MODELLING STOCK PRICES AS A DISCRETESELF-EXCITING PROCESS: HAWKES VS VAR

Gorka González de Mendivil Grau

Trabajo de investigación 20/005

Master en Banca y Finanzas Cuantitativas

Directores: Dr. José S. Penalva

Universidad Complutense de Madrid

Universidad del País Vasco

Universidad de Valencia

Universidad de Castilla-La Mancha

www.finanzascuantitativas.com

Modelling stock prices as a discrete self-exciting process: Hawkes vs VAR

Gorka González de Mendívil Grau

Máster en Banca y Finanzas Cuantitativas
Trabajo de Investigación

Director: Dr. José S. Penalva

Universidad de Valencia
Universidad Complutense de Madrid
Universidad del País Vasco
Universidad de Castilla la Mancha

www.finanzascuantitativas.com

6 de julio de 2020

Modelling stock prices as a discrete self-exciting process: Hawkes vs VAR

Gorka González de Mendívil Grau
gorkamendivil@gmail.com

Facultad de Economía y Empresa
Universidad del País Vasco/Esukal Herriko Universitatea
Campus de Sarriko, 48015 Bilbao (Bizkaia)

Dr. José S. Penalva
jpenalva@emp.uc3m.es

Departamento de la Economía de la Empresa
Universidad Carlos III de Madrid
C/ Madrid, 126 - 28903 Getafe (Madrid)

July 6, 2020

Abstract

Automated trading, executed on computers via pre-programmed algorithms, constitutes a very large percentage of price movements that are observed during trading hours. Our main interest is to provide accurate price dynamics that allows us to make predictions on price movements or their trends. Clearly, price anticipation drives profits for High Frequency trading. Relatively recent contributions develop, both theoretically and through aplications, models based on discrete price events: jump processes. In particular, we consider multivariate Hawkes process as a prism through which we can study and observe the market reaction to different order types and their mutual interactions. Our main objective is to evaluate the appropriateness of this model to describe and predict price movements in electronic exchanges. To evaluate the performance of the model we will compare it to the standard benchmark, the VAR process.

Contents

1 Introduction

Today's electronic markets move very quickly. A very large percentage of price movements we see during trading hours is driven by automated trading, executed on computers via preprogrammed algorithms (see [1]). Another property of electronic markets is that trading occurs on a discrete grid of prices. In the US, where our data comes from, prices are set in cents. An asset can trade at \$10.02, but not at \$10.0178. Activity occurs on a very small time scale and on a fixed price grid.

We are interested in price dynamics. The main reason for this interest is that having a model of price dynamics allows us to make predictions on price movements or their trends, which is advantageous on trading strategies. In particular, price anticipation and short-term price deviations from the fundamental value of an asset are important drivers of profits for High Frequency trading strategies [11][2][1].

In this paper, the main tool used to study price dynamics is the VAR and related models that are based on a theoretical structure with discrete time steps and continuous prices. However, time is continuous and the price grid discrete. The continuity of time is some times addressed by changing the time scale, for example, using event time rather than clock time, and the discreteness of the grid is considered a sufficiently small problem that can be ignored. But, there are other tools that can be used instead of the VAR approach to study the price dynamics. Relatively recent contributions [12][3][7] develop, both theoretically and through aplications, models based on discrete price events: jump processes. In particular, [12] provides a very elegant model of a process, termed a Hawkes process, that has three distinct properties: (i) it is a continuous time process, (ii) it is discrete in nature: changes occur rarely and in discrete sized chunks, and (iii) the future movements of the process are affected by its own recent history. The main objective of the current paper is to evaluate the appropriateness of this model to describe price movements in electronic exchanges. To evaluate the performance of the model we will compare it to the standard benchmark, the VAR process.

The study proceeds as follows: we will first describe the data we use (section 2), then, in section 3, we will describe the Hawkes process, use it to build a model for our data, and justify the need from the additional complexity of the Hawkes process relative to a simpler standard Poisson process. This is followed by the modeling of the same data but with the VAR model (section 4). The study concludes by reviewing the results from both analyses and comparing the performance of the two models.

2 Data description

2.1 The Trades and Quotes (TAQ) dataset

Data used in this paper are obtained from the Trades and Quotes (TAQ) dataset which was purchased directly from an authorized commercial data provider. TAQ consists of high frequency recordings of trades and quotes for two derivative assets on twelve US exchanges. Table 1 presents the list of exchanges for the two particular assets SPY and SDS considered in this manuscript. The SPY Exchange Traded Fund (ETF) objective is to track the Standard & Poor's 500 Index (ticker: SPY), and the ProShares UltraPro Short S&P 500, which is a leveraged in-

[1] High Frequency algorithmic trading is studied in the book [2]. The proposed models that are developed look at how the strategies depend on different factors including the market maker's aversion to inventory risk, adverse selection, and short-term lived trends in the dynamics of the midprice.

verse ETF, aims a return twice the inverse of the daily performance of the S&P 500 Index (ticker: SDS).

Exchange	SDS	SPY
NASDAQ	15.61 %	21.37 %
BATS	11.47 %	16.30 %
NYSE-ARCA	16.64 %	22.18 %
NSE	–	0.15%
FINRA	0.23 %	0.30 %
CSE	6.76 %	4.69 %
EDGA	10.38 %	4.16 %
EDGX	9.96 %	12.39 %
NASDAQ-BX	8.09 %	4.19 %
NASDAQ-PSX	6.61 %	6.78 %
BATS-Y	12.96 %	5.10 %
IEX	1.28 %	2.37 %

Table 1: Exchanges for SDS and SPY in the TAQ

Table 1 indicates the percentage of regular trades and quotes for each asset per exchange in the TAQ. The arbitrage opportunities and the *National Best and Best Offer* rule make that order books of different exchanges are not significatively divergent. In this manuscript, we have selected the Nasdaq exchange during the complete observation period: December 14th, 2016.

The TAQ dataset consists of two main data tables: the Trades table and the Quotes table. The Trades table has a record for each transaction that took place on any of the participating exchanges. This table includes information about the volume, price and exchange for each the transaction. The Quotes table contains information on the best bid price and best ask price. Quotes table has a record for each update on either the price or volume of the bid side or ask side in any one of the exchanges. It is relevant to indicate that each record in the Quotes table represents a new state in the Order Book at Level I of the concrete exchange and asset just after a change has been produced by a particular order operation. Those orders that cause changes in the Quotes tables are Limit Orders (LO), Cancellation Ordes (CO) and Market Orders (MO) matching stored outstanding LOs. The time stamps in both tables are accurate to 1 millisecond.

Table 2 shows an extract of the TAQ dataset for SPY obtained on December 14th, 2016. Let us observe that we do not know the details about what is the concrete message that produces such information in the TAQ.

Exchange	Timestamp	Price	Quantity	bBidSide	bQuote
1	14916700	227.73	11	1	1
1	14916700	227.78	28	0	1
3	14916700	227.74	30	1	1
3	14916700	227.78	28	0	1
1	14916700	227.74	20	1	1

Table 2: Sample of TAQ dataset for SPY

The extraction of information from TAQ gives us the capacity to differentiate three types of files classified by the two assets, SPY and SDS, for the exchange considered in this paper, Nasdaq. For brevity, we omit the reference to 'Nasdaq' in the rest of the manuscript. Each type of file stores the Trades table, the Quotes table on the bid side, and the Quotes table on the ask side.

—-	Trades table	——	——	Quotes Table ask side	—-
Timestamp	Price	Quantity	Timestamp	Price	Size
34500597	227.74	10	34500027	227.74	15
34500597	227.74	1.7	34500028	227.74	15
34500597	227.74	1	34500028	227.74	18
34500597	227.74	2	34500028	227.74	19
34500597	227.74	1	34500031	227.74	19
34505203	227.73	1	34500031	227.74	18
34505203	227.73	1	34500031	227.74	18

Table 3: Trades table (on the left) and Quotes table of the ask side for SPY.

As an example of these tables, Table 3 shows the structure of Trades table (left) and Quotes table (right) on an actual extract of the dataset for SPY on December 14th, 2016. The first entry in Trades table indicates that at time 34500597 msg (9:35:0.597 a.m.) there has been a transaction of volume 10 at price 227.74$. Each entry in the Trades table can be translated into a MO that triggers a transaction. Unfortunately, the Trades table does not indicate if the order was a buy or a sell MO. The structure of the Quotes table on the ask side is also shown in Table 3 [2].

Note that multiple entries in those tables can have the same time stamp. We do not have unique time stamps for each entry in the tables. In reality, there are many message arrivals in 1-millisecond intervals, but we do not have a more refined time stamp. However, in the Trades table, MOs that appear with the same time stamp have been recorded in the same order as they appear in the table. Similarly, entries with the same time stamp in the Quotes table show the changes at level I of the Order Book in the order that they have occurred. For a group of entries with same time stamp in a Quotes table (just before the apparition of an entry with a new time stamp) we retain the price after the last entry as the best price for that millisecond (one for each side of the book, the bid and the ask). In addition, the volume indicated in that last entry corresponds to the volume outstanding in the queue of level I at the ask and bid sides respectively.

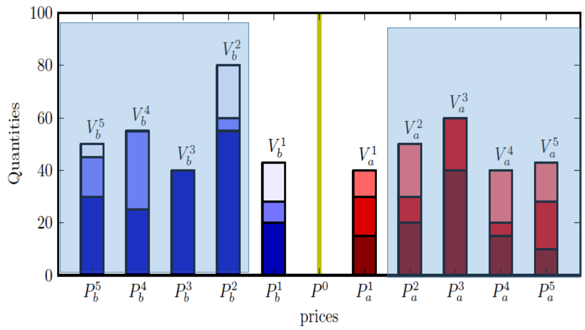

Figure 1: Order Book example.

Figure 1 is a graphical view of the potential information available in the LOB. Blue bars on the left half represents the queues of LOs with the same buy price. These buy orders form the bid side with prices P_b^i and quantities V_b^i. The right hand bars represent the queues of LOs with the same sell price. They form the ask side with prices P_a^i and quantities V_a^i. The level I of the book only represents the activity at the queues of first level. We do not have available all the activity in the book, and only observe the activity affecting $P_b^1(t)$, $V_b^1(t)$, $P_a^1(t)$, $V_a^1(t)$ at

[2] The structure of the Quotes table for the bid side is similar.

timestamp t. The line in the middle corresponds to the *mid-price* where $m(t) \overset{\text{def}}{=} \dfrac{P_a^1(t)+P_b^1(t)}{2}$; and $s(t) \overset{\text{def}}{=} P_a^1(t) - P_b^1(t)$ is the *spread* at time stamp t.

2.2 Data preprocessing and Order Flow reconstruction

Data in the TAQ are neither perfect nor complete and do not contain all the information to reconstruct each Order Book with full accuracy. Some kind of data preprocessing is necessary. We remove entries with price equal to 0. In addition, we use data for the trading time from 10:00 to 15:00 instead of the complete normal trading hours (9:30 a.m. - 4:00 p.m.). We have removed the first thirty minutes and the last hour of trading for parameter estimation and prediction purposes, since the dynamics of the Order Book close to the market open and market close may differ significantly to the behaviour throughout the remainder of the trading day.

We consider only *regular messages* on the TAQ for data processing. For each concrete exchange and asset, we can make a partial reconstruction of the order flow by identifying (i) LOs that establish a new best ask price (LO_{pa}); (ii) LOs that establish a new bid price (LO_{pb}); (iii) LOs that change the volume on the best ask price (LO_{qa}); (iv) LOs that change the volume on the best bid price (LO_{qb}); (v) COs on the ask side (CO_a); and (vi) COs on the bid side (CO_b). However, those last orders could also be MOs on one of the sides of the book. The identification between COs and MOs is not perfect, and requires some heuristic approach. We find that for some markets we observe no movement in the bid and ask price or quantity for significant periods of time, so even though there are many markets, the main activity is concentrated in a small subset of them. In Appendix A, we give a more detailed explanation of how we reconstruct the order flow.

2.3 State of the Order Book

Order flow reconstruction from TAQ is imprecise since we have only the information about the Order Book at Level I. For this reason, we use a more direct information that is obtained from the TAQ through the Quotes tables. Basically, we reconstruct the state of the Order Book for each asset at unique time stamps, the end of each millisecond (with some activity). Thus, we can merge information on the ask and bid sides for any asset to form a Unique Quotes Table (UQT) for that asset, where each entry contains the information at the end of each time stamp. In other words, we construct for each asset the state of the Order Book at unique time stamps reflecting the achieved state at that time. In the rest of this paper, we reserve the letter 'a' for 'ask' and 'b' for 'bid'. We construct two tables, UQT^{sds} and UQT^{spy}, with the following variables:

$$\begin{aligned}
\text{UQT}^{sds} &= [ts^{sds}, P_b^{sds}, V_b^{sds}, P_a^{sds}, V_a^{sds}] \\
\text{UQT}^{spy} &= [ts^{spy}, P_b^{spy}, V_b^{spy}, P_a^{spy}, V_a^{spy}]
\end{aligned}$$

Recall that the sets of time stamps are not the same on both tables, i.e., it is possible to find different time stamps in both tables. Table 4 shows an extract of the UQT^{spy}. Each entry represents the state of the Order Book at time stamp ts^{spy}. In both tables, entries are ordered: k-th and r-th at UQT^{spy} satisfies that $ts_k^{spy} > ts_r^{spy}$ if and only if $k > r$.

