y_{n-1}), (t_{n-2}, y_{n-2}), \ldots\} \), the algorithms can be modified accordingly.

##### 4.1.1 Inverse Transform

The first simulation algorithm for Hawkes processes appears in Ozaki [29]. Suppose the intensity is governed by the univariate Hawkes model in (3). Let \( t_n \) be the arrival time and \( \tau_n = t_{n+1} - t_n \) be the inter-arrival time. By (63), \( \lambda(t) = \frac{h_n(t) - h_n(t_{n-1})}{\tau_n} \) for \( t \in (t_{n-1}, t_n) \) where \( h_n(t) = g_n(t)/(1 - G_n(t^-)) \) and \( g, G \) are the conditional pdf, cdf of \( \tau_n \) given \( \mathcal{F}_{t_{n-1}} \). If \( G_n(t) \) is continuous, \( h_n(t) \) is simply the hazard function, and it can be shown that

\[
G_n(\tau_n) = 1 - \exp \left( -\int_0^{\tau_n} h_n(s) ds \right) = 1 - \exp \left( -\int_{t_{n-1}}^{t_{n-1}+\tau_n} \lambda(s) ds \right)
\]

(18)

Given \( \tau_{n-1} \), we can generate \( t_n = t_{n-1} + \tau_n \) by inverse cdf transform \( \tau_n = G_n^{-1}(U), \ U \sim \text{Unif}(0,1) \). However, the inversion needs to be done numerically, so this method is largely superseded by Ogata’s modified thinning which we now discuss.

##### 4.1.2 Ogata’s Modified Thinning

Ogata [30] introduces the modified thinning method which does not require numerical inversion. The algorithm is based on the following theorem. Let \( N = (N_1, \ldots, N_d) \) be a multivariate point process with intensity \((\lambda_1, \ldots, \lambda_d)\) such that \( \sum_{i=1}^d \lambda_i(t) \leq \lambda^*(t) \) \( \forall t \) a.s. \( \lambda^*(t) \) is an exogenously chosen deterministic rate function and \( N^v \) is the univariate...
non-homogeneous Poisson process with intensity $\lambda^*(t)$. If each point $t_n$ in $N^*$ is given a mark $y_n$ such that $P(y_n = i) = \lambda_i(t_n)/\lambda^*(t_n)$, $i = 1, \ldots, d$, then $(N_1^*, \ldots, N_d^*)$ has the same distribution as $(N_1, \ldots, N_d)$.

The following algorithm generates a $d$-dimensional multivariate Hawkes process such that $\lambda_i(t)$ is decreasing between points and $|\lambda_i(t) - \lambda_i(t^-)| \leq \alpha_i \forall t$.

\textit{Ogata’s Modified Thinning [30]}
1. $n = 1$, $t_0 = 0$
2. Generate $\tau_n \sim \text{Exp}(\lambda_n^*)$ for some $\lambda_n^* \geq \sum_{i=1}^{d} (\lambda_i(t_{n-1}) + \alpha_i)$
3. Let $t_n = t_{n-1} + \tau_n$
4. Generate $U_n \sim \text{Unif}(0,1)$
5. if $U_n \in \left(\frac{\sum_{i=1}^{d} \lambda_i(t_n) / \lambda_n^*, \sum_{i=1}^{d} \lambda_i(t_n) / \lambda_n^*}\right)$ for some $k \in \{1, \ldots, d\}$ return $t_n$ and the point is of type $k$, else discard $t_n$ (but keep the value for use in next generation step 2, 3)
6. $n = n + 1$, goto step 2

4.1.3 Simulation by Branching Structure

This method generates points using the branching structure representation of linear marked Hawkes processes. Type-$j$ immigrants arrive according to a non-homogeneous Poisson process with rate $\mu_j(t)$. Next, the type-$j$ parent arriving at $t_n$ with mark $y_n$ produces type-$i$ descendants according to non-homogeneous Poisson process with rate $\gamma_j(t-t_n,y_n)$ and the generation is repeated for each descendant until all of them exceed the pre-defined time $T$. Since all the non-homogeneous Poisson processes are independent, the generations can be done in parallel.

\textit{Simulation by the Branching Structure [31]}
1. Generate non-homogeneous Poisson processes with intensities $\mu_i(t)$, $i = 1, \ldots, d$ on $[0, T]$
2. For each points $t_n$, generate $y_n|t_n$
3. Suppose $t_n$ is of type $j$, generates type-$i$ descendants according to non-homogeneous Poisson process with intensity $\gamma_j(t-t_n,y_n)$ on $[t_n, T]$, $i = 1, \ldots, d$
4. repeat step 2, 3 for all descendants

The non-homogeneous Poisson process with intensity $\mu(t)$ on $[0, T]$ can be generated using Lewis’ thinning algorithm [32]
1. generate $N \sim \text{Poisson}(\mu^*)$ for some $\mu^* \geq \max_{t \in [0,T]} \mu(t)$
2. generate $U_n \sim \text{Unif}(0,1)$, $n = 1, \ldots, N$
3. $T_n = U_n \cdot T$, $n = 1, \ldots, N$ ($\{U_n\}$ is the order statistics of $\{U_n\}$)
4. generate $V_n \sim \text{Unif}(0,1)$, $n = 1, \ldots, N$
5. return $T_n$ if $V_n \leq \mu(T_n)/\mu^*$, $n = 1, \ldots, N$, otherwise discard $T_n$

4.2 Estimation

Suppose we observe a point process on $[0, T]$ and collect the event times and marks $\{(t_1, y_1), \ldots, (t_N, y_N)\}$, now we would like to estimate the functions $\mu(t)$ and $\gamma(t, y)$ in the intensity $\lambda(t)$ which drives the process $N(t)$. We will summarize the various methods appearing in the literature, but so far the focus is on unmarked processes. In the special case where the marks are independent and identically distributed (IID), the mark distribution can be estimated separately from the point process.

If we assume $\mu(t)$ and $\gamma(t)$ have some parametric representations, we can use Maximum Likelihood Estimation (MLE), Expectation Maximization (EM), or Generalized Method of Moments (GMM) to estimate the parameters. Otherwise, we need to rely on some advanced non-parametric techniques to estimate the whole function curves.

