Tilburg School of Economics and Management

Finance Department

Analysis of Order Clustering Using High Frequency Data: A Point Process Approach

F. Lorenzen - ANR. 188798

Supervisor: F.C.J.M. de Jong

August - 2012

Acknowledgements

The accomplishment of this work was made possible because of many people. First, I would like to thank my supervisor Frank de Jong, whose experience as an academic in Finance, helped me to turn a simple research idea into the present work. Without Frank's support this work would not exist.

I would also like to thank all second year research master students in Finance. With them I had the opportunity to share my views and ideas about several different topics. A special thanks to Andreas Rapp who helped me a lot while I was writing this work. I also thank Matjaž Maletič and Andreas for the endless discussions about Finance and Econometrics.

A very special thanks to my family that, apart from the distance, managed to help me, in many different situations, every time I needed them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The amount of transactions observed in financial markets has increased dramatically over the last decade. Not only has the volume traded in exchanges around the world seen unprecedent levels but also the frequency at which these transactions occur have increased (Lo and Wang, 2001 analyze several interesting time series properties related to the volume of trades in financial markets). The technological improvement experienced by exchanges and traders has brought trading activity to a new standard. This new standard is often referred to as High Frequency Trading (HFT), a new kind of trading strategy whose trademarks are low latency and high volume trading. HFT accounted for a relatively small amount of trading activity in equity markets during the beginning of 2000 but has nowadays grown to be the dominant force in these markets (Zhang, 2010; Hendershott, Jones and Menkveld, 2011).

The massive presence of HFT has a growing impact on microstructure aspects of financial markets. Recently, the SEC has proposed to conduct a review of the equity markets microstructure to check whether "*market structure rules have kept pace with, among other things, changes in trading technology and practices*", SEC (2010, p.1). The SEC (2010) release also reports how drastically the "speed" of financial markets has increased: in 2005 the average speed of execution for small orders in the NYSE was 10.1 seconds, this number was reduced to 0.7 seconds in 2009. At the same time the consolidate number of trades in the NYSE jumped from 2.9 millions trades in 2005 to 22.1 millions trades in 2009. These impressive figures have put financial markets into a new paradigm and set new challenges

to the academic literature in finance. Taking a very detailed look into the market one may wonder how to analyze the huge amount of noisy, fine-grained data that is now available.

Trading activity is not evenly distributed during the trading day. It is well known that trades occur more frequently during the beginning and the end of the trading day. Moreover, trades tend to occur in clusters. If one observes the intensity of trades during the day it is easy to see that there are short periods of very intense trading and periods of very low trading intensity. This makes the duration between trades irregular and poses a challenge to standard econometric techniques. One of the first attempts to model irregularly spaced data was the Autoregressive Conditional Duration (ACD) model of Engle and Russel (1998). The ACD assumes that arrival times (of trades, quotes or some other event) are stochastic variables that follow a point process.

Point processes are a class of stochastic processes in which one realization of the process is characterized by a point in time or in some other space. The simplest case of a point process is the homogenous Poisson process that describes the rate of arrival of new events using a constant rate μ. The Poisson process has found application in many different situations. It can be used to model the rate of radiation arriving in a Geiger counter or the rate of arrival of clients in a shop. A slightly more sophisticated point process is a non-homogenous Poisson process, where the rate of arrival can vary as a deterministic function of time, i.e. $\mu = \mu(t)$. The non-homogenous Poisson process is also widely applied to model arrival rates of aircrafts, containerships and telephone calls. Generally speaking, point processes are the mathematical foundation of many different theories like Renewal Theory, Reliability Theory and Queueing Theory to name a few. Nevertheless, the analysis of more complex random signals like that descendant from earthquakes or from the stock market requires models that

are more sophisticated than the simple (non-)homogenous Poisson process. Earthquakes and stock market data present endogenous clustering effects that cannot be captured by the simple Poisson process. The occurrence of an earthquake is usually followed by aftershocks and the arrival rate of buy or sell orders to the stock market usually occurs in bursts where one order is followed by many other orders. The ACD model is an early attempt to incorporate such stylized facts into a parsimonious model.

Another popular model that is able to reproduce some of the stylized facts related to stock market data, like the clustering of order arrivals, is the Hawkes' model (Hawkes, 1971). The Hawkes' model is a self-exciting point process that has found application in many different fields like seismology, neurophysiology, epidemiology and finance. The popularity of the Hawkes' model is explained by its ability to model clustering effects in a parsimonious way maintaining a linear representation for its conditional intensity (Daley and Vere-Jones, 2003). The term self-exciting stems from the fact that in the Hawkes' model events that arrive at a rate μ, which is possibly time-varying, can give rise to second order events, that in turn can give rise to third order events and so on. In this work we estimate a Hawkes' model using high-frequency stock market data on the durations of trades and quotes. Our work follows closely the estimation performed by Filimonov and Sornette (2012) (FS, hereafter) who estimated a Hawkes' process using data on the duration of mid-price changes of the E-mini S&P 500 contract. The main difference from FS (2012) is that we use data on equity markets and do not fully rely on the randomization of timestamps, as used by FS (2012). The randozimation of timestamps is necessary because the data used by FS (2012) has a timestamp that is rounded to the nearest second. Given the high-activity of markets is quite common to have several events within one second. In order to distinguish between those

events, FS (2012) make use of a random number in the interval $[0, 1)$ that is added to the original timestamp. Instead of imposing this strong assumption, we try to assess the impact of the randomization of timestamps on the estimates and the fit of the Hawkes' model.

In the next section we present a brief introduction to the mathematics of point processes. We then review the simulation and the estimation of Hawkes' process as well as some applications in finance. Our last step is to estimate the parameters and the fit of the process on stock data using a sample for the U.S. market and a sample for the European market.

Chapter 2

Point Processes

Let t_i be some random variable that satisfies $t_1 < t_2 < ... < t_N$. These variables may be used to identify the time epochs in which a given event occurs. We focus here on the unidimensional problem so that the variables t_i can be arranged on a line. We call the stochastic process defined by the variables t_i a point process. The counting process associated with the set of time epochs $\{t_i\}$ is an alternative description of the point process and is given by

$$N(t) = \sum_i 1_{t_i \leq t}. \tag{2.1}$$

Note that the definition above excludes the possibility of more than one event occurring at any time t_i. Let the duration between two consecutive events be defined as

$$\tau(t_i) = t_i - t_{i-1}. \tag{2.2}$$

The definitions so far are quite general and can be used to describe many different phenomena. In order to construct a model that explains the intra-event durations we have to add more structure to the point process. Consider then the counting function $N(t)$. A point process can be defined in terms of $N(t)$ using the equation below

$$\begin{aligned} P[N(t+h) - N(t) = 1] &= \lambda(t)h + o(h), \\ P[N(t+h) - N(t) > 1] &= o(h). \end{aligned} \tag{2.3}$$

Equation (2.3) reflects the fact that no more than one event occur at a single time t and that events occur with a time-varying intensity $\lambda(t)$. The researcher has the freedom to choose an

intensity that describes the data well. Note that $\lambda(t)$ can be taken either as a deterministic or a stochastic function of time.

