risks

MDPI

Article

Quantitative and Comparative Analyses of Limit Order Books with General Compound Hawkes Processes

Qiyue He * and Anatoliy Swishchuk

Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary, AB T2N1N4, Canada; aswish@ucalgary.ca

* Correspondence: qiyue.he@ucalgary.ca

Received: 30 June 2019; Accepted: 25 October 2019; Published: 1 November 2019

Abstract: In this paper, we solve the problem of mid price movements arising in high-frequency and algorithmic trading using real data. Namely, we introduce different new types of General Compound Hawkes Processes (GCHPDO, GCHP2SDO, GCHPnSDO) and find their diffusive limits to model the mid price movements of 6 stocks-EBAY, FB, MU, PCAR, SMH, CSCO. We also define error rates to estimate the models fitting accuracy. Maximum Likelihood Estimation (MLE) and Particle Swarm Optimization (PSO) are used for Hawkes processes and models parameters' calibration.

Keywords: general compound Hawkes process; diffusive limit; maximum likelihood estimation (MLE); particle swarm optimization (PSO); error measurements

1. Introduction

Big data has now become a driver of model building and analysis in a number of areas, including finance. More than half of the markets in today's highly competitive and relentlessly fast-paced financial world now use a limit order book (LOB) mechanism to facilitate trade. The main problem here is how to deal with big data arising in electronic markets for algorithmic and high-frequency (milliseconds) trading (HFT).

The present paper introduces new different types of general compound Hawkes processes (GCHP)—namely GCHPDO (two fixed ticks, $+\delta$, $-\delta$, dependent orders), GCHP2SDO (two non-fixed ticks, $(a(1), a(2))$, dependent orders) and GCHPnSDO (n non-fixed ticks, $(a(1), \ldots, a(n))$, dependent order)—to model the mid prices S_t dynamics of the assets in HFT, namely, EBAY, FB, MU, PCAR, SMH, CSCO, (provided by Reference Cartea et al. (2015)).

As we mentioned, high-frequency trading happens in milliseconds, as order arrivals and cancellations are very frequent. How we can study and model the dynamics of the mid-prices? One of the ways is to look over a larger time scale, for example, 5, 10 or 20 min, that is, consider time scale nt instead of t, then n could be $n = 100, 1000, \ldots$, etc. It means that we consider the dynamics of order flow over large time scales. Thousands of order book events may occur over such large time scales (e.g., for CISCO data on one day in 2014 it is around 500,000, see Reference Cartea et al. (2015)). Thus, we can use asymptotic methods to study the link between order flow and price volatility by considering the diffusive limit of the mid-price processes. More precisely, we use the functional central limit theorems for above-mentioned GCHP and present the volatility of price changes in terms of the parameters of initial models.

To estimate the models fitting accuracy we define an error rate for each model and set the threshold value as 15%. The model with error rate less than the threshold is considered as well fitted. Thus, we define which of our models is the best fit for our real data.

We also use Maximum Likelihood Estimation (MLE) and Particle Swarm Optimization (PSO) for Hawkes processes and our parameters' calibration.

There are many papers devoted to the modelling of HFT and applications of Hawkes processes in finance. See Reference Bacry et al. (2015) for more details. Below we give an overview of the most relevant literature.

Reference Cont and De Larrard (2013) proposed a simple Markovian stochastic model for the dynamics of a limit order book, in which arrivals of market orders, limit orders and order cancellations are independent and inter-arrival times have exponential distribution. They also studied the diffusion limit of the price process and expressed the volatility of price changes in terms of parameters describing the arrival rates of buy and sell orders and cancellations.

As suggested by empirical observations, Reference Swishchuk and Vadori (2017) extended their framework to (1) arbitrary distributions for book events inter-arrival times (possibly non-exponential) and (2) both the nature of a new book event and its corresponding inter-arrival time depend on the nature of the previous book event. They did so by resorting to Markov renewal processes to model the dynamics of the bid and ask queues. They kept analytical tractability via explicit expressions for the Laplace transforms of various quantities of interest. They also justified and illustrated their approach by calibrating their model to the five stocks Amazon, Apple, Google, Intel, Microsoft on 21 June 2012, to the 15 stocks from Deutsche Boerse Group (23 September 2013) and to CISCO asset (3 November 2014). As in Reference Cont and De Larrard (2013), the bid-ask spread remains constant equal to one tick, only the bid and ask queues are modelled (they are independent from each other and get reinitialized after a price change) and all orders have the same size. We discussed possible extensions of our model for the case when the spread is not fixed, including the diffusion limit of the price dynamics in this case and we also discussed stochastic optimal control and market making problems.

The paper by Swishchuk and Hofmeister Swishchuk et al. (2017) considered a general semi-Markov model for limit order books with two states that incorporate price changes that are not fixed to one tick. Furthermore, even more general cases of the semi-Markov model for limit order books was introduced that incorporates an arbitrary number of states for the price changes. For both cases the justifications, diffusion limits, implementations and numerical results were presented for different limit order book data—Apple, Amazon, Google, Microsoft, Intel on 21 June 2012 and Cisco, Facebook, Intel, Liberty Global, Liberty Interactive, Microsoft, Vodafone from 3 November 2014 to 7 November 2014. Reference Chavez-Casillas et al. (2019) proposed a simple stochastic model for the dynamics of a limit order book, extending the recent work of Cont and De Larrard (2013), where the price dynamics are endogenous, resulting from market transactions. They also showed that the conditional diffusion limit of the price process is the so-called Brownian meander.

Trading activity is not a completely random and memoryless process. That is why the Poisson process is not suitable for modelling trade arrival times. Trading activity also shows clustering behaviour. These properties suggest the use of the Hawkes process, a point process mathematically defined by Reference Hawkes (1971), which is an extension of the classical Poisson process that possesses this clustering property. It explains the large number of works on trading activity and more generally high-frequency econometrics based on this process as a modelling framework. (See Reference (Bacry et al. 2015) for applications of Hawkes processes in finance).

Reference Da Fonseca and Zaatour (2013) provided explicit formulas for the moments and the autocorrelation function of the number of jumps over a given interval for a self-excited Hawkes process. The estimation strategy was applied to trade arrival times for major stocks that show a clustering behaviour, a feature the Hawkes process can effectively handle. As the calibration is fast, the estimation was rolled to determine the stability of the estimated parameters. Also, the analytical results enable the computation of the diffusive limit in a simple model for the price evolution based on the Hawkes process. It determines the connection between the parameters driving the high-frequency activity to the daily volatility.

Reference Swishchuk et al. (2019) introduced two new Hawkes processes, namely, compound and regime-switching compound Hawkes processes, to model the price processes in limit order books. They proved Law of Large Numbers and Functional Central Limit Theorems (FCLT) for both processes. The two FCLTs were applied to limit order books, where they used these asymptotic methods to study the link between price volatility and order flow in these two models by using the diffusion limits of these price processes. The volatilities of price changes were expressed in terms of parameters describing the arrival rates and price changes. They also presented some numerical examples based on CISCO data (3–7 November 2014).

Reference Swishchuk and Huffman (2018) studied various new Hawkes processes. Specifically, they constructed general compound Hawkes processes and investigate their properties in limit order books. With regards to these general compound Hawkes processes, they proved a Law of Large Numbers (LLN) and a Functional Central Limit Theorems (FCLT) for several specific variations. Then they applied several of these FCLTs to limit order books in Lobster data (21 June 2012) 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. Quantitative and comparative analyses were performed for different models.

When it comes to multivariate cases of Hawkes Processes, Reference Bacry et al. (2013) proved a law of large numbers, a functional central limit theorem, and the asymptotic behaviour. Reference Embrechts et al. (2011) derived the maximum likelihood estimation and goodness-of-fit. As an Application, they analyzed the data sets from finance and fit the multivariate Hawkes process to daily closing values from the Dow Jones Industrial Average from 1 January 1994 to 31 December 2010. Reference Zhou et al. (2013) applied multivariate Hawkes processes to social network analysis and showed that the proposed method performed significantly better on both synthetic and real world datasets.

The rest of the paper is organized as follows. Different type of Hawkes processes and diffusive limits for them are introduced in Section 2. Applications to limit order books are considered in Section 3, including data and their clustering features descriptions, QQ-plots and autocorrelations. Section 4 presents Hawkes process and different models' parameters calibration's results. Error measurements and comparative analysis are considered in Section 5. Section 6 concludes the paper.

2. Theoretical Analysis

2.1. One-Dimensional Hawkes Process

2.1.1. Definition

The one-dimensional Hawkes Process is a point process $N(t)$ which is characterized by its intensity function $\lambda(t)$

$$\lambda(t) = h(\lambda + \int_0^t \mu(t-s)dN(s)) \tag{1}$$

where the constant λ is called the background intensity and the function $\mu(\cdot)$ is called the excitation function that satisfies $\int_0^\infty \mu(s)ds < 1$. In Equation (1), if $h(\cdot)$ is a non-linear function, then we call this Hawkes Process as One-dimensional Non-linear Hawkes Process.