In addition, we classify each transition between two consecutive states into *types of events*. In this study, we consider transitions indicating a change in the price. Notation for the types of events are given in the next Table 5, where the notation used is: $asset_{side}^{move}$, $asset$ is the ticker

timestamp	price bid	volume bid	price ask	volume ask
36002574	227.55	1	227.56	60
36004084	227.55	1	227.56	60
36004377	227.55	2	227.56	60
36004833	227.55	4	227.56	60
36004834	227.55	10	227.56	59

Table 4: An extract of UQT^{spy}

of the asset (*sds* or *spy*), *side* is the side of the order book: a : ask, b : bid, and, *move* describes the change in the price (*up* or *down*).

Type of event	Notation	Type of event	Notation
sds_a^{down}	$e1$	spy_a^{down}	$e5$
sds_a^{up}	$e2$	spy_a^{up}	$e6$
sds_b^{down}	$e3$	spy_b^{down}	$e7$
sds_b^{up}	$e4$	spy_b^{up}	$e8$

Table 5: Notation for the types of events

Consecutive entries without any change are of the type *unchanged* events. However, we do not include events of this type in the analysis. Thus, we have a total of eight types of different events. This events are also timestamped, e.g, if the transition $UQT_{k-1}^{spy} \rightarrow UQT_k^{spy}$, for two consecutive entries $k-1$ and k, is an event of type spy_a^{up} then the pair (ts_k^{spy}, spy_a^{up}) defines the time stamped event. We consider the time stamp of the event when it takes place.

Furthermore, data in UQT^{sds} and UQT^{spy} allow us to track different signals for each asset $x \in \{sds, spy\}$: the spread signal $s^x(t)$, the mid-price signal $m^x(t)$, and the imbalance $imb^x(t)$[3] for each asset $x \in \{sds, spy\}$.

Figure 2 illustrates the price evolution for the two assets, $P_a^{sds}(t)$, $P_b^{sds}(t)$, $P_a^{spy}(t)$, and $P_b^{sds}(t)$, from 10:00 to 15:00. It is worth mentioning that SDS prices move in the opposite direction to SPY prices.

3 Multivariate Hawkes Process

In order to model the data we have just described, we consider using a Hawkes process. Certain physical systems exhibit temporal behaviours in which the arrival of events modifying the system state tend to cluster. These behaviors cannot be modeled by homogeneous Poisson process, due the fact that the arrival of each event depends on the recent history of arrivals. One of the simplest model that accouns for the dependency between events is the Hawkes process. This process generalizes the notion of homogeneous Poisson process model. Hawkes introduced such a model as a self-exciting point process [4]. This model has been useful for studying and modeling earthquarkes [9], and, more recently, in the study of spike trains in neurons [8].

Due to its ability for modeling clustering of events in temporal series, Hawkes processes have also been applied in different areas of finance. A particular application of interest is the modelling of the dynamics of Limit Order Books (LOB). In [5], the authors propose a mathematical

[3] *Imbalance* value was introduced by Cartea [6] to measure the imbalance between the bid-ask volumes as a predictor for price movement $imb^x(t) = V_b^x - V_a^x / V_b^x + V_a^x$

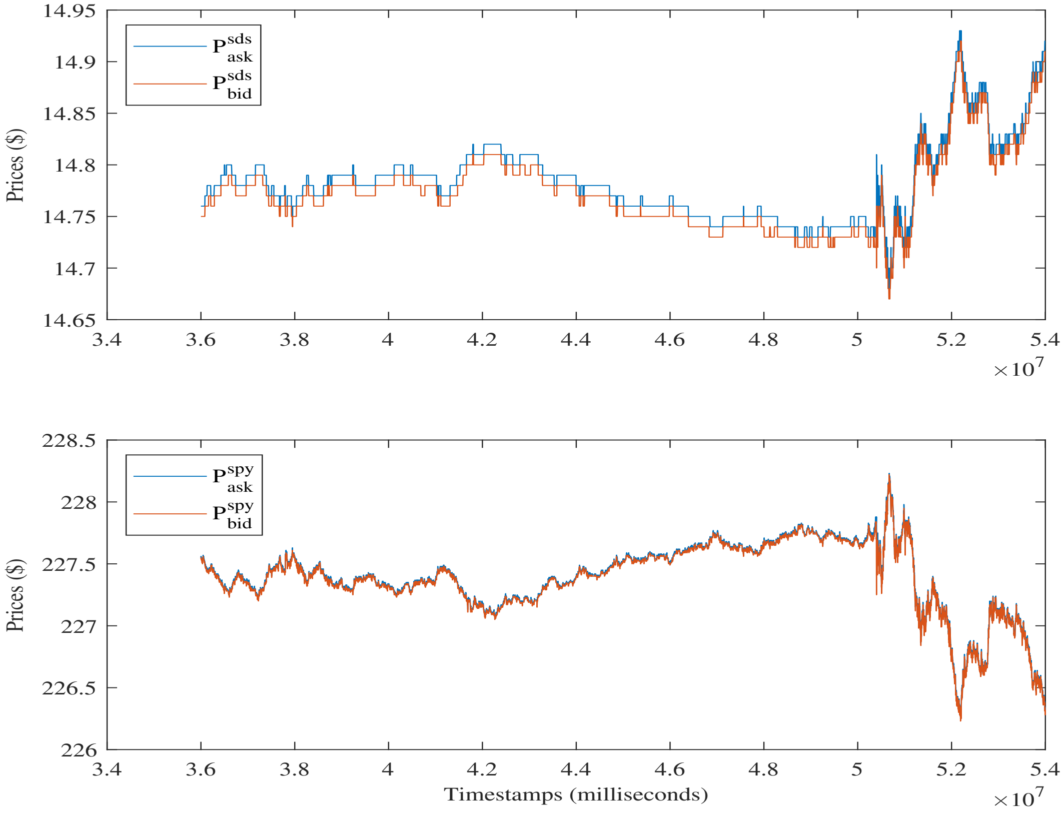

Figure 2: Movements of ask and bid prices of SDS and SPY in Nasdaq.

structure in which the temporal series of LOs, MOs an COs arriving at the LOB are modeled using multivariate Hawkes processes. Multivariate-Hawkes processes have the characteristic that they can include contagion effects among the temporal series. Mutual influence between series can be also interpreted as causal relations between them, see [10].

In a similar way, Y. Chen [7], introduces a framework based on multivariate Hawkes processes with exponential decays for estimating the direction of the next price movement in a LOB. Chen demonstrates that Hawkes processes perform better than Poisson processes in terms of both model fitting and performance prediction. In addition, Cartea et. al. [11] have developed high frequency trading strategies to post limit sell and buy orders taking into account multi-factor mutually exciting processes modeled via Hawkes processes. In this study, we consider the Hawkes process as a prism through which we can study and observe the market reaction to different order types and their interactions. From that point of view, we will use Hawkes processes for modeling the sequences of events that modify the state of the Order Book.

This section is devoted to apply estimated multivariate Hawkes processes as a tool for predicting the price movements of the assets considered in this manuscript, SDS and SPY in Nasdaq. The rest of this section is organized as follows. Subsection 3.1 provides a short presentation of the multivariate Hawkes process and its main properties. Subsection 3.2 describes the data preparation for the different estimations and predictions we are going to do in the study. Subsection 3.3 illustrates the goodness of fit of the estimated models in order to validate our initial assumption about the appropriate selection of Hawkes processes. In subsection 3.4, we explain the prediction process using the estimated Hawkes model and the results obtained for the prediction tests. Finally, a subsection of discussion end the section.

3.1 Notion of multivariate Hawkes process

This subsection includes a short presentation of the notion of multivariate Hawkes process and some well-known results. We start with the original definition of a D-variate Hawkes process [4] (in short form D-variate-HP). The reader may consult [14] for a short introduction to Hawkes processes and some of their statistical properties. For a complete revision of point processes see Daley and Vere-Jones's book [15].

A D-dimensional multivariate point process consists of an increasing sequence of positive random times $\langle T_k \rangle_{k\in\mathbb{N}}$ and a matching sequence of random marks $\langle E_k \rangle_{k\in\mathbb{N}}$ where each mark E_k is an element of a set of marks, i.e., $E_k \in \mathcal{E}$ with $\mathcal{E} = \{1, 2, ..., D\}$[4].

For each $k \in \mathbb{N}$, the pair (T_k, E_k) is an *event* of type E_k arriving at time T_k. Each finite concrete realization of the random process is a finite sequence of events that will be denoted by lower case letters, $\langle (t_k, i_k) \rangle_{1\leq k \leq K}$. In addition, we denote by t_k^j the time of the k-th event of type j in the sequence. In the following, i, and j denote types of events and k is used to denote an index of order of the elements in the sequence.

For a D-dimensional multivariate point process, $\langle (T_k, E_k) \rangle_{k\in\mathbb{N}}$, we can provide an equivalent counting process representation given by $\mathbf{N}(t) \overset{\text{def}}{=} (N_1(t), N_2(t), ..., N_D(t))$ where

$$N_i(t) \overset{\text{def}}{=} \sum_{k\in\mathbb{N}} \mathbb{1}_{\{T_k \leq t, E_k = i\}}, \qquad t \geq 0, \qquad i \in \mathcal{E} \tag{1}$$

The process $N_i(t)$ counts the number of events of type $i \in \mathcal{E}$ that have occurred until time t. The counting processes $\mathbf{N}(t)$ is called *non-explosive* if $\Pr\{Lim_{k\to\infty} T_k = \infty\} = 1$. Intuitively: a counting process is non-explossive if no realizations have a finite time interval with an infinite number of events in it.

The next definition presents the notion of a D-variate-HP as a mutually exciting process. It captures the self-excitation feature from the univariate case, and it also has cross-exciting effects between different dimensions.

Definition 1 *(Hawkes 1971 [4]) Let* $\mathbf{N}(t) \overset{def}{=} (N_1(t), N_2(t), ..., N_D(t))$ *be a D-dimensional multivariate point process satisfying*

(i) $\mathbf{N}(0) = \mathbf{0}$*; and, for each* $i \in \mathcal{E}$*,*

(ii) $\lambda_i(t)$ *is a left continuous stochastic process given by the Stieltjes integral*

$$\begin{aligned} \lambda_i(t) &\overset{def}{=} \mu_i + \sum_{j\in\mathcal{E}} \int_0^t \alpha_{ij}\ e^{-\beta_{ij}(t-s)} dN_j(s) && (2) \\ &= \mu_i + \sum_{j\in\mathcal{E}} \sum_{\{k:t_k^j < t\}} \alpha_{ij}\ e^{-\beta_{ij}(t-t_k^j)} && (3) \end{aligned}$$

[4] The label assigned to a mark denotes the type of event.

where $\mu_i > 0$, $\alpha_{ij} \geq 0$ and $\beta_{ij} \geq 0$, for i, $j \in \mathcal{E}$.

(iii) $\lambda_i(t)$, independently for each $i \in \mathcal{E}$, is the stochastic intensity of the marginal point process $N_i(t)$

$$P\{N_i(t+h) - N_i(t) = 1 | \mathcal{F}^{\mathbf{N}}_{t^-}\} = \lambda_i(t)h + o(h) \tag{4}$$

(iv) the point process is orderly,

$$P\{N_i(t+h) - N_i(t) \geq 2 | \mathcal{F}^{\mathbf{N}}_{t^-}\} = o(h) \tag{5}$$

where $\mathcal{F}^{\mathbf{N}}_{t^-}$ is the natural filtration of the process, is called a D-variate Hawkes process with exponential decays on $[0, \infty)$.

By (4), the intensity $\lambda_i(t)$ in (2) is interpreted as the infinitesimal rate of events of type i at time t. The μ_i parameters are called the *base intensities* and can be viewed as background intensities. Whenever an event occurs, the intensities increase, making subsequent events arrive at a higher frequency. Such effects are controlled by the kernels $k_{ij}(t)$. The original definition [4] is for *exponential kernels* where an exponential kernel is parameterized as:

$$k_{ij}(t) \stackrel{\text{def}}{=} \alpha_{ij} \exp(-\beta_{ij} \cdot t), \qquad t \geq 0, \qquad i, j \in \mathcal{E} \tag{6}$$

with $\alpha_{ij} \geq 0$ and $\beta_{ij} \geq 0$. These parameters are called the *impact coefficient* and *decay coefficient* between dimensions j and i respectively.

Kernel functions control the instantaneous increases and the relaxation speeds of the intensities in response to excitations. For a D-variate-HP, $k_{ii}(t)$ describe the self-excitation, while $k_{ij}(t)$ for different events, $i \neq j$, may be viewed as a measure of the cross-excitation, that is, the impact of an event of type j on the arrival of an event of type i.