A very simple case of a point process is the homogeneous Poisson process. It is a process such that the probability that one event occurs in the next (small) time interval h is proportional to a constant μ ($\lambda(t) = \mu$). The simple structure of the Poisson process makes it a very popular model. Nevertheless, this simplicity makes it unable to reproduce some of the stylized facts observed in the stock market, like the clustering of order arrivals. The main issue with the homogeneous Poisson process is that it is a process that has no memory, i.e. the intra-event duration does not depend on previous events and is thus i.i.d. If one wants to reproduce some of the stylized facts of order flows, like clustering of order arrivals, then some correlation structure must be incorporated into the Poisson process. The next section presents the Hawkes' model, which is a model that incorporates an additive correlation structure to the Poisson process.

2.1 Hawkes' process

The Hawkes' process is a point process that has a response function (or kernel) $h(t - t_i)$ which takes into account the influence of past events on the current conditional intensity. It was introduced by Hawkes (1971) and is a more general model than the Poisson process discussed before; it has the potential to explain some of the stylized facts related to quote and trade dynamics. As explained by Daley and Vere-Jones (2003, p.183), the Hawkes' process: "*comes closest to fulfilling, for point processes, the kind of role that the autoregressive model plays for conventional time series*".

The Hawkes' process is easily described in terms of its conditional intensity function, given

by

$$\lambda_t(t) = \mu(t) + \sum_{t_i<t} h(t-t_i), \tag{2.4}$$

where the functional form of the response function used in many applications (FS, 2012; Shek, 2011; Hewlett, 2006; and Bowsher, 2007) is exponential, $h(t-t_i) = \alpha e^{-\beta(t-t_i)}$, leading to a conditional intensity

$$\lambda_t(t) = \mu(t) + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}. \tag{2.5}$$

The first term in the conditional intensity is the "base" intensity of the model that determines the rate of arrival of first order[1] events per unit of time. The response function controls then how offsprings are generate by first order events and is the source of clustering in the model. As it will become clear later, the simple conditional intensity representation of the Hawkes' model given by Equation (2.5) and the fact that the model is described in event time (in contrast to wall-clock time) are two important advantages of the Hawkes' model over models that describe directly durations, such as the ACD model of Engle and Russel (1998). To see how the Hawkes' process resembles an autoregressive model we rewrite the conditional intensity as

$$\lambda(t) - \mu(t) = \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}. \tag{2.6}$$

Consider now the intensity of the process at some given epoch t_i that is in the past with respect to t. This intensity can be written as

$$\lambda(t_i) - \mu(t_i) = \sum_{t_k<t_i} \alpha e^{-\beta(t_i-t_k)}. \tag{2.7}$$

[1] In earthquake terminology, the events of first order are named main events while the second order events are named aftershocks. Immigrants for the first order events and descendants or offspring for the second order events are also a common terminology.

If we multiply both sides of the previous equation by $\exp\left[-\beta\left(t-t_i\right)\right]$ we get

$$\left[\lambda\left(t_i\right)-\mu\left(t_i\right)\right]e^{-\beta(t-t_i)}=\sum_{t_k<t_i}\alpha e^{-\beta(t-t_k)}, \tag{2.8}$$

where t_k represents all events that occurred before the event t_i that is itself in the past with respect to t. Now note that the response function $\sum_{t_i<t}\alpha e^{-\beta(t-t_i)}$ can be decomposed as

$$\sum_{t_i<t}\alpha e^{-\beta(t-t_i)}=\sum_{t_k<t_i}\alpha e^{-\beta\left(t-t_k\right)}+\sum_{t_k>t_i<t}\alpha e^{-\beta(t-t_k)}. \tag{2.9}$$

If we now combine the last two equations we can write $\lambda\left(t\right)-\mu\left(t\right)$ as

$$\lambda\left(t\right)-\mu\left(t\right)=\left[\lambda\left(t_i\right)-\mu\left(t_i\right)\right]e^{-\beta(t-t_i)}+\sum_{t_k>t_i}\alpha e^{-\beta(t-t_k)}. \tag{2.10}$$

Equation (2.10) resembles the continuous time form of an autoregressive model, given by,

$$X_t-\mu=e^{-\beta(t-s)}\left(X_s-\mu\right)+\text{sum of innovations}, \tag{2.11}$$

where the term $\left[\lambda\left(t_i\right)-\mu\left(t_i\right)\right]e^{-\beta(t-t_i)}$ is the autoregressive term and the term $\sum_{t_k>t_i<t}\alpha e^{-\beta(t-t_k)}$ represents the sum of the innovations in the AR process.

The unconditional expectation of the intensity of the Hawkes' process is a measure of the trading intensity in a given day and it is given by

$$E\left(\lambda\right)=E\left(\mu\right)+E\left(\int_{-\infty}^{t}\alpha e^{-\beta(t-s)}dN\left(s\right)\right). \tag{2.12}$$

Assuming stationarity we get the expected intensity as

$$E\left(\lambda\right)=\frac{\mu}{1-\frac{\alpha}{\beta}}. \tag{2.13}$$

From the expression above we check that the stationarity condition for the process is that $\alpha/\beta<1$. Two other important quantities in the context of a Hawkes' process are the clustering size c of the process

$$c=\frac{1}{1-\frac{\alpha}{\beta}}, \tag{2.14}$$

and the branching ratio n

$$n = \int_0^\infty \alpha e^{-\beta t} dt = \frac{\alpha}{\beta}, \tag{2.15}$$

where both expressions are valid for an exponential kernel. The branching ratio will be of special interest in this work. Whenever the intensity μ is a constant and the process is in the sub-critical ($n < 1$) or in the critical ($n = 1$) regime the branching ratio can be used as a measure of the proportion of events that are generated inside the model (by the presence of the exponential kernel, i.e. endogenously generated events) to all events (FS, 2012). To gain some more insight into the Hawkes' model we describe in the next section how the Hawkes' model relates to discrete time point processes.

