Abstract: In this paper, we study various new Hawkes processes. Specifically, we construct general compound Hawkes processes and investigate their properties in limit order books. With regard to these general compound Hawkes processes, we prove a Law of Large Numbers (LLN) and a Functional Central Limit Theorems (FCLT) for several specific variations. We apply several of these FCLTs to limit order books to study the link between price volatility and order flow, where the volatility in mid-price changes is expressed in terms of parameters describing the arrival rates and mid-price process.

Keywords: Hawkes processes; general compound Hawkes processes; limit order books; functional central limit theorems; LOBSTER data

1. Introduction

The Hawkes process (HP) is named after its creator, Hawkes (1971); Hawkes and Oakes (1974). The HP is a simple point process equipped with a self-exciting property, clustering effect and long run memory. Through its dependence on the history of the process, the HP captures the temporal and cross sectional dependence of the event arrival process as well as the ‘self-exciting’ property observed in our empirical data on limit order books. Self-exciting point processes have recently been applied to high frequency data for price changes Bacry et al. (2011) or order arrival times Embrechts et al. (2011). HPs have seen their application in many areas, like genetics Carstensen (2010), occurrence of crime Mohler et al. (2011), bank defaults Azizpour et al. (2018) and earthquakes Ogata (1988).

Point processes gained a significant amount of attention in statistics during the 1950s and 1960s. Cox (1955) introduced the notion of a doubly stochastic Poisson process (called the Cox process now) and Bartlett (1963) investigated statistical methods for point processes based on their power spectral densities. Lewis (1964) formulated a point process model (for computer power failure patterns) which was a step in the direction of the HP. A nice introduction to the theory of point processes can be found in Daley and Vere-Jones (2003). The first type of point process in the context of market microstructure is the autoregressive conditional duration (ACD) model introduced by Engle and Russell (1998).

A recent application of HP is in financial analysis, in particular limit order books. In this paper, we study various new Hawkes processes, namely general compound Hawkes processes to model the price process in limit order books. We prove a Law of Large Numbers (LLN) and a Functional Central Limit Theorem (FCLT) for specific cases of these processes. Several of these FCLTs are applied to limit order books where we use asymptotic methods to study the link between price volatility and order flow in our models. The volatility of the price changes is expressed in terms of parameters describing the arrival rates and price changes. We also present some numerical
examples. The general compound Hawkes process was first introduced in Swishchuk (2017) to model the risk process in insurance and studied in detail there. In the paper Swishchuk et al. (2019), we obtained functional CLTs and LLNs for general compound Hawkes processes with dependent orders and regime-switching compound Hawkes processes. Bowsher (2007) was the first one who applied the HP to financial data modelling. Cartea et al. (2014) applied HP to model market order arrivals. Filimonov et al. (2014) and Filimonov and Sornette (2012) applied the HPs to estimate the percentage of price changes caused by endogenous self-generated activity rather than by the exogenous impact of news or novel information. Bauwens and Hautsch (2009) used a five-dimensional HP to estimate multivariate volatility between five stocks, based on price intensities. Hewlett (2006) used the instantaneous jump in the intensity caused by the occurrence of an event to qualify the market impact of that event, taking into account the cascading effect of secondary events causing further events. Hewlett (2006) also used the Hawkes model to derive optimal pricing strategies for market makers and optimal trading strategies for investors given that the rational market makers have the historic trading data. Large (2007) applied a Hawkes model for the purpose of investigating market impact, with a specific interest in order book resiliency. Specifically, he considered limit orders, market orders and cancellations on both the buy and sell side, and further categorizes these events based on their level of aggression, resulting in a ten-dimensional Hawkes process. Other econometric models based on marked point processes with stochastic intensity include autoregressive conditional intensity (ACI) models with the intensity depending on its history. Hasbrouck (1999) introduced a multivariate point process to model the different events of an order book but did not parametrize the intensity. We note that Brémaud and Massoulié (1996) generalized the HP to its nonlinear form. In addition, a functional central limit theorem for nonlinear Hawkes processes was obtained in Zheng et al. (2013). The ‘Hawkes diffusion model’ introduced in Aït-Sahalia et al. (2015) attempted to extend previous models of stock prices to include financial contagion. Chavez-Demoulin and McGill (2012) used Hawkes processes to model high-frequency financial data. An application of affine point processes to portfolio credit risk may be found in Errais et al. (2010). Some applications of Hawkes processes to financial data are also given in Embrechts et al. (2011). Cohen and Elliott (2013) derived an explicit filter for Markov modulated Hawkes processes. Vinkovskaya (2014) considered a regime-switching Hawkes process to model its dependency on the bid-ask spread in limit order book. Regime-switching models for pricing of European and American options were considered inBuffington and Elliott (2002b) and Buffington and Elliott (2002a), respectively. Semi-Markov processes were applied to limit order books in Swishchuk and Vadori (2017) to model the mid-price. We also note that level-1 limit order books with time dependent arrival rates \( \lambda(t) \) were studied in Chávez-Casillas et al. (2019), including the asymptotic distribution of the price process.

The paper by Bacry et al. (2015) proposes an overview of the recent academic literature devoted to the applications of Hawkes processes in finance. It is a nice survey of applications of Hawkes processes in finance. In general, the main models in high-frequency finance can be divided into univariate models, price models, impact models, order-book models and some systemic risk models, models accounting for news, high-dimensional models and clustering with graph models. The book by Cartea et al. (2015) developed models for algorithmic trading such as methods for executing large orders, market making, trading pairs of collections of assets, and executing in the dark pool. This book also contains a link from which several datasets can be downloaded, along with MATLAB code to assist in experimentation with the data.

A detailed description of the mathematical theory of Hawkes processes is given in Liniger (2009). The paper by Laub et al. (2015) provides background, introduces the field and historical development, and touches upon all major aspects of Hawkes processes. The results of the current paper were first announced in Swishchuk and Huffman (2018). The main contribution and novelty of He and Swishchuk (2019) paper consists of considering different types of general compound Hawkes processes and their diffusive limits to model the mid-prices of six different stocks. Namely, EBAY, FB,
MU, PCAR, SMH, and CSCO were used for quantitative and comparative analysis, and for defining the error rates to estimate the models fitting accuracy. The paper Swishchuk et al. (2017) deals with compound and regime-switching Hawkes processes to model the mid-price processes in limit order books. Diffusive limits were used for both models to study the link between the price volatility and the order flow. Numerical examples were presented using CISCO data dated from 3 November to 7 November 2014. Thus, the present paper contains not only new results with proofs for different general compound Hawkes processes, but also another data set where those results were applied compared with previous papers.

The paper is organized as follows. A definition of a Hawkes process and a description of its properties are given in Section 2. Law of Large Numbers (LLN) and Functional Central Limit Theorems (FCLT) for various general compound Hawkes processes, including nonlinear, in limit order books are proved in Section 3. Descriptions of the data set and empirical results are presented in Section 4. Section 5 gives a quantitative analysis of our results and Section 6 concludes the paper.

2. Definition and Some Properties of Hawkes Processes (HP)

2.1. Definition of Hawkes Processes (HPs)

In this section, we give various definitions and some properties of Hawkes processes which can be found in the existing literature (see, e.g., Hawkes (1971), Hawkes and Oakes (1974), Embrechts et al. (2011) and Zheng et al. (2013), to name a few). They include in particular one-dimensional and nonlinear Hawkes processes.

**Definition 1 (Counting Process).** A counting process is a stochastic process \( N(t) \) with \( t \geq 0 \), where \( N(t) \) takes positive integer values and satisfies \( N(0) = 0 \). It is almost surely finite and a right-continuous step function with increments of size +1.