Typically, in this paper, the point process $N(t)$ is represented as the cumulative number of mid-price changes up to current time t in the trading system. $h(\cdot)$ is set to be a linear function, that is, $h(x) = x$, and $\mu(\cdot)$ is set to be an exponential function, that is, $\mu(t) = \alpha e^{-\beta t}$. This typical setting is widely used in most of the academic research (see Laub et al. 2015). Therefore, in this paper, the intensity function of the One-dimensional Hawkes Process is

$$\lambda(t) = \lambda + \int_0^t \alpha e^{-\beta(t-s)}dN(s) \tag{2}$$

2.1.2. Calibration

In this paper, Maximum Likelihood Estimation (MLE) is used for calibration of the One-dimensional Hawkes Process. Suppose that the Hawkes process is observed over some time period $[0, T] \supset [0, t_k]$. Then, the log-likelihood function of Hawkes process is (see References Laub et al. 2015; Lorenzen 2012),

$$l = \sum_{i=1}^{k} \log[\lambda + \alpha \sum_{j=1}^{i-1} e^{-\beta(t_i - t_j)}] - \lambda t_k + \frac{\alpha}{\beta} \sum_{i=1}^{k} [e^{-\beta(t_k - t_i)} - 1] \tag{3}$$

If we let $A(i) = \sum_{j=1}^{i-1} e^{-\beta(t_i - t_j)}$, for $i = \{2, 3, \ldots, k\}$, and $A(1) = 0$, then Equation (3) can be transformed into

$$l = \sum_{i=1}^{k} \log[\lambda + \alpha A(i)] - \lambda t_k + \frac{\alpha}{\beta} \sum_{i=1}^{k} [e^{-\beta(t_k - t_i)} - 1] \tag{4}$$

However, in this case, the log-likelihood function is difficult to be solved manually and thus we need to use some computational methods (e.g., Partical Swarm Optimization (PSO)) to find the global optimization of the parameters λ, α, β.

2.2. General Compound Hawkes Process

2.2.1. Definition

Compound Poisson Process is a widely used model in risk management and finance, with some independently arriving jumps that follow the Poisson Process,

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k). \tag{5}$$

In this model, $N(t)$ is a Poisson Process, X_k is a continuous time n-state Markov Chain; $a(X_k)$ is a continuous and bounded mapping function on the state space $X := \{1, 2, \ldots, n\}$. If we change the model a little bit, we can get our new General Compound Hawkes Process

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k) \tag{6}$$

where $N(t)$ is a one-dimensional Hawkes Process with the intensity function mentioned in Equation (2). In our case, this model is used for mid-price modelling and thus $N(t)$ represents the number of mid-price changes up to time t; $a(X_k)$ represents the price movements interval. As we can see, models (5) and (6) look the same at first sight but they are different with respect to $N(t)$: in (5) it is the Poisson process and in (6) it is the Hawkes process.

From general to typical cases, if X_k is an n-state Markov Chain, this model is called the General Compound Hawkes Process with n-state Dependent Orders (GCHPnSDO); if X_k is a 2-state Markov Chain, this model is called the General Compound Hawkes Process with 2-state Markov Chain (GCHP2SDO); if in the 2-state Markov Chain, the value of $a(X_k)$ is set to be the most typical case, that is, the value of $a(X_k)$ is fixed, $a(X_k) = \delta$ or $a(X_k) = -\delta$, then this model is called the General Compound Hawkes Process with Dependent Orders (GCHPDO).

2.2.2. Diffusive Limit

Let X_k be an ergodic n-state Markov Chain with state space $X := \{1, \ldots, n\}$, that is, $X_k = \{1, 2, \ldots, n\}$ and with ergodic probabilistic $(\pi_1^*, \pi_2^*, \ldots, \pi_n^*)$; S_t is defined in GCHPnSDO model. The diffusive limit of GCHPnSDO is derived in Reference Swishchuk and Huffman (2018),

$$\frac{S_{nt} - N(nt)\hat{a}^*}{\sqrt{n}} \overset{n\to\infty}{\longrightarrow} \hat{\sigma}^* \sqrt{\lambda/(1-\hat{\mu})} W(t) \tag{7}$$

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

$$0 < \hat{\mu} := \int_0^\infty \mu(s)ds = \frac{\alpha}{\beta} < 1 \tag{8}$$

$$\int_0^\infty s\mu(s)ds < \infty \tag{9}$$

$$\hat{a}^* := \sum_{i\in X} \pi_i^* a(i) \tag{10}$$

$$\begin{gathered} \hat{\sigma}^* := \sum_{i\in X} \pi_i^* v(i) \\ v(i) = b(i)^2 + \sum_{j\in X} (g(j) - g(i))^2 P(i,j) - 2b(i) \sum_{j\in X} (g(j) - g(i)) P(i,j) \\ b(i) := a(i) - a^* \\ b = (b(1), b(2), \ldots, b(n))' \\ g := (P + \Pi - I)^{-1} b \end{gathered} \tag{11}$$

In Equation (11), P denotes the transition probability matrix of the Markov Chain, that is, $P(i,j) = P(X_{k+1} = j \mid X_k = i)$; Π^* denotes the matrix of stationary distribution of the Markov Chain. When the power n of the transition probability matrix P is large enough, $P^n \overset{n\to\infty}{\longrightarrow} \Pi^*$, and all rows of Π^* will be the same.

The diffusive limit of GCHP2SDO and GCHPDO model can still use Equation (7) since they are the typical cases of GCHPnSDO model. In Section 3, we will further introduce how to use the diffusive limit to estimate the correctness of the mid-price modelling by the General Compound Hawkes Process.

3. Application

3.1. Limit Order Book

A Limit Order Book (LOB) is just like a book that records financial big data in the market. Here, we consider two types of tradings, one is called market order (buy/sell) and the other one is called limit order (buy/sell). In the trading system, market order will be executed immediately when it arrives; limit order will wait for later execution and while waiting it can be cancelled. Therefore, limit order book is a collection of queued active limit orders awaiting execution or cancellation (see Gould et al. 2013).

Then we will give some definitions in Limit Order Book. Lot size means the smallest amount of the asset that can be traded within LOB. Tick size means the smallest permissible price interval between different orders within LOB (e.g., Suppose that the tick size in the trading system is 1 cent. Then, all the orders should be submitted in the exactly two decimal places). Bid price means the highest stated price among active buy orders at time t. Ask price means the lowest stated price among active sell orders at time t. Mid price is the price between the bid and ask price, that is,

$$Mid\ Price = \frac{Bid\ Price + Ask\ Price}{2} \tag{12}$$

3.2. Data

One remark should be made with respect to the choice of real data. In different papers we used different types of real data—LOBster (five stocks Amazon, Apple, Google, Intel, Microsoft on 21 June 2012), 15 stocks from Deutsche Boerse Group (23 September 2013) and CISCO asset (3 November 2014) in References Chavez-Casillas et al. (2019); Swishchuk and Vadori (2017); Swishchuk et al. (2017), respectively, to justify and illustrate our approach.

In the present paper, we decide to use another set of real data, namely EBAY, FB, MU, PCAR, SMH, CSCO, (provided by Reference Cartea et al. (2015), book), to check our methods and to justify and illustrate our approach, and as in our previous papers, the outcomes of the present paper show that the method is right and gives very good results.

Thus, in this paper, the real LOB data of EBAY, FB, MU, PCAR, SMH is downloaded from http://sebastian.statistics.utoronto.ca/books/algo-and-hf-trading/data/. What we get from this link is a zip file with matlab data from November 2014 for the following tickers—AMZN, EBAY, FB, GOOG, INTC, MSFT, MU, PCAR, SMH, VOD for every second of the trading day. We reject the stocks that we tested in the previous papers. Each file contains a structure called LOB where each field has entries corresponding to one second of the trading day—NumberMO, volumeMO, EventTime, BuyPrice, SellPrice, BuyVolume, SellVolume, MO. Moreover, the real LOB data of CSCO is downloaded from https://sites.google.com/site/algorithmictradingbook/website-builder. What we get from this link is a zip file with matlab data from 3 November 2014 to 7 November 2014. Each file contains a structure called data with five elements—Event, SellVolume, SellPrice, BuyVolumne, BuyPrice.

We randomly choose 1 day in November to see how our models work on these stocks. For the mid price modelling in Section 4.2, the data files we use in this paper are EBAY_20141110, FB_20141110, MU_20141110, PCAR_20141110, SMH_20141110, CSCO_20141107. Consider for each trading day, it starts from 9:30 and ends at 16:00. In order to avoid the open and close auctions, we ignore the first and last 15 min, and thus we end with 6 trading hours per day, between 9:45 and 15:45. The number of mid price changes in a day for EBAY, FB, MU, PCAR, SMH, CSCO are 1255, 3988, 1756, 1008, 379, 943.

3.3. Descriptive Data Analysis

Before we will work with our new General Compound Hawkes Process Model, we need to check the following questions first. Comparing with the Compound Poisson Process, we want to know whether we could reject Poisson process; furthermore, we want to know whether there is a correlation between the mid-price changes. After that, we plot the number of arrivals under a fixed time window to explore these features in the real data.