2.2 Relation to Discrete Time Models

Modelling of irregularly spaced data is an econometric challenge that was first tackled using discrete time models. The Autoregressive Conditional Duration (ACD) model, proposed by Engle and Russel (1998), is a discrete time stochastic process that models the duration of events conditional on past durations. The ACD model can be thought as a GARCH model for the expectation of the conditional duration. Let the duration be defined as $x_i = t_i - t_{i-1}$ and let $\psi_i = E(x_i|x_{i-1}, ..., x_1)$ be the conditional expected duration. Then the ACD(p,q) model can be written as

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j}, \tag{2.16}$$

and the conditional intensity of the model is given by

$$\lambda\left(t|x_{N(t)}, ..., x_1\right) = \lambda_0 \left(\frac{t - t_{N(t)}}{\psi_{N(t)+1}}\right) \frac{1}{\psi_{N(t)+1}}, \tag{2.17}$$

which is further simplified in the case that the durations are conditionally exponential, leading to a unitary baseline hazard

$$\lambda\left(t|x_{N(t)},...,x_1\right) = \frac{1}{\psi_{N(t)+1}}. \tag{2.18}$$

For an ACD(p,q) model in Equation (2.16) the unconditional expected duration is given by

$$E(x_i) = \frac{\omega}{1-\sum(\alpha_j+\beta_j)}, \tag{2.19}$$

which is very similar to the unconditional expectation of the intensity of a Hawkes' process. The first disadvantage of the ACD model when compared to the Hawkes' model is that while the latter is described in wall-clock time the former is described in event time (see Easley and O'Hara, 1992; Hasbrouck, 1999 and Dufour and Engle, 2000 for the relevance of wall-clock time models). The second disadvantage is that while the Hawkes' model provides a simple description of the conditional intensity of the model this is not true for the ACD model (Russel, 1999 proposes to model directly the intensity of the process, the model is known as Autoregressive Conditional Intensity). This advantage of the Hawkes' model makes it easier to treat multivariate point processes (see Russel, 1999 and Bowsher 2007, for a more detailed discussion of the disadvantages of the ACD model).

2.3 Estimation of Hawkes' Process via Maximum Likelihood

Having established the main properties of the Hawkes' process and compared it to competing models, we now turn to the estimation of the Hawkes' process using the Maximum Likelihood Estimation. The estimation procedure of the Hawkes' self-exciting process presented here

builds on the work by Ozaki (1979). Ogata (1978) established the asymptotic properties of the ML estimator of the Hawkes' process. The general form of the log-likelihood function of a Hawkes' process with an arbitrary response fuction is given by

$$\log L(t_1, ..., t_N) = -\int_{-\infty}^{t_N} \lambda\left(t|\theta\right) dt + \int_{0}^{t_N} \log \lambda\left(t|\theta\right) dN\left(t\right), \tag{2.20}$$

where $\lambda(t|\theta)$ is the conditional intensity of the process. The likelihood function analyzed by Ozaki (1979) is valid for a Hawkes' process with intensity given by Equation (2.5), where the response function is exponential

$$h\left(t\right) = \alpha e^{-\beta t}, \tag{2.21}$$

which represents also the particular choice made by FS (2012) who apply the process given by Equation (2.5) to study endogenous price formation during market crashes. The point process with intensity given by (2.5) is a sample of events characterized by the times when each event occurs. The events can represent times when a transaction takes place, an order arrives or there is a change in the mid-price of a given stock, to name a few examples. We label each event by an index i which runs from 1 to N. The times when an event takes place must satisfy $t_1 < t_2 < ... < t_N$ as already discussed. Ozaki (1979) shows that the log-likelihood function for the Hawkes' process described by Equation (2.5) is given by

$$\log L\left(t_1, ..., t_N|\theta\right) = -\mu t_N + \sum_{i=1}^{N} \frac{\alpha}{\beta}\left(e^{-\beta(t_N - t_i)} - 1\right) + \sum_{i=1}^{N} \log\left\{\mu + \alpha A\left(i\right)\right\}, \tag{2.22}$$

where $A\left(i\right)$ is given by

$$A\left(i\right) = \begin{cases} \sum\limits_{t_j < t_i} e^{-\beta(t_i - t_j)}, \text{ for } i \geq 2, \\ \\ 0, \text{ otherwise,} \end{cases} \tag{2.23}$$

and can be rewritten using a recursive formula as $A(i) = e^{-\beta(t_i - t_{i-1})}A(i-1)$. The derivation of the log-likelihood function given by 2.22 is presented in Appendix 1. Note that the log-likelihood function is non-linear in the parameters of the model. Also note that it is quite easy to rewrite (2.22) in terms of the the branching ratio $n = \alpha/\beta$. This re-parametrization is presented below.

$$\log L\left(t_1, ..., t_N|\theta\right) = -\mu t_N + \sum_{i=1}^{N} n\left(e^{-\beta(t_N - t_i)} - 1\right) + \sum_{i=1}^{N} \log\left\{\mu + n\beta A\left(i\right)\right\}. \tag{2.24}$$

Before we turn to more specific details of the estimation of the Hawkes' process using equations (2.22) or (2.24) we discuss in the next section an important property of point processes. Namely, we discuss the time change theorem and how one can use it to construct a goodness-of-fit measure of point processes. A more formal treatment of the time change theorem can be found in Daley and Vere-Jones (2003, Section 7.4).

2.4 Compensator of Hawkes' Process and Random Time Change Theorem

Generally speaking the compensator of a stochastic process is a deterministic function that is subtracted from the process to make it a local martingale. Mathematically, it can be defined as the integral of the intensity over the whole history of the process

$$\Lambda\left(t\right) = \int_0^t \lambda(s)ds. \tag{2.25}$$

For the simple case of a Poisson process with intensity λ the compensator can be written as $\Lambda\left(t\right) = \lambda t$, because for the (homogenous) Poisson process the intensity λ is a constant. For a general point process the compensator defined by the Equation (2.25) takes the point

process with intensity $\lambda(s)$ to a unit-rate Poisson process. Therefore the durations as defined bellow are exponentially distributed

$$\Lambda\left(t_i, t_{i+1}\right) = \int_{t_i}^{t_{i+1}} \lambda(s) ds, \tag{2.26}$$

and can be used to test the simulated process via a quantile-quantile plot. Equation (2.26) performs a random change in the time-scale of the process. The resulting process from this random time change is called the residual process. Inserting the intensity for a Hawkes' process in (2.26) we get

$$\Lambda\left(t_i, t_{i+1}\right) = \int_{t_i}^{t_{i+1}} \mu(s) ds + \int_{t_i}^{t_{i+1}} \sum_{t_k<s} \alpha \exp(-\beta\left(s - t_k\right)) ds, \tag{2.27}$$

and noting that the summation is over the (discrete) event times that are smaller or equal to t_i we get (assuming that $\mu(s) = \mu$)

$$\Lambda\left(t_{i+1}, t_i\right) = \mu(t_i - t_{i+1}) + \sum_{k=1}^{i} \int_{t_i}^{t_{i+1}} \alpha \exp(-\beta\left(s - t_k\right)) ds, \tag{2.28}$$

which leads to

$$\Lambda\left(t_i, t_{i+1}\right) = \mu(t_{i+1} - t_i) - \sum_{k=1}^{i} \frac{\alpha}{\beta} \left[\exp(-\beta\left(t_{i+1} - t_k\right)) - \exp(-\beta\left(t_i - t_k\right))\right], \tag{2.29}$$