4.2.1 Maximum Likelihood Estimation (MLE)

The log-likelihood of a Hawkes process is given by [29]

$$
\log(L(\theta)) = \sum_{i=1}^{d} \left( - \int_0^T \lambda_i(t; \theta) dt + \int_0^T \log(\lambda_i(t; \theta)) dN_i(t) \right) 
$$

(19)
In the case of multivariate linear Hawkes process, it becomes

$$\log(L(\theta)) = - \sum_{n=1}^{N} \sum_{i=1}^{d} \int_{t_n}^{T} \left( \sum_{i=1}^{d} \mu_i(t; \theta) \right) dt - \sum_{n=1}^{N} \int_{t_n}^{T} \left( \sum_{i=1}^{d} \gamma_{i,w_n}(t - t_n; \theta) \right) dt + \sum_{n=1}^{N} \log \left( \mu_{w_n}(t_n; \theta) + \sum_{t_{m} < t_n} \gamma_{w_n,w_m}(t_n - t_m; \theta) \right) \quad (20)$$

The parameters $\theta$ can be estimated by maximizing the log-likelihood. However, the numerical optimization is problematic as the log-likelihood function is usually quite flat (see [33, fig.2.3]) and may have a lot of local maxima (see [33, fig.4]).

### 4.2.2 Expectation Maximization (EM)

For linear Hawkes process, the estimation can also be done via Expectation Maximization (EM) [34, 35] as in [33, 36–39]. EM is a variant of MLE where part of the data is missing. In the branching structure representation, the missing data is the parents which produce the descendants. Let $z_n$ denotes the index of the parent of $t_n$ and $w_{z_n}$ represents the type of the parent of $t_n$. If $z_n = m$ and $w_{z_n} = j$, that means $t_n$ is produced by the type $j$ point $t_m$. When $z_n = 0$, $t_n$ is an immigrant. Also we define $w_0 = 0$, $\gamma_0(t) = \mu(t)$ and $t_0 = 0$ to simplify the expression. Suppose $\{t_n, w_n, z_n\}$ are known, since each generation is an independent Poisson process, the complete data log-likelihood is

$$\log(L(\theta)) = \sum_{n=0}^{N} \left\{ - \int_{t_n}^{T} \gamma_{w_n}(t - t_n; \theta) dt + \sum_{t_{m} > t_n} \log(\gamma_{w_n}(t_m - t_n; \theta)) \mathbb{I}(z_m = n) \mathbb{I}(w_m = i) \right\} \quad (21)$$

$$Q(\theta|\theta^{(k)}) = \mathbb{E}^{\theta^{(k)}}(\log(L(\theta)) | \{(t_k, w_k)\})$$

$$= \sum_{n=0}^{N} \sum_{i=1}^{d} \left\{ - \int_{t_n}^{T} \gamma_{w_n}(t - t_n; \theta) dt + \sum_{t_{m} > t_n} \log(\gamma_{w_n}(t_m - t_n; \theta)) \mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i) \right\} \quad (22)$$

$$\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i) = \frac{\gamma_{w_n}(t_m - t_n; \theta^{(k)}) \mathbb{I}(w_m = i)}{\sum_{m=1}^{M} \gamma_{w_n}(t_m - t_n; \theta^{(k)})} \quad (23)$$

The EM algorithm can be implemented as follows:

1. $k = 0$ and choose an initial guess $\theta^{(0)}$
2. E-step: compute $Q(\theta|\theta^{(k)}) = \mathbb{E}^{\theta^{(k)}}(\log(L(\theta)) | \{(t_k, w_k)\})$
3. M-step: compute $\theta^{(k+1)} = \text{argmax}_\theta Q(\theta|\theta^{(k)})$
4. $k = k + 1$, repeat E-step and M-step until $\theta^{(k)}$ converges (e.g. $||\theta^{(k+1)} - \theta^{(k)}|| < \varepsilon$)

In general, the optimization in M-step need to be solve numerically but when the decay kernel has the exponential form $\alpha_{ij} \beta_{ij} \exp(-\beta_{ij}t)$, Olson and Carley [39] suggest a closed form approximate iteration.

$$\mu_{i}^{(k+1)} = \sum_{m=1}^{M} \mathbb{P}^{\theta^{(k)}}(z_m = 0|\{(t_k, w_k)\}) \mathbb{I}(w_m = i) \quad \frac{T}{T}$$

$$\alpha_{ij}^{(k+1)} = \sum_{m=1}^{M} \sum_{m+n+1}^{N} \mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i, w_n = j) \quad \sum_{n=1}^{N} \mathbb{I}(w_n = j) \quad (25)$$

$$\beta_{ij}^{(k+1)} = \frac{\sum_{n=1}^{N} \sum_{m+n+1}^{N} \mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i, w_n = j)}{\sum_{n=1}^{N} \sum_{m+n+1}^{N} \mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i, w_n = j)} \quad (26)$$

$$\mathbb{P}^{\theta^{(k)}}(z_m = n|\{(t_k, w_k)\}) \mathbb{I}(w_m = i, w_n = j) = \frac{\alpha_{ij}^{(k)} \beta_{ij}^{(k)} \exp(-\beta_{ij}^{(k)}(t_m - t_n)) \mathbb{I}(w_m = i, w_n = j)}{\mu_{i}^{(k)} + \sum_{l=1}^{M} \alpha_{i,l}^{(k)} \beta_{i,l}^{(k)} \exp(-\beta_{i,l}^{(k)}(t_m - t_l))} \quad (27)$$

$$\mathbb{P}^{\theta^{(k)}}(z_m = 0|\{(t_k, w_k)\}) \mathbb{I}(w_m = i) = \frac{\mu_{i}^{(k)} \mathbb{I}(w_m = i)}{\mu_{i}^{(k)} + \sum_{l=1}^{M} \alpha_{i,l}^{(k)} \beta_{i,l}^{(k)} \exp(-\beta_{i,l}^{(k)}(t_m - t_l))} \quad (28)$$
In addition, the summation \( \sum_{i=1}^{m-1} \alpha_i^{(k)} \beta_i^{(k)} \exp(-\beta_i^{(k)}(t_m-t_i)) \) can be truncated after \( \exp(-\beta_i^{(k)}(t_m-t_i)) \) has decayed to a small value. The speed of EM is reported to be 10-100 times faster than MLE and more importantly, MLE does not converge within 500 iterations in practically all test cases while EM does [39].

4.2.3 Generalized Method of Moments (GMM)

Another method for statistical estimation apart from MLE is the **Generalized Method of Moments** [40]. The idea is to find the parameters which minimize the difference between theoretically moments (see Appendix B) in terms of the unknown parameters and the empirical moments computed directly from the data. If we have more moments than the number of parameters, the method involves solving a weighted least squares problem.