Denote by \( F_N(t) \), \( t \geq 0 \), the history of the arrival up to time \( t \); that is, \( F_N(t) \), \( t \geq 0 \), is a filtration (an increasing sequence of \( \sigma \)-algebras).

A counting process \( N(t) \) can be interpreted as a cumulative count of the number of arrivals into a system up to the current time \( t \). The counting process can also be characterized by the sequence of random arrival times \( (T_1, T_2, \ldots) \) at which the counting process \( N(t) \) has jumped. The process defined by these arrival times is called a point process (see Daley and Vere-Jones 2003).

**Definition 2 (Point Process).** If a sequence of random variables \( (T_1, T_2, \ldots) \), taking values in \([0, \infty)\) has \( P(0 \leq T_1 \leq T_2 \leq \ldots) = 1 \), and the number of points in a bounded region is almost surely finite, then \((T_1, T_2, \ldots)\) is called a point process.

**Definition 3 (Conditional Intensity Function).** Consider a counting process \( N(t) \) with associated histories \( F_N(t) \), \( t \geq 0 \). If a non-negative function \( \lambda(t) \) exists such that

\[
\lambda(t) = \lim_{h \to 0} \frac{E[N(t + h) - N(t) | F_N(t)]}{h}
\]

Then, it is called the conditional intensity function of \( N(t) \) (see Laub et al. 2015). We note that originally this function was called the hazard function (see Cox 1955).

**Definition 4 (One-dimensional Hawkes Process).** The one-dimensional Hawkes process (see Laub et al. 2015; Hawkes and Oakes 1974) is a point process \( N(t) \) which is characterized by its intensity \( \lambda(t) \) with respect to its natural filtration:

\[
\lambda(t) = \lambda + \int_0^t \mu(t - s) dN(s)
\]

where \( \lambda > 0 \), and the response function \( \mu(t) \) is a positive function that satisfies \( \int_0^\infty \mu(s) ds < 1 \).
The constant $\lambda$ is called the background intensity and the function $\mu(t)$ is sometimes called the excitation function. To avoid the trivial case of a homogeneous Poisson process, we assume $\mu(t) \neq 0$. Thus, the Hawkes process is a non-Markovian extension of the Poisson process.

With respect to the Definitions of $\lambda(t)$ in 3 and $N(t)$ in 4, it follows that

$$P(N(t + h) - N(t) = m \mid F^N(t)) = \begin{cases} 
\lambda(t)h + o(h) & m = 1 \\
o(h) & m > 1 \\
1 - \lambda(t)h + o(h) & m = 0
\end{cases}$$

The interpretation of Equation (2) is that the events occur according to an intensity with a background intensity $\lambda(t)$ which increases by $\mu(0)$ at each new event, eventually decaying back to the background intensity value according to the evolution of the function $\mu(t)$. Choosing $\mu(0) > 0$ leads to a jolt in the intensity at each new event, and this feature is often called the self-exciting feature. In other words, if an arrival causes the conditional intensity function $\lambda(t)$ in Equations (1) and (2) to increase, then the process is called self-exciting.

We would like to mention that the conditional intensity function $\lambda(t)$ in Equations (1) and (2) can be associated with the compensator $\Lambda(t)$ of the counting process $N(t)$, that is,

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

We note that $\Lambda(t)$ is the unique non-decreasing, $\mathcal{F}^N(t)$, $t \geq 0$, predictable function, with $\Lambda(0) = 0$ such that

$$N(t) = M(t) + \Lambda(t) \text{ a.s.},$$

where $M(t)$ is an $\mathcal{F}^N(t)$, $t \geq 0$, local martingale (existence of which is guaranteed by the Doob–Meyer decomposition).

A common choice for the function $\mu(t)$ in Equation (2) is the one of exponential decay (see Hawkes 1971)

$$\mu(t) = ae^{-\beta t}$$

with parameters $a, \beta > 0$. In this case, the Hawkes process is called the Hawkes process with exponentially decaying intensity.

In the case of Equation (4), Equation (2) becomes

$$\lambda(t) = \lambda + \int_0^t ae^{-\beta(t-s)}dN(s)$$

We note that, in the case of Equation (4), the process $(N(t), \lambda(t))$ is a continuous-time Markov process, which is not the case for a general choice of excitation function in Equation (1).

With some initial condition $\lambda(0) = \lambda_0$, the conditional intensity $\lambda(t)$ in Equation (5) with exponential decay in Equation (4) satisfies the SDE

$$d\lambda(t) = \beta(\lambda - \lambda(t))dt + adN(t), t \geq 0$$

which can be solved using stochastic calculus as

$$\lambda(t) = e^{-\beta t}(\lambda_0 - \lambda) + \lambda + \int_0^t ae^{-\beta(t-s)}dN(s),$$

which is an extension of Equation (5).
Another choice for $\mu(t)$ is a power law function

$$\lambda(t) = \lambda + \int_0^t \frac{k}{(c + (t - s))^p} \, dN(s)\tag{8}$$

with positive parameters $(c, k, p)$. This power law form for $\lambda(t)$ in Equation (8) was applied in the geological model called Omori’s law, and used to predict the rate of aftershocks caused by an earthquake.

**Definition 5** (D-dimensional Hawkes Process). The D-dimensional Hawkes process (see Embrechts et al. 2011) is a point process $\tilde{N}(t) = (N^i(t))_{i=1}^D$ which is characterized by its intensity vector $\tilde{\lambda}(t) = (\lambda^i(t))_{i=1}^D$ such that:

$$\lambda^i(t) = \lambda^i + \int_0^t \mu^{ij}(t - s) \, dN^j(s)\tag{9}$$

where $\lambda^i > 0$, and $M(t) = (\mu^{ij}(t))$ is a matrix-valued kernel such that:

1. it is component-wise non-negative: $(\mu^{ij}(t)) \geq 0$ for each $1 \leq i, j \leq D$
2. it is component-wise $L^1$-integrable

In matrix-convolution form, Equation (9) can be written as

$$\tilde{\lambda}(t) = \tilde{\lambda} + M \ast d\tilde{N}(t)\tag{10}$$

where $\tilde{\lambda}(t) = (\lambda^i)_{i=1}^D$.

**Definition 6** (Nonlinear Hawkes Process). The nonlinear Hawkes process (see, e.g., Zheng et al. 2013) is defined by the intensity function in the following form:

$$\lambda(t) = h\left(\lambda + \int_0^t \mu(t - s) \, dN(s)\right)\tag{11}$$

where $h(\cdot)$ is a nonlinear function with support in $\mathbb{R}^+$. Typical examples for $h(\cdot)$ are $h(x) = 1_{x \in \mathbb{R}^+}$ and $h(x) = e^x$.

**Remark 1.** Many other generalizations of Hawkes processes have been proposed. They include mixed diffusion–Hawkes models Errais et al. (2010), Hawkes models with shot noise exogenous events Dassios and Zhao (2011), and Hawkes processes with generation dependent kernels Mehrdad and Zhu (2014), to name a few.

2.2. Compound Hawkes Processes

In this section, we define nonlinear compound Hawkes process with N-state dependent orders. The dependent orders means the dependency of both, the type of a book event and its corresponding inter-arrival times, on the type of the previous book event. We also consider special cases of this general compound Hawkes process.