From (2), each D-variate-HP is fully determined by their baseline intensity vector $\mu \stackrel{\text{def}}{=} [\mu_i]_{i \in \mathcal{E}}$ and the matrix $\mathbf{K} \stackrel{\text{def}}{=} [k_{ij}(t)]_{i,j \in \mathcal{E}}$ of kernel functions. The choice of exponential kernel functions (6) presents the additional advantage that the joint process $(\mathbf{N}, \lambda)$, with $\lambda \stackrel{\text{def}}{=} [\lambda_i]_{i \in \mathcal{E}}$, is a Markov model as was proven by Massoulié [16]. Massoulié also studied general conditions for *stability* and *uniqueness* of a D-variate-HP with arbitrary kernels. One of Massoulié's results is that the components of the kernel matrix $\mathbf{K}$ have to be formed by bounded functions. This condition is satisfied by exponential kernels. In fact, the original paper of Hawkes [4] indicates that a sufficient condition for *stationarity* is that the matrix $\Gamma = [\frac{\alpha_{ij}}{\beta_{ij}}]_{i,j \in \mathcal{E}}$ has a spectral radius[5] $\rho(\Gamma) < 1$.

Simulation algorithms for D-variate-HPs have been provided by different authors. In particular, Chen [7] presents an algorithm extending the previous one proposed by Ogata [9].

Likelihood function. Parameters of a D-variate-HP can be estimated via maximum likelihood estimation as it was proposed by Ozaki (1979) [18]. Let us consider a marked point process $\langle (T_k, E_k) \rangle_{k \in \mathbb{N}}$ and a given realization $\langle (t_k, i_k) \rangle_{1 \leq k \leq K}$ over a time horizon $[0, \tau]$. By grouping the parameters of the model in a vector θ where $\theta \stackrel{\text{def}}{=} [\mu, \alpha, \beta]$ with $\mu \stackrel{\text{def}}{=} [\mu_i]_{i \in \mathcal{E}}$, $\alpha \stackrel{\text{def}}{=} [\alpha_{ij}]_{i,j \in \mathcal{E}}$, and $\beta = [\beta_{ij}]_{i,j \in \mathcal{E}}$, the *log likelihood* function can be expressed as:

[5] The maximum of the set formed by the modules of its eigenvalues.

$$\text{Ln } \mathcal{L}(\theta) = \sum_{k=1}^{K} \text{Ln } \lambda_{i_k}(t_k) - \int_0^{\tau} \sum_{i \in \mathcal{E}} \lambda_i(t)dt \tag{7}$$

Equation (7) is obtained from Daley and Vere-Jones' book [15] in subchapter 7.3, for log likelihood functions using conditional intensities of marked point processes. The details on the computation of such a function is not provided in this manuscript. Its computation requires numerical methods of minimization and the execution time is expensive. However, the $\sum_{i \in \mathcal{E}}$ appearing explicitly in the second term of (7) and implicitly in the first term, allow a parallel computation for the log likelihood function. The resulting estimated parameters are denoted by $\hat{\theta}$.

Goodness of fit. Once the parameters $(\hat{\theta})$ of a D-variate-HP have been estimated for a given realization $\langle (t_k, i_k) \rangle_{1 \leq k \leq K}$ of the process $\langle (T_k, E_k) \rangle_{k \in \mathbb{N}}$, it is required to study the adequacy of the estimated model to the observed data. For such a purpose, we use a well-known *goodness-of-fit* criterion for the distribution of forward recurrence times. For notation convenience, we denote by k_i the index k in the sequence that exactly corresponds to the event of type i. This notation allows us to use them like ordinal indexes for i, e.g., $(k-1)_i$ is the previous index of i to k_i, or 1_i is the first index for i.

Given the sequence $\langle T_{k_i} \rangle_{1_i \leq k_i \leq K_i}$ of times at which an event of type i occurred in $\langle (T_k, E_k) \rangle_{k \in \mathbb{k}}$, we have the following classic result: the transformed durations (*time residuals*) $\{\tau_{k_i}\}_{1_i \leq k_i \leq K_i}$ are defined by

$$\tau_{k_i} \overset{\text{def}}{=} \int_{T_{(k-1)_i}}^{T_{k_i}} \lambda_i(t)dt \qquad k_i = 1_i..K_i, \quad i \in \mathcal{E} \tag{8}$$

which are i.i.d. *exponential random variables with parameter* 1, see for example the paper of Bowsher [17].

This property is used to test the goodness-of-fit of the estimated model by using the estimated intensity $\hat{\lambda}_i(t)$ for each type of event i in the realization $\langle (t_k, i_k) \rangle_{1 \leq k \leq K}$. The Q-Q plots of the empirical quantiles with respect to the theoretical exponential distribution quantiles are used to asses the goodness of fit.

Other exponential kernels. In this paper, we also use for estimation of a D-variate-HP a kernel of the following form:

$$\phi_{ij} = \sum_{u=1}^{U} \alpha_{ij}^{u} \beta^{u} exp(-\beta^{u} \cdot t) \qquad i, j \in \mathcal{E} \tag{9}$$

This kind of exponential kernel is called exponential kernel (with fixed decays) of order U. In this case, the decay coefficients are fixed *a priori*. These kernels maintain the property of a fast computation of the intensity function via a recurrence equation. As decay coefficients are fixed, the computation of the maximum Likelihood function is faster than in the case of a simple exponential kernel.

In Appendix B, we provide a fast form to compute the intensity function using the kernel (9). In Appendix C, we illustrate the algorithm for the simulation of the next event based on Ogata's simulation algorithm [9]. We have implemented those programs in Matlab R2017 since they are needed for forecasting based on estimated D-variate-HPs.

3.2 Data preparation

To prepare data for parameter estimation and prediction we need to define time intervals and establish the sequence of events within those intervals. Each interval $[\tau_r, \tau_r']$ is denoted by Tr, i.e., $Tr \overset{\text{def}}{=} [\tau_r, \tau_r']$. The sequence of events within interval Tr is denoted by $S^r = \langle (t_k, i_k) \rangle^r_{1 \leq k \leq K^r}$, where $i_k \in \{1..8\}$ and K^r denotes the total number of events in the sequence. We recall that there are eight types of events. The reader is referred to Table 5 for the types of events and the notation, e.g., an event $e1$ (or simply 1) is an event of type sds_a^{down}.

We do not study each asset in isolation. Instead we consider the behaviour of events of the two assets as a whole. In order to build the sequences of events, we have to use the state Order Book as indicated in section 2.3. Each time stamped event in a sequence corresponds to a state transition in the Order Book that is produced by such an event.

Each sequence S^r has associated to it a counting process: $\mathbf{N}^r(t) = (N_1^r(t), ..., N_8^r(t))$. For each process $\mathbf{N}^r(t)$ we will estimate the parameters of an 8-variate-HP. Thus, the estimated 8-variate-HP will be used to make predictions about the evolution of the counting process. In order to evaluate such predictions, we consider the previous time ranges extend about 30 minutes. We use the notation ΔS^r to denote the events within that time interval. Table 6 shows the proposed periods for estimation and prediction. The whole period [10:00, 15:00] is simply denoted by T.

Counting Process	Time Range	Extended Range	Sequence of events	Sequence extensions
$\mathbf{N}(t)$	T = [10:00, 15:00]	—	S	—
$\mathbf{N}^1(t)$	$T1$ = [10:00, 12:00]	(12:00, 12:30]	S^1	ΔS^1
$\mathbf{N}^2(t)$	$T2$ = [11:00, 13:00]	(13:00, 13:30]	S^2	ΔS^2
$\mathbf{N}^3(t)$	$T3$ = [12:00, 14:30]	(14:30, 15:00]	S^3	ΔS^3

Table 6: Periods for estimation and prediction

We have selected these periods based on the results in Figure 2. In Table 6, the first period $T1$ = [10:00, 12:00] starts near of the market open (9:30) and the second period $T2$ = [11:00, 13:00] shows a behavior similar to the first one with respect to price movements. The last period $T3$ = [12:00, 14:30] presents a different behavior as it approaches the market closing time (16:00). Activity in the number of orders received is significantly increased, and this fact is also reflected in the state of the Order Book.

Type of event	S	S^1	S^2	S^3
$e1$: sds_a^{down}	294	30	21	198
$e2$: sds_a^{up}	305	33	18	206
$e3$: sds_b^{down}	260	31	23	164
$e4$: sds_b^{up}	276	34	20	177
$e5$: spy_a^{down}	13035	2030	1248	7436
$e6$: spy_a^{up}	12789	1992	1270	7250
$e7$: spy_b^{down}	13031	2138	1322	7259
$e8$: spy_b^{up}	13228	2109	1346	7434

Table 7: Number of events of each type for the considered periods.

To check the different activity patterns, Table 7 shows the total number of events of each type for each period. Though periods are overlapping, periods $T1$ and $T2$ show slight differences.

In the second period, as we move away from the opening time, the number of events decreases and the activity is more relaxed than in the first period near the opening. In the last period $T3$ the number of events increases significantly as we get closer to the market closing time (16:00). In all the periods the activity in SDS is notably lower than SPY.

3.3 Parameter estimation and goodness of fit

Let us consider the whole period of observation $T = [10{:}00, 15{:}00]$ and the complete sequence of events $S = \langle (t_k, i_k) \rangle_{1 \leq k \leq K}$. Each recorded event time is an observation of points $\{t_{k_i}\}_{1_i \leq k_i \leq K_i}$ in the counting process $N_i(t)$ for each $i \in \{1..8\}$. We assume that the observations are from a realization of an 8-variate-HP with exponential kernels as defined in Equation (6). The log-likelihood function for $\theta = [\mu_{1\times 8}, \alpha_{8\times 8}, \beta_{8\times 8}]$ is Ln $\mathcal{L}(\theta | \{t^i_k\}^{i:1..8}_{1\leq k\leq K})$. By eq. (7), the log likelihood function is nonlinear with respect to the parameters θ and its maximization has to be performed numerically by using nonlinear optimization techniques. For such a purpose, we use the Python library `mpoints`[6] developed by Patrichi and Pakkanen [13]. The main goal of this initial parameter estimation for the whole counting process $\mathbf{N}(t)$ is to provide a significant evidence that the process under analysis is derived from an 8-variate-HP. At this point of the analysis the values of the estimated parameters $\hat{\theta}$ are not relevant since they are only useful to evaluate the goodness of fit of the estimated model.

For such purpose, we examine the time residuals τ_{k_i} (see (8)) obtained with the estimated intensity $\hat{\lambda}(t)$ for the sequence S. For comparison, we use an 8-variate Poisson processes as the baseline model. An 8-variate Poisson process can be considered as a special case of an 8-variate-HP, with the constraint that $\alpha_{ij} = 0$ for $i, j \in \{1..8\}$. The parameters to estimate are $\theta_{poisson} = [\nu_1, .., \nu_8]$ and whose values are calculated using $\nu_i = \dfrac{K_i}{\tau' - \tau}$ where K_i are the total number of events of type i and $\tau' - \tau$ is the duration of the period T.

In Figure 3 red colour lines are for the residuals of the estimated 8-variate-HP, line black is the exponential distribution with parameter 1, and the green colour lines are for the 8-variate Poisson process. The Q-Q plots in Figure 3 show that the unconditional distribution of residuals vary among the type of events. *SPY* events are well fitted to the exponential distribution and *SDS* events are poorly adjusted. However, in comparison with the Poisson process, the provided estimated Hawkes model can be considered as an 8-variate-HP.

As the reader can see in Table 7, although the total number of events of SDS is much smaller than the total number of events of SPY, there is convergence when searching for the maximum of the log likelihood function. However, using an exponential kernel as the one used in the parameter estimation with the algorithm provided in [13], the result for the maximum of the log likelihood function is a negative value. We reevaluate the estimation using adjustments in the gradient descent technique used in the Python library to improve the results. Unfortunately, the adjustments lead to very high execution times.

Due to timing constraints, we consider an exponential kernel of order U with fixed decays (see eq. (9)). The program used for the estimation of the multivariate Hawkes process with kernel functions of this type is the Python library `Ticks`[7]. Using exponential kernels of order U, the model parameters to be estimated are $\mu \overset{\text{def}}{=} [\mu_i]_{i\in\mathcal{E}}$, $\kappa \overset{\text{def}}{=} [\kappa^u_{ij}] = [\alpha^u_{ij} \cdot \beta^u]_{i,i\in\mathcal{E},u\in U}$ given the fixed decays $\beta \overset{\text{def}}{=} [\beta^u]_{u\in U}$. We recall that our set of events is $\mathcal{E}$ is $\{1, ..., 8\}$.

[6] `httpps://mpoints.readhedocs.io`

[7] `https://x-datainitiative.github.io/tick/modules/api.html#api-datasets`

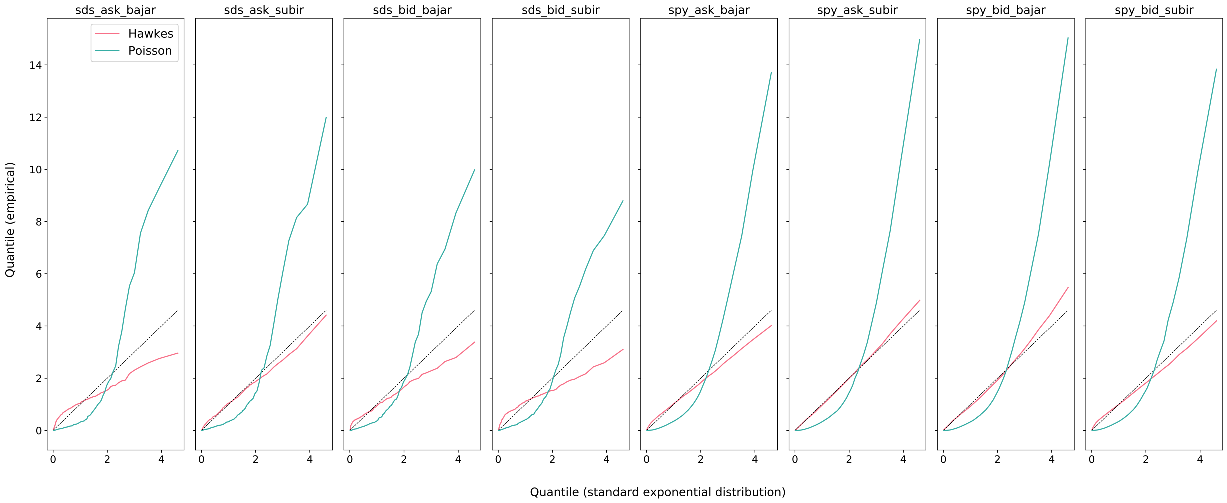

Figure 3: Q-Q plot for the estimated 8-variate-HP (red), 8-variate Poisson process (green) against the theoretical quatiles of a standard exponential distribution (black).