Da Fonseca and Zaatour [41] obtain the analytic moment expressions by restricting the process to be univariate with exponential kernel and making use of the Markov property in this special case. The authors claim that this method is extremely fast but no speed comparison result is provided.

4.2.4 Nonparametric Estimation

Without assuming any parametric form for \( \mu(t) \) or \( \gamma(t) \), some nonparametric methods are developed recently to estimate the whole base rate and decay kernel functions. Similar to parametric estimation, penalized MLE or GMM is used to find the function with desirable characteristics (e.g. smooth functions, sparse coefficients). Nonetheless, the nonparametric method, which involves finding the unknown functions in infinite-dimensional spaces, requires extensive computational effort and the underlying statistical construction is usually much more involved than the parametric counterpart.

To the best of our knowledge, the first attempt in nonparametric estimation of Hawkes process is by Gusto and Schbath [42] in 2005. The authors express the kernel function of the multivariate Hawkes process using B-splines [43] with equally spaced knots. The log-likelihood function involving the basis coefficients are then maximized numerically to find the parameters which minimize the difference between theoretically moments (see Appendix A) in terms of the empirical moments computed directly from the data. If we have more moments than the number of parameters, the method involves solving a weighted least squares problem.

Another method for statistical estimation apart from MLE is the **Generalized Method of Moments** [40]. The idea is to find the parameters which minimize the difference between theoretically moments (see Appendix A) in terms of the unknown parameters and the empirical moments computed directly from the data. If we have more moments than the number of parameters, the method involves solving a weighted least squares problem.

4.3 Hypothesis Testing

4.3.1 Random Time Change

The classical method to test the goodness-of-fit of a point process model on \( \mathbb{R}_+ \) is Ogata’s residual analysis [14]. The residual process is defined as the difference between the observed events and the expected events under the assumed model. The residual process can be used to test the goodness-of-fit of a point process model. The residual process is defined as the difference between the observed events and the expected events under the assumed model. The residual process can be used to test the goodness-of-fit of a point process model.

The residual process is defined as the difference between the observed events and the expected events under the assumed model. The residual process can be used to test the goodness-of-fit of a point process model. The residual process is defined as the difference between the observed events and the expected events under the assumed model. The residual process can be used to test the goodness-of-fit of a point process model. The residual process is defined as the difference between the observed events and the expected events under the assumed model. The residual process can be used to test the goodness-of-fit of a point process model.

\[ \hat{\lambda}(t) = \frac{dN(t)}{dt} \]

where \( \lambda(t) \) is the intensity function and \( N(t) \) is the number of events up to time \( t \).

Ogata calls \( \{ \tilde{t}_n = \int_0^{t_n} \hat{\lambda}(s)ds \} \) the residual process and according to the random time change theorem (see Appendix A), the residual process should be close to a standard Poisson process if the estimated intensity \( \hat{\lambda}(t) \) is close to the true intensity \( \lambda(t) \). The hypothesis that \( \{ \tilde{t}_n \} \) is a standard Poisson process can be tested by the following methods:

1. QQ Plot [56] of \( \{ \tilde{t}_n - \tilde{t}_{n-1} \} \) vs \( \text{Exp}(1) \).
2. Kolmogorov-Smirnov Test [57–59] to test \( \tilde{t}_n \sim \text{Exp}(1) \).
3. Ljung-Box Test [60] to test the lack of serial correlation of \( \{ \tilde{t}_n \} \).

---

14 Although GMM is consistent under some mild regularity conditions, unlike MLE, it is not asymptotic efficient among the class of consistent estimators.

15 The terminology is not standard, Baddeley et al. [55] refer \( \{ N(t_n) - \int_0^{t_n} \hat{\lambda}(s)ds \} \) as residual in order to extend the concept to higher dimension.
4.3.2 Approximate Thinning

Another method to test goodness-of-fit is by thinning, which doesn’t require integration of the intensity function. It is useful if the intensity function is estimated non-parametrically. However, the thinned residual process is only approximately a Poisson process.

By Ogata’s modified thinning [30], we know that if there exists \( b > 0 \) such that \( b \leq \lambda(t) \forall t \) and we keep point \( t_n \) with probability \( b/\lambda(t_n) \), the thinned point process is a homogeneous Poisson process with rate \( b \). However, the infimum \( b \) of the intensity function is often close to 0, making the number of points in the thinned process very small and the test to have little power. A remedy is to use approximate thinning [61] as follows: choose an integer \( k \ll N \), select one point from \( \{t_1, \ldots, t_N\} \) with probability of selecting \( t_n \) proportional to \( \lambda(t_n)^{-1} \). Repeat the selection (without replacement) until \( k \) points are selected. The resulting \( k \) points will be approximately a homogeneous Poisson process.

5 Applications of Hawkes processes

After the groundwork of basic theory and statistical inference for Hawkes processes, we now unleash their power to model HF data. First, the readers are reminded how diverse the notion of stock trading frequency can be. According to the Trade And Quote database (TAQ), between 9:30am to 4:00pm on May 2, 2014, there were 11 million quote changes (limit + cancellation + market orders) and 0.3 million trades (market orders) for SPDR S&P 500 ETF (SPY). In other words, on average there are 460 quote changes and 13 trades per second. If we take a snapshot every 5 minutes as in [4, 5], we will only use 0.03% of trade data and 0.0007% of quote data. In comparison, Pathfinder Bancorp (PBHC) only has 306 quote changes and 11 trades on the the same day, which means there is a 35 minutes lag between trades on average and thus the 5 minutes snapshots will just give a series of repeated information. Regardless of the sampling frequency, we are likely to get some misleading result if we analyze the asynchronous data from a portfolio of liquid and illiquid stocks using models with fixed intervals.

The construction of multivariate point processes shows that each variate can have a completely different arrival intensity \( \lambda_i(t) \). Nonetheless, the multivariate Hawkes process can still model the dependence structure easily via the \( \gamma_{ij}(t) \)'s, which are estimated by duly considering all the asynchronous data in the highest frequency without any re-sampling.

Order arrivals and price changes are unarguably two of the most important elements in HF trading. Using Hawkes processes, we can estimate their conditional distributions based on all the historical HF asynchronous data, enabling us to give a more accurate real time prediction of future event occurrences. In the following subsections, we are going to highlight some of the literature which take advantage of Hawkes processes to model HF data.