**Definition 7** (Nonlinear Compound Hawkes Process with n-state Dependent Orders (NLCHPnSDO) in Limit Order Books). Consider the price process $S_t$

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)\tag{12}$$
where $X_k$ is a continuous time $n$-state Markov chain, $a(x)$ is a continuous and bounded function on the state space $X := \{1, 2, ..., n\}$, $N(t)$ is the nonlinear Hawkes process (see, e.g., Zheng et al. (2013) defined by the intensity function in the following form (see Equation (11)):

$$\lambda(t) = h \left( \lambda + \int_0^t \mu(t - s)dN(s) \right)$$

where $h(\cdot)$ is a nonlinear increasing function with support in $R^+$. We note that in Brémaud and Massoulié (1996) it was shown that, if $h(\cdot)$ is $\alpha$-Lipschitz (see Brémaud and Massoulié 1996) such that $\alpha||h||_{L^1} < 1$, then there exists a unique stationary and ergodic Hawkes process satisfying the dynamics of Equation (11). We shall refer to the process in Equation (12) as a Nonlinear Compound Hawkes Process with $n$-State Dependent Orders (NLCHPnSDO).

This nonlinear compound Hawkes process will be the foundation for our studies throughout this paper. In the following subsection, we will introduce four specific examples, which will be used for our empirical investigations of the mid-price processes.

2.3. General Compound Hawkes Processes

**Definition 8** (General Compound Hawkes Process With $n$-state Dependent Orders (GCHPnSDO)). Suppose that $X_k$ is an ergodic continuous-time Markov chain, independent of $N(t)$, with state space $X = \{1, 2, ..., n\}$, $N(t)$ is a one-dimensional Hawkes process defined in Definition 4 and $a(x)$ is any bounded and continuous function on $X$. We define the General Compound Hawkes process with $n$-state Dependent Orders (GCHPnSDO) by the following process:

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

(13)

Note that this process can be recovered from Equation (12) by letting $h(x) = x$.

**Definition 9** (General Compound Hawkes Process with Two-State Dependent Orders (GCHP2SDO)). Suppose that $X_k$ is an ergodic continuous time Markov chain, independent of $N(t)$, with two states $\{1, 2\}$. Then, Equation (13) becomes

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

(14)

where $a(X_k)$ takes only the values $a(1)$ and $a(2)$. Of course, we can view this as a special case of the $n$-state case, where $n = 2$. This model was used in Swishchuk et al. (2017) for the mid-price process in limit order books with non-fixed tick $\delta$ and two-valued price changes.

**Definition 10** (General Compound Hawkes Process with Dependent Orders (GCHPDO)). Suppose that $X_k \in \{-\delta, \delta\}$ and that $a(x) = x$, then $S_t$ in Equation (13) becomes

$$S_t = S_0 + \sum_{k=1}^{N(t)} X_k$$

(15)

This type of process can be a model for the mid-price $= (\text{bidprice} + \text{askprice})/2$ in limit order books, where $\delta$ is a fixed tick size and $N(t)$ is the number of order arrivals up to time $t$. We shall call this process a General Compound Hawkes Process with Dependent Orders (GCHPDO). This is a generalization of the previous process, obtained by letting $a(1) = -\delta$ and $a(2) = \delta$.

Having defined several modifications of Hawkes processes, we now prove diffusion limit theorems and LLNs for each price process in the following section. These diffusion processes will be used for our exploration of the applicability of this model to real world limit order book data.
Since order arrivals and cancellations are very frequent, with many of the applications and practical needs occurring at the millisecond time scale (e.g., order liquidations, order executions, etc.), we are interested in the dynamics that occur on a large time scale, such as tens of second or minutes. Therefore, in the following section, we consider the scale $nt$ instead of $t$, where $t$ is in milliseconds and $n$ can be 1000, 10,000, making $nt$ large with respect to $t$.

3. Diffusion Limits and LLNs for GCHP

3.1. Diffusion Limit and LLN for NLCHP_{nSDO}

We consider the mid-price process $S_t$ defined in Definition 7, namely

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

where $X_k$ is a continuous time $N$-state Markov chain and $a(x)$ is a continuous bounded function on the state space $X = \{1, 2, ..., n\}$. $N_t$ is the number of price changes up to time $t$, described by the nonlinear Hawkes process given in Equation (11).

**Theorem 1** (Diffusion Limit for NLCHP_{nSDO}). Let $X_k$ be an ergodic Markov chain with $N$-states $\{1, 2, ..., n\}$ and with ergodic probabilities $(\pi_1^*, \pi_2^*, ..., \pi_n^*)$. Let also $S_t$ be as defined in Definition 7, then

$$S_{nt} = N(nt) \cdot \hat{\sigma}^* \frac{\sqrt{nt}}{\sqrt{n}} \xrightarrow{n \to \infty} \hat{\sigma}^* \sqrt{E[N[0,1]]} W_t$$

where $W_t$ is a standard Wiener process and $E[N[0,1]]$ is the mean of the number of arrivals on a unit interval under the stationary and ergodic measure. Furthermore,

$$0 < \hat{\mu} := \int_0^\infty \mu(s) ds < 1 \text{ and } \int_0^\infty s\mu(s) ds < \infty$$

$$(\hat{\sigma}^*)^2 := \sum_{i \in X} \pi_i^* v(i)$$

$$v(i) = b(i)^2 + \sum_{j \in X} (g(j) - g(i))^2 P(i,j) - 2b(i) \sum_{j \in X} (g(j) - g(i)) P(i,j)$$

$$b = (b(1), b(2), ..., b(n))'$$

$$b(i) := a(X_i) - a^* := a(i) - a^*$$

$$g := (P + \Pi^* - I)^{-1} b$$

$$\hat{a}^* := \sum_{i \in X} \pi_i^* a(X_i)$$

$P$ is the transition probability matrix for $X_k$, i.e., $P(i,j) = P(X_{k+1} = j \mid X_k = i)$. $\Pi^*$ denotes the matrix of stationary distributions of $P$ and $g(j)$ is the $j$th entry of $g$.

**Proof.** From Equation (12), we have

$$S_{nt} = S_0 + \sum_{k=1}^{N(nt)} a(X_k)$$

and

$$S_{nt} = S_0 + \sum_{i=1}^{N(nt)} (a(X_k) - \hat{a}^*) + N(nt)\hat{a}^*.$$
Therefore,

\[
\frac{S_{nt} - N(nt)\hat{a}^*}{\sqrt{n}} = \frac{S_0 + \sum_{k=1}^{N(nt)} (a(X_k) - \hat{a}^*)}{\sqrt{n}}. \tag{21}
\]

As long as \(S_0 \frac{n}{\sqrt{n}} \xrightarrow{n \to \infty} 0\), we need only find the limit for

\[
a(X_k) - \hat{a}^* \quad \sqrt{n}
\]

when \(n \to +\infty\). Consider the following sums

\[
\hat{R}_n^* := \sum_{k=1}^{n} (a(X_k) - \hat{a}^*) \tag{22}
\]

and

\[
\hat{U}_n^*(t) := n^{-1/2} \left[ (1 - (nt - [nt])) \hat{R}_{[nt]}^* + (nt - [nt]) \hat{R}_{[nt] + 1}^* \right] \tag{23}
\]

where \([\cdot]\) is the floor function. Following the martingale method from Vadori and Swishchuk (2015), we have the following weak convergence in the Skorokhod topology (see Getoor 1967):

\[
\hat{U}_n^*(t) \xrightarrow{n \to \infty} \hat{\sigma}^* W(t) \tag{24}
\]

We note that the results from Brémaud and Massoulié (1996) imply by the ergodic theorem that

\[
\frac{N_t}{T} \xrightarrow{t \to \infty} E[N[0,1]] \tag{25}
\]

or

\[
\frac{N_{nt}}{nt} \xrightarrow{n \to \infty} tE[N[0,1]] \tag{26}
\]

Using the change of time \(t \to N_{nt}/n\), we find that

\[
\hat{U}_n^*(N_{nt}/n) \xrightarrow{n \to \infty} \hat{\sigma}^* W(tE[N[0,1]]) \tag{27}
\]

or

\[
\hat{U}_n^*(N_{nt}/n) \xrightarrow{n \to \infty} \hat{\sigma}^* \sqrt{E[N[0,1]]} W(t) \tag{28}
\]

The result now follows from Equations (19)–(21). \(\square\)

**Lemma 1** (LLN for NLCHPnSDO). The process \(S_{nt}\) in Equation (19) satisfies the following weak convergence in the Skorokhod topology (see Getoor 1967):

\[
\frac{S_{nt}}{n} \xrightarrow{n \to \infty} \hat{\sigma}^* E[N[0,1]] t \tag{29}
\]

where \(\hat{\sigma}^*\) is defined in Equation (18), respectively.