Thus, we start the estimation of our 8-variate-HP by selecting a set of fixed decays. We use an heuristic procedure with the target of obtaining a better value for the maximum log likelihood for the whole sequence S and its associated counting process $\mathbf{N}(t)$. We have obtained that with four decays ($U = 4$) defined as $\beta^1 = 0.1$, $\beta^2 = 2$, $\beta^3 = 32$, and $\beta^4 = 256$, the estimated parameters are obtained with a maximum log likelihood value of 16.3016104. Varying the order U and values of fixed decays do not provide any significant improvement in the value of the maximum log likelihood function. The base intensities obtained are $\mu = [0.0000, 0.0000, 0.0000, 0.0000, 0.0114, 0.0213, 0.0286, 0.00445]$. The base intensities for the SDS events are negligible by which we conclude that their conditional intensities functions are purely endogenous. However, for the SPY events there appear to be contributions to conditional intensities functions coming from outside the process considered.

We now turn to the matrix $\hat{\Gamma}$ (kernel norm) with the estimated parameters given by the equation,

$$\|\hat{\Gamma}_{ij}\| \overset{\text{def}}{=} \int_0^\infty \hat{\phi}_{ij}(s)ds \qquad i, j \in \mathcal{E} \tag{10}$$

Table 8 illustrates the values in the kernel norm matrix. Each value $\hat{\Gamma}_{ij}$ may be interpreted as the average number of events of type i that have been directly triggered by an event of type j within t seconds of its occurrence (in practice its not necessary to do the integral up to ∞). Thus, we may view those values as a measure of the Granger causality among the types of events. We find cross-effects: SDS events trigger SPY events and viceversa. Zeros in the kernel norm matrix indicate a lack of Granger causality between event types. The fact that SPY aims to track the return of the SPX, and the SDS, the twice inverse of daily performance of SPX, may justify such zero values in the kernel norm matrix in those entries. For example, when an spy_a^{up} event happens increasing the SPY price ask, events sds_a^{up} or sds_b^{up} increasing SDS price ask or bid do not occur.

To check stationarity we calculated the spectral radius $\rho(\hat{\Gamma})$. Its value is greater than 1: $\rho(\hat{\Gamma}) = 1.8$. This does not prevent that the estimated process is a stationary process since the

——	sds_a^{down}	sds_a^{up}	sds_b^{down}	sds_b^{up}	spy_a^{down}	spy_a^{up}	spy_b^{down}	spy_b^{up}
sds_a^{down}	0.2446	0.3628	0.8281	0.2423	0	0.0080	0	0.0026
sds_a^{up}	0.3772	0.3053	0.3605	0.3125	0.0054	0	0.0132	0
sds_b^{down}	0.3271	0.3377	0.3622	0.4219	0	0.0100	0	0.0039
sds_b^{up}	0.3438	0.7897	0.4013	0.2471	0.0040	0	0.0064	0
spy_a^{down}	0.4462	0.9399	0.3807	0.6578	0	0.5376	0.3774	0
spy_a^{up}	0.7143	0.3137	0.5055	0.0602	0.6138	0.1191	0	0.1383
spy_b^{down}	0.2965	0.8630	0.2441	0.3515	0.1887	0	0.1194	0.5528
spy_b^{up}	0.9156	0.6263	0.9244	0.4230	0	0.3832	0.5469	0

Table 8: Kernel norm of the estimated 8-variate-HP for the counting process $\mathbb{N}(t)$

condition $\rho(\hat{\Gamma}) < 1$ is a sufficient condition for stationarity. However, we think that this value (1.8) may be due to difference in the number of events in the joint estimation–the number of events for SDS is very low as compared with those of the SPY. There are 52.083 SPY events and only 1.135 SDS events in the entire period. This difference affects the estimation of the Hawkes process. If we compare this estimate with the ones when we fit the models individually, we find that the spectral radius of the kernel norms for the estimated 4-variate-HP of SPY events only and the estimated 4-variate-HP of SDS events only are 0.9673 and 0.8619 respectively. The point process for SDS events and SPY events taking in an isolated way are stationary 4-variate-HPs.

3.4 Prediction via a Multivariate Hawkes process

Our main goal is to fit the dynamics of SDS and SPY prices. We now look at how well the Hawkes process helps predict the price movements for given an estimated 8-variate-HP for a given time period. Table 6 indicates estimation periods and the periods used to predict the data. For each estimation period $T_r = [\tau_r, \tau_r']$ given in that table, we can provide a price model as follows.

We start from the assumption that the data is generated by the counting process $\mathbf{N}^r(t)$, with $t \in [\tau_r, \tau_r']$. The counting process is an 8-dimensional point process $S^r = \langle (t_k, i_k) \rangle_{1 \leq k \leq K^r}$ is formed by the marginal counting processes $(N_1^r(t), ..., N_8(t))$. Each $N_i^r(t)$ corresponds to an event of type $i \in \{1, .., 8\}$. In our case, the eight types of events are $1 : sds_a^{down}$, $2 : sds_a^{up}$,$3 : sds_b^{down}$, $4 : sds_b^{up}$, $5 : spy_a^{down}$, $6 : spy_a^{up}$, $7 : spy_b^{down}$, and $8 : spy_b^{up}$. At this point, we make a simplifying assumption, namely, that the jump of the best bid or ask price following an event of each type is a constant value, one tick (one cent). This approximation is consistent with the real behaviour of the chosen data set, for which the average jump size of the best bid and ask prices is about \$0.01. Under this assumption, we can reconstruct the price dynamics in a simple way as a by-product of event arrivals:

$$\begin{array}{rcl} P_a^{(r)sds}(t) & = & P_a^{(r)sds}(\tau_r) + (N_2^r(t) - N_1^r(t)) \cdot \eta_a^{(r)sds} \\ P_b^{(r)sds}(t) & = & P_b^{(r)sds}(\tau_r) + (N_4^r(t) - N_3^r(t)) \cdot \eta_b^{(r)sds} \\ P_a^{(r)spy}(t) & = & P_a^{(r)spy}(\tau_r) + (N_6^r(t) - N_5^r(t)) \cdot \eta_a^{(r)spy} \\ P_b^{(r)spy}(t) & = & P_b^{(r)spy}(\tau_r) + (N_8^r(t) - N_7^r(t)) \cdot \eta_b^{(r)spy} \end{array} \qquad (11)$$

for any $t \in [\tau_r, \tau_r']$.

As we have estimated an 8-variate-HP in that period T_r, we need a method in order to predict those prices in the test period of prediction. The idea of the method is based on the prediction method proposed by Daley and Vere-Jones [15]. Briefly, the prediction algorithm is based on a simulation algorithm of the Hawkes process [15] (Chapter 7):

For a given realization of the point process on some finite interval $[t1, t2]$*, let us suppose that our aim is to predict a particular quantity* V *that can be represented as a functional of a finite segment of the future of the process. To fulfil our aim, we estimate the distribution of* V*. The prediction procedure is as follows*

1. *Choose a time horizon* $(t2, t3)$ *sufficient to encompass the predicted quantity of interest.*
2. *Simulate the process forward over* $(t2, t3)$ *using the estimated structure of the conditional intensity function and initial history* $H0$*.*
3. *Extract from the simulation the value* V *of the functional that it is required to predict.*
4. *Repeat steps 2 and 3 sufficiently often to obtain the required precision for the prediction.*
5. *The output consists of the empirical distribution of the values of* V *obtained from the successive simulations.*

In the previous procedure the quantity of interest may be for example the time interval for the next event, or the next four events. In our case, we are interested on predicting a sequence of events of certain length since that predicted sequence can be used to predict the marginal counting processes and, thus, to have a prediction of the price movements. Given equation (11), the accuracy in the predicting price changes depends on an accurate prediction of the differences $(N^r_j(t) - N^r_i(t))$. Therefore, we describe how to predict such differences.

1. Let us consider an estimation period $T_r = [\tau_r, \tau'_r]$. Let $S^r = \langle (t_k, i_k) \rangle_{1 \leq k \leq K^r}$ be the point process in that period and $\mathbf{N}^r(t)$ with $t \in [\tau_r, \tau'_r]$ its counting process. The 8-variante-HP, with exponential kernel of order U and fixed decays, is estimated for that period with parameters $\hat{\theta}^r = [\mu_i, \kappa_{i,j}, \beta^u]$. Its estimated intensity is $\hat{\lambda}^r(t) = [\hat{\lambda}^r_i(t)]$. The intensity is calculated using the procedure described in the Appendix B. From this we have the matrix $A^r(t_{K^r}) = [A^u_j]$ for the last point of S^r.
2. Let $\Delta S^r = \langle (t_p, i_p) \rangle_{K^r < p \leq L^r}$ the test point process for $(\tau'_r, \tau''_r]$[8]. The counting process $\mathbf{N}^r(t)$ is calculated for $t \in [\tau_r, \tau''_r]$, i.e., it is updated with the extended point process ΔS^r.
3. The prediction process is based on the idea of 'evaluation on a rolling forecasting origin'.
4. (**Initial Step**): Starting at the last event of S^r with index K^r, we select a number δ of new events to be predicted. Set a variable *origin* $:= K^r$ and $\mathbf{N}^r(t_{origin})$.
5. (**Iteration Step**): We use the simulation algorithm based on next event computation provided in Appendix C to calculate such new events from *origin* $+ \delta$. Due to the random nature of that algorithm, we follow the advice of Daley and Vere-Jons, we repeat this computation up to 10 times. Thus, we have ten different sequences of new events. For each sequence of new events, we calculate the predicted increment in the counting process $\Delta\mathbf{N}_{ne}$, with $ne : 1..10$. Finally, they are aggregated to form the predicted increment $\Delta\mathbf{N}$.
6. (**Result**): The counting process $\mathbf{N}^r(t_{origin}) + \Delta\mathbf{N}$ is the prediction for the test counting process $\mathbf{N}^r(t_{origin+\delta})$.
7. (**Exit condition**): Since, this procedure is iterative: (1) as our simulation algorithm requires the matrix A, we update it for $t_{origin+\delta}$, i.e., recalculate $A(t_{origin+\delta})$; (2) we update *origin* $:=$ *origin* $+ \delta$ and (3) repeat (**Iteration Step**) until the length of the test sequence L^r has been reached.

[8] τ''_r is approximately $\tau'_r + 1800$ sg.

Remark: $t_{origin+\delta}$ is always the time of an event at $\Delta S^r = \langle(t_p, i_p)\rangle_{K^r<p\leq L^r}$, this also implies that $A(t_{origin+\delta})$ is well calculated. In (**Iteration Step**), the aggregation may be done on different ways. We have selected the simplest one: $\Delta\mathbf{N} = round(\frac{1}{10}\sum_{ne=1}^{10}\Delta\mathbf{N}_{ne})$. In addition, we can select a greater number of trials in the (**Iteration Step**). However, ten trials has been sufficient to obtain reasonable results.

As the reader can see, the proposed prediction procedure uses the test point process as feedback for the next prediction once the previous prediction has been obtained. This is a legitimate prediction procedure since we never use the future knowledge of the test point process to get a prediction. Note that the possible errors in the predictions are maintained throughout the forecast period. Having the results $\mathbf{N}^r(t_{origin}) + \Delta\mathbf{N}$ and $\mathbf{N}^r(t_{origin+\delta})$ we can trivially calculate the differences that are required in equation (11) to obtain the price prediction.

In order to evaluate prediction accuracy, we use the Mean Absolute Error,

$$\text{MAE}_y^{(r)x}(\delta) = \frac{1}{M}\sum_{m=1}^{M}|\hat{P}_y^{(r)x}(t_{K^r+m\delta}) - P_y^{(r)x}(t_{K^r+m\delta})| \tag{12}$$

where $x \in \{sds, spy\}$; $y \in \{a, b\}$; δ is the number of predicted events; $t_{K^r+m\delta}$, with $m : 1..M$, are time stamps of the test point process $\Delta S^r = \langle(t_p, i_p)\rangle_{K^r<p\leq L^r}$; and M is the last value satisfying $M \leq \dfrac{L^r - K^r}{\delta}$.

The following Tables show the results obtained with the proposed prediction procedure.