5.1 Modeling Order Arrivals

Bowsher [22]\textsuperscript{16} is the first to use Hawkes processes to model order arrivals. He uses nonlinear Hawkes processes to allow for inhibitory effect and he considers two rate functions \( \Phi_i(x,t) = \mu_i(t) + \exp(x) \) and \( \Phi_j(x,t) = \max(\mu_j(t) + x, \varepsilon_j) \), \( \varepsilon_j > 0 \), where both of them guarantee that the stochastic intensity will be strictly positive at all times. For the deterministic base rate \( \mu_i(t) \), he exploits a piecewise linear function with knots at 9:30, 10:00, 11:00,...,16:00 while the decay kernel is the exponential function without marks. In addition, an extra term is included to represent the spillover effects from the previous trading day.

Bowsher uses Maximum Likelihood Estimation (MLE) to estimate the parameters for the bivariate point process of trade and quote of General Motor (GM), trading on NYSE between 5 July 2000 to 29 August 2000. The model is found to be decent according to the goodness-of-fit test using random time change.

Instead of modeling arrivals of all trades and quotes, Large [62] uses Hawkes processes to model only the arrivals of aggressive orders, which are market orders depleting the queue and limit orders falling inside the bid-ask spread, in order to study the resiliency of the LOB. A LOB is called resilient if it reverts to its generic shape promptly after large trades. The idea is that when a large trade causes the bid-ask spread to widen, the arrival intensity of aggressive limit orders in a resilient LOB will surge so that the gap will be filled very quickly. In order words, the cross excitation effect \( \gamma_{ij}(t) \) from aggressive market orders to aggressive limit orders should be reasonably large in a resilient LOB.

In addition to market orders and limit orders, Large also includes the cancellations of limit orders as well as limit orders falling outside the best quotes. Therefore, he builds a 10-variate linear marked Hawkes process with exponential

\textsuperscript{16}Though Bowsher’s paper was published in 2007, the first draft appeared in 2002.
decay and mark impact kernel to fit the HF data of Barclays (BARC), trading on LSE between 2 Jan 2002 to 31 Jan 2002. The result shows that the widening of bid-ask spread indeed pumps up the intensities of aggressive limit orders, causing the gap to be filled very quickly and hence making the LOB resilient.

More examples of applications of Hawkes processes to order arrivals include the following papers: Muni Toke and Pomponio [63] use similar approach as Large [62] to model trades-through, namely market orders which deplete the best queues and consume at least one share in the second best. Muni Toke [64] designs a more realistic market simulator using Hawkes processes with exponential kernel for order arrivals. Hewlett [65] models the arrival of market orders with Hawkes processes for single period market making. Finally, Alfonsi and Blanc [66], Jaisson [67] tackle the problem of optimal execution with market orders following multivariate Hawkes processes.

5.2 Modeling Price Jumps

5.2.1 Single Asset

Traditionally the events of price jumps are modeled by Poisson processes, which suffer from the drawbacks mentioned in the introduction section. Again, Hawkes processes can be applied to model price jumps, which often delineate clustering, self and cross excitation behaviors.

Bacry et al. [68] use Hawkes processes to model the price jumps, resulting in a model which can reproduce the microstructure noise [69], Epps effect [70] and jump clustering, while maintaining the coarse scale limit of Brownian diffusion. In their model, the trade price $X(t)$ has the dynamics

$$X(t) = N_1(t) - N_2(t)$$

where $N(t) = (N_1(t), N_2(t))$ is a bivariate linear Hawkes process with exponential decay kernel. $N_1(t), N_2(t)$ represents the total number of upward and downward jumps respectively. The authors make additional assumptions that the Hawkes process $N$ has only cross excitation and coefficients are symmetric in order to simplify computation.

$$\lambda_1(t) = \mu + \int_{(0,t]} \gamma(t-s) dN_2(t), \quad \lambda_2(t) = \mu + \int_{(0,t]} \gamma(t-s) dN_1(t), \quad \gamma(t) = \alpha e^{-\beta t}$$

According to the model, when $X$ jumps up (down), $\lambda_2$ (resp. $\lambda_1$) increases, causing the probability of jumping down (resp. up) to increase. Such a cross linkage generates the effect of microstructure noise where the trade price is bouncing between best bid and best ask.

Due to the bid-ask bounce, it is well-known that the realized variance (annualized) increases when the sampling frequency increases [71].

$$V(\tau) = \mathbb{E} \left( \frac{1}{\tau} \sum_{n=0}^{T/\tau} (X((n+1)\tau) - X(n\tau))^2 \right)$$

$$= \frac{2\mu}{1 - \alpha/\beta} \left( \frac{1}{1 + \alpha/\beta} \frac{1 - \frac{1}{(1 + \alpha/\beta)^2}}{1 - \frac{1}{(1 + \alpha/\beta)^2}} e^{-(\alpha + \beta)\tau} \right)$$

$$\frac{\alpha}{(\alpha + \beta)\tau}$$

Such an effect can be easily demonstrated by computing the expected realized variance (32) of the jump model (29) and the result with $\mu = 0.16, \alpha = 0.024, \beta = 0.11$ is shown in Fig.1. The authors apply the model to Euro-Bund futures and find a very good fit between the observed and theoretical realized variance under this highly simplified model.

Let $Y(t) = X(nt)$, then $Y(t)$ is a coarse scale version of $X(t)$. For example, if $t$ in $X$ is in micro second and $n = 60000, \tau$ in $Y$ will be in minute. When we look at the trade price in a low frequency setting, Bacry et al. [24] show that the macroscopic Hawkes jump model goes back to the classical model of Brownian motion due to the functional central limit theorem for linear Hawkes process (7). Assuming that $\int_0^\infty \gamma(t)dt < 1$, then

$$\frac{X(n\bullet)}{\sqrt{n}} \xrightarrow[n \to \infty]{\text{weak}} \sigma W(\bullet), \quad \sigma^2 = \frac{2\mu}{(1 - \int_0^\infty \gamma(t)dt)(1 + \int_0^\infty \gamma(t)dt)^2}$$