**Proof.** From Equation (12), we have

\[
\frac{S_{nt}}{n} = \frac{S_0}{n} + \sum_{k=1}^{N(nt)} \frac{a(X_k)}{n} \tag{30}
\]
The first term goes to zero when \( n \to \infty \). On the right-hand side, with respect to the strong LLN for Markov chains (see, e.g., Norris 1997),

\[
\frac{1}{n} \sum_{k=1}^{n} a(X_k) \to^{\text{a.s.}} \hat{a}^*
\]  
(31)

Then, taking into account Equation (26), we obtain

\[
\sum_{i=1}^{N(t)} \frac{a(X_k)}{n} = \frac{N(t)}{n} \sum_{k=1}^{N(nt)} a(X_k) \to^{\text{a.s.}} \hat{a}^* \mathbb{E}[N[0, 1)] t
\]  
(32)

from which the desired result follows. □

3.2. Diffusion Limit and LLN for GCHPnSDO

We consider here the mid-price process \( S_t \) which was defined in Definition 8, namely

\[
S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)
\]  
(33)

where \( X_k \) is a continuous time n-state Markov chain, \( a(x) \) is a continuous and bounded function on the state space \( X = \{1, 2, ..., n\} \), and \( N(t) \) is the number of price changes up to time \( t \), described by a one-dimensional linear Hawkes process defined in Definition 4. This can be interpreted as the case of non-fixed tick sizes, n-valued price changes and dependent orders.

**Theorem 2** (Diffusion limit for GCHPnSDO). Let \( X_k \) be an ergodic Markov chain with \( N \)-states \( \{1, 2, ..., n\} \) and with ergodic probabilities \( (\pi_1^*, \pi_2^*, ..., \pi_n^*) \). Let also \( S_t \) be as defined in Definition 8, then

\[
\frac{S_{nt} - N(nt)\hat{a}^*}{\sqrt{n}} \to^{\text{a.s.}} \hat{a}^* \sqrt{\lambda/(1 - \hat{\mu})} W(t)
\]  
(34)

where \( W(t) \) is a standard Wiener process,

\[
0 < \hat{\mu} := \int_0^\infty \mu(s) ds < 1 \text{ and } \int_0^\infty s\mu(s) ds < \infty
\]  
(35)

\[
(\hat{a}^*)^2 = \sum_{i \in X} \pi_i^* v(i)
\]

\[
v(i) = b(i)^2 + \sum_{j \in X} (g(j) - g(i))^2 P(i, j) - 2b(i) \sum_{j \in X} (g(j) - g(i)) P(i, j)
\]

\[
b = (b(1), b(2), ..., b(n))^t
\]

\[
b(i) := a(X_i) - a^* := a(i) - a^*
\]

\[
g := (P + \Pi^* - I)^{-1} b
\]

\[
\hat{a}^* := \sum_{i \in X} \pi_i^* a(X_i)
\]

\( P \) is the transition probability matrix for \( X_k \), i.e., \( P(i, j) = P(X_{k+1} = j | X_k = i) \). \( \Pi^* \) denotes the matrix of stationary distributions of \( P \) and \( g(j) \) is the \( j \)th entry of \( g \).

**Proof.** The proof follows from Theorem 1 above, taking into account that \( \mathbb{E}[N(0, 1)] = \lambda/(1 - \hat{\mu}) \) for a one-dimensional linear Hawkes process \( N(t) \). □
Lemma 2 (LLN for GCHPnSDO). The process $S_{rt}$ defined in Definition 8, satisfies the following weak convergence in the Skorokhod topology (see Getoor 1967)

$$\frac{S_{nt}}{n} \xrightarrow{n \to \infty} \hat{a}^* \frac{\lambda}{1 - \hat{\mu}}$$

where $\hat{\mu}$ and $\hat{a}^*$ are defined in Equations (35) and (36), respectively.

Proof. The proof follows from Lemma 1 above, taking into account that $E[N(0, 1)] = \lambda/(1 - \hat{\mu})$. □

Diffusion Limits and LLNs for Special Cases of GCHPnSDO

Theorem 2 can be reduced to some of the special cases we outlined previously in Section 2.3. Specifically in Definitions 9 and 10, we consider the case of a 2-state Markov chain for which we provide the diffusion limit and LLN result as Corollaries below.

We begin by considering the mid-price process $S_t$ (GCHP2SD) which was defined in Definition 9, namely

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

where $X_k$ is a continuous-time 2-state Markov chain, $a(x)$ is a continuous and bounded function on $X = \{1, 2\}$ and $N(t)$ is the number of price changes up to moment $t$, described by a one-dimensional Hawkes process defined in Definition 4. This can be interpreted as the case of non-fixed tick sizes, two-valued price changes and dependent orders.

Corollary 1 (Diffusion Limit for GCHP2SD). Let $X_t$ be an ergodic Markov chain with two states $\{1, 2\}$ and with ergodic probabilities $(\pi_1^*, \pi_2^*)$. Furthermore, let $S_t$ be defined as in Definition 9. Then,

$$\frac{S_{nt} - N(nt)a^*}{\sqrt{n}} \xrightarrow{n \to \infty} \sigma^* \sqrt{\lambda/(1 - \hat{\mu})}W(t)$$

where $W(t)$ is standard Wiener process,

$$0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1 \text{ and } \int_0^\infty \mu(s)ds < \infty$$

$$(\sigma^*)^2 := \pi_1^*(1 - p) + \pi_2^*(1 - p')(a_1 - a_2)^2 + a_1 a_2 (1 - p') - 2a_1 \pi_2^* - (\pi_1^* a_1 + \pi_2^* a_2) (\pi_1^* + \pi_2^*)$$

$$+ \pi_1^*(1 - p) + \pi_2^*(1 - p')(a_1 - a_2)^2$$

$$+ 2(a_2 - a_1) \left[ \frac{\pi_2^* a_2 (1 - p') - \pi_1^* a_1 (1 - p)}{p + p' - 2} \right]$$

$$+ \frac{\pi_1^* a_1 + \pi_2^* a_2 (\pi_1^* - p \pi_1^* + \pi_2^* + p' \pi_2^*)}{p + p' - 2}$$

$$a^* := a_1 \pi_1^* + a_2 \pi_2^*$$

where $(p, p')$ are the transition probabilities of the Markov chain.