	$\delta = 1$	$\delta = 2$	$\delta = 4$	$\delta = 8$
$\text{MAE}_a^{(1)sds}(\delta)$	0.0001	0.0002	0.0004	0.0008
$\text{MAE}_b^{(1)sds}(\delta)$	0.0001	0.0002	0.0004	0.0011
$\text{MAE}_a^{(1)spy}(\delta)$	0.0049	0.0066	0.0078	0.0123
$\text{MAE}_a^{(1)spy}(\delta)$	0.0054	0.0066	0.0083	0.0133

Table 9: Results obtained in the test period T1

	$\delta = 1$	$\delta = 2$	$\delta = 4$	$\delta = 8$
$\text{MAE}_a^{(2)sds}(\delta)$	0.0001	0.0001	0.0002	0.0008
$\text{MAE}_b^{(2)sds}(\delta)$	0.0001	0.0001	0.0004	0.0015
$\text{MAE}_a^{(2)spy}(\delta)$	0.0053	0.0078	0.0095	0.0147
$\text{MAE}_a^{(2)spy}(\delta)$	0.0053	0.0072	0.0091	0.0141

Table 10: Results obtained in the test period T2

As we can see in Tables 9 to 11, MAE values for the prediction of the next δ events are very similar across sample periods. There are no great differences between the different periods despite the fact that the last period near the closing market has a different behavior (see Figure 2). Best predictions are obtained when $\delta = 1$, i.e., when we attempt to predict the next event. As we try to predict more new events, the prediction results worsen. The error when predicting $\delta = 8$ events is highest and yet the prediction error grows less than linearly. In all cases, since

	$\delta = 1$	$\delta = 2$	$\delta = 4$	$\delta = 8$
$\mathrm{MAE}_a^{(3)sds}(\delta)$	0.0001	0.0002	0.0003	0.0007
$\mathrm{MAE}_b^{(3)sds}(\delta)$	0.0001	0.0002	0.0003	0.0009
$\mathrm{MAE}_a^{(3)spy}(\delta)$	0.0052	0.0080	0.0083	0.0124
$\mathrm{MAE}_a^{(3)spy}(\delta)$	0.0054	0.0074	0.0084	0.0125

Table 11: Results obtained in the test period T3

the greater frequency of SPY events is associated with with greater MAE. However, in most cases, the MAE falls below the tick (one cent).

Given that in period $T3$, the number of events in the test is much greater than in the other periods (see Table 7), we comment in more detail prediction test $T3$ with $\delta = 1$. In the prediction made, 99.08% of ask prices $\hat{P}_a^{(3)sds}(t)$ and $P_a^{(3)sds}(t)$ have the same value. Similar result of 99.11% is found for bid prices $\hat{P}_b^{(3)sds}(t)$ and $P_b^{(3)sds}(t)$. In the case of SPY, 49.9% of the ask prices $\hat{P}_a^{(3)spy}(t)$ and $P_a^{(3)spy}(t)$ are the same; and 48.39% for the case of bid prices $\hat{P}_b^{(3)spy}(t)$ and $P_b^{(3)spy}(t)$. For the rest of SPY prices wrongly predicted in the test, 48.1% of them shows a difference, $|\hat{P}_a^{(3)spy}(t) - P_a^{(3)spy}(t)|$, about \$0.01; and 1.99% about \$0.02. There are no ask price differences greater than \$0.03. For SPY bid prices, the results are 49.3% for \$0.01, 2.34% for \$0.02, and again, there are no bid price differences greater than \$0.03 between predicted price and test price.

Based on the above results, we can conclude that the estimated 8-variate-HP models in the proposed periods have a good behavior when making predictions about price dynamics for the two assets.

4 VAR model

In this section, we are going to consider an N-dimensional Vector AutoRegressive Model (VAR). The VAR model is a widespread model introduced in [19] and customarily used in econometrics in order to describe the dynamic relationship among the different components of an economic time-series. The model is defined by a set of N processes $Y(t) \overset{\text{def}}{=} \{Y_i(t) \mid i : 1..N\}$ evolving in discrete time. In vector notation, the model is defined by the equation:

$$Y(t) = A_0 + \sum_{k=1}^{\ell} A_k Y(t-k) + \varepsilon(t) \tag{13}$$

where t is a discrete time, i.e., $t \in \{t_1, t_2, ..\}$. In (13), A_0 is a vector $(N \times 1)$ of *constant terms*; ℓ denotes the maximum number of lags; and, for each lag $k \in \{1, .., \ell\}$, A_k is a matrix $(N \times N)$ of *coefficients*. Finally, $\varepsilon(t)$ is a vector $(N \times 1)$ of innovations, i.e., non autocorrelated processes, $E(\varepsilon_t \varepsilon'_{t+h}) = \mathbf{0}_{N\times N}$ for $h \neq 0$, with covariance matrix $E(\varepsilon_t \varepsilon'_t) = \Sigma_\varepsilon$.

In the VAR model all the variables are treated symmetrically, being explained by the past of all variables taken together. The number of equations in the model coincides with the number of variables, and the lagged values of those variables in all the equations enter as explanatory variables. we use the VAR model as the standard benchmark with which to compare the performance of the Hawkes process estimated with our data, in particular, we want to compare their predictions for evolution of the bid and ask prices for the two assets SDS and SPY. Figure 2 shows the price movements for SDS and SPY in Nasdaq. The VAR model allows us to analyze jointly the bid and ask prices as a whole system. We analyze how the two quotes prices for SDS and SPY interact with each other in order to predict the future movement of the whole price series.

This section is organized as follows. In subsection 4.1, we explain the process of parameter selection and variable definition for our VAR model. Subsection 4.2 is devoted to parameter estimation of the proposed VAR models. In subsection 4.3, we study the impulse response function of estimated VAR models. Finally, subsection 4.4 presents the procedure of prediction via VAR model and the results obtained in the test periods. In this section, we have mainly used the Matlab Econometrics Toolbox[9].

4.1 VAR model selection

In general, VAR models assume that the variables $y_i(t)$ are *stationary.* Because the variables (prices) in our model are not stationary, we follow standard procedure and look at first differences, which removes trends in the price series. We apply the Augmented Dickey-Fuller test [20] to determine whether we could reject the null hypothesis of a unit root in the level of the quote prices. Table 12 shows an example of results of the ADF tests for the ask price of SDS. Former result illustrates that the original price process it is non-stationary; the next result obtained for the difference price process indicates that the transformed series is stationary.

```
Augmented Dickey-Fuller Test on "SDS ask" --------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -2.8321
Critical value 1% = -3.43
Critical value 5% = -2.862
Critical value 10% = -2.567 => P-Value = 0.0538.
Weak evidence to reject the Null Hypothesis. => Series is Non-Stationary.

Augmented Dickey-Fuller Test on "SDS ask" --------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level = 0.05
Test Statistic = -376.2572
Critical value 1% = -3.43
Critical value 5% = -2.862
Critical value 10% = -2.567 => P-Value = 0.0.
Rejecting Null Hypothesis. => Series is Stationary.
```

Table 12: ADF tests for the original ask price of SDS and the first difference of prices respectively for the period (10:00-12:00). The rest of tests for other periods and prices have shown the same results.

Our VAR model is then fitted on the first differences in prices. Using the following notation for prices: $P_a^x(t)$, $P_b^x(t)$ with $x \in \{sds, spy\}$, we use $\Delta P_a^x(t)$, $\Delta P_b^x(t)$ to denote their first differences. We propose the following VAR model:

[9] https://es.mathworks.com/products/econometrics.html

$$
\begin{aligned}
\Delta P_a^{sds}(t) &= c_1 + \sum_{i=1}^{\ell} \beta_1^i \Delta P_a^{sds}(t-i) + \sum_{i=1}^{\ell} \beta_2^i \Delta P_b^{sds}(t-i) + \\
&+ \sum_{i=1}^{\ell} \beta_3^i \Delta P_a^{spy}(t-i) + \sum_{i=1}^{\ell} \beta_4^i \Delta P_b^{spy}(t-i) + \varepsilon_1(t) \\
\\
\Delta P_b^{sds}(t) &= c_2 + \sum_{i=1}^{\ell} \beta_5^i \Delta P_a^{sds}(t-i) + \sum_{i=1}^{\ell} \beta_6^i \Delta P_b^{sds}(t-i) + \\
&+ \sum_{i=1}^{\ell} \beta_7^i \Delta P_a^{spy}(t-i) + \sum_{i=1}^{\ell} \beta_8^i \Delta P_b^{spy}(t-i) + \varepsilon_2(t) \\
\\
\Delta P_a^{spy}(t) &= c_3 + \sum_{i=1}^{\ell} \beta_9^i \Delta P_a^{sds}(t-i) + \sum_{i=1}^{\ell} \beta_{10}^i \Delta P_b^{sds}(t-i) + \\
&+ \sum_{i=1}^{\ell} \beta_{11}^i \Delta P_a^{spy}(t-i) + \sum_{i=1}^{\ell} \beta_{12}^i \Delta P_b^{spy}(t-i) + \varepsilon_3(t) \\
\\
\Delta P_b^{spy}(t) &= c_4 + \sum_{i=1}^{\ell} \beta_{13}^i \Delta P_a^{sds}(t-i) + \sum_{i=1}^{\ell} \beta_{14}^i \Delta P_b^{sds}(t-i) + \\
&+ \sum_{i=1}^{\ell} \beta_{15}^i \Delta P_a^{spy}(t-i) + \sum_{i=1}^{\ell} \beta_{16}^i \Delta P_b^{spy}(t-i) + \varepsilon_4(t)
\end{aligned}
\tag{14}
$$

We denote by $\text{VAR}^{(r)}(\ell, \theta^r)$ the VAR model proposed in the previous equations (14) for each estimation period. The superscript r identifies that estimation period T_r (see Table 6). The parameters of each VAR model, θ^r comprises the coefficients $\theta^r = [c_j, \beta_m^i]$ with ranges defined in equation (14), $j : 1..4$, $i : 1..\ell$, and $m : 1..16$. Hence, we are going to construct three VAR models, one for each estimation period. The next step, prior to model estimation, consists on an adequate selection of the number of lags, the parameter ℓ.

Lag order selection. One of the most common contrasts in a VAR model is the one related to the number of lags (ℓ) to be included as explanatory variables. Let us observe that in each equation there is a block of delays of all dependent variables. We have already seen that the number of parameters to estimate increases quickly with the number of variables in the model. This makes that the VAR model quickly become over-parameterized, making reliable estimations difficult and preventing its adoption as a prediction tool in high dimensional settings. Several authors have sought to address this issue by incorporating regularized approaches, e.g, the lasso regularization method [21]. This regularization imposes sparse or low-rank structures on the estimated coefficient of the VAR. The traditional approach, which we use in this study, attempts to eliminate the over-parametrization by selecting a universal low lag order for all components, based on the assumption that dynamic dependencies among components are short-range [22]. This strategy allows us to remove the autocorrelation of the error terms. We propose to use the Bayesian Information Criterion (BIC) that was developed by Schwarz [23]. The use of information criteria as a model selection method became popular from the work of Akaike [26]. BIC is a penalized-likelihood criteria, it penalizes model complexity more heavily, i.e., it provides a lesser number of lags, than other information criteria proposed in the literature. The number of lags proposed by BIC resolves the trade-off between the elimination of autocorrelation of the residu-

	— Lag order selection —		
Information criteria	T1	T2	T3
Akaike Information Criterion	35	29	35
Akaike's Final Prediction Error	35	29	35
Hannan-Quinn Criterion	25	19	35
Bayesian Information Criterion	11	11	24

Table 13: Information criteria for lag order selection

als and model complexity. Let us observe that keeping the order of complexity of the model as low as possible favors a lower computational time-cost for adjusting the VAR model's coefficients.

The BIC equation for each period is $BIC^r = |\theta^r| \cdot log(K^r) - 2log(\hat{\mathcal{L}}^r)$, where K^r is the number of observations in T_r, $|\theta^r|$ is the number of model coefficients, and $\hat{\mathcal{L}}^r$ is the maximized value of the likelihood function of the model VAR$^r(\ell, \theta^r)$. We have used the `statsmodels v0.12.0.dev0` Python library to compute the BIC. In our proposed models the number of lags selected for all periods is $\ell = 11$. Table 13 shows other information criteria.

4.2 VAR parameter estimation

Equations in (14) define the structure of the VAR models proposed in this study with $\ell = 11$ lags. The VAR system of equations can also be expressed in matrix form. For notation brevity, we consider the general VAR equation (13) from which our models are particular cases. Each equation in (13) can be rewritten as:

$$y_i(t) = a_i^0 + \sum_{k=1}^{11} \sum_{j=1}^{4} a_{ij}^k \cdot y_j(t-k) + \varepsilon_i(t) \qquad i \in \{1, .., 4\} \tag{15}$$

where i is referred to the i-th price difference serie in reference to our original VAR in equation (14). By considering all discrete times $t \in \{t_1, t_2, .., t_K\}$[10]; for each $i \in \{1, .., 4\}$, all resulting set of equations $\{y_i(t_1), y_i(t_2), ..., y_i(t_K)\}$, allow us to form the vector $\mathbf{y}i = [y_i(t_1), y_i(t_2), ..., y_i(t_K)]'$ and to produce a compact matrix form for all resulting set of equations for i:

$$\mathbf{y}i = \mathbf{X}i \cdot \pi i + \varepsilon i \tag{16}$$

where $\mathbf{X}i$ is the matrix of the variables at each lag, πi is the matrix of all coefficients and εi is the vector of innovations. Let us observe that equation (16) does not have 'time' as free variable because we have used the history of the process to characterze the process as described in the equations. In this system of equations the only free variables are the coefficients that we want to estimate, which we do using Ordinary Least Squares (OLS). Basically,

$$\hat{\pi} i = ((\mathbf{X}i)' \cdot \mathbf{X}i)^{-1} \mathbf{X}i' \mathbf{y}i \tag{17}$$

We will use Ordinary Least Squares (OLS) to estimate the unknown parameters of the VAR models in the three estimation periods. For each VAR$^r(\ell, \theta^r)$, the number of parameters to be estimated is 180. In order to study the consistency of the VAR models, we consider the innovations terms. If the model is fitted correctly, then innovations should be noise processes uncorrelated with the explanatory variables. The absence of autocorrelation in the innovation

[10] We abuse the notation and we do not write the letter r that is associated in the text with the period of estimation T_r. We think it is clear that discrete time values for t are in the range $[\tau_r, \tau_r']$ for each estimation period.

terms in all equations is thus a relevant property of the estimated parameters. In order to check this we use the Ljung–Box test and the ACF.