It is interesting to see how the macroscopic variance $\sigma^2$ is related to the microscopic base rate $\mu$ and cross excitation $\gamma(t)$. As $\int_0^\infty \gamma(t)dt$ approaches 1, the variance goes to $\infty$.
Jaisson and Rosenbaum [25] extend the model of Bacry et al. [68] to the case of nearly unstable Hawkes process, where \( \int_0^\infty \gamma(t)dt \approx 1 \), by constructing a sequence of kernel functions whose integrals converge to one at the speed of \( n^{-1} \). They show that the properly scaled price process converges to Brownian diffusion with Heston stochastic volatility [72]. The full result is stated below.

\[
X^{(n)}(t) = N_1^{(n)}(t) - N_2^{(n)}(t) \tag{34}
\]

\[
\lambda_1^{(n)}(t) = \mu + \int_0^t \gamma_1^{(n)}(t-s)dN_1^{(n)}(s) + \int_0^t \gamma_2^{(n)}(t-s)dN_2^{(n)}(s) \tag{35}
\]

\[
\lambda_2^{(n)}(t) = \mu + \int_0^t \gamma_2^{(n)}(t-s)dN_1^{(n)}(s) + \int_0^t \gamma_1^{(n)}(t-s)dN_2^{(n)}(s) \tag{36}
\]

\[
\left( \int_0^t \gamma_1^{(n)}(t)dt + \int_0^\infty \gamma_2^{(n)}(t)dt \right) < 1, \quad \gamma_n^{(n)}(t) = \alpha(n) \gamma(t) \tag{37}
\]

\[
\gamma : \mathbb{R}_+ \rightarrow \mathbb{R}_+, \quad \int_0^\infty \left( \gamma_1(t) + \gamma_2(t) \right)dt = 1, \quad \int_0^\infty t(\gamma_1(t) + \gamma_2(t))dt = m < \infty \tag{38}
\]

\[
\int_0^\infty |\gamma(t)|dt < \infty, \quad \sup_{t \in [0,\infty)} |\gamma(t)| < \infty \tag{39}
\]

\[
\alpha(n)^{(n)} \in (0,1), \quad \lim_{n \to \infty} \alpha(n)^{(n)} = 1, \quad \lim_{n \to \infty} n(1 - \alpha(n)^{(n)}) = c > 0 \tag{40}
\]

\[
\psi^{(n)}(t) = \sum_{k=1}^\infty \left( \gamma_1^{(n)} + \gamma_2^{(n)} \right)^{\otimes k}(t), \quad \rho^{(n)}(t) = \frac{n\psi^{(n)}(mt)}{\int_0^\infty \psi^{(n)}(t)dt}, \quad |\rho^{(n)}(t)| \leq M \forall n \forall t \tag{41}
\]

Under the conditions of (34 - 41),

\[
\frac{X^{(n)}(t)}{n} \xrightarrow{\text{weak}} \frac{Y(t)}{n} \tag{42}
\]

\[
dY_t = \frac{\sqrt{\psi^2}}{1 - \int_0^\infty |\gamma(t)|dt}dW^1_t, \quad Y_0 = 0 \tag{43}
\]

\[
dV_t = \frac{c}{m} \frac{2\mu - V_t}{V_t} dt + \sqrt{\psi^2} \frac{1}{m} dW^2_t, \quad V_0 = 0 \tag{44}
\]

Conditions (34 - 36) is just a bivariate Hawkes model (with both self and cross excitation) but now we have a different \( \gamma^{(n)}(t) \) for each \( n \) that use to scale the time. The rest are the regularity conditions similar to the univariate.
nearly unstable Hawkes process (17) and the most important one is (40) which states that \( \alpha^{(n)} \) converges to one at the speed of \( n^{-1} \). However, the interesting result is that instead of converging to an integrated CIR, the price dynamics formed by the difference between two Hawkes processes converges to a stochastic volatility model.

### 5.2.2 Two Assets

To model the Epps effect, Bacry et al. [68] consider the two-asset case with prices \( (X_1(t), X_2(t)) \) given by

\[
X_1(t) = N_1(t) - N_2(t), \quad X_2(t) = N_3(t) - N_4(t)
\]

\[
\lambda_i(t) = \mu_i + \sum_{j=1}^{4} \alpha_{ij} \exp(-\beta(t-s))dN_j(s), \quad i = 1, \ldots, 4
\]

\( (N_1(t), \ldots, N_4(t)) \) is a 4-variate Hawkes process with exponential kernel where \( \beta_{ij} = \beta \). The coupling of excitation effects is constrained to have the form

\[
\alpha = \begin{pmatrix}
0 & \alpha_{12} & \alpha_{13} & 0 \\
0 & \alpha_{22} & 0 & \alpha_{23} \\
0 & \alpha_{33} & 0 & \alpha_{34} \\
0 & \alpha_{43} & \alpha_{44} & 0
\end{pmatrix}
\]

In this case, there is a closed form representation for the realized correlation, which vanishes when the sampling interval goes to zero (Epps effect).

If we assume \( \mu_1 = \mu_2, \mu_3 = \mu_4, \alpha_{12} = \alpha_{34} = 0, (\int_0^\infty \gamma_1(t)dt)(\int_0^\infty \gamma_1(t)dt) < 1 \), then the macroscopic bivariate asset prices converges to correlated Brownian diffusion [24]

\[
\frac{1}{\sqrt{n}} \begin{pmatrix} X_1(n) \\ X_2(n) \end{pmatrix} \xrightarrow{\text{weak}} \sqrt{2} \begin{pmatrix}
\sqrt{\int_0^\infty \gamma_1(t)dt} + \sqrt{\int_0^\infty \alpha_{13}(t)dt}W_1(t) \\
\sqrt{\int_0^\infty \alpha_{31}(t)dt}W_1(t) + \sqrt{\int_0^\infty \gamma_1(t)dt}
\end{pmatrix}
\]

\[
\nu_1 = \mu_1 + \left( \int_0^\infty \gamma_1(t)dt \right) \mu_3, \quad \nu_2 = \mu_3 + \left( \int_0^\infty \gamma_1(t)dt \right) \mu_1
\]

This convergence result gives us an explicit formula to estimate the macroscopic correlation from the asynchronous HF data.

As a final remark, under this jump representation, the observed trade price is not some hidden continuous fair value process plus some microstructure noise as in [73]. It is the result of the trading interactions between buyers \( N_1, N_3 \) and sellers \( N_2, N_4 \) on a fixed price grid. There is no such thing as HF volatility or correlation since prices are not diffusions but pure jump processes in the HF scale. Volatility and correlation are only meaningful when we look at the coarse scale diffusion approximation, but those low frequency representation parameters can be computed directly from the HF jump model characteristics.