Corollary 2 (LLN for GCHP2SDO). The process $S_{nt}$ defined in Definition 9 satisfies the following weak convergence in the Skorokhod topology (see Getoor 1967)

$$\frac{S_{nt}}{n} \xrightarrow{n \to \infty} \hat{a}^* \frac{\lambda}{1 - \hat{\mu}}$$
where $\hat{\mu}$ and $a^*$ are defined in Equations (40) and (41), respectively.

Now, let us consider the process $S_t$ defined in Definition 10, specifically

$$S_t = S_0 + \sum_{k=1}^{N(t)} X_k$$

where $X_k \in \{-\delta, \delta\}$ is a continuous-time 2-state Markov chain, $\delta$ is the fixed tick size, and $N(t)$ is the number of price changes up to moment $t$, described by a one-dimensional Hawkes process defined in Definition 4. In this case, we have a fixed tick size, two-valued price changes and dependent orders.

**Corollary 3** (Diffusion Limit for CHPDO). Let $X_k$ be an ergodic Markov chain with two states $\{\delta, -\delta\}$ and with ergodic probabilities $(\pi^*, 1 - \pi^*)$. Taking $S_t$, as defined in Definition 10, we have

$$\frac{S_{nt} - N(nt)a^*}{\sqrt{n}} \xrightarrow{n \to \infty} \sigma \sqrt{\frac{\lambda}{1 - \hat{\mu}}} W(t),$$

where $W(t)$ is a standard Wiener process,

$$0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1 \text{ and } \int_0^\infty \mu(s)ds < \infty$$

$$a^* := \delta(2\pi^* - 1)$$

$$\sigma^2 := 4\delta^2 \left( \frac{1 - p' + \pi^*(p' - p)}{(p + p' - 2)^2} - \pi^*(1 - \pi^*) \right),$$

and $(p, p')$ are the transition probabilities of the Markov chain $X_k$. We note that $\lambda$ and $\mu(t)$ are defined in Equation (2).

We also note that the LLN for both GCHP2SDO and CHPDO is identical to the result given in Lemma 2, after some simplification. Furthermore, the result of Corollary 3 was first stated in Swishchuk et al. (2017), and the CHPDO model was justified there by using CISCO data dated from 3 November to 7 November 2014.

4. **Empirical Results**

4.1. **Description of the Data Set**

In order to test the validity of our models and determine which best fits empirical data, we have considered level 1 LOB data (meaning the limit orders sitting at the best bid and ask) for Apple, Amazon, Google, Microsoft and Intel on 21 June 2012. We first verify that the data are reasonable for our model by checking its liquidity. That is illustrated in Table 1.

**Table 1.** Stock liquidity of AAPL, AMZN, GOOG, MSFT, and INTC for 21 June 2012.

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Avg # of Orders per 10 s</th>
<th>Price Changes in 1 Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>51</td>
<td>64,350</td>
</tr>
<tr>
<td>AMZN</td>
<td>25</td>
<td>27,557</td>
</tr>
<tr>
<td>GOOG</td>
<td>21</td>
<td>24,084</td>
</tr>
<tr>
<td>MSFT</td>
<td>173</td>
<td>3217</td>
</tr>
<tr>
<td>INTC</td>
<td>176</td>
<td>4060</td>
</tr>
</tbody>
</table>
The high number of daily price changes motivates the idea that we can use asymptotic analysis in order to approximate long-run volatility using order flow by finding the diffusion limit of the price process. Because we do not want to include opening and closing auctions, we omit the first and last fifteen minutes of our data. We motivate the arrival process by analyzing the inter-arrival times and clustering to ensure the arrival process is not Poisson and exhibits the characteristics of a Hawkes Process. This is illustrated in Figures 1 and 2.

Figure 1. Above, we provide a quantile-quantile plot of our empirical inter-arrival times against a Poisson process for each of the five stocks. We see that the inter-arrival data does not fit the expected curve, providing evidence that the underlying arrival process is not Poisson.
Figure 2. Each plot shows the number of arrivals for a moving one minute window. From this, we can conclude that there is a significant amount of clustering in the arrival of mid-price changes, motivating the Hawkes model of our arrival process.

Furthermore, the relationship between our diffusion coefficient and arrival process is limited to the expected number of arrivals on a unit interval. This implies that results for a simple exponential model can be easily generalized to a nonlinear one. This makes it possible to work with a simplified model for our Hawkes process, which is still rich enough to capture our observations. Keeping this in mind, we restrict ourselves to an exponential kernel and estimate parameters using an MLE Laub et al. (2015). We provide these estimates in Table 2 and compare the empirical expected number of arrivals and compare with the MLE estimate in Table 3.
Table 2. Each parameter was estimated using a particle swarm optimization method in an attempt to globally optimize the negative log-likelihood function. The values for $\lambda$, $\alpha$ and $\beta$ for each data set are as provided.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>1.4683</td>
<td>1045.2676</td>
<td>2556.1844</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.6443</td>
<td>653.7524</td>
<td>1556.1702</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.4985</td>
<td>865.8553</td>
<td>1980.4409</td>
</tr>
<tr>
<td>MSFT</td>
<td>0.0659</td>
<td>479.3482</td>
<td>908.0032</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0471</td>
<td>399.6389</td>
<td>760.4991</td>
</tr>
</tbody>
</table>

Obviously, the MLE method accurately estimates the expected number of arrivals on the unit interval. This means we can confidently say for our data that our parameters will work reasonably with our models. We provide the estimated parameters in Table 2. We note that the novelty of the empirical results here compared with He and Swishchuk (2019) lies not only in the different stocks but also in considering the nonlinear Hawkes process as the model for a stock price.

Table 3. Expected number of arrivals on a unit interval using estimated parameters from an MLE method is compared against the empirical arrivals.

<table>
<thead>
<tr>
<th></th>
<th>Emp. $E[N([0, 1])]$</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2.4840</td>
<td>2.4841</td>
</tr>
<tr>
<td>AMZN</td>
<td>1.1110</td>
<td>1.1110</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.8857</td>
<td>0.8857</td>
</tr>
<tr>
<td>MSFT</td>
<td>0.1395</td>
<td>0.1396</td>
</tr>
<tr>
<td>INTC</td>
<td>0.0991</td>
<td>0.0992</td>
</tr>
</tbody>
</table>

Now that we have approximated the underlying arrival process, we quickly review our process for obtaining our empirical results:

1. Estimate the parameters of the arrival process using particle swarm optimization to maximize the negative loglikelihood function, as above.
2. Model the price change caused by an event using an $\mathcal{N}$-state Markov chain, with probabilities estimated by the relative frequencies.
3. Use the diffusion limits determined previously to calculate the theoretical volatility and compare against what is observed.

4.2. Compound Hawkes Process with Dependent Orders (CHPDO)

We first consider a compound Hawkes process with dependent orders defined in Definition 10, namely

$$S_t = S_0 + \sum_{k=1}^{N(t)} X_k$$

(47)

where $X_k \in \{+\delta, -\delta\}$ is a continuous-time two state Markov chain, $\delta$ is of fixed size and $N(t)$ is the number of mid-price movements up to time $t$ described by a Hawkes process.

We have opted to study the mid-price changes of our model. Thus, $S_t$ can be computed by averaging the best bid and ask price. Noting that the price is recorded in cents, the smallest possible jump in the mid-price is a half a cent, which we will use as $\delta$. Furthermore, in order to estimate the transition matrix for the Markov chain $X_k$, we count the absolute frequencies of upward and downward price movements and from this calculate the relative frequencies giving an estimate for $p_{uu}$ and $p_{dd}$, which represent the conditional probabilities of an up/down movement given an up/down movement. Later, we will consider several different sizes of mid-price movements and will work with the convention that each movement will be assigned a state based off of its ordering in the reals.
In this case, $-\delta$ will be state one, and $\delta$ will be state two. This results in the transition matrix $P$ given below:

$$
P = \begin{bmatrix}
p_{dd} & 1 - p_{dd} \\
1 - p_{uu} & p_{uu}
\end{bmatrix}
$$

After determining our parameters and transition probabilities, we calculate $a^*$ and $\sigma$ in Table 4 together with $p_{uu}$ and $p_{dd}$.