3.3.1. QQ-Plot

In Figure 1, the QQ-plot rejects Poisson Process since the inter-arrival time does not follow the exponential distribution. In Figure 1, the two black straight lines are the confidence boundary under 95% confidence level; the red straight line is $y = x$. If the inter-arrival time of the mid price changes follow the exponential distribution, all the black points should approximately be on the red line, within the confidence boundary. However, from the Figure 1, it is obvious to reject this assumption.

Furthermore, if the exponential distribution does not fit the inter-arrival times well, we want to know which distribution will fit better. We tried normal distribution, Gamma distribution, Weibull distribution, and then compared the theoretical Cumulative Density Function (CDF) with the empirical CDF (See Figure 2). From Figure 2, it is obvious that in most cases, Gamma or Weibull distribution performs much more better than the others since the curve of the theoretical CDF nearly coincides with that of empirical CDF.

Of course, quantitative measures can be used as well to compare the different models, such as MLE of the Weibull and Gamma distributions' parameters as we did in Reference Swishchuk and Vadori (2017), for LOBster data or in Reference Chavez-Casillas et al. (2019), for Facebook data. However, we

gave here two different visualized comparisons, namely *QQ*-plots and CDFs, for 6 different stocks to illustrate our approach which proved to be right.

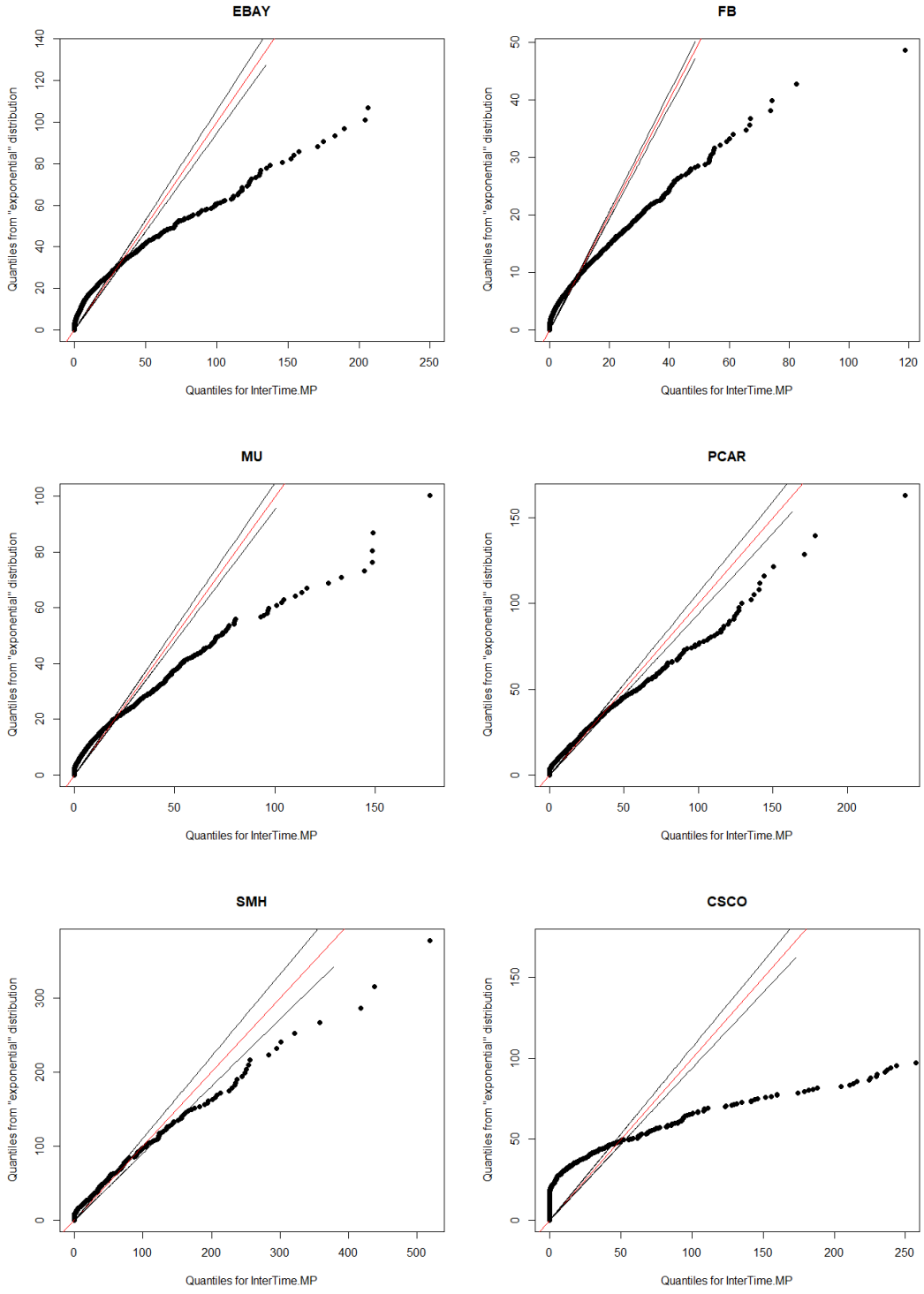

Figure 1. QQ-plot for 6 different stocks.

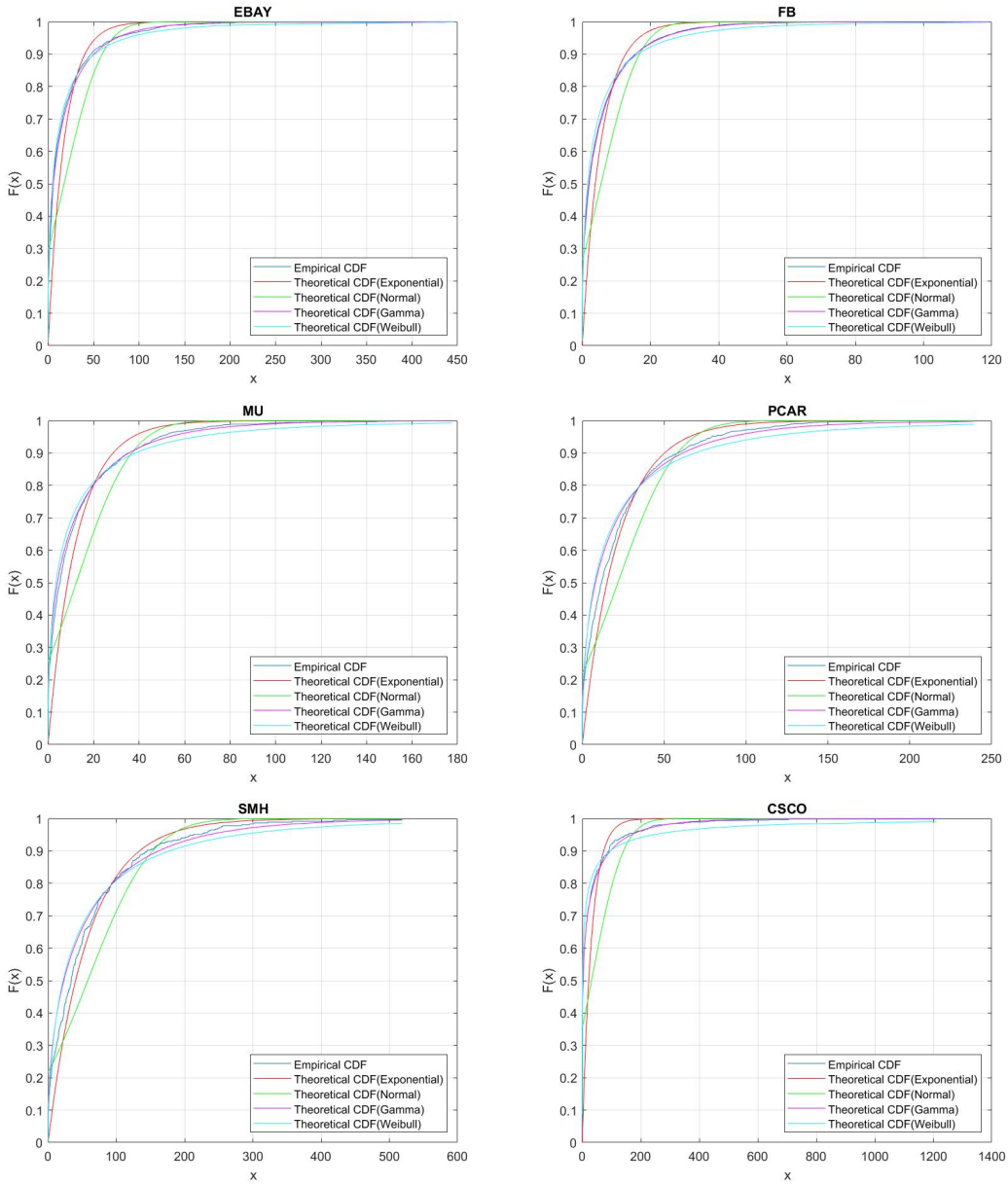

Figure 2. CDF comparison between different distributions.