A common way to measure the goodness-of-fit of the Hawkes' model is making use of the residual process derived from the model. The time change property of point processes assures that the integrated Hawkes' process is a Poisson process with unit rate. Therefore, the durations of the integrated process are exponentially distributed with unit rate. Following Ogata (1988), it is possible to make one-to-one transformation of the point process described by the events $\{t_i\}$ to the random time changed set $\{\xi_i\}$ by making use of Equation (2.26) and letting $\xi_i = \Lambda(t_{i-1}, t_i)$. The set of times $\{\xi_i\}$ is the residual process and, using the time

change property, follows a unit rate Poisson process. If we now take $U_k = 1 - \exp \Lambda(t_{i-1}, t_i)$, then U_k is distributed as a uniform random variable in the range $[0, 1)$.

Therefore, a very simple way of assessing the goodness of fit of Hawkes' model is to calculate the estimated $\hat{\Lambda}(t_i)$ making use of the vector of parameters $\hat{\theta} = (\hat{\mu}, \hat{\alpha}, \hat{\beta})$ that we have previously estimated, to obtain the estimated residual process $\left\{\hat{\xi}_i\right\}$. Then one can calculate the $U_k's$ and compare it with random uniform variables in the range $[0, 1)$. The general idea is that if the Hawkes' model is a good description of our data then we might expect that the estimated residual process follows a unit rate Poisson process, or equivalently, that the durations of the estimated residual process have a unit rate exponential distribution. To test the hypothesis that the estimated residual process comes from a unit rate Poisson process we can make use of the Kolmogorov-Smirnov (KS) statistic to draw confidence bounds for the process. Daley and Vere-Jones (2003, p.262) describe an algorithm that can be used to assess the goodness-of-fit of a point process using the KS statistic.

Chapter 3

Literature Review

Point processes are a class of stochastic processes applied to model many different phenomena (see Thompson, 1988 for applications in safety and reliability; Sornette et al., 2004 applies a Hawkes' process to model book sales using Amazon data; Crane and Sornette, 2008 use the Hawkes' process to study the dynamics of views of YouTube videos; Ogata, 1988 uses point processes to analyze earthquake data).

In finance, point processes are applied mainly to explain some of the stylized facts related to the microstructure of financial markets. More specifically, point processes offer a parsimonious way to model the duration between events and have been extensively used to model the arrival rate of quotes and prices in different markets. The applications can be distinguished between those that make use of discrete point processes, like the ACD and the ACI models, and those that make use of continuous time point processes, like the Hawkes' model. Here we review the most relevant works that applied the Hawkes' process to transaction data.

Even though the Hawkes' model was proposed in the 1970's its applications in finance are relatively recent. Bowsher (2007) was one of the first to consistently apply the Hawkes' process to describe events related to financial markets. In his work, Bowsher develops a generalized Hawkes' model, described in terms of its vector conditional intensity, and applies the bivariate version of it to explain the durations of trades an mid-quote changes using data on one stock (General Motors Corporation) and 40 trading days in the year 2000. Bowsher (2007) shows that there is a two-way interaction between trades and changes in

mid-prices where the occurrence of a trade increases the intensity of mid-price changes and mid-price changes increases trade intensity. Bowsher (2007) uses data from the Trades an Quotes (TAQ) database with timestamps of one second precision. The issue of multiple events within a single timestamp, common to high frequency data that is timestamped to the nearest second, is solved by adding a uniform random component that distinguishes between equal timestamps. This strong assumption related to the ordering of events may not be a big issue in Bowser's (2007) paper that uses data from 2000 with a small number of simultaneous events per time stamp (Bowsher, 2007 explains that the number of events that occurred within same timestamps represents only 0.26% of all events for trades and only 0.14% for mid-price changes). Nevertheless, if one uses more recent data, the presence of High Frequency Traders increases drastically the number of events within a given second. For instance, using TAQ data on trades for the Yahoo stock in one trading in February of 2010 roughly 30% of all events (i.e. price changes) occurred within the same second.

Another early application of the Hawkes' process in finance was made by Hewlett (2006). In his work, Hewlett (2006) used a bivariate Hawkes' process to model order flow in the FX market. Hewlett proposed a model that predicts future trading intensity conditional on the pattern of past trades that is modelled via a Hawkes' process. In the market, liquidity takers that need to fill a large order are faced with the dilemma whether they should try to fill the order at once or split the large order in small tranches. While filling the large order at once will influence the market price, splitting the order into tranches is subject to front running of market-makers if they are capable to identify the pattern of buy and sell order arrivals. Hewlett's (2006) model tries to tell how the liquidity taker should behave, given the reaction of the market-maker, assuming that the process of order arrival follows a bivariate Hawkes'

process.

In a recent paper Filimonov and Sornette (2012) proposed a measure of market endogeneity (termed reflexivity, as proposed by George Soros) that measures whether price changes are driven by exogenous events like fundamental news related to a firm or the economy, or endogenously by market movements that emerge through positive feedback mechanisms that induce correlation among price changes. In their work, Filimonov and Sornette (2012) use the branching ratio of a self-excited conditional Hawkes' model as a proxy for market endogeneity. Using quote data on the E-mini S&P 500 futures that spans the years from 1998 to 2010 they analyze the dynamic behavior of the branching ratio, estimated via Maximum Likelihood, and find that while before 2000 the market endogeneity was relatively low (with branching ratio of $\simeq 0.3$), after 2004 it has reached levels close to 0.9, being consistently above 0.6. Filimonov and Sornette (2012) then show that the branching ratio is fairly stable even during periods of market stress as long as they are justified by some exogenous news. In their sample, they use the downgrading of Greece and Portugal on April 27, 2010 as evidence that supports this fact. Nevertheless, there is a large increase in the branching ratio during the crash of May 6, 2010 - popularly known as Flash-Crash - when stock markets fell without any relevant exogenous news. Filimonov and Sornette (2012) also noticed that the increase in the branching ratio coincides with the rise in activity by High Frequency Traders. The flash-crash itself, even if there is no evidence that it was triggered by High Frequency Traders, was to some extent associated with the presence of high-speed automated trading systems that might have exacerbated the extreme market movements observed on that day (see the SEC report on the subject as well as Kirilenko et al., 2011).