### 5.3 Modeling Jump-Diffusion

Duffie et al. [74, 75] propose the affine jump-diffusion \( X(t) \), which has the following structure.\(^{17}\)

\[
dX(t) = (k_0(t) + k_1(t)X(t))dt + (h_0(t) + h_1(t)X(t))dW(t) + \zeta dN(t)
\]

\[
\lambda(t) = a_0(t) + a_1(t)X(t)
\]

The jump intensity \( \lambda(t) \) of \( N(t) \) is an affine function of \( X(t) \), which depends on the Brownian motion \( W(t) \) and the jump process \( N(t) \), with jump size \( \zeta \) drawn from a fixed distribution. When \( k_0 = \beta \theta, k_1 = -\beta, h_0 = h_1 = 0, a_0 = 0, a_1 = 1, \zeta = \alpha \), we can see that \( \lambda(t) = X(t) \) and \( d\lambda(t) = \beta(\theta - \lambda(t))dt + \alpha dN(t) \). Hence in this case, \( N(t) \) is the Hawkes process with exponential kernel.

\(^{17}\)We only show the one dimensional case for simplicity.
Zhu [76] derives some convergence results when $\zeta$ is a constant and the diffusion part is a CIR process.

$$dX(t) = \beta(\mu - X(t))dt + \sigma\sqrt{X(t)}dW(t) + \alpha dN(t)$$

$$\lambda(t) = a_0 + a_1 X(t)$$

Aït-Sahalia et al. [77] model the contagion of financial crisis with the Hawkes jump-diffusion where the price dynamic $X_i(t)$ is given by

$$dX_i(t) = \mu_i dt + \sqrt{V_i(t)}dW_i(t) + Z_i(t)dN_i(t)$$

$$dV_i(t) = \kappa_i(\theta_i - V_i(t))dt + \eta_i \sqrt{V_i(t)}dW_i(t)$$

The diffusion part is the Heston stochastic volatility model and the jump part is a multivariate Hawkes process modeling the clustering and propagation of jumps among multiple assets. $Z_i(t)$ corresponds to the jump size and direction.

### 5.4 Measuring Endogeneity (Reflexivity)

In term of the Hawkes branching structure representation of events arrivals, Filimonov et al. [78, 79] portray immigrants as exogenous news whereas the descendants are endogenous incidents. In the context of price movements in the stock market, immigrants are the price discovery due to orders from informed traders, who react to external information, whereas the descendants are the destabilizing ripples created by noise traders, who engage in herding [80], momentum trading [81] and parasite trading [82] etc.

Under the univariate linear Hawkes model with exponential decay kernel and constant base rate, the expected number of direct descendants per individual (branching coefficient) is given by

$$n = \int_0^\infty \gamma(s)ds = \int_0^\infty \alpha e^{-\beta t}dt = \frac{\alpha}{\beta}$$

For a given immigrant, the expected number of descendants in all generations is $n + n^2 + n^3 + ... = n/(1-n)$ if $n < 1$, so the ratio of descendants (non-immigrants) vs total population is

$$\frac{\text{descendants}}{\text{descendants} + \text{immigrant}} = \frac{n/(1-n)}{n/(1-n) + 1} = n$$

Therefore, the branching coefficient $n$ characterizes the amount of endogenous feedback activities while the base rate $\mu$ measures the arrival rate of exogenous information.

Using E-mini S&P futures as proxy, Filimonov and Sornette [78] find that the level of endogeneity (reflexivity) $n$ in the US market has gone from 0.3 in 1998 to 0.7 in 2007. Moreover, in the flash crash of May 6, 2010, $n$ reached a peak of 0.95.

Nonetheless, using the power law decay kernel, Hardiman et al. [84] challenge the result of Filimonov and Sornette [78] by reporting that the branching ratio $n$ has always been close to one since 1998 and that the market could be a critical Hawkes process [23], but Filimonov and Sornette [85] refute that the power law kernel is sensitive to outliers in addition to other counter arguments. Later Hardiman and Bouchaud [86] devised a nonparametric estimation of the branching ratio in term of moments, but the result depends heavily on the window size used in the empirical moment computation.

### Appendix A: Point Processes

**Definition**

Let $X$ (state space) be a locally compact Hausdorff second countable topological space, $B_X$ be the Borel sets on $X$ and $\mathcal{B}$ be the collection of bounded (relatively compact) borel sets on $X$. A Borel measure $\mu$ on $(X, B_X)$ is

---

18Filimonov and Sornette [78] borrow this term from Soros [83].

19Some textbooks use complete separable metric space, but locally compact Hausdorff second countable space has a complete separable metrization and all the results here do not depend on any particular choice of metric [8, p.11]. In most cases, $X = \mathbb{R}^m$. 

---
called locally finite if $\mu(B) < \infty \forall B \in \mathcal{B}$. Let $\Omega(X)$\textsuperscript{20} be the set of (positive) locally finite Borel counting (integer-valued) measure on $(X, \mathcal{B}_X)$ and $\mathcal{M}(X)$\textsuperscript{21} be the $\sigma$-algebra of $\Omega(X)$ generated by the set of evaluation functionals $\{\Phi_B : \Omega(X) \to \mathbb{N} | B \in \mathcal{B}\}$ where $\Phi_B(\mu) = \mu(B)$ and $\mathbb{N} = \{0, 1, 2, \ldots\}$.

A point process $N$ on $X$ is defined as a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Omega(X), \mathcal{M}(X))$; thus a point process is formally a measure-valued random element. However, for any point process $N$, there exists random variables $b_i \in \mathbb{Z}_+ = \{1, 2, \ldots\}$, $x_i \in X$, $n \in \mathbb{Z}_+ = \mathbb{Z}_1 \cup \{\infty\}$ such that $N(\bullet) = \sum_{i=1}^n b_i \delta_i(\bullet)$ where $\delta_i$ is the Dirac measure $\delta_i(x) = \mathbb{I}(x \in A)$ \[8, p.20\]. If we think of $b_i$ as the number of points at $x_i$, we can see that the point process $N$ is indeed the random counting measure showing the total number of points in any given region and this matches our intuition that a point process is a random set of points $\{x_i\}$ on $X$.

The point process $N$ is called simple if $\mathbb{P}(N(\{x\}) > 1) = 0 \forall x \in X^{22}$; that is, each location has at most one point. In this case, $N(\bullet) = \sum_{i=1}^n \delta_i(\bullet)$.