3.3.2. Autocorrelation

To further confirm that the mid price changes are not independent, we calculated the autocorrelation by using equation

$$C(\tau,\delta) = \frac{E[(N_{t+\tau} - N_t)(N_{t+2\tau+\delta} - N_{t+\tau+\delta})] - E[(N_{t+\tau} - N_t)]E[(N_{t+2\tau+\delta} - N_{t+\tau+\delta})]}{\sqrt{var(N_{t+\tau} - N_t)var(N_{t+2\tau+\delta} - N_{t+\tau+\delta})}} \quad (13)$$

where τ is the length of the time interval, δ is the time lag, that is, the length between two consecutive time intervals. In Equation (13), if τ is fixed, then the autocorrelation function $C(\tau,\delta)$ will become a function that is respective to δ, that is, $C(\delta)$. Take FB20141110 data as example. In the four plots shown in Figure 3, the length of the time interval is changed, (a) τ = 20 s, (b) τ = 30 s, (c) τ = 60 s, (d) τ = 90 s. In each graph, the X-coordinate represents the time lag δ which is changed from 1 s to 30 min by a step of 1 s.

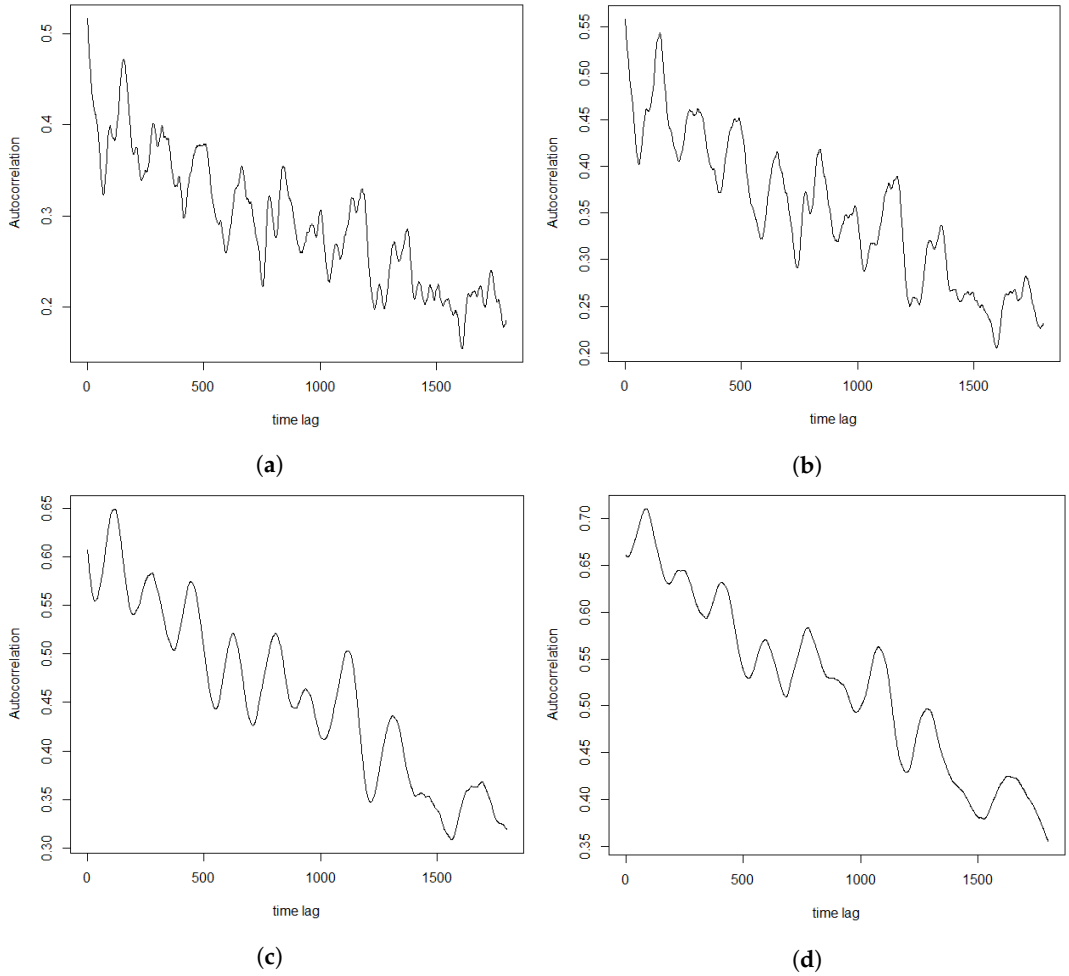

Figure 3. Autocorrelation under different τ. (**a**) $\tau = 20$ s. (**b**) $\tau = 30$ s. (**c**) $\tau = 60$ s. (**d**) $\tau = 90$ s.

From Figure 3, it is obvious to see that the shape of the autocorrelation function are nearly identical even if the plot is noisier when τ decreases. Moreover, something more interesting could also be discovered from Figure 3. Even if δ = 30 min, the value of the correlation function in each plot is still above 0.2 which is not neglectable. This means that suppose there is an arrival of mid price change at current time t and then the impact of this arrival will last more than 30 min in the trading system, which is much longer than we can imagine.

3.3.3. Clustering Feature

At the end of the Descriptive Data Analysis, we plot the number of mid price changes occurring in every minute during the trading day. The clustering feature is obvious to be discovered in Figure 4.

From Sections 3.3.1–3.3.3, the main conclusion that we can make is trading activities are not a completely random and memoryless process.

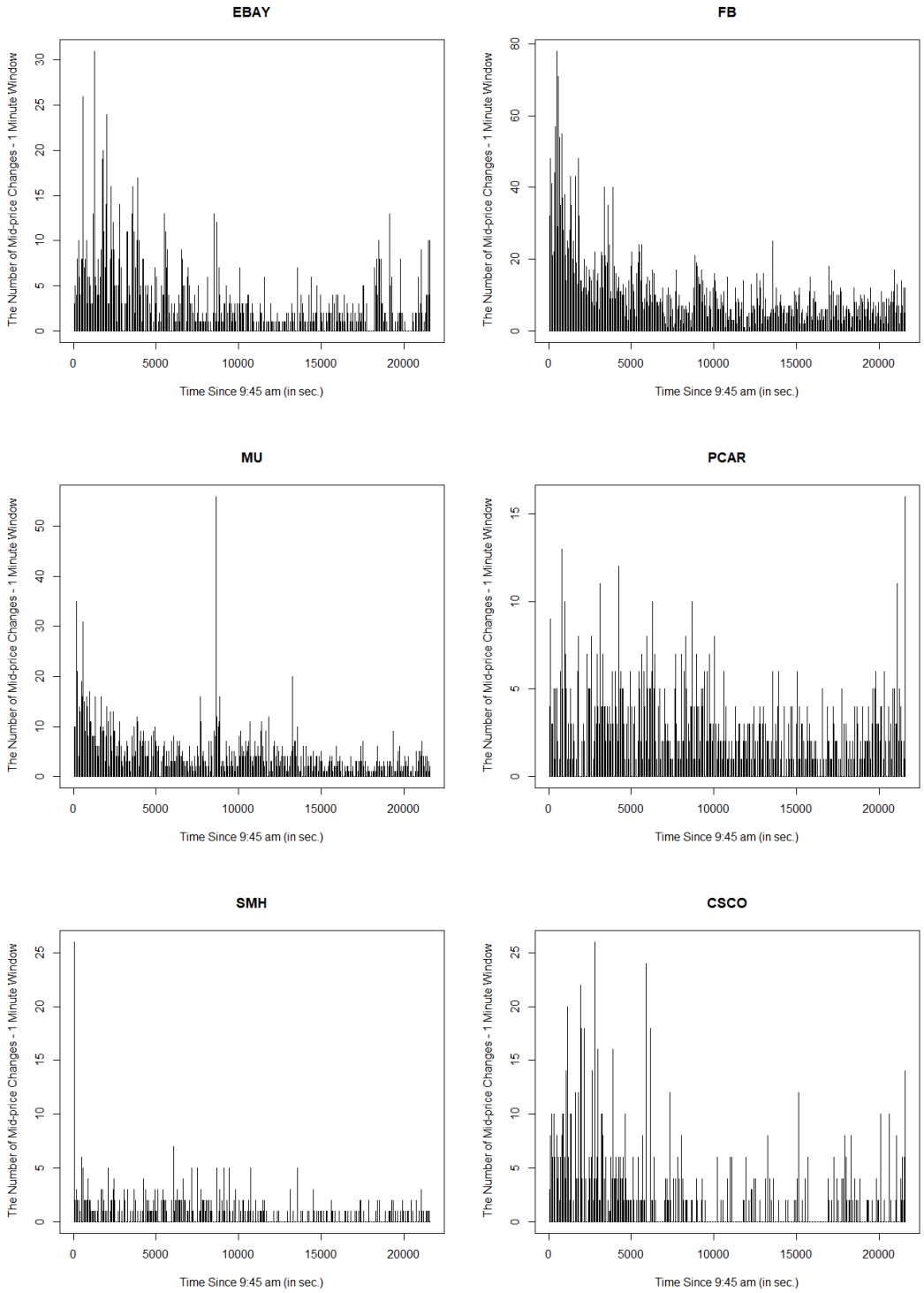

Figure 4. Clustering features for 6 different stocks.