Some other works that used the Hawkes' process to model financial market phenomena

are Large (2007) that uses a Hawkes' process to measure the resiliency of a limit order book, defined as the speed to which prices recover from large trades that disturb the bid-ask spread, and the paper of Bacry et al. (2011) that uses the process to model the Epps effect[1].

[1]The Epps effect relates to the decrease in the correlation among price changes of stocks as the interval on which the price changes are measured decreases (the sampling frequency increases). Epps (1979) related this effect with the non-stationarity of price changes.

Chapter 4

Simulation of a Hawkes' Process

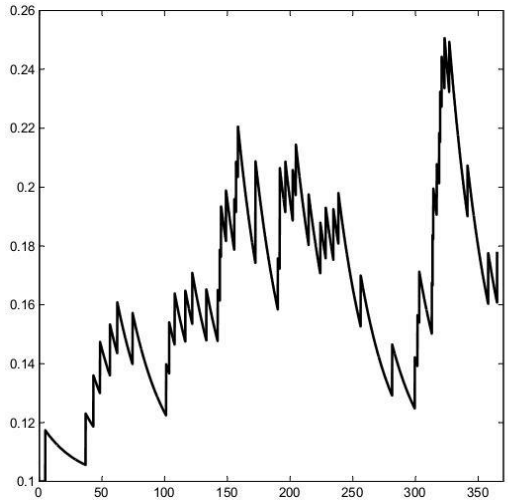
Before turning to more specific issues related to the estimation and the fit of the model it is worth to note that Ozaki (1979) presents an algorithm which can be used to generate Hawkes' process data. The generation of simulated data is interesting because knowledge of the true values of the vector of parameters θ enables us to calibrate the estimation procedure and check how the log-likelihood function behaves. The algorithm for generation of Hawkes' process data follows the steps outlined below (as described by Ozaki, 1979):

1. Generate a uniform random number U on $[0, 1]$;
2. Let $t_1 = -\log(U)/\mu$;
3. Generate a uniform random number U on $[0, 1]$;
4. Solve $\log(U) + \mu(u - t_k) + \frac{\alpha}{\beta} A(k)\left(1 - e^{-\beta(u-t_k)}\right) = 0$;
5. Let $t_{k+1} = u$ and $A(k+1) = e^{-\beta(t_{k+1}-t_k)} S(k) + 1$;
6. Go back to step 3 and increase k by one.

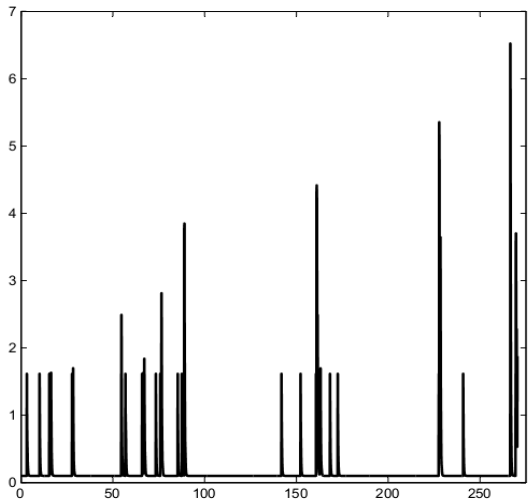
Note that

$$A(k) = \begin{cases} e^{-\beta(t_k - t_{k-1})} A(k-1) + 1, \text{ for } i \geq 2, \\ 1, \text{ otherwise.} \end{cases} \tag{4.1}$$

The above described algorithm was implemented using Matlab. After the data was generated it is possible to calculate the intensity of the process using Equation (2.5). Note that the



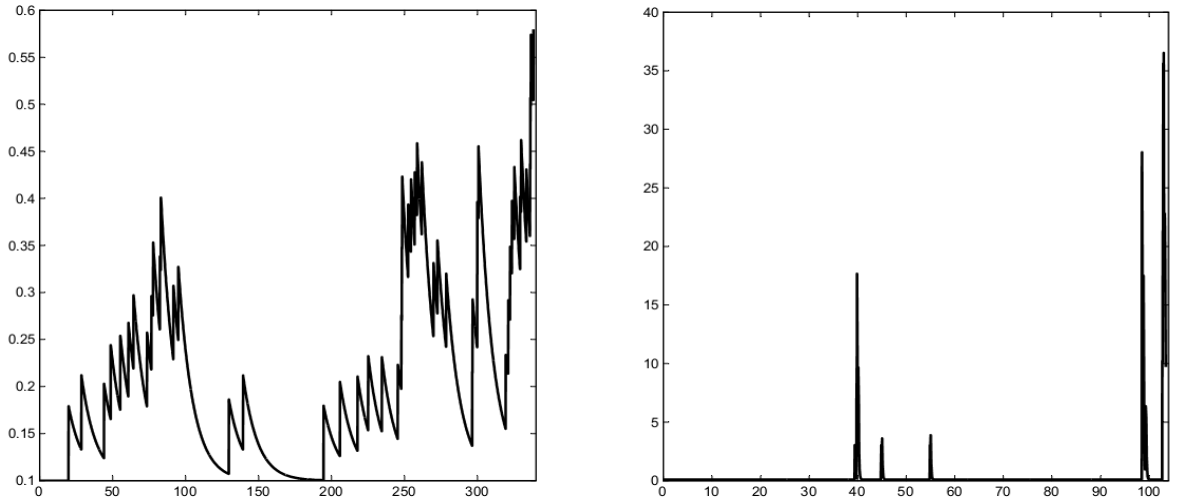
(a) $\theta = (0.1, 0.0175, 0.035)$



(b) $\theta = (0.1, 2.5, 5)$

Figure 4.1: Intensity: $n = 0.5$

parameter μ determines the intensity of exogenous events (roughly, how many events occur per unit of time) and that α and β determine the clustering of the process and the intra-event dynamics. A higher branching ratio increases clustering and, keeping μ and n constant, smaller values for α and β decreases clustering. To visualize the individual effect of α and β we make two different plots keeping μ and n constant, but varying α and β. The first plot has $\theta = (\mu = 0.1, \alpha = 0.0175, \beta = 0.035)$, while the second has $\theta = (0.1, 2.5, 5)$. Both have a branching ratio $n = 0.5$ and a sample size of 50 events. It is possible to observe that there is some clustering in the intensity of Figure 4.1 as displayed by the occurrence of spikes in the graph. The most striking difference of Figure 4.1b from Figure 4.1a is that for smaller values of β the intra-event intensity decays much more slowly than for higher values of β. Note that in both graphs the "base" intensity is 0.1 but the spikes are much less intense in Figure 4.1a than in Figure 4.1b. Obviously, since both graphs were constructed with the same branching ratio the mean intensity is also the same.