Suppose $X$ is also a topological vector space (e.g. $\mathbb{R}^m$), the shift operator $S_t : \mathcal{B}_X \to \mathcal{B}_X$ is defined as $S_t(A) = A + t = \{x + t \in X | x \in A\}$. A point process $N$ is called stationary if the shifted process $N \circ S_t^{23}$ has the same distribution as $N \forall t \in X$.

**Moments**

Let $k \in \mathbb{Z}_+$, the $k^{th}$ moment measure\textsuperscript{24} $M^k : \mathcal{B}_X \to [0, \infty]$ of a point process $N$ is defined as

$$M^k(A_1, \ldots, A_k) = \mathbb{E}(N(A_1) \ldots N(A_k)) = \mathbb{E} \left( \sum_{x_1} \cdots \sum_{x_k} \delta_{x_1,\ldots,x_k}(A_1 \times \ldots \times A_k) \right)$$

(58)

The first moment measure is also called mean (intensity) measure and denoted as $M(\bullet)$. The covariance measure is defined as

$$C^2(A_1, A_2) = \text{Cov}(N(A_1), N(A_2)) = M^2(A_1, A_2) - M(A_1)M(A_2)$$

(59)

The second and higher moment measures have concentration along diagonals, so we also have the $k^{th}$ factorial moment measure.

$$M^{(k)}(A_1, \ldots, A_k) = \mathbb{E} \left( \sum_{x_1 \neq \ldots \neq x_k} \sum_{x_1,\ldots,x_k} \delta_{x_1,\ldots,x_k}(A_1 \times \ldots \times A_k) \right)$$

(60)

The name factorial comes from the fact that $M^{(k)}(A_1, \ldots, A_k) = \mathbb{E}(N(A) - (N(A) - 1))\ldots(N(A) - k + 1))$. Obviously, $M(A) = M^{(1)}(A)$ and for $k = 2$, we have $M^2(A_1, A_2) = M^{(2)}(A_1, A_2) + M(A_1 \cap A_2)$.

If $X = \mathbb{R}^m$ and $N$ is stationary, it can be shown that $M(A) = \lambda \mathcal{L}(A)$ where $\lambda = M([0, 1]^m)$ and $\mathcal{L}$ is Lebesgue measure. That implies the mean measure $M$ of a stationary point process is absolutely continuous with respect to Lebesgue measure with constant density $M([0, 1]^m)$. If the covariance factorial moment measure $C^{(2)}(\bullet)$ is also absolutely continuous, we denote its density function as $c^{(2)}(x, y)$. Since $N$ is stationary, $c^{(2)}(x, y) = c^{(2)}(y - x)$ and $c^{(2)}(\bullet)$ is called reduced covariance density. The covariance measure $C^2$ is usually not absolutely continuous but for simple point process $N$ on $\mathbb{R}_+$, the quantity below is still called (reduced) covariance density, and is useful in estimation:

$$c^2(dx) = \mathbb{E}(N(x + dx)N(x))/dx^2 - \lambda^2 = \lambda \delta(dx) + c^{(2)}(dx)$$

(61)

**Marked Point Processes**

When an event happens, it may carry an additional information (mark). For instance, each order arrival is associated with an order quantity (volume) and each earthquake is reported with a magnitude. A point process with marks is called marked point process.

Let $Y$ (mark space) be a locally compact Hausdorff second countable space, $(Y, \mathcal{B}_Y)$ be a measurable space and $\nu$ (mark distribution) be a probability measure on $(Y, \mathcal{B}_Y)$. A marked point process (MPP) $N$ is a measurable mapping

---

\textsuperscript{20}On locally compact Hausdorff second countable space, all locally finite Borel measures are Radon measures.

\textsuperscript{21} $\mathcal{M}(Y)$ is the same as the Borel $\sigma$-algebra generated by the vaguely topology of $\Omega(Y)$ \[8, p.32\].

\textsuperscript{22}$X$ is Hausdorff, so all singletons are closed and thus measurable.

\textsuperscript{23}$N \circ S_t : \Omega \to (\Omega(X), \mathcal{M}(X)), ((N \circ S_t)(\omega))(A) = (N(\omega))(S_t(A))$, that is $N$ is shifted $t$ unit to the left when $X = \mathbb{R}$.

\textsuperscript{24}The notations of moment, covariance, factorial moment, reduced moment vary between authors.
$N : \Omega \rightarrow \{0,1\}^X$ such that the ground measure $N_\pi(\bullet) = N(\bullet \times Y)$ is a process (i.e. locally finite)\(^{25}\). Hence a marked point process is nothing but a point process on a product space, but usually we treat the location $x$ and mark $y$ differently and we have a few more definitions.

$N$ is called a multivariate point process if $Y = \{1, \ldots, d\}$. In this case, $N_i(\bullet) = N(\bullet \times \{i\})$ is called the marginal process of type $i$ points. A MPP $N$ is called simple if $N_\pi$ is simple\(^{26}\). The marks of a MPP are called unpredictable if $y_n$ is independent of $\{(x_i,y_i)\}_{i<n}$ and they are called independent if $y_n$ is independent of $\{(x_i,y_i)\}_{i\neq n}$\(^{27}\).

**Stochastic Intensity**

In this section, $X = \mathbb{R}_+ \times \mathbb{R}$ and $N_\pi = N((0,t])$. Let $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ be a filtered complete probability space. A stochastic process $Z : \mathbb{R}_+ \times \Omega \rightarrow \mathbb{R}$ is called $\mathcal{F}_t$-predictable if it is measurable with respect to the predictable $\sigma$-algebra $\mathcal{P} = \sigma\{\{s,t] \times A \mid 0 \leq s < t, A \in \mathcal{F}_t\}$. If $Z_t$ is adapted and left-continuous, then $Z_t$ is predictable [10, p.9]. In practice, all the predictable processes we use are in this category. Also if $Z_t$ is predictable, then $Z_t \in \mathcal{F}_t$; in other words, the value of the predictable process $Z_t$ at time $t$ is "known" just before time $t$.