4. Hawkes Process and Models Calibrations

In this section we calibrate Hawkes process' parameters λ, α, β (Section 4.1) and our mid-price S_t model's parameters $a(1), \ldots, a(n)$ (Section 4.2) for different GCHP (see (Sections 4.2.1–4.2.3) to find the price's volatility $\sigma^* \sqrt{\lambda/(1-\alpha/\beta)}$, obtained in (7). To estimate the models fitting accuracy, we define in Section 5 an error rate for each model, and set the threshold value as 15%. The model with error rate less than the threshold is considered as well fitted. Thus, we define which of the models is the best fit for which real data.

4.1. Hawkes Process' Parameters Calibration

The Particle Swarm Optimization (PSO) method is used to solve the log-likelihood function in Equation (4) and find the global optimization of the parameters $\hat{\lambda}$, $\hat{\alpha}$, $\hat{\beta}$. If the calibration result is appropriate, $E[N([0,1])] \to \frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}$ (see Laub et al. 2015), where $E[N([0,1])]$ is the empirical expectation of the number of mid price changes arriving in a unit time interval.

From Table 1, it is obvious to see that the values in the last two columns are super close, which means that our parameter estimation is appropriate.

Table 1. Calibration Result of Hawkes Process.

Stock	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$E[N([0,1])]$	$\frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}$
EBAY	0.05142964	24.16091427	208.361411	0.058169396	0.05817548
FB	0.1460757	7.0368564	33.6354758	0.184708815	0.184721079
MU	0.06685776	8.49267842	47.38655865	0.081446361	0.081456495
PCAR	0.04105567	28.42153021	234.1878409	0.046720101	0.046726496
SMH	0.0157736	10.6972559	102.6908889	0.017602398	0.017607795
CSCO	0.0223269	642.6391460	1311.8684573	0.043657407	0.043766646

4.2. Mid Price Modelling and Calibration

After Section 4.1, we can successfully model $N(t)$, the cumulative number of mid price changes up to current time t. Then, in this section, we will move forward to the mid price modelling by General Compound Hawkes Process. We will start from the most typical GCHPDO model and then move forward to the most general GCHPnSDO model. Under each model, the most important thing to figure out is the value of $a(X_k)$.

4.2.1. GCHPDO

In the case of GCHPDO model, the value of $a(X_k)$ can only take the fixed number. Suppose that $X_k \in \{-\delta, \delta\}$ and $a(x) = x$, that is, $a(X_k) = -\delta$ or $a(X_k) = \delta$, where δ is the tick size of mid price in the trading system. In our case, suppose the tick size in the trading system is 1 cent. Whenever the ask/bid price goes up/down by 1 cent, the mid price will go up/down by 0.5 cent. Therefore, the tick size of the mid price movement is 0.5 cent.

Rearrange the Equation (7),

$$S_{nt} - N(nt)\hat{a}^* \xrightarrow{n\to\infty} \hat{\sigma}^* \sqrt{n} \sqrt{\frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}} W(t) \tag{14}$$

From 9:45 a.m. to 15:45 p.m., cut this 6 trading hours into some disjoint windows of size n, with $t = 1$, that is, the disjoint time intervals are $[in, (i+1)n]$. Then, the left hand side of Equation (14) can be discretized into

$$S_i^* = (S_{(i+1)n} - S_{in}) - (N((i+1)n) - N(in))\hat{a}^* \tag{15}$$

Combining Equations (14) and (15), a converging formula between the standard deviation and its theoretical counterpart can be received,

$$std\{S_i^*\} \xrightarrow{n\to\infty} \hat{\sigma}^* \sqrt{nt} \sqrt{\frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}},\ t = 1 \tag{16}$$

By (16), we can use this to estimate the correctness of the model fitting. If our model fitting is perfect, then

$$std\{S_i^*\} \approx \hat{\sigma}^* \sqrt{n} \sqrt{\frac{\hat{\lambda}}{1 - \frac{\hat{\alpha}}{\hat{\beta}}}} \tag{17}$$

Figure 5 shows the empirical standard deviation against its theoretical counterpart for various window sizes, starting from 10 s to 20 min by the step of 10 s. Except for the CSCO result, it is obvious to see that the differences between the empirical standard deviation and its theoretical counterpart are kind of large in the rest of five stocks. The reason for these unsatisfying results is that GCHPDO model assumes at each change, the mid price can only go up/down by tick size = 0.05 cent. However, in reality, only a small percentage of mid price movements change in this way (see Table 2).

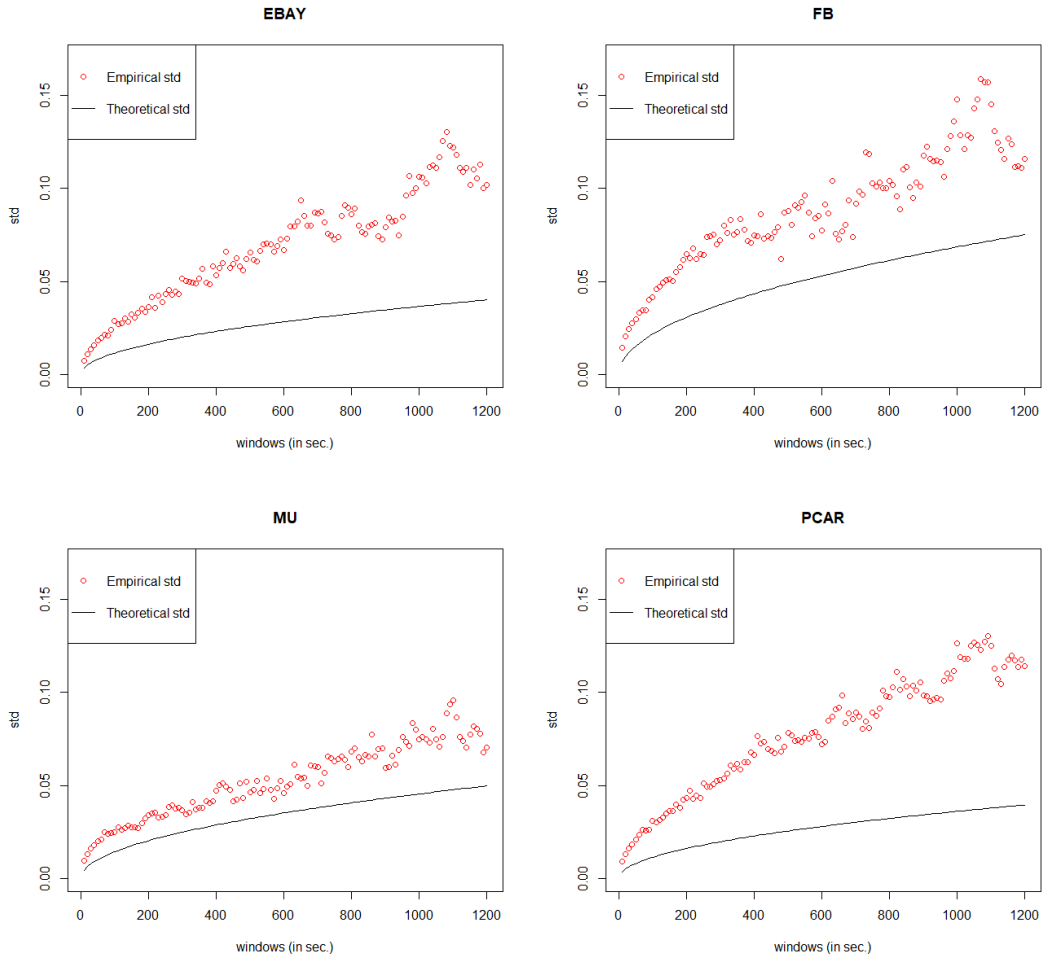

Figure 5. *Cont.*

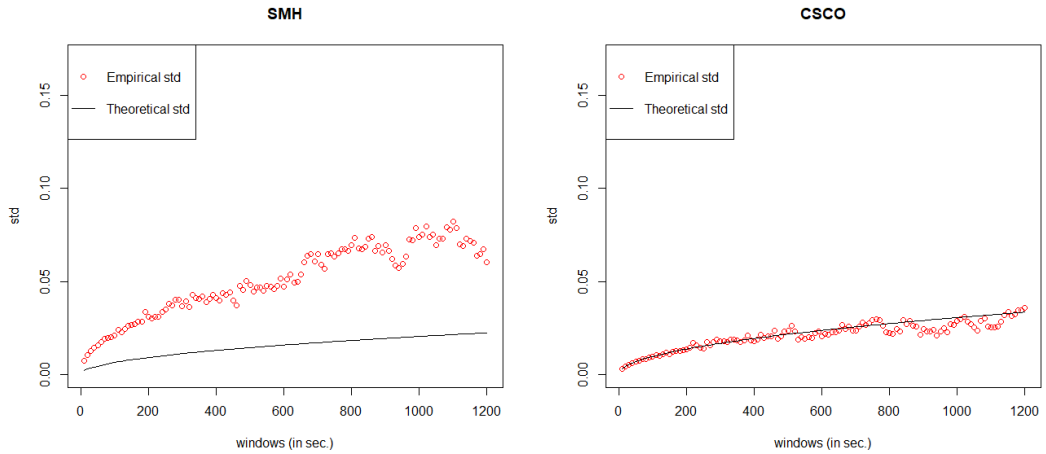

Figure 5. General Compound Hawkes Process with 2 fixed Dependent Orders model fitting results.