(a) $\theta = (0.1, 0.08, 0.1)$ (b) $\theta = (0.1, 8, 10)$

Figure 4.2: Intensity: $n = 0.8$

If we now increase the branching to $n = 0.8$ we get more clustering. We first make α and β low $\theta = (0.1, 0.08, 0.1)$ and then make both large $\theta = (0.1, 8, 10)$. The results are presented in Figure 4.2.

4.1 Consistency of Estimates and Comparison with Ozaki (1979)

Here we conduct two different exercises using simulated data. First, in a simple consistency exercise we create 30 different samples, where each sample has a different size. We begin with a small sample of 100 events and increase the sample size by 100 observations every time a new sample is simulated, the last sample we draw has 3000 observations. For this exercise we use parameters values $\theta = (\mu = 0.1, \alpha = 1.0, \beta = 2.0)$. In Figure 4.3f we plot the value of the estimated parameters on the vertical axis and the size of the sample used in the estimation on the horizontal axis, the horizontal lines around the parameters estimates represent the

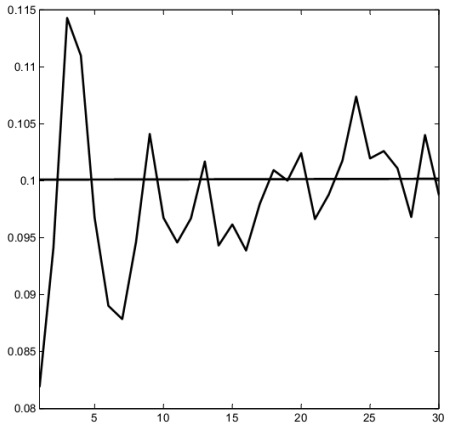
true value of the parameters. We also plot the standard errors of each coefficient next to the parameter estimates. Standard errors were constructed by calculating the inverse of the Hessian and multiplying it by minus 1.

The second exercise we perform is a resampling exercise. Using a sample size of 1000 events we generate 50 different samples and calculate the estimates for each sample. Descriptive statistics for the results obtained are shown in the Table 4.1. Variances and standard errors are calculated in two different ways: first we calculate the sample variance using the 100 different estimates for each sample we generated (denoted by Sample Variance in the table), second we calculate the variance for each estimate by calculating the inverse of the Hessian and multiplying it by minus 1 (denoted by Estimated Variance in the table below). The obtained values (which are estimates of the variance-covariance matrix of the parameter estimates) were then averaged over the 100 different variance-covariance matrices we obtained. The small difference in the values of the variance calculated using these different approaches is due to the small size properties of the sample. The estimations are carried out with true parameters $\theta = (0.1, 1, 2)$.

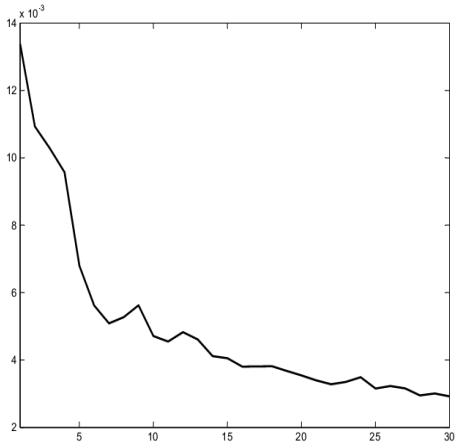
	μ	α	β
Average	0.0997	1.0038	2.0084
Minimum	0.0871	0.8357	1.7353
Maximum	0.1098	1.2523	2.3889
Sample Standard Errors	0.0054	0.0961	0.1661
Estimated Standard Errors	0.0050	0.0800	0.1478

Table 4.1: Statistics of Estimated Parameters

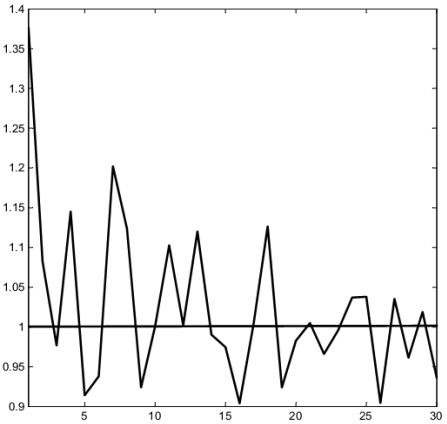
Figure 4.3: Consistency of Estimates



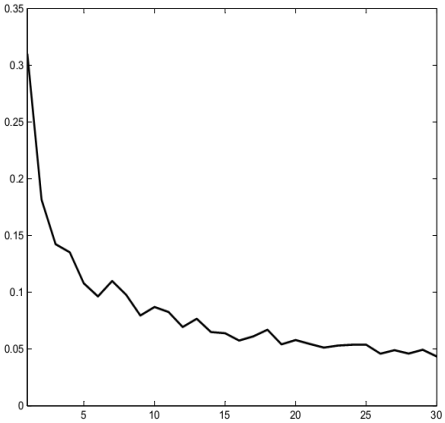
(a) Estimates of μ



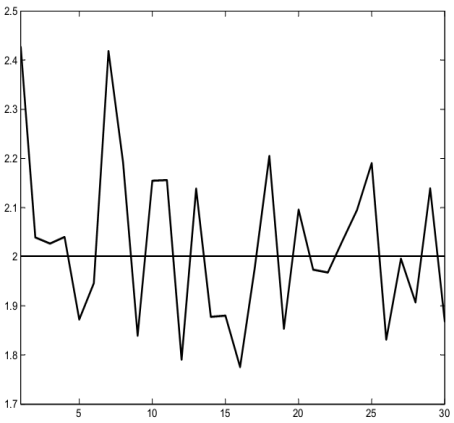
(b) Standard Deviation of μ



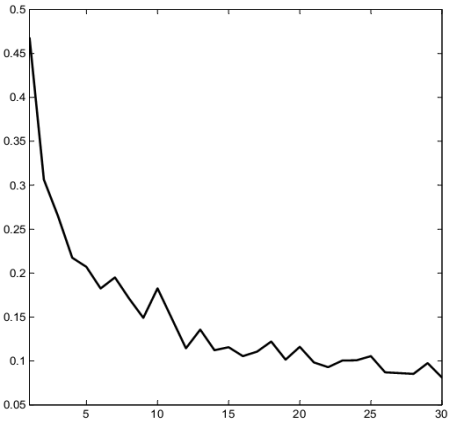
(c) Estimates of α



(d) Standard Deviation of α



(e) Estimates of β



(f) Standard Deviation of β

Table 4.2 presents the estimation results using a single sample of 500 events, the true parameters are chosen as $\theta = (\mu = 0.5, \alpha = 0.8, \beta = 1.0)$, both the sample size an the parameters are the same as used in the estimation of Ozaki (1979). Table 4.2 shows the results obtained for this sample and also present the results obtained by Ozaki (1979) as a comparison.