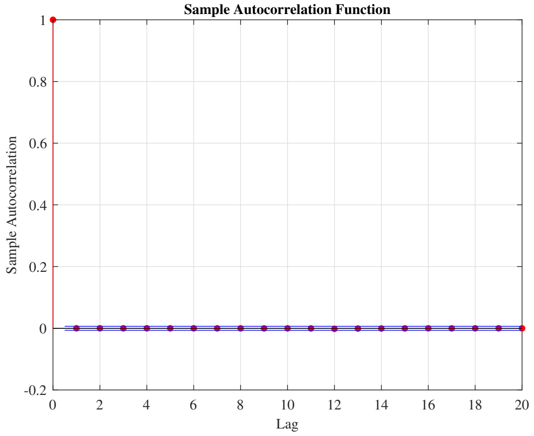

Figure 4: ACF for estimation SDS in period $T2$

Figure 4 shows the ACF graph of innovation $\varepsilon_2(t)$ for the estimation period $T2$ (see equation (14)). Relevant delays do not have structure; however, to make sure of this fact, we apply the Ljung–Box test [25] to check the non-autocorrelation in the innovations. The test results also indicate that there is no autocorrelation in innovations. Therefore, we have confidence that the model is well estimated. Once the model is estimated, we could proceed to exclude explanatory variables based on their statistical significance. However, there are reasons not to do it. In particular, if the same set of explanatory variables is maintained in all equations, then Ordinary Least Squares is an efficient estimator.

4.3 Impulse response function

After parameter estimation for the VAR models, $\text{VAR}^{(r)}(11, \hat{\theta}^r)$ with $r \in \{T1, T2, T3\}$, we turn to study the General Impulse Response Function (GIRF) of each VAR model. GIRF is used to describe how time series react to exogenous impulses or shocks [27] [28]. Let us denote the Vector Moving Average (VMA) representation of the VAR model in equation (13) as

$$\mathbf{Y}_t = \mathbf{c}_t + \sum_{h=0}^{\infty} B_h \varepsilon_{t-h} \tag{18}$$

where $\mathbf{c}$ is the VMA constant vector, and $B_0, B_1, ...$ are the VMA coeffcient matrices with $B_0 = \mathbf{I}_4$. The GIRF is defined as,

$$\text{GIRF}(h, \delta_j, \Omega_{t-1}) = E(\mathbf{Y}_{t+h} | \varepsilon_{j,t} = \delta_j, \Omega_{t-1}) - E(\mathbf{Y}_{t+h}, \Omega_{t-1}) \tag{19}$$

where δ_j is the size of the shock from variable j and $\Omega_{t-1} = \{\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, ...\}$ is the total information set known at time $t-1$. By setting $\delta_j = \sqrt{\sigma_{jj}}$ then the size of the shock is one standard deviation. Thus, the scaled GIRF is [27][28],

$$\text{GIRF}(h, \delta_j = \sqrt{\sigma_{jj}}, \Omega_{t-1}) = B_h \Sigma_\varepsilon e_j (\sigma_{jj})^{-\frac{1}{2}} \tag{20}$$

The GIRF is interpreted as the effect of a one-standard-deviation shock in the innovation (δ_j) from variable j at a given time horizon H. That shock causes significant increases (decreases)

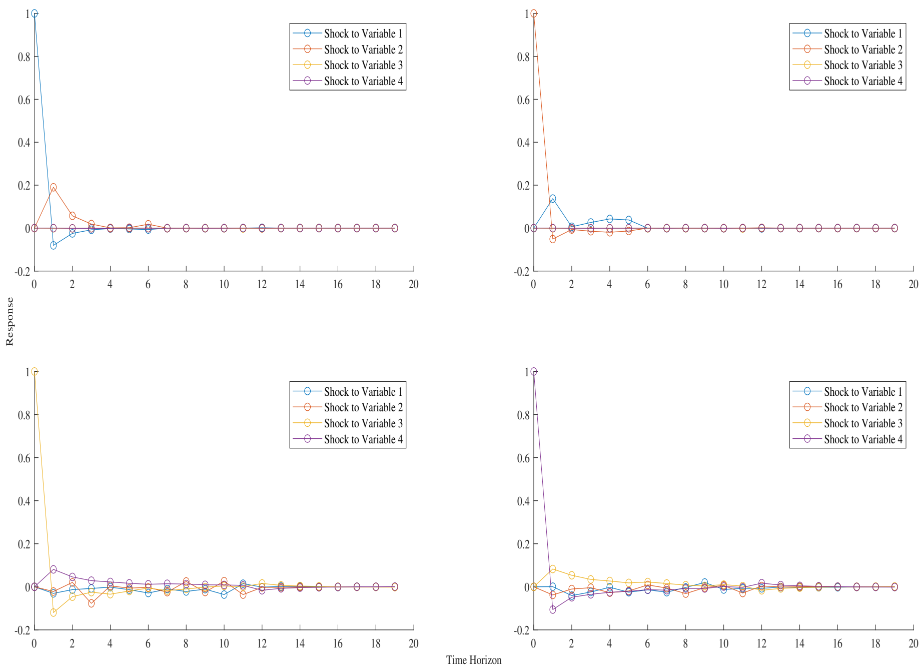

Figure 5: Plot of GIRF for the estimated model at period T1.

in the others variable for a time horizon after which its effect dissipates. Let us consider the estimated VAR models $VAR^1(11, \hat{\theta}^1)$ and $VAR^2(11, \hat{\theta}^2)$ for the periods $T1$ and $T2$ respectively. Each variable shock only affects the opposite side of the asset (ask or bid) to which the shock refers, e.g., if we generate a one standard deviation shock to the innovation error of $\Delta P_a^{(r)sds}$ at time 0, then this shock affects positively to $\Delta P_b^{(r)sds}$ at $h = 2$. The study of the possible cases provide us with information about the relationships between the behaviour on both sides of the order book for both the assets. This indicates that the market is keeping the spread for each asset constant and it is not moving away from its tick size (which for these two assets is one cent). These effects are shown in Figure 5. This Figure contains several lines, one for a shock in one of the variables: the GIRF from applying a one-standard-deviation shock on variable j at time 0, and its effect on all variables in the system over the forecast horizon, in this case $H = 20$.

Let us consider the estimated VAR model $VAR^3(11, \hat{\theta}^3)$ for the period $T3$. We observe in Figure 6 that a shock in the fourth variable, $\Delta P_b^{(3)spy}$, affects all other variables since it is also transmitted to the other variables of the system. In Figure 6, an increase in innovation of $\Delta P_b^{(3)spy}$ is followed by an increase in $\Delta P_a^{(3)spy}$. However, it also goes up $\Delta P_a^{(3)sds}$ at $h = 2$ when $\Delta P_b^{(3)spy}$ decreases but $\Delta P_a^{(3)sds}$ goes down very quickly $h = 3$ and finally, it returns to its normal state. In a sense, the behavior described by the GIRF reflects the fact that SPY aims to track the return of the SPX, and the SDS, the twice inverse of daily performance of SPX.

4.4 Prediction via VAR model

As we said in subsection 3.4, our main goal is to predict the price movements for SDS and SPY. In this subsection the prediction of price movements is done using an estimated VAR model for a given time period. Let $T_r = [\tau_r, \tau_r']$ be the period under study. In this period, we know the

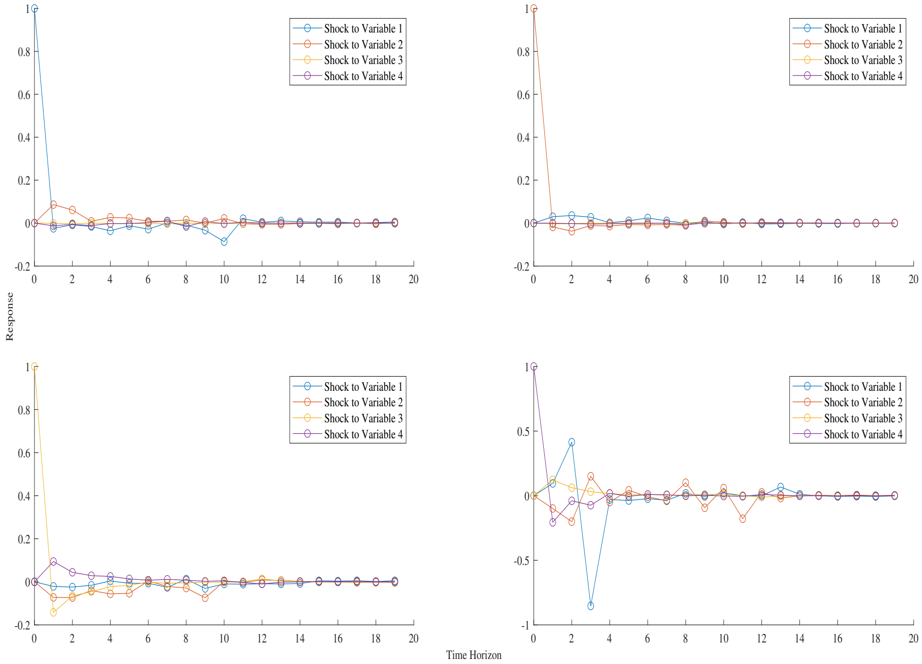

Figure 6: Plot of GIRF for the estimated model at period T3.

series $\Delta P_y^{(r)x}(t)$ and $P_y^{(r)x}(t)$ where $x \in \{sds, spy\}$, $y \in \{a, b\}$, and $t \in [\tau_r, \tau_r']$. For the period T_r, the extended period for the prediction is the time range $(\tau_r', \tau_r'']$. In this extended period the differences of prices and the prices are also available: $\Delta P_y^{(r)x}(t)$ and $P_y^{(r)x}(t)$ for $t \in (\tau_r', \tau_r'']$. The details of those time ranges are shown in Table 6. The procedure for making predictions is as follows.

1. For the considered period $T_r = [\tau_r, \tau_r']$, given $\Delta P_y^{(r)x}(t)$ for $t \in [\tau_r, \tau_r']$, we estimate the VAR model for that period, $\text{VAR}^r(11, \hat{\theta}^r)$ using OLS as we have indicated in subsection 4.4.

2. Given the estimated VAR model $\text{VAR}^r(11, \hat{\theta}_r)$, by equation (14) with innovations setting to value 0, we proceed to predict $\Delta \hat{P}_y^{(r)x}(t)$ with t within the period test $(\tau_r', \tau_r'']$. The initial values for starting the prediction are the values $\Delta P_y^{(r)x}(\tau_r' - 11)$ to $\Delta P_y^{(r)x}(\tau_r')$ where time is in milliseconds.

3. Iteratively compute $\Delta \hat{P}_y^{(r)x}(\tau_r' + h)$ via $\text{VAR}^r(11, \hat{\theta}^r)$ where $h \in \{1, 2, ..., H\}$, with $\tau_r' + H = \tau_r''$. That is, the iteration ends when τ_r'' is reached.

4. Finally, we reconstruct the predicted prices by reverting the differences of prices, $\hat{P}_y^{(r)x}(t)$ with $t \in (\tau_r', \tau_r'']$.

As we can see, the prediction procedure does not make use of the information available on prices or their differences that are available in the test period. Each new prediction point is based exclusively on the values of the previous predictions and their delays. In all cases the predictions are made approximately during the half hour after the estimation period. We consider it to be a long enough time range to adequately evaluate the predictability of the estimated VAR models in each period.

	$T1$	$T2$	$T3$
$\text{MAE}_a^{(r)sds}$	0.0000	0.0000	0.0000
$\text{MAE}_b^{(r)sds}$	0.0000	0.0000	0.0000
$\text{MAE}_a^{(r)spy}$	0.0003	0.0003	0.0010
$\text{MAE}_b^{(r)spy}$	0.0003	0.0003	0.0010

Table 14: MAE values for the three test periods.

In order to evaluate prediction accuracy, we use the Mean Absolute Error,

$$\text{MAE}_y^{(r)x} = \frac{1}{H}\sum_{h=1}^{H} |\hat{P}_y^{(r)x}(\tau_r' + h) - P_y^{(r)x}(\tau_r' + h)| \tag{21}$$

where $x \in \{sds, spy\}$, $y \in \{a, b\}$, and $\tau_r' + h \in (\tau_r', \tau_r'']$ (test period for T_r) with $\tau_r' + H = \tau_r''$.