We assume the filtration $\{\mathcal{F}_t\}$ satisfies the usual condition (complete and right-continuous) and $\{N_t\}$ is adapted and simple. A stochastic process $\lambda : \mathbb{R}_+ \times \Omega \rightarrow \mathbb{R}$ is called $\mathcal{P}$-compensator of a point process $N$ if $N_t$ is increasing, right-continuous, $\mathcal{P}$-predictable, $A_0 = 0$ a.s. and $(N_t - A_t)$ is a $\mathcal{P}$-local martingale. If $A_t = \int_0^t \lambda_s ds$ a.s., $\lambda_t$ is non-negative and $\mathcal{F}_t$-predictable, then $\lambda_t$ is called the stochastic or conditional intensity of $N$\(^{29,30,31}\). A defining properties of $\lambda_t$ is

$$\mathbb{E}\left(\int_s^t \lambda_u du \mid \mathcal{F}_s\right) = \mathbb{E}(N_t - N_s \mid \mathcal{F}_s) \quad \text{a.s.} \quad \forall s < t \tag{62}$$

When $s \rightarrow t$, this becomes $\lambda_t dt = \mathbb{E}(N(dt) \mid \mathcal{F}_t)$. We can see that the stochastic intensity $\lambda_t$ is the instantaneous rate of arrival conditioned on all information just before time $t$. For a multivariate point process, $\lambda_i(t)$ is the intensity of the marginal process $N_i(t)$.

From the definition, we notice that intensity exists if and only if the compensator is absolutely continuous. In fact, the compensator of a point process can be expressed in terms of the conditional inter-arrival time $\lambda(t) = \int_0^t \lambda_u du \mid \mathcal{F}_t$; if the conditional distribution has support over $\mathbb{R}_+$. Under this condition, the intensity exists if and only if the conditional inter-arrival time is absolutely continuous. In this case, the intensity is given by [9, p.70]

$$\lambda_t = h_n(t-t_{n-1}) \text{ if } t \in (t_{n-1}, t_n] \tag{63}$$

$$h_n(t) = \frac{g_n(t)}{1-G_n(t^{-})}, \quad (t_n-t_{n-1}) \mid \mathcal{F}_{t_{n-1}} \sim G_n \tag{64}$$

Once we know the intensity, we know the conditional distributions of all inter-arrival times and hence the complete distribution of the point process [11, p.233].

$A : \mathbb{R}_+ \times \mathcal{B}_Y \times \Omega \rightarrow \mathbb{R}_+$ is called a compensator of the MPP $N$ if $A(\bullet, B)$ is a compensator of $N(\bullet \times B) \forall B \in \mathcal{B}_Y$ and $A(t, \bullet)$ is a measure on $(Y, \mathcal{B}_Y) \forall t \in \mathbb{R}_+$. If $A(t,B) = \int_0^t \int_B \lambda(s) f(s,dy) ds$ a.s. where $\lambda(t)$ is non-negative and predictable, then $\lambda(t)$ is called the stochastic intensity of the MPP $N$ and $v(t_n,dy) = \mathbb{P}(y_n \in dy \mid \mathcal{F}_{t_n})$ is called the conditional mark distribution.

**Random Time Change**

If the filtration is usual, a point process $N$ on $\mathbb{R}_+$ is simple and adapted, its intensity $\lambda(t)$ exists and $\int_0^\infty \lambda(s)ds = \infty$ a.s., then $\{J_\lambda = \int_0^\infty \lambda(s)ds\}$ is a standard Poisson process (rate=1). The above theorem is called random time change theorem [88, 89] and is extremely useful in testing the goodness-of-fit of a stochastic intensity model.

---

\(^{25}\)From this definition, Poisson random measure $N$ on $\mathbb{R}^2$ is not MPP on $\mathbb{R} \times \mathbb{R}$ as $N(A \times \mathbb{R}) = \infty$.

\(^{26}\)Any point process can be treated as simple MPP with the mark being the number of points at $x_i$.

\(^{27}\)Notice that if marks are independent, future location $x_{n+1}$ cannot depends on previous mark $y_n$.

\(^{28}\)The stochastic intensity of point process on $\mathbb{R}_+$ is extended to higher dimension in [87].

\(^{29}\)Stochastic intensity is unique up to modification [10, p.31].

\(^{30}\)Notice that stochastic intensity depends on the underlying filtration, so some text use the notation $\lambda(t \mid \mathcal{F}_t)$ but we will simply use $\lambda(t)$ and call it stochastic intensity or intensity when there is no confusion about the filtration.

\(^{31}\)If $A_t$ is absolutely continuous with respect to Lebesgue measure and $\lambda_t$ is the Radon-Nikodym derivative (may not be predictable) then $\mathbb{E}(\lambda_t \mid \mathcal{F}_t)$ is a version of the stochastic intensity. Some authors only require the intensity to be adapted, but using the conditional expectation, one can always find a predictable version of intensity provided that the intensity has finite first moment.
Appendix B: A Brief History of Hawkes processes

Hawkes processes are proposed by Hawkes [13] in 1971 in order to model contagious processes like epidemics, neuron firing and particle emission, where the occurrences of events trigger further events. Although the intensity of Cox processes [90], introduced in 1955, are stochastic, they are determined before the events are unfolded\(^{32}\). In order to portray the excitation behavior in contagious processes, Hawkes extends the model in such a way that the intensity is a predictable stochastic process with an intuitive autoregressive form, which allows it to adapt to events that happen over time.


Although the first application of Hawkes processes to earthquake occurrences appeared in 1982 [15], it was not until 1988 [14] that Hawkes processes received much attention. Since then, the versatility of Hawkes model is leveraged in seismology [14, 91], finance (risk and credit default modeling) [92, 93], social networks [94, 95], neuroscience [96, 97] etc. (see [98, 99] for more applications).

The use of Hawkes processes in HF financial data modeling starts with Bowsher [22] in 2007\(^{33}\) and then Large [62] in the same year. Both authors exploit Hawkes processes so as to describe the interactions among different types of order arrivals. Later, Bacy et al. [68] and Aït-Sahalia et al. [77] employ Hawkes processes to reproduce jump clustering in pure jump and jump diffusion representation of stock prices in 2013 and 2015 respectively. On the other hand, an interesting idea from Filimonov and Sornette [78] appears in 2012 which utilizes the branching coefficient of linear Hawkes model to measure the level of endogenous activities in the US stock market, though the debate about the validity of the result is still on going [84–86].

References


\(^{32}\)For processes on \(\mathbb{R}_+\), it means intensity of Cox process is \(\mathcal{F}_0\) measurable.

\(^{33}\)Though Bowsher’s paper was published in 2007, the first draft appeared in 2002.