Table 2. Tick-size rate.

Stock	Tick-Size Rate
EBAY	21.35%
FB	18.18%
MU	9.85%
PCAR	22.22%
SMH	12.96%
CSCO	100%

4.2.2. GCHP2SDO

According to Section 4.2.1, GCHPDO is too special to model the mid price movements of stocks EBAY, FB, MU, PCAR, and SMH, and thus we need to consider some more general models. In GCHP2SDO model, the way we determine the value of $a(X_k)$ depends on the real data rather than just on the trading system. Take the mean of the upward and downward mid price movements and then assign them to $a(1)$ and $a(2)$. The result in Table 3 shows that in GCHP2SDO model, the mid price does not go up/down by exactly 0.5 cent at each change, which leads the model fitting to become much better than the GCHPDO model (see Figure 6).

Table 3. $a(X_k)$ for GCHP2SDO.

Stock	$a(1)$	$a(2)$
EBAY	0.0091730474	−0.0092358804
FB	0.0095449949	−0.0093280239
MU	0.0098359729	−0.0096846330
PCAR	0.0116390041	−0.0108269962
SMH	0.0152972973	−0.0134196891

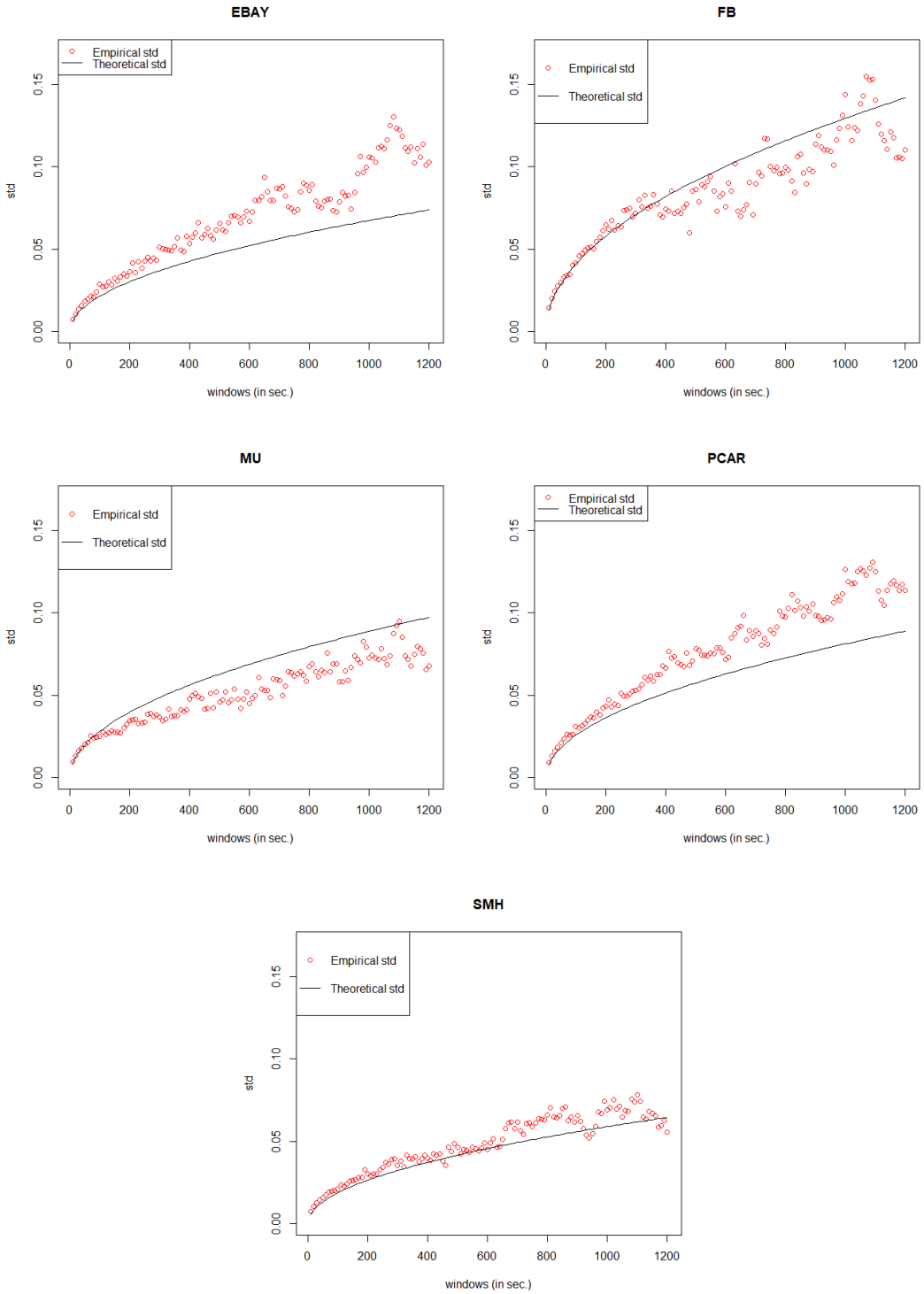

Figure 6. General Compound Hawkes Process with 2 non-fixed Dependent Oorder model fitting results.

4.2.3. GCHPnSDO

For upward or downward mid price movement, the above GCHPDO and GCHP2SDO models only consider 1 possible price interval since X_k is only a 2-state Markov Chain. In this section, we are trying to figure out whether there is more than 1 possible price interval that could better fit the mid price movements than GCHP2SDO model.

Firstly, we need to figure out a way to determine the value of $a(X_k)$. Step1: After calculating the mid price changes, we separate the data into upward and downward movements. Step 2: we calculate the (evenly/unevenly) distributed quantiles for both data sets. Depending on the data, several quantiles may be identical and then we reject any duplicates. At the end of this step, we have gotten a list of bounds B_k, which we will use for determine the value of $a(X_k)$. Step 3: $a(X_k)$ = the average of all mid-price changes located between two neighbouring boundary values, that is, $[B_k, B_{k+1})$. Step 4: There is an exception for the largest upper boundary, that is, after step 3, we will not get the value for $a(X_n)$. And thus, we need to calculate the mean of the rest data manually and assign it to $a(X_n)$.

Secondly, we need to determine the best n for GCHPnSDO model. For different stocks, the Mean Square Errors (MSE) between the empirical standard deviation and its theoretical counterpart under varied number of states are shown in Tables 4–8. We tried limited n, and then pick up the best n under the smallest MSE. For EBAY, the best GCHPnSDO model is the GCHP8SDO model; for FB, the best GCHPnSDO model is GCHP4SDO model; for MU, the best GCHPnSDO model is the GCHP4SDO model; for PCAR, the best GCHPnSDO model is the GCHP15SDO model; for SMH, the best GCHPnSDO model is the GCHP9SDO model.

Table 4. n Selection for EBAY.

n	**4**	**5**	**6**	**7**	**8**
MSE	0.00046701	0.000418703	0.000399311	0.000397898	0.000393683

Table 5. n Selection for FB.

n	**4**	**5**	**6**	**7**	**8**
MSE	0.000317633	0.000330058	0.000344656	0.000349087	0.000361214

Table 6. n selection for MU.

n	**4**	**6**	**8**
MSE	0.000263588	0.000272237	0.000275675

Table 7. n selection for PCAR.

n	**6**	**9**	**12**	**15**
MSE	0.000238185	0.000132224	0.000129238	0.000124427

Table 8. n selection for SMH.

n	**6**	**7**	**9**	**10**
MSE	0.00003853463	0.00003436168	0.00002395992	0.00002475711

In the end, compare the MSE of the best GCHPnSDO models with that of the previous GCHPDO and GCHP2SDO models. Again, pick up the best model for mid price movements under the smallest MSE (see Table 9). Under the best model, the model fitting results can be seen in Figure 7.