Table 14 shows the results obtained with the proposed prediction procedure. SDS prices are predicted without errors and the errors in SPY prices are below the tick size. In conclusion, the proposed VAR model has a very high predictive capacity as is evident from the results in Table 14. To graphically illustrate this fact, we show in the Figure 7 the predicted price charts (red lines) versus current (blue lines) in the third period. Figure 8 is a zoom of the Figure 7 for the SPY prices.

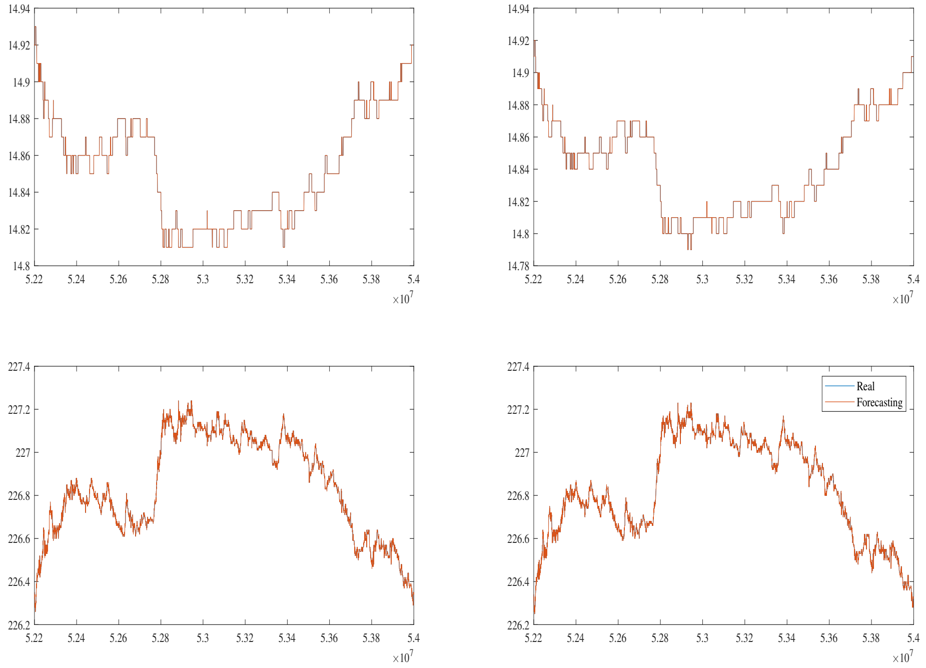

Figure 7: Prediction period T3. SDS and SPY prices are in top and bottom panels respectively. Red and blue lines are for predicted and current prices respectively

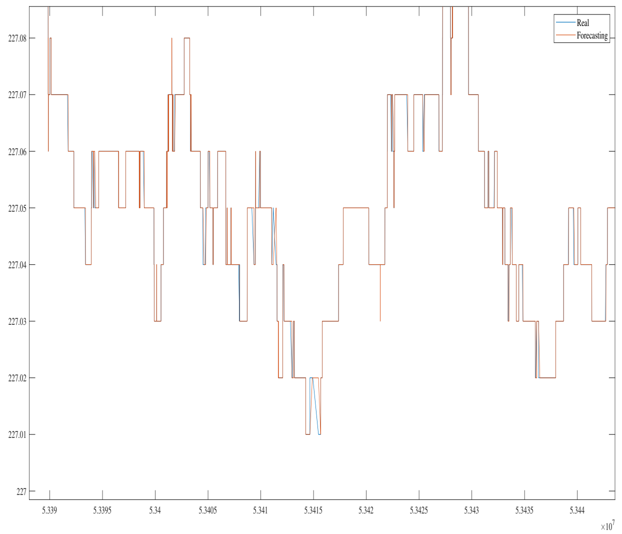

Figure 8: An enlargement of the graph in Figure 7 that corresponds to SPY bib prices $\hat{P}_b^{(3)spy}(t)$ (red line) and $P_b^{(3)spy}(t)$ (blue line)

5 Conclusions and future work

As we have indicated in the introductory section, the main objective of the current paper is to evaluate the appropriateness of multivariate Hawkes process for modelling and predicting the price movements in electronic exchanges. We have used the high frequency data provided in the TAQ for SDS and SPY assets in Nasdaq. To evaluate the performance of multivariate Hawkes process, we have compared prediction results to the standard benchmark, the VAR model.

We have used an 8-variate-HP for modeling the sequences of events that modify the state of the Order Book at Level I. The eight types of events (Table 5) correspond with transitions that modify the prices in the Order Book. In a first analysis, we have found sufficient evidence to justify the use of the 8-variate point process to described the Order Book dynamics. This is reflected in th QQ-plots (Figure 3) for the estimated 8-variate-HP (with exponential kernels (6)) as compared with a homogeneous Poisson process. After that, we have studied another kind of exponential kernel in order to (a) improve the parameter estimation, and (b) reduce the time computation of the estimator. The usage of exponential kernels with fixed decays (9) has shown to be adequate for these goals.

By using the exponential kernels with fixed decays, we have estimated an 8-variate-HP for the entire period of observation considered in this paper: 10:00 - 15:00. The kernel norm matrix obtained (Table 8) may be interpreted as an estimate of the Granger causal relationships between the different types of events. There are natural cross-exciting effects: SDS events trigger SPY events and viceversa; however, there are some types of events that display no Granger causal relationships (as concluded from the zeros in the kernel norm matrix). In fact, the absence of those relations may be interpreted in terms of the fundamental relationship between the SPY and SDS: both aim to mimick the S&P500 index index in different ways; the SPY aims to replicate the index returns, while the SDS tries to obtain a return equal to twice the inverse of the daily performance of the index. A point not note about the estimation is that SDS events appear to have an endogenous dynamic driven by SPY events, and these latter ones appear to be driven by events outside our analysis.

In order to evaluate the prediction capability of the proposed 8-variate-HP, we have considered three time ranges along the trading day for capturing different patterns of activity. For each period, we have estimated an 8-variate-HP as base model for making predictions. As our main objective is price prediction, we have proposed a price model based on the counting process that is generated by the sequence of events (see equation (11)). The fact that the current ask and bid price differences is about the tick size (one cent) allows us to consider the proposed price model very adjusted to the current data. Then, we have proposed a prediction procedure (subsection 3.4 that is based on a general prediction schema proposed by Vere-Jones [15]. Briefly, we use the algorithm of the 'next event simulation', which is based on the Ogatas's simulation algorithm [9], to predict the following δ next events before the δ real events occur in the test period. Tables 9, 10, and 11 show that the proposed prediction process is well defined and accurate when making those predictions.

At this point on the discussion, we add the following comment: the 8-variate-HP proposed in this paper provides an explanation for the price dynamics using the internal information about how the prices have been formed trough events happened along the traiding day. This is an important difference with other prediction models as the one used in this paper as a comparative benchmark: VAR model.

In order to compare the prediction results obtained with the 8-variate-HP, we have defined a VAR model which, described briefly, (a) has four main variables, the ask and bid price first differences for SDS and SPY, for modeling the sequences of price differences at discrete times; and (b) its equations are formed by considering 11 lags which have been selected acording to BIC [23]. VAR models have been estimated for the different periods under consideration using standard OLS. In a similar way as we have done with the Hawkes process, we have used the estimated VAR models for predicting the prices in the test periods (subsection 4.4). Table 14 shows prediction results using the VAR model. From that table, we conclude that the proposed VAR models have better performance than the 8-variate-HPs for predicting SPY prices and for SDS prices, the results are very close in both models. Although the VAR model undoubtedly has a greater ability to predict the dynamics of prices, it is of less use as an explanatory tool to describe the way in which prices are formed since it only model the price signals and their lags over time.

As future works, we propose the following: 1) complete the price prediction study proposed in this work for the different exchanges indicated in Table 1, 2) apply other Hawkes process models proposed in the literature, specifically the model of regimes introduced by Vinkovskaya [12] and state-dependent Hawkes process proposed by Morariu-Patrichi and Pakkanen [13]. Both models are an extension of the Hawkes process in which system state is included in the Hawkess process to refine the type of events. In particular, we propose to include the imbalance measure [6] as a dependent state in the multivariate Hawkes process due to its good properties for price prediction.

References

[1] J. Brogaard, et. al.. Price discovery without trading: evidence from Limit Orders. The Journal of Finance, LXXIV(4): 1621–1657, 2019.

[2] Á. Cartea, S. Jaimungal, J. Penalva. Algorithm and High-Frequency Trading. Cambridge University Press, 2015.

[3] A. Fulop, et. al.. Self-Exciting jumps, learning, and asset pricing implications. The Review of Financial Studies, 28(3): 876–912, 2015.

[4] A.G. Hawkes. Spectra of some mutually exciting point processes. Biometrika 58(1), 83–90, 1971.

[5] F. Abergel, et. al. Limit Order Books. Cambridge University Press, 2016.

[6] Á. Cartea, Donnelly and S. Jaimungal. Enhancing trading strategies with order book signals. Applied Mathematical Finance, 25(1):1–35, 2018.

[7] Y. Chen. Modelling Limit Order Books dynamics using Hawkes processes. PhD Thesis, Florida State University, 2017.

[8] M. Krumin, et. al.. Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. Frontiers in Computational Neuroscience, 4 (article 147), 2019.

[9] Y. Ogata. Statistical models for earthquakes occurrences and residual analysis for point processes. Journal of the American Statistical Association 83(401):9—27, 1988.

[10] J. Etesami, et. al.. Learning Network of Multivariante Hawkes Processes: A time series approach. arXiv:1603.04319v1, 2016.

[11] Á. Cartea, et. al.. Algorithmic Trading, Stochastic Control, and Mutually Exciting Processes. SIAM Review, 60(3), 673–703, 2018.

[12] E. Vinkovskaya. A Point Process Model for the Dynamics of Limit Order Books. PhD Thesis, Columbia University, 2014.

[13] M. Morariu-Patrichi and M. S. Pakkanen. State-Dependent Hawkes processes and their application to limit order book modeling. arXiv:1809.08060v2, 2018.

[14] P. Laub, et. al.. Hawkes processes. arXiv:1507.02822v1, 2015.

[15] D. J. Daley and D. Vere-Jones. An Introduction to the Theory of Point Processes. Vol. I. Springer, New York, second edition, 2003.

[16] L. Massoulié. Stability results for a general class of interacting point processes dynamics, and applications. Stochastic Processes and their Applications, 75(1), 1–30, 1998.

[17] C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. Journal of Econometrics, 141(2):876–912, 2007.

[18] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. Annals of the Institute of Statistical Mathematics, 31(1):145–155, 1979.

[19] C. A. Sims. Macroeconomics and reality. Econometrica: Journal of the Econometric Society, 1–48, 1980.

[20] D. A. Dickey, W. A. Fuller. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74(366): 427-–431, 1979.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Socienty, Series B: Statistical Methodology, 58(1):267-–288, 1996.

[22] W. B. Nicholson, et. al.. High Dimensional Forecasting via Interpretable Vector Autoregression. arXiv:1412.5250, 2014.

[23] G. E. Schwarz. Estimating the dimension of a model. Annals of Statistics, 6(2): 461-–464, 1978.

[24] Alfonso Novales. Modelos vectoriales autoregresivos (VAR). Course Notes, Universidad Complutense, 2017.

[25] G. E. P. Box and D. A. Pierce. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. Journal of the American Statistical Association, 65: 1509–1526, 1970.

[26] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., International Symposium on Information Theory, 267–281, 1973.

[27] G. Koop et. al.. Impulse response analysis in nonlinear multivariate models. Journal of Econometrics, 74(1):119—147, 1996.

[28] H. Pesaran and Y. Shin. Generalized impulse response analysis in linear multivariate models. Economics Letters, 58(1):17-–29, 1998.

A Order flow reconstruction

For the goal of reconstructing the precise order flow, i.e. the complete sequence of MOs, LOs and COs along with their price and volume that lead to the generation of given trades and quotes tables, is necessary an algorithm for the determination of that order flow. The algorithm would be straightforward if the database were perfectly coherent, but inconsistencies exist, mainly because of the non-synchronization of the timestamps of the trades and quotes files, and maybe because of probable inaccuracies in the data building process. In addition, in our case, we only have the information at Leve I of the book. The procedure of reconstruction basically consists on scan line by line the quotes tables. In the following, we explain the procedure with a simple example:

Timestamp	Price	Size
34524356	227.71	60
34524357	227.70	1
34524360	227.70	1
34524374	227.70	2
34524376	227.70	1
34524376	227.71	60

Table 15: An extract of the Quotes table of ask side for SPY in Nasdaq

We start at the first entry on the quotes table (see Table 15). At 34524356 ms, there is an ask price of 227.71 and a quantity of 60. In the next entry, at 34524357 ms, we observe how the price is 227.70 and a quantity of 1 unit. This is clearly explained by the arrival of a LO that affects the price (LO_{pa}) at that timestamp. The quote table does not change at the next time. At 34524374 ms a new unit appears over the established quantity, this is because a new LO has entered that adds a unit at the same price (LO_{qa}). Two ms later a unit has disappeared, due to a CO (cancellation) or MO (trade). At the last entry, a possible CO or MO returns to the initial state. Situations without any change are identified as *unchanged* (U) in the order flow. They corresponds with hidden orders unknown al Leve I.