Table 4. Provided above are the values for $s^*$, $\sigma$ as well as the probabilities of an upward/downward movement given an upward/downward movement for each of the five stocks in question.

<table>
<thead>
<tr>
<th>Stock</th>
<th>$p_{dd}$</th>
<th>$p_{uu}$</th>
<th>$\sigma$</th>
<th>$a^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>0.4956</td>
<td>0.4933</td>
<td>0.0049</td>
<td>$-1.1463 \times 10^{-5}$</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.4635</td>
<td>0.4576</td>
<td>0.0046</td>
<td>$-2.7373 \times 10^{-5}$</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.4769</td>
<td>0.4461</td>
<td>0.0046</td>
<td>$-1.4301 \times 10^{-4}$</td>
</tr>
<tr>
<td>MSFT</td>
<td>0.6269</td>
<td>0.5827</td>
<td>0.0062</td>
<td>$-2.7956 \times 10^{-4}$</td>
</tr>
<tr>
<td>INTC</td>
<td>0.6106</td>
<td>0.5588</td>
<td>0.0059</td>
<td>$-3.1185 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Provided these values, the asymptotic method from Section 3 and its Corollaries can be used to study the link between order flow in our model and price volatility. The estimated volatility of price changes is expressed in terms of parameters describing arrival rate of limit orders. Allowing us to test our claim that our model accurately describes the mid-price process. We recall that the mid-price $= (\text{bidprice} + \text{askprice})/2$, and the following diffusion limit

$$
\frac{S_{nt} - N(nt)a^*}{\sqrt{n}} \xrightarrow{n \to \infty} \sigma \sqrt{\frac{\lambda}{1 - \alpha/\beta}} W(t).
$$

(48)

If the data satisfy our proposed model, then, after considering large windows of time (5 min, 10 min, 20 min), we would expect to see the empirical and theoretical standard deviations to follow each other closely. To test this, we compare the equivalent process, constructed by multiplying the LHS and RHS by $\sqrt{n}$. Then, cutting our data into disjoint windows of size $n$, specifically $[in, (i + 1)n]$ with $t = 1$ and by setting the left bound as our starting time, we can calculate $S_{nt} - N(nt)a^*$ for each individual window and give a generalized formula for this below:

$$
S_{i}^* = S_{(i+1)nt} - S_{nt} - (N(((i + 1)nt) - N(nt))a^*
$$

(49)

This gives a collection of values $\{S_i^*\}$ over which we compute the standard deviation. If our model is accurate, we would would expect that

$$
\text{std}\{S_i^*\} \approx \sqrt{n}t\sigma \sqrt{\frac{\lambda}{1 - \alpha/\beta}}, \text{ where } t = 1.
$$

(50)

We plot the empirical standard deviation against the theoretical one for various window sizes starting at 10 s and increasing in steps of 10 s until we reach 20 min; this is illustrated in Figure 3.

Several important remarks should be made at this point. It is clear that, while the model accurately predicts the overall trend for MSFT and INTC, we severely underestimate the variability in the mid-price process for APPL, AMZN and GOOG. Furthermore, as the window size increases, the overall spread in the data increases. We attribute this to the decreasing sample size imposed on us as we increase the window size. For example, when we consider a 20 min window, we can only construct 27 disjoint windows in the 9 h trading day forcing us to deal with the problem of predicting a ‘population’ standard deviation from a increasingly small sample. We remedy this by using a variance stabilizing transformation in later sections. Specifically, a popular method for a Poisson process is to
take the square root of our empirical and theoretical standard deviations. This makes it possible to qualitatively view the overall trend in the data, gaining a clearer idea of goodness of fit from there.

![Figure 3](image)

**Figure 3.** Each figure compares the empirical standard deviation for a fixed window size to the theoretical standard deviation. We have plotted an empirical standard deviation for all \( n \) from 10 s to 20 min in step sizes of 10 s. Each empirical standard deviation corresponds to a single point in the scatter plot, and the plotted curve corresponds to the predicted theoretical value.

### 4.3. General Compound Hawkes Process with Two Dependent Orders

As of now, we have considered a fixed \( \delta \) related to the trading tick size. However, if we consider the mid-price changes for APPL, AMZN and GOOG the assumption of a fixed tick size is violated. In fact, we observe that approximately 61%, 53% and 71% of all mid-price changes are larger than half
a tick size, which is opposed to what we observe for MSFT and INTC where all mid-price changes occur at the half tick size, we illustrate this for AAPL, AMZN and GOOG in Figure 4.

![Figure 4](image_url)

**Figure 4.** We can see clearly that the change in the mid-price is often larger than a half tick. These mid-price changes make up a significant portion of the actual data, contradicting the assumption needed for the CHPDO model that the mid-price changes occur on average at a half tick size.

It is clear in Figure 4 that additional considerations need to be made. A simple way to include the variability in mid-price movements in our model is to introduce \( a(X_i) \) as described in Definition 9. It is of course necessary to determine the values of \( a(\cdot) \) for each state of our Markov chain. A naive method is to take the mean of the downward and upward mid-price movements and assign them to \( a(1) \) and \( a(2) \), respectively. We provide these values in Table 5.

<table>
<thead>
<tr>
<th></th>
<th>( a(1) )</th>
<th>( a(2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>-0.0172</td>
<td>0.0170</td>
</tr>
<tr>
<td>AMZN</td>
<td>-0.0134</td>
<td>0.0133</td>
</tr>
<tr>
<td>GOOG</td>
<td>-0.0302</td>
<td>0.0308</td>
</tr>
</tbody>
</table>

Table 5. \( a(i) \) is the average of the upward or downward mid-price movements. Following our previous convention, the first state will be associated with the mean of all downward mid-price movements and the second state will be associated with the mean of all upward mid-price movements.

In this step, we have only endeavoured to better realize the actual price movements in our data. Therefore, when we observe a downward mid-price movement, we continue to assign it to state one and similarly for an upward price movement we continue to assign it to state two. It follows that our transition matrix will remain the same. Then, using these new state values, we recalculate \( \alpha^* \) and \( \sigma \),
providing them in Table 6. The effect of these changes is investigated in Figure 5. Note that in Figure 5 we have used the variance stabilizing transformation discussed earlier in order to better visualize the overall trend in our data.

Table 6. Above, we have the values for $a^*$, $\sigma$, as well as the probabilities of an upwards/downwards movement, given an upwards/downwards movement for the three stocks of interest.

<table>
<thead>
<tr>
<th></th>
<th>$p_{dd}$</th>
<th>$p_{uu}$</th>
<th>$\sigma$</th>
<th>$a^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>0.4956</td>
<td>0.4933</td>
<td>0.0169</td>
<td>$-1.5624 \times 10^{-4}$</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.4635</td>
<td>0.4576</td>
<td>0.0123</td>
<td>$-1.0475 \times 10^{-4}$</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.4769</td>
<td>0.4461</td>
<td>0.0282</td>
<td>$-5.5095 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Notice that there is a significant qualitative improvement in the fits for AAPL and GOOG in Figure 5, but the variability in mid-price movements for AMZN is still clearly underestimated by our model. The unexplained variance may be captured by investigating an $N$-state Markov chain since the additional transition probabilities could explain the variability missing in the 2-state case.