	μ	α	β
True Values	0.50000	0.80000	1.00000
Estimates	0.66200	1.04418	1.32240
Estimates of Ozaki (1979)	0.67200	0.68400	1.01800
Estimated Standard Errors	0.15205	0.16198	0.22149
Standard Errors of Ozaki (1979)	0.12369	0.11832	0.19313

Table 4.2: Comparison with Ozaki (1979) Estimates - $\theta = (0.5, 0.8, 1.0)$

4.2 Goodness-of-Fit

In order to validate the simulated process we calculate the compensator using simulated data that follows a Hawkes' process with exponential response function. The durations given by (2.29) are then exponentially distributed. Figure 4.4 shows the behavior of the integrated simulated process with respect to a unit-rate exponentially distributed process. The fact that both, the simulated process and the theoretical process lies on the same 45° line indicates that both processes have the same distribution. In the plot, the empirical and theoretical axis represent the values for the durations of the simulated and theoretical processes respectively. The simulated process is just the integral in (2.29) of the generated

Hawkes' process and the theoretical process is a unit rate exponentially distributed process. From Figure 4.4 it is possible to see that most of the observations are clustered at very small values of the duration while only a small proportion of the sample have durations above 5 units (approximately 0.7% of all durations are above 5 units). Note that each of the crosses in the plot represent one observation of the simulated process and that the simulated process has 5000 observations. A more formal way of assessing the goodness-of-fit of the

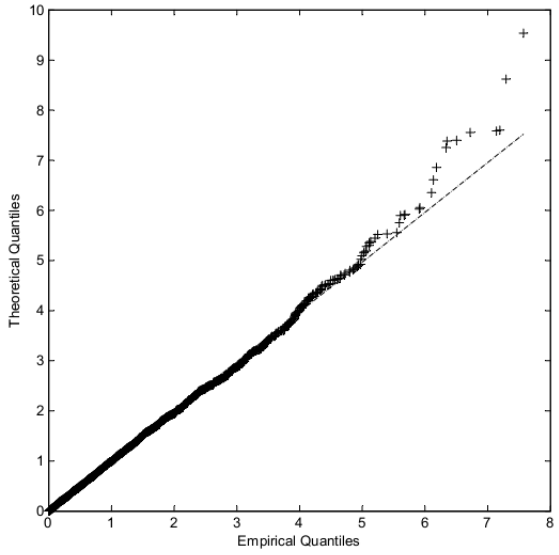
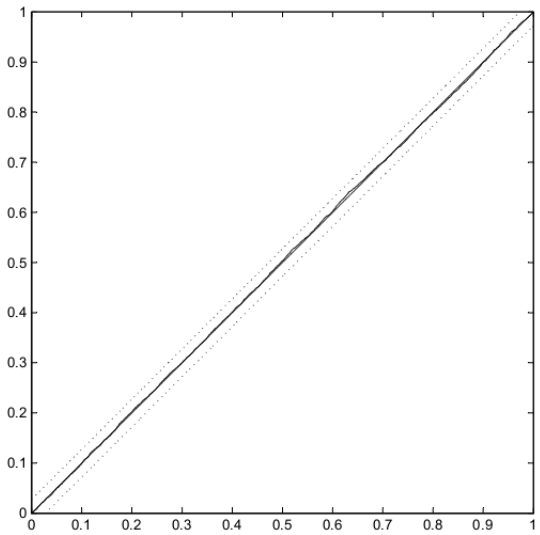
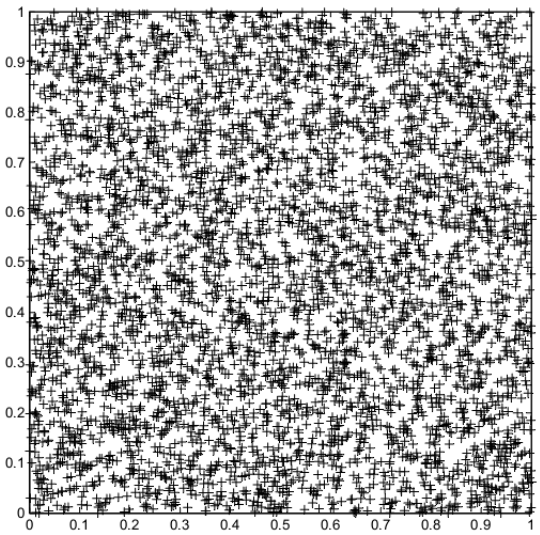


Figure 4.4: Quantile-Quantile Plot for the Simulated Process

model is by making use of the Kolmogorov-Smirnov statistic to draw confidence bounds for the process. To test this approach we generate a process with 5000 observations and true parameter $\theta = (\mu = 0.1500, \alpha = 1.0000, \beta = 2.0000)$. This choice of parameters results in a branching ratio of 0.5. The estimated parameter was $\hat{\theta} = (\hat{\mu} = 0.1519, \hat{\alpha} = 0.9996, \hat{\beta} = 1.9576)$. Figure 4.5a illustrates the fit of the Hawkes' model for the simulated data. The horizontal axis is composed of the CDF of an exponential distribution with unit rate, i.e. $U_k = 1 - exp(\xi_k - \xi_{k-1})$ and the vertical axis is the CDF of a $U(0,1)$ distribution. The vertical line $y = x$ is the line of best fit, the line around the line of best fit is the fit of the



(a) Goodness of Fit - Simulated Data



(b) Durations Independence - Simulated Data

Figure 4.5: Goodness of Fit

model and the dashed lines represent 95% confidence bounds. We can see that the model and theoretical lines are almost indistinguishable. We therefore validate both our algorithm that generates data and the procedure to assess the goodness-of-fit. A complementary test of the validity of the model is to check whether there is serial correlation in neighbouring intervals. Berman (1983) proposes to plot U_k against U_{k+1} to check for the presence of autocorrelation. Figure 4.5b shows this plot.

Chapter 5

Data and Estimation

We now turn to the description of the procedure we use to estimated the Hawkes' process using real transaction data. Our results are obtained making use of two different databases. The Trades and Quotes (TAQ) database is composed of stocks traded in exchanges in the U.S. and the Thomson Reuters database is composed of stocks traded in exchanges in Europe. For both databases we have high-frequency data on all trades and quotes on each trading day. For quote data, every time there is a bid without a corresponding ask (or vice-versa) we take the missing value as equal to the last observed value. For trades, whenever there are trades with a missing price we delete that observation from our dataset. The main technical difference between the two databases is that the timestamps provided by the TAQ database have a precision of one second while the timestamps provided by Thomson Reuters have a precision of one millisecond. Since it is very common to have more than one trade/quote being filled at the same second, the Thomson Reuters database gives a much more precise view of the process we want to estimate (roughly 30% of all trades in the TAQ database, 22% of all trades and up to 45% of all mid-price changes in the Thomson Reuters database occur at the same second). To deal with the issue of multiple events per timestamp a common procedure is to disentangle equal timestamps by adding a uniform random variable as an artificial precision component to the one second precision timestamp (as in Bowsher, 2007 and FS, 2012). We estimate the model on TAQ data using this procedure. Later, we check how this procedure affects the estimates using rounded timestamps of the Thomson Reuters

database.