Order Type	Timestamp	Price	Size
LOpa	34524357	227.70	1
U	34524360	0	0
LOqa	34524374	227.70	1
COa	34524376	227.70	1
COa	34524376	227.70	1
U	34524376	0	0

Table 16: Order flow reconstruction for the example in Table 15

The second step of the order flow reconstruction deals with the parsing of the trades table. Each entry on the trade table should represent a MO, however we do not know if it is a sell or buy MO. In a perfect case, at the exact same timestamp, we should see in the quotes tables of ask or bid a decrease of the quantity at the same price and with the same volume. In other words, each line of the trades table should match a CO of the same volume, at the same price and timestamp in the order flow. However, it is very rare to exactly match a corresponding MO read in the trades table to a CO read in the quotes tables. Therefore, the approximation to identify a MO consists on looking for not only at the exact MO timestamp, but within $-\delta$ and $+\delta$ milliseconds around this MO timestamp in both quotes tables. For the possible CO candidates, we use a scoring method to search for the maximum score CO candidates. We may identify up to 20% of MOs with maximum score. In Table 17, we show the number of types of orders without considering MOs. There are 259 MOs in Nasdaq for SDS (0.123% of the total number of orders) and 14365 MOs (0.530%) in the Nasdaq for SPY. Thus, we do not consider relevant MOs identification. This is a common practice in other papers using Order Books at Level I [12]. Let us also observe in Table 17 an important difference in activity at SDS and SPY.

SDS	ask	bid	SPY	ask	bid
U	52532	56012	U	753664	860882
LOq	22220	20376	LOq	302169	243419
LOp	306	352	LOp	15624	17008
COq	24116	23045	COq	275895	243965
COp	321	335	COp	15556	16691

Table 17: Number of types of orders in the considered period 10:00-15:00 at Nasdaq

B Computation of the intensity function

Let us consider a D-dimensional point process $\langle (T_k, E_k) \rangle_{k \in \mathbb{N}}$ where $E_k \in \mathcal{E}$, $\mathcal{E} = \{1, 2, .., D\}$. Let us assume that the point process is modelling via a D-variate-HP with exponential kernels of fixed decays defined as in eq. (9),

$$\phi_{ij}(t) = \sum_{u=1}^{U} \alpha_{ij}^u \beta^u \exp(-\beta^u t) \qquad i, j \in \mathcal{E}$$

By (4) and (5), the conditional intensity is defined by

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{\{t_k^j < t\}} \phi_{ij}(t - t_k^j) \qquad i \in \mathcal{E}$$

where t_k^j denotes the time of the k-th event of type j in the given realization $\mathcal{H}(t) = \langle (t_k^j, j_k) \rangle_{1 \leq k \leq K}$ of the D-dimensional point process up to time t.

Thus, for each $i \in \mathcal{E}$,

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} S_{ij}(t) \text{ with}$$

$$S_{ij}(t) = \sum_{\{t_k^j < t\}} \phi_{ij}(t - t_k^j) = \sum_{\{t_k^j < t\}} \sum_{u=1}^{U} \kappa_{ij}^u \exp(-\beta^u (t - t_k^j))$$

$$S_{ij}(t) = \sum_{u=1}^{U} \kappa_{ij}^u \sum_{\{t_{r_j} < t\}} \exp(-\beta^u (t - t_{r_j})) \qquad j \in \mathcal{E}$$

where $\kappa_{ij}^u = \alpha_{ij}^u \beta^u$ for any $i, j \in \mathcal{E}$ and $u \in \mathcal{U}$, $\mathcal{U} = \{1, .., U\}$. In the previous equation r_j denotes the index in $\mathcal{H}(t)$ of the r-th apparition of an event of type $j \in \mathcal{E}$. For a fixed $j \in \mathcal{E}$, $\{t_k^j < t\} = \{t_{r_j} < t\}$ and the previous equalities hold.

Then, for each $i, j \in \mathcal{E}$,

$$S_{ij}(t) = \sum_{u=1}^{U} \kappa_{ij}^{u} B_j^u(t) \text{ with}$$

$$B_j^u(t) = \sum_{\{t_{r_j} < t\}} \exp(-\beta^u(t - t_{r_j})) \qquad u \in \mathcal{U}$$

We provide the computation of $B_j^u(t)$ for each $j \in \mathcal{E}$ and $u \in \mathcal{U}$:

- By convention, if $\{t_{r_j} < t\} = \emptyset$ then $B_j^u(t) = 0$, in particular $B_j^u(0) = 0$; otherwise,
- $B_j^u(t) = \exp(-\beta^u(t - t_{1_j}))$ if $t_{1_j} < t \leq t_{2_j}$; and, in general,
- $B_j^u(t) = \sum_{r=1_j}^{f_j} \exp(-\beta^u(t - t_r))$ if $t_{1_j} < t_{2_j} < ... < t_{f_j} < t \leq t_{(f+1)_j}$

In the following, we provide a recurrence for a fast computation of $B_j^u(t)$. For each $j \in \mathcal{E}$ and $u \in \mathcal{U}$:

$$B_j^u(t) = \sum_{r=1_j}^{f_j} \exp(-\beta^u(t - t_r)), \text{ as } t - t_r = t - t_{f_j} + t_{f_j} - t_r$$

$$B_j^u(t) = \exp(-\beta^u(t - t_{f_j})) \sum_{r=1_j}^{f_j} \exp(-\beta^u(t_{f_j} - t_r))$$

$$B_j^u(t) = \exp(-\beta^u(t - t_{f_j})) \cdot [1 + \sum_{r=1_j}^{(f-1)_j} \exp(-\beta^u(t_{f_j} - t_r))], \text{ since } t_{f_j} - t_{f_j} = 0$$

Define, $A_j^u(f_j) = \sum_{r=1_j}^{(f-1)_j} \exp(-\beta^u(t_{f_j} - t_r))$, with $A_j^u(1_j) = 0$

Thus, for $f_j \geq 2_j$, as $t_{f_j} - t_r = t_{f_j} - t_{(f-1)_j} + t_{(f-1)_j} - t_r$

$$A_j^u(f_j) = \exp(-\beta^u(t_{f_j} - t_{(f-1)_j})) \sum_{r=1_j}^{(f-1)_j} \exp(-\beta^u(t_{(f-1)_j} - t_r))$$

$$A_j^u(f_j) = \exp(-\beta^u(t_{f_j} - t_{(f-1)_j})) \cdot [1 + \sum_{r=1_j}^{(f-2)_j} \exp(-\beta^u(t_{(f-1)_j} - t_r))]$$

Therefore, $A_j^u(f_j) = \exp(-\beta^u(t_{f_j} - t_{(f-1)_j})) \cdot (1 + A_j^u((f-1)_j))$

Below we present a summary of the intensity function computation:

For each $i \in \mathcal{E}$, $\quad \lambda_i(t) = \mu_i + \sum_{j=1}^{D} S_{ij}(t)$

For each $i, j \in \mathcal{E}$, $\quad S_{ij}(t) = \sum_{u=1}^{U} \kappa_{ij}^{u} B_j^u(t)$

For each $j \in \mathcal{E}, u \in \mathcal{U}$,

$$\begin{array}{ll} B_j^u(t) = 0 & \text{if } t \leq t_{1_j} \\ B_j^u(t) = \exp(-\beta^u(t - t_{f_j}))\ (1 + A_j^u(f_j)) & \text{if } t_{1_j} < ... < t_{f_j} < t \end{array}$$

where

$$\begin{array}{l} A_j^u(1_j) = 0, \text{and, otherwise} \\ A_j^u(f_j) = \exp(-\beta^u(t_{f_j} - t_{(f-1)_j})) \cdot (1 + A_j^u((f-1)_j)),\ \text{with, } 2_j \leq f_j \end{array}$$

C Simulation algorithm of the Next Event in a D-variate Hawkes Process

Let us consider an estimated D-variate Hawkes process with exponential kernel of fixed decays. Estimated parameters of the process, $\hat{\theta} = [[\hat{\alpha}^u_{ij}\hat{\beta}^u]_{i,j\in\mathcal{E},u\in\mathcal{U}}, [\hat{\beta}^u]_{u\in\mathcal{U}}, [\hat{\mu}_i]_{i\in\mathcal{E}}]$, have been learned for a given realization $\mathcal{H}(t) = \langle(t^j_k, j_k)\rangle_{1\leq k\leq K}$ in the period $[t_1, t_K]$, where t_K is the time of the last event in $\mathcal{H}(t)$. Let $\mathbf{A}(t_k)$ be the matrix $\mathbf{A}(t_k) = [A^u_j(t_{f_j} \leq t_k)]$. This matrix is needed to the intensity function computation as we have indicated in Appendix B. The matrix is calculated at the times t_k (event occurrence times) in $\mathcal{H}(t)$. Let us recall that $A^u_j(t_{1_j}) = 0$ for each $j \in \mathcal{E}$. Thus, we assume that $\mathcal{H}(t)$ has a sufficient number of events of each type in order to obtain non null values for $\mathbf{A}$.

The simulation of a D-variate Hawkes process can be performed by thinning as proposed by Ogata [**?**][9]. The Ogata's simulation algorithm can be modified to simulate the process in an event-by-event fashion. This is convenient in practice, for example, when our target is to simulate a given number of events rather than all the events from a Hawkes process over a specific interval. We assume that $\mathcal{H}(t) = \langle(t^j_k, j_k)\rangle_{1\leq k\leq K}$ has been generated by the simulation algorithm up to t_K, but the algorithm has not finished yet. Then, we obtain the next event as an additional iteration of the simulation algorithm. Simulation algorithm of the Next Event is detailed in Algorithm 1.

Algorithm 1 can be used as a procedure in order to generate a simulation of δ next events. We provide an algorithm to simulate the increment of the counting process $\mathbf{N}(t) = (N_1(t), ..., N_D(t))$, $t \in [t_1, t_K]$, from the last instant of time t_K in history $\mathcal{H}(t)$. This Algorithm 2 is applied in subsection 3.4 for predicting via Hawkess process.

Algorithm 1 Simulation algorithm of the Next Event in a D-variate Hawkes Process

```
Input: $\hat{\theta} = [[\hat{\alpha}^u_{ij} \cdot \hat{\beta}^u]_{i,j \in \mathcal{E}, u \in \mathcal{U}}, [\hat{\beta}^u]_{u \in \mathcal{U}}, [\hat{\mu}_i]_{i \in \mathcal{E}}], t_K, \mathbf{A}(t_K)$

Output: $(s, e)$ the next event is of type $e$, at time $s$

$s := t_K + 0.0001$;                      ▷ Initialize $s$ with the last time $t_K$ plus 0.0001 sg.
                                          ▷ to take into account in the intensity the occurrence of the last event at $t_K$
Calculate $[\hat{\lambda}_i(s)]_{i \in \mathcal{E}}$ using $\mathbf{A}(t_K)$ and $\hat{\theta}$;      ▷ see Appendix B
$exit := false$;
while not $exit$ do
    Set $\bar{\lambda} := \sum_{i=1}^{D} \hat{\lambda}_i(s)$;
    Generate $p \sim$ uniform$(0,1)$;
    Let $\omega := -\frac{\mathrm{Ln}(p)}{\bar{\lambda}}$;              ▷ thus, $\omega \sim$ exponential$(\bar{\lambda})$
    Set $s := s + \omega$;
    Generate $d \sim$ uniform$(0,1)$;
    Calculate $[\hat{\lambda}_i(s)]_{i \in \mathcal{E}}$ using $\mathbf{A}(t_K)$ and $\hat{\theta}$;
    if $d \cdot \bar{\lambda} \leq \sum_{i=1}^{D} \hat{\lambda}_i(s)$ then
        $e := 1$;                         ▷ search for the first $e$ such that $d \cdot \bar{\lambda} \leq \sum_{i=1}^{e} \hat{\lambda}_i(s)$
        while $d \cdot \bar{\lambda} > \sum_{i=1}^{e} \hat{\lambda}_i(s)$ do
            $e := e + 1$;
        fwhile
        $exit := true$;
    fif
fwhile
return $(s, e)$                           ▷ the next event is of type $e$, at time $s$

```

Algorithm 2 Simulation algorithm of δ Next Events in a D-variate Hawkes Process

```
**Input**: $\hat{\theta}$, $t_K$, $\mathbf{A}(t_K)$, $\mathbf{N}(t_K) = (N_1(t_K), .., N_D(t_K))$, $\delta$.

**Output**: $\Delta\mathbf{N} = \mathbf{N}(t_{K+\delta}) - \mathbf{N}(t_K)$    ▷ return the increment of the counting process after $\delta$ new events

$h := 0$;

**repeat**
    $(s, e) :=$ Algoritmo_1($\hat{\theta}$, $t_{K+h}$, $\mathbf{A}(t_{K+h})$);    ▷ simulate next event

    $t_{K+h+1} := s$;
    $N_e(t_{K+h+1}) = N_e(t_{K+h}) + 1$;    ▷ Update the counting process for the event of type $e$

    Update $\mathbf{A}(t_{K+h+1})$;

    $h := h + 1$;
**until** $h = \delta$

**return** $(\Delta\mathbf{N} = \mathbf{N}(t_{K+\delta}) - \mathbf{N}(t_K))$

```