Figure 5. A comparison of the empirical standard deviation for a fixed window size $n$ to the theoretical standard deviation for AAPL, AMZN and GOOG using the 2-state dependent order model. We have plotted the empirical standard deviation for all $n$ from 10 s to 20 min in step sizes of 10 s. Each empirical standard deviation corresponds to a single point in the scatter plot and the plotted curve corresponds to the predicted theoretical value. Visually, there is a significant improvement for all stocks, although the theoretical standard deviation for AMZN is still underestimating the empirical variability.
4.4. General Compound Hawkes Process with \( N \) Dependent Orders

We recall the \( N \)-state model described in Definition 8. The immediate question becomes how best to choose the state values. We modify the quantile based approach from Swishchuk et al. (2017). After calculating the mid-price changes, we separate the data into upward and downward price movements. Then, we calculate evenly distributed quantiles for both data sets. Depending on the data, several quantiles may be identical, we reject any duplicates. We thus obtain a list of bounds which we complete by adding the minimum observed value if necessary.

To determine the state values \( a(X_i) \), we take the average of all mid-price changes located between two neighbouring boundary values. Furthermore, we assign a mid-price change to state \( i \) if it is greater than or equal to the \((i - 1)\)th boundary and strictly less than the \(i\)th boundary. An exception is made for the largest upper bound where equality is permitted at both ends.

As we could not capture the full variability of mid-price changes for AMZN in the previous method, we investigate it for this case. Furthermore, for tractability, we only consider 14 boundary values from which we obtain a 12-state Markov chain. Instead of providing the transition matrix, we provide the ergodic probabilities for the transition matrix and the associated states in Table 7.

| Table 7. Above, we have provided the state, associated ergodic probabilities and state values \( a(i) \) for AMZN, given a 12-state Markov chain that was obtained from choosing a 16 quantile method. |
|---|---|---|
| AMZN | \( \pi_i^* \) | \( a(X_i) \) |
| 1 | 0.0275 | -0.0524 |
| 2 | 0.0281 | -0.0318 |
| 3 | 0.0264 | -0.0250 |
| 4 | 0.0382 | -0.0200 |
| 5 | 0.0576 | -0.0150 |
| 6 | 0.3249 | -0.0064 |
| 7 | 0.2321 | 0.0050 |
| 8 | 0.0923 | 0.0100 |
| 9 | 0.0578 | 0.0150 |
| 10 | 0.0353 | 0.0200 |
| 11 | 0.0412 | 0.0271 |
| 12 | 0.0387 | 0.0476 |

In order to compare the two state and \( N \)-state approaches, we first take a qualitative approach and plot the two theoretical and empirical standard deviations against each other in Figure 6. When we compare the mean squared residuals, the 2-state model discussed before has mean squared error 0.0208 while our 12-state Markov chain brings that down to 0.0125. Considering an even larger Markov chain with 24-states, we are only able to obtain a meager improvement to 0.0123, suggesting that there is some underlying variance in the mid-price process not captured by our model. We investigate these more quantitative measures in the following subsection. First, comparing the models against a numerical best fit, and then investigating the mean squared error, we gain a better quantitative understanding of the overall improvement obtained from each model increasing the number of quantiles. While we do not propose a method for selecting an \( N \)-state model in general, a possible piece of criteria is to halt when there is no appreciable gain in the MSE for more information (see Swishchuk et al. 2017, p. 17). In light of this, in the following section, we use k-fold cross-validation to make a determination.
Figure 6. We consider the N-state model for AMZN discussed previously in the paper. While there is a slight improvement against the original fit, the model still struggles to perfectly predict the variability in the mid-price changes of our data.

5. Quantitative Analysis

While our model does visually appear to fit the expected variability in four of the five cases, it still fails to capture the complete dynamics of mid-price changes seen in our AMZN data. We investigate the mean square error of our models with a varying number of quantiles in Table 8. This gives a good indication as to whether the $N$-state model is a better fit for our data. If we look closely at AAPL and GOOG, we see that the $N$-state case can still improve our results from the 2-state case. For AAPL, we constructed a 17-state Markov chain by taking 16 quantiles on the downward movements and 16 quantiles upward movements. This resulted in a mean squared error of 0.0036 that is approximately a 28% improvement to the 2-state case where the mean squared error was 0.0050. Even more extreme, using a 25-state Markov chain for GOOG which was constructed similarly, we observed a mean squared error of 0.0046, which is a 60% improvement from the 2-state case with a mean squared error of 0.0115. We conclude with Table 8, which provides the mean residuals for AAPL, AMZN and GOOG with several Markov chains constructed from various numbers of quantiles. We also include mean residuals for INTC and MSFT for comparison.

Another quantitative measure of our fit would be to compare them to the one which best minimizes the residuals. Notice that each of our models assumes that the standard deviation proportionally to the square root of the time step. Therefore, we can estimate the best possible coefficient by minimizing $L^2$-norm or equivalently the mean squared error. We provide plots of these hypothetical best fits against the empirical data and theoretical fits in Figure 7. We also provide the coefficients for the theoretical fits and regression in Table 9 in order to have a more quantitative comparison. They are calculated using least-square regression.
Table 8. We list the mean residuals for several Markov chains with varying numbers of states. These were generated using our modified quantile approach choosing to start with 2, 8, 16 or 32 quantiles. We see that, in general, the mean residual decreases to some lower limit where we can no longer perform any better. Recall that the only observed mid-price changes for INTC and MSFT were of a half tick size, and any increase in the number of quantiles will result in the same performance.

<table>
<thead>
<tr>
<th></th>
<th>CHPDO</th>
<th>2</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>0.2679</td>
<td>0.0050</td>
<td>0.0036</td>
<td>0.0036</td>
<td>0.0036</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.1122</td>
<td>0.0208</td>
<td>0.0131</td>
<td>0.0124</td>
<td>0.0123</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.4036</td>
<td>0.0115</td>
<td>0.0048</td>
<td>0.0045</td>
<td>0.0047</td>
</tr>
<tr>
<td>INTC</td>
<td>1.7917×10^{-5}</td>
<td>1.7917×10^{-5}</td>
<td>1.7917×10^{-5}</td>
<td>1.7917×10^{-5}</td>
<td>1.7917×10^{-5}</td>
</tr>
<tr>
<td>MSFT</td>
<td>1.0586×10^{-4}</td>
<td>1.0586×10^{-4}</td>
<td>1.0586×10^{-4}</td>
<td>1.0586×10^{-4}</td>
<td>1.0586×10^{-4}</td>
</tr>
</tbody>
</table>

We notice that in each case the errors are close to, or under five percent, with AMZN being the biggest offender. This is consistent with the discussion provided throughout our analysis and highlights the general applicability of our model.

Table 9. The coefficients calculated for AAPL, AMZN and GOOG are generated using a Markov chain created by 16 quantiles on the upward and downward movements, while the coefficients for INTC and MSFT are obtained from the CHPD0 case.