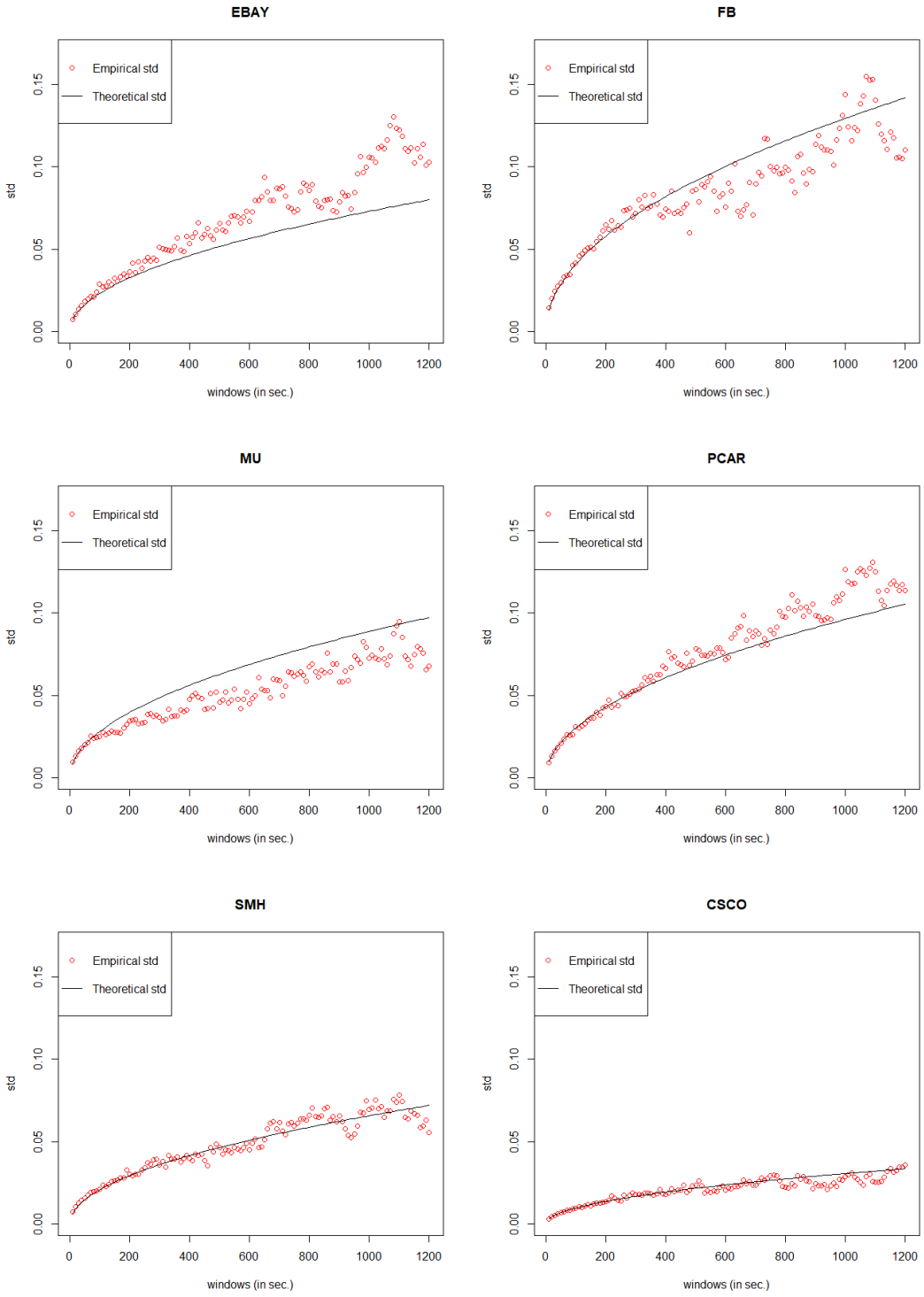

Figure 7. Best model fitting results for 6 stocks.

Table 9. Model comparison.

Stock	GCHPDO (MSE)	GCHP2SDO (MSE)	GCHPnSDO (MSE)	Best Model
EBAY	0.002200996	0.000566305	0.000393683	GCHP8SDO
FB	0.00173283	0.000241354	0.000317633	GCHP2SDO
MU	0.000450242	0.000227604	0.000263588	GCHP2SDO
PCAR	0.003129217	0.000468104	0.000124427	GCHP15SDO
SMH	0.001477308	0.000053474	0.00002395992	GCHP9SDO

5. Error Measurement

Since the value of MSE is always very small under GCHPDO, GCHP2SDO, and GCHPnSDO models, it makes no sense to measure the correctness of model fitting under the best model only by MSE. Therefore, in this paper, we proposed a new way to measure the correctness. If we take the square of both sides in Equation (16), we can get

$$(std\{S_i^*\})^2 \overset{n\to\infty}{\longrightarrow} ((\hat{\sigma}^*)^2 \frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}})nt,\ t=1 \tag{18}$$

Equation (18) means that when n goes to infinity, the square of $std\{S_i^*\}$ converges to a linear regression function $\mathscr{L}_1 = ((\hat{\sigma}^*)^2 \frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}})n$ with respective to n. Therefore, based on the real points $(std\{S_i^*\})^2$ under different n, we can estimate the linear regression function $\mathscr{L} = cn$ by least square method, where c is the estimated slope. Then, we take the square root of estimated function $\mathscr{L}$, we can get our estimated curve of $std\{S_i^*\}$. The comparison between the empirical curve and the theoretical curve is shown in Figure 8.

Under the best model, the error rate of each stock is listed in Table 10. If we set the threshold value as 15%, the model with error rate less than threshold can be considered as well fitted. Table 10 shows that our General Compound Hawkes Process correctly models the mid price movements of 4 stocks among 6.

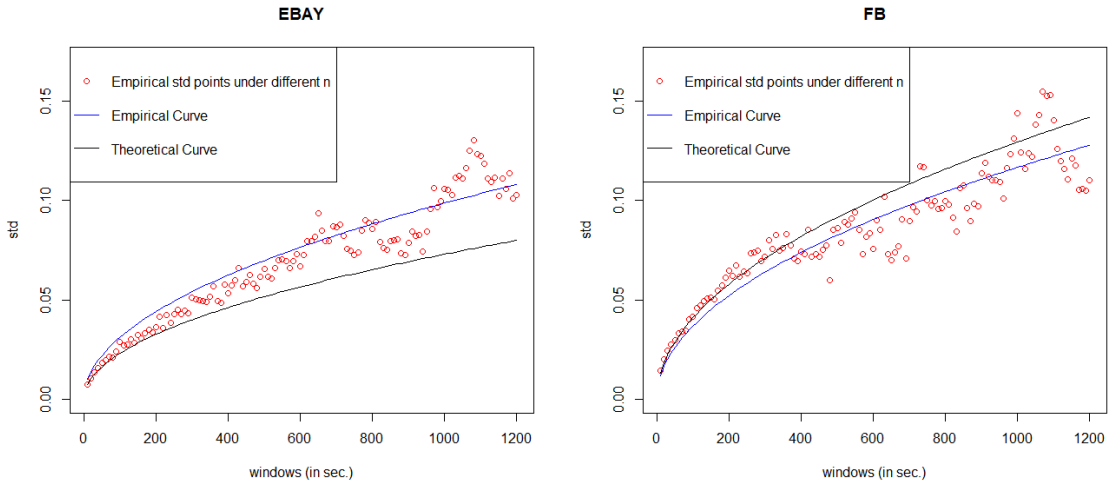

Figure 8. *Cont.*

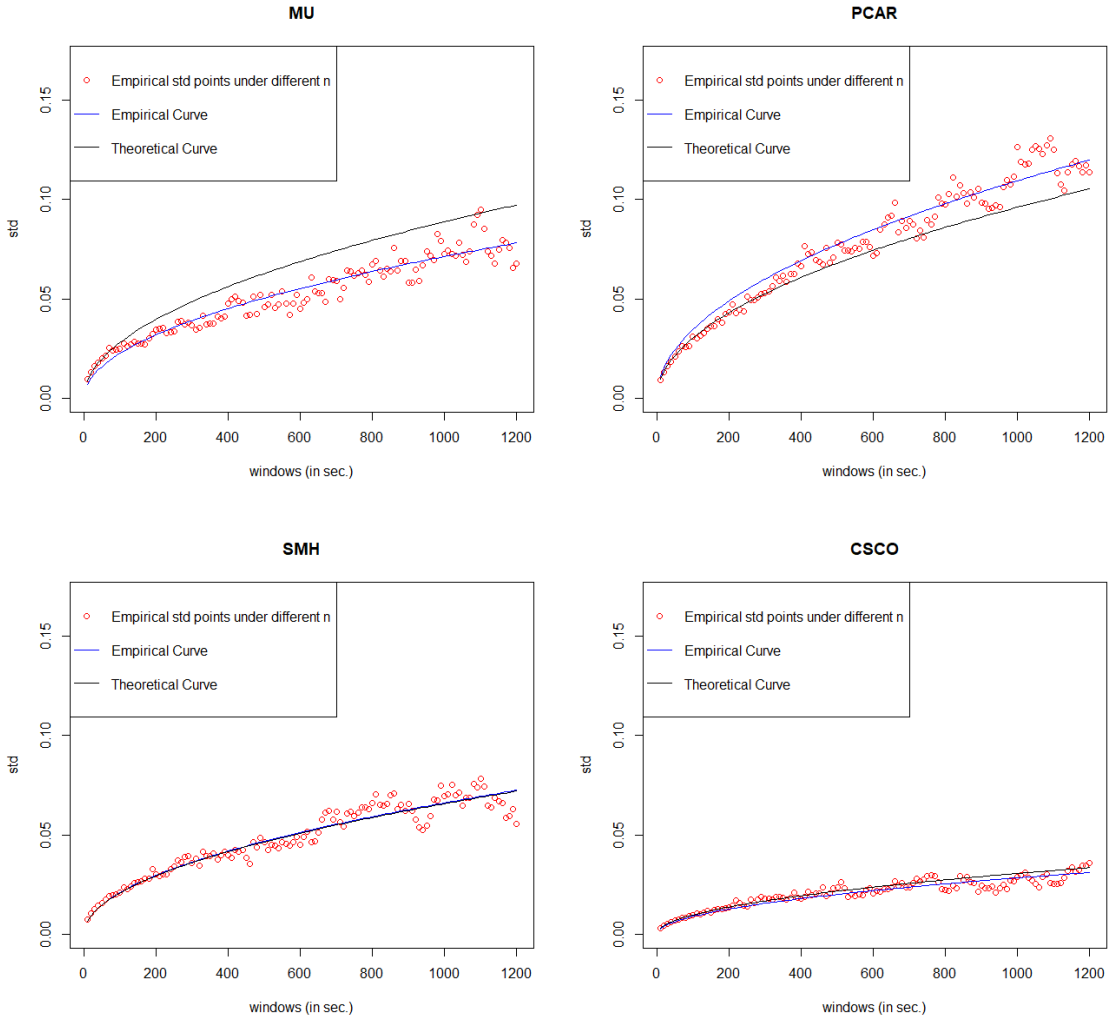

Figure 8. Error measurements for 6 stocks.