For data on U.S. stocks we estimate the model using data on trades for one trading day and one stock (Yahoo Inc. - ticker YHOO)[1]. For data on European stocks we estimate the model using data on trades and quotes for one trading day and one stock (Vodafone - ticker VOD). An important aspect of the estimation is the definition of the event. In our work, we define one event as the change in the price of a trade relative to the previous price for data on trades and as a change in the mid-price relative to the previous mid-price for data on quotes. We begin our estimation 5 minutes after the beginning of the regular trading day (09:30 AM for the U.S. market and 08:00 AM for the European market - both times are local times) and end it 5 minutes before the closing of the trading day (04:00 PM for the U.S. market and 04:30 PM for the European market). In order to assess the intra-day dynamics of the parameters, we estimate the model using overlapping rolling windows of 20, 30 and 40 minutes, with a time step of 5 minutes. We also estimate the model using data on one full trading day at once. Both estimation procedures proved to give similar results.

[1]We also used three additional trading days in the estimation and obtained similar results. We do not present these results here.

Chapter 6

Results

We first present the results when the estimation is conducted using TAQ data. Next, we turn to the results using Thomson Reuters data.

6.1 Results - TAQ Database

6.1.1 Timestamps with One Second Precision

We begin by estimating the Hawkes' process using TAQ data on trades with timestamps of one second precision. We solve the issue of multiple events per second by keeping only one event per second (note that since we have an unmarked Hawkes' process it does not matter whether we keep the first or the last event). We present our results using data for one trading day (February, $1^{st}, 2010$) on the stock of Yahoo Inc. (ticker - YHOO). The estimation was carried out using all observations for a whole trading day at once. Table 6.1 shows the results obtained, standard errors are presented below the estimates. The first four columns present estimates for the parameters of the Hawkes' process, where $n = \alpha/\beta$ is the branching ratio. The fifth column presents the unconditional expected intensity calculated as the average of the estimated intensity, as given by Equation (2.5). The sixth column presents the mean duration between events. The standard error of $n = \alpha/\beta$ was calculated using the delta method [1].

[1]The formula for the variance of n is given by $\text{Var}(n) = \frac{\hat{\alpha}^2}{\hat{\beta}^4}\text{Var}(\hat{\beta}) + \frac{1}{\hat{\beta}^2}\text{Var}(\hat{\alpha}) - 2\frac{\hat{\alpha}}{\hat{\beta}^3}\text{Cov}(\hat{\alpha}, \hat{\beta})$.

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.0407	0.8244	0.0094	0.0114	0.2246	4.4012	5180
Standard Error	0.0102	0.0474	0.0018	0.0013	0.0613	4.9878	-

Table 6.1: Estimates Using TAQ Data on Trades

The branching ratio is very high. Its value means that roughly 82% of all price changes are driven by endogenous feedback mechanisms in the market and less than 20% of price movements are driven by exogenous events. This high value for the branching ratio is in concordance with the values calculated by FS (2012) for the E-mini S&P 500 future. The value of μ is smaller than the one obtained by FS (2012). The estimated intensity tells us that, on average, 0.2 events occur per second. The estimated value for the intensity has a close relationship with the average duration. If we calculate the average duration between events we get $\tau_m = 4.4012$ seconds and a standard deviation of $\sigma_\tau = 4.9878$ seconds. The inverse of the average duration gives approximately the average intensity. Note that we can check the half-life of a shock using the formula $t_{1/2} = \ln(2)/\beta$, which in this case leads to a relatively high value $t_{1/2} \sim 61$ seconds.

To give a first hint whether the estimates of the model are reasonable we perform a very simple check. Using the algorithm that generates data that follows a Hawkes' process, we generate a sample with 5180 observations and parameters as given in Table 6.1. The results are very far from what one would expect if the model was a good description of the data. The average duration of the simulated data is 0.3976 second with a standard deviation of 1.2327 second. It is not so unexpected that we obtain such different results. The fact that

the durations have a lower bound of 1 second introduces some sort of bias in the model.

In order to give a more formal assessment of the performance of the model we have calculated the fit of the Hawkes' process, presented in Figure 6.1. We can see from the figure that the model fits relatively well when $(x, y) \gtrsim (0.3, 0.3)$. Obviously, the durations of trades here are bounded below by one. The data cannot produce any durations that are smaller than one, since the precision of the timestamps is limited to one second. Since the x-axis of the graph is constructed by taking $(1 - \exp[-(\xi_i - \xi_{i-1})])$, where ξ_i is the residual process, the lack of very small durations is introducing the large deviation from the 45° line in the lower region of the graph.

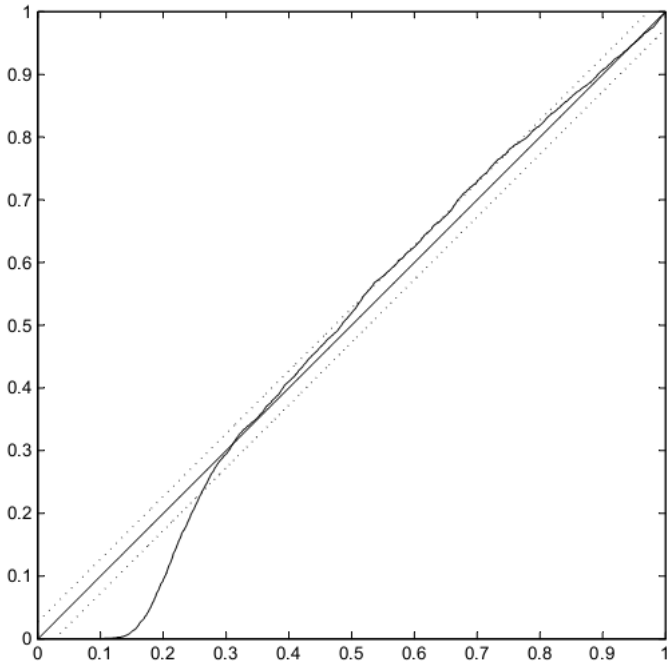


Figure 6.1: Fit of the Model Using TAQ Data

6.1.2 Timestamps with Randomized Precision

In order to disentangle events that occurred at the same second we test a randomization of timestamps within the same second. This randomization is carried out by introducing,