<table>
<thead>
<tr>
<th></th>
<th>Theoretical Coefficient</th>
<th>Regression Coefficient</th>
<th>Percent Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>0.02868</td>
<td>0.02828</td>
<td>1.42%</td>
</tr>
<tr>
<td>AMZN</td>
<td>0.01450</td>
<td>0.01831</td>
<td>20.8%</td>
</tr>
<tr>
<td>GOOG</td>
<td>0.02883</td>
<td>0.03023</td>
<td>4.63%</td>
</tr>
<tr>
<td>INTC</td>
<td>0.00186</td>
<td>0.00193</td>
<td>3.4%</td>
</tr>
<tr>
<td>MSFT</td>
<td>0.00231</td>
<td>0.00246</td>
<td>6.4%</td>
</tr>
</tbody>
</table>

Cross-Validation

We conclude this subsection using a method similar to k-fold cross-validation for model selection. Generally, the data are shuffled randomly and then partitioned into k sets of approximately equal size. From these k sets, one is selected for testing the model while the remaining are kept for training it. For our data, we are unable to perform this shuffling as there is an important ordering imposed by the arrival times. Fortunately, since we are considering high-frequency models for the mid-price changes, we expect a minimal amount of correlation in the volatility of the price process over large windows of time. Therefore, dividing the data into twelve 30 min windows selecting them one at a time should still be meaningful.

The training data will need to be glued back together after the thirty-minute window is removed. This is done by assuming the events never occurred. We provide an example below for how one would remove the second thirty-minute window in Tables 10 and 11.

This introduces a small error into the arrival times, but only at a single point in a data set of several thousand arrival times. Had we randomly removed points, or shuffled the data, we could introduce many errors which could significantly change the underlying arrival process. At this point, we perform the same fitting procedure on the training data and after the fitting is completed, we can compute the empirical standard deviations on the testing data to obtain a mean squared error. This determines how well the model was able to predict the training data, calculating this for each window of time and averaging provides a score for how well the model was able to make predictions.

We recall that, for MSFT and INTC, the only observed prices changes were a single tick, and it is not possible to outperform the original CHPD0 model. Instead, we consider AAPL, AMZN and GOOG calculating scores for CHPD0, GCHP2SDO and various GCHPnSDO models. Specifically the set of GCHPnSDO models are constructed using the quantile method, allowing for 2–14 quantiles.
This was chosen only to make the recording more tractable; the resulting scores are provided in Table 12.

**Figure 7.** A qualitative comparison of the regression to the theoretical model. For APPL, AMZN and GOOG, we have used a Markov chain generated from 16 quantiles taken on the upward movements and downward movements. For INTC and MSFT, we have taken the CHPDO coefficient since a different coefficient is not possible with the other models.
Table 10. An imagined sequence of price change events before the second thirty minute window has been removed.

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>···</th>
<th>1789</th>
<th>1795</th>
<th>1803</th>
<th>···</th>
<th>3193</th>
<th>3601</th>
<th>3608</th>
<th>···</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>···</td>
<td>n − 1</td>
<td>n</td>
<td>n + 1</td>
<td>···</td>
<td>n + k</td>
<td>n + k + 1</td>
<td>n + k + 2</td>
<td>···</td>
</tr>
</tbody>
</table>

Table 11. An imagined sequence of price change events after the second thirty minute window has been removed.

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>···</th>
<th>1789</th>
<th>1795</th>
<th>1801</th>
<th>1808</th>
<th>···</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>···</td>
<td>n − 1</td>
<td>n</td>
<td>n + k + 1</td>
<td>n + k + 2</td>
<td>···</td>
</tr>
</tbody>
</table>

Table 12. A table of testing scores for various models. Computations were performed for up to the 40-quantile case for each stock, and this table represents a sample of that data. We note that the models stop obtaining appreciable performance gains at around 4–7 quantiles, fluctuating around the same scores.

<table>
<thead>
<tr>
<th></th>
<th>AAPl</th>
<th>AMzn</th>
<th>GOog</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHPDO</td>
<td>0.0348</td>
<td>0.0142</td>
<td>0.0436</td>
</tr>
<tr>
<td>GCHP2SDO</td>
<td>0.0049</td>
<td>0.0048</td>
<td>0.0044</td>
</tr>
<tr>
<td>2-Quantiles</td>
<td>0.0042</td>
<td>0.0041</td>
<td>0.0036</td>
</tr>
<tr>
<td>3-Quantiles</td>
<td>0.0041</td>
<td>0.0042</td>
<td>0.0035</td>
</tr>
<tr>
<td>4-Quantiles</td>
<td>0.0040</td>
<td>0.0040</td>
<td>0.0035</td>
</tr>
<tr>
<td>5-Quantiles</td>
<td>4.0341 × 10⁻³</td>
<td>3.9878 × 10⁻³</td>
<td>3.4593 × 10⁻³</td>
</tr>
<tr>
<td>6-Quantiles</td>
<td>4.0362 × 10⁻³</td>
<td>3.8971 × 10⁻³</td>
<td>3.4168 × 10⁻³</td>
</tr>
<tr>
<td>7-Quantiles</td>
<td>4.0184 × 10⁻³</td>
<td>3.8971 × 10⁻³</td>
<td>3.3977 × 10⁻³</td>
</tr>
<tr>
<td>8-Quantiles</td>
<td>4.0097 × 10⁻³</td>
<td>3.8656 × 10⁻³</td>
<td>3.3881 × 10⁻³</td>
</tr>
<tr>
<td>9-Quantiles</td>
<td>4.0026 × 10⁻³</td>
<td>3.8124 × 10⁻³</td>
<td>3.3976 × 10⁻³</td>
</tr>
<tr>
<td>10-Quantiles</td>
<td>3.9984 × 10⁻³</td>
<td>3.8124 × 10⁻³</td>
<td>3.425 × 10⁻³</td>
</tr>
<tr>
<td>11-Quantiles</td>
<td>3.9971 × 10⁻³</td>
<td>3.81594 × 10⁻³</td>
<td>3.3867 × 10⁻³</td>
</tr>
<tr>
<td>12-Quantiles</td>
<td>3.9826 × 10⁻³</td>
<td>3.81463 × 10⁻³</td>
<td>3.3906 × 10⁻³</td>
</tr>
<tr>
<td>13-Quantiles</td>
<td>3.9804 × 10⁻³</td>
<td>3.81517 × 10⁻³</td>
<td>3.3562 × 10⁻³</td>
</tr>
<tr>
<td>14-Quantiles</td>
<td>3.9725 × 10⁻³</td>
<td>3.81427 × 10⁻³</td>
<td>3.3549 × 10⁻³</td>
</tr>
</tbody>
</table>

Overall, there is a meaningful gain in performance for the \( N \)-state models, especially when we take into account the previous discussion. The cross-validation we’ve performed provides a reasonable criteria for model rejection, similar to when we stop observing appreciable gains in the mean squared error. From our analysis, the best model would be one constructed from 4–7 quantiles, depending on the stock of interest.

6. Conclusions and Future Work

Overall, the \( N \)-state model outperforms the others when the number of states is kept small. It generates fits for four out of the five datasets that are reasonable and providing reductions in the mean squared error by upwards of 25%. While not able to capture the full dynamics observed in AMZN, it appears to be a strong candidate for a simple model of price dynamics observed in our data. Further investigation would be necessary to determine what causes the additional volatility observed in AMZN, and potentially implement a more robust model which captures this. The potential users of our models are practitioners working in the financial industry who wish to implement our results in high-frequency and algorithmic trading combined with their industrial knowledge.

Author Contributions: Both authors have contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSERC Grant No. RT732266.
Conflicts of Interest: The authors declare no conflict of interest.

References


He, Qiyue, and Anatoliy Swishchuk. 2019. Quantitative and comparative analyses of limit order books with general compound hawkes processes. Risks 7: 110. [CrossRef]
Vadori, Nelson, and Anatoliy Swishchuk. 2015. Strong law of large numbers and central limit theorems for functionals of inhomogeneous semi-markov processes. Stochastic Analysis and Applications 33: 213–43. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).