Table 10. Error rate under the best model.

Stock	Error Rate
EBAY	26.05%
FB	10.95%
MU	24.42%
PCAR	12.18%
SMH	0.67%
CSCO	7.99%

We define the error rate as

$$\left| \frac{\sqrt{c} - \hat{\sigma}^* \sqrt{\frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}}}{\sqrt{c}} \right| \times 100\% \tag{19}$$

In order to avoid any bias of data selection, we do the simulation again by following the same process mentioned in Section 3 to Section 5 to justify our General Compound Hawkes Processes work well not just for one day but also for other days. In the data set of EBAY, FB, MU, PCAR, SMH, we randomly choose the date 18 November 2014; for the data set of CSCO, we randomly choose the date 6 November 2014.

For descriptive data analysis, we have gotten the quite similar results as those shown in Figures 1–4. Hawkes Process Calibration results are shown in Table 11.

Table 11. Calibration Result of Hawkes Process.

Stock	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$E[N([0,1])]$	$\frac{\hat{\lambda}}{1-\frac{\hat{\alpha}}{\hat{\beta}}}$
EBAY	0.037386079	11.06718675	77.58564954	0.04356481	0.043606288
FB	0.16361531	9.49962096	42.48237019	0.210416667	0.210739442
MU	0.059890034	9.767710137	65.90938964	0.070277777	0.070309895
PCAR	0.03032757	10.88944631	82.44806627	0.034861111	0.034942673
SMH	0.01580581	41.96535566	351.0221707	0.01787037	0.017952006
CSCO	0.022357326	677.5504008	1385.215404	0.04337963	0.04376324

Similarly, for CSCO data, since the intervals of price movements are always \$0.5, the GCHPDO model should be the best fit. To find the best n of the GCHPnSDO model for EBAY, FB, MU, PCAR, SMH, we have tried limited n and then picked up the best n under the smallest MSE. From Tables 12–16, we can see that for EBAY, the best GCHPnSDO model is GCHP5SDO; for FB, the best GCHPnSDO model is GCHP4SDO; for MU, the best GCHPnSDO model is GCHP4SDO; for PCAR, the best GCHPnSDO model is GCHP4SDO; for SMH, the best GCHPnSDO model is GCHP6SDO.

Table 12. *n* Selection for EBAY.

n	**4**	**5**	**6**	**7**
MSE	0.000154926	0.000151477	0.000182662	0.000193251

Table 13. *n* Selection for FB.

n	**4**	**5**	**6**	**7**
MSE	0.000204729	0.000211706	0.000241339	0.000249355

Table 14. *n* Selection for MU.

n	**4**	**6**	**8**
MSE	0.000031205	0.000032304	0.000038158

Table 15. *n* Selection for PCAR.

n	**4**	**6**	**8**	**10**
MSE	0.000143116	0.000154922	0.000162016	0.00032631

Table 16. *n* Selection for SMH.

n	**4**	**6**	**7**	**9**	**10**
MSE	0.000063641	0.000051274	0.000077954	0.000102804	0.00010632

Compare the MSE of the best GCHPnSDO models with that of GCHPDO and GCHP2SDO models. Again, pick up the best model for mid price movements under the smallest MSE and then calculate the error rate following the same method derived in Equation (19). Within the error rate threshold of 15%, in Table 17, we could see that our General Compound Hawkes Processes successfully track the mid price movements of FB, MU, PCAR, SMH, and CSCO.

Table 17. Model comparison.

Stock	GCHPDO (MSE)	GCHP2SDO (MSE)	GCHPnSDO (MSE)	Best Model	Error Rate
EBAY	0.000395775	0.000070119	0.000151477	GCHP2SDO	15.67%
FB	0.002979562	0.000164628	0.000204729	GCHP2SDO	4.15%
MU	0.000691422	0.000027226	0.000031205	GCHP2SDO	1.52%
PCAR	0.001220103	0.000075708	0.000143116	GCHP2SDO	6.75%
SMH	0.001058399	0.000076867	0.000051274	GCHP6SDO	4.21%
CSCO	0.000020788	N/A	N/A	GCHPDO	13.29%

6. Conclusions and Future Work

The main contribution and novelty of the paper consists in introducing different new types of General Compound Hawkes Processes (GCHPDO, GCHP2SDO, GCHPnSDO) and their diffusive limits to model the mid price movements. Based on our previous research, we further expand our data sets to justify whether our models still work well on the stocks of EBAY, FB, MU, PCAR, SMH, CSCO (provided by Reference Cartea et al. (2015) book). We define the error rates to estimate the models fitting accuracy and set the threshold to 15%. Based on the data sets named EBAY_20141110, FB_20141110, MU_20141110, PCAR_20141110, SMH_20141110, CSCO_20141107, the best model for EBAY is GCHP8SDO; for FB, the best model is GCHP2SDO; for MU, the best model is GCHP2SDO; for PCAR, the best model is GCHP15SDO; for SMH, the best model is GCHP9SDO; for CSCO, the best model is GCHPDO. Under the best model, our General Hawkes Processes successfully track the mid price movements of 4 stocks out of 6. To further avoid any bias of data selection process, again, we randomly choose the data from another date in our data sets, and gain the quite similar results. Those results justify that our General Compound Hawkes Processes could be right and could be applied for the data not just from a specific day but also from other days.

The future work will be devoted to the prediction analysis with different models introduced in this paper, and justifications of those model using real data.

Author Contributions: Q.H.: software, validation, data curation, visualization, writing—original draft preparation. A.S.: project administration, supervision, writing—review and editing, conceptualization, methodology.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Bacry, Emmanuel, Sylvain Delattre, Marc Hoffmann, and Jean-Francois Muzy. 2013. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications* 123: 2475–99.

Bacry, Emmanuel, Iacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes processes in finance. *arXiv* arXiv:1502.04592v2.

Cartea, Álvaro, Sebastian Jaimungal, and José Penalva. 2015. *Algorithmic and High-Frequency Trading*. Cambridge: Cambridge University Press.

Chavez-Casillas, Jonathan, Robert J. Elliott, Bruno Remillard, and Anatoliy V. Swishchuk. 2019. A level-1 limit order book with time dependent arrival rates. *Methodology and Computing in Applied Probability* 21: 699–719. [CrossRef]

Cont, Rama, and Adrien De Larrard. 2013. Price dynamics in a Markovian limit order markets. *SIAM Journal on Financial Mathematics* 4: 1–25.

Da Fonseca, José, and Riadh Zaatour. 2013. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets* 34: 548–79.

Embrechts, Paul, Thomas Liniger, and Lu Lin. 2011. Multivariate hawkes processes: An application to financial data. *Journal of Applied Probability* 48: 367–78.

Gould, Martin D., Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. 2013. Limit order books. *Quantitative Finance* 13: 1709–42.

Hawkes, Alan G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58: 83–90.

Laub, Patrick J., Thomas Taimre, and Philip K. Pollett. 2015. Hawkes Processes. *arXiv* arXiv:1507.02822v1.

Lorenzen, Florian. 2012. Analysis of Order Clustering Using High Frequency Data: A Point Process Approach. Ph.D. thesis, Swiss Federal Institute of Technology Zurich (ETH Zurich), Zurich, Switzerland.

Swishchuk, Anatoliy, and Aiden Huffman. 2018. General Compound Hawkes Process in Limit Order Books. *arXiv* arXiv:1812.02298v1.

Swishchuk, Anatoliy, and Nelson Vadori. 2017. A semi-Markovian modeling of limit order markets. *SIAM Journal on Financial Mathematics* 8: 240–73.

Swishchuk, Anatoliy, Tyler Hofmeister, Katharina Cera, and Julia Schmidt. 2017. General semi-Markov model for limit order books. *International Journal of Theoretical and Applied Finance* 20: 1750019.

Swishchuk, Anatoliy, Bruno Remillard, Robert Elliott, and Jonathan Chavez-Casillas. 2019. Compound Hawkes processes in limit order books. In *Financial Mathematics, Volatility and Covariance Modelling*, 1st ed. Edited by Julien Chevallier, Goutte Stéphane, Guerreiro David, Saglio Sophie and Bilel Sanhaji. Abingdon: Routledge, vol. 2.

Zhou, Ke, Hongyuan Zha, and Le Song. 2013. Learning triggering kernels for multi-dimensional hawkes processes. Paper presented at the 30th International Conference on Machine Learning, PMLR, Atlanta, GA, USA, June 16–21; vol. 29, pp. 1301–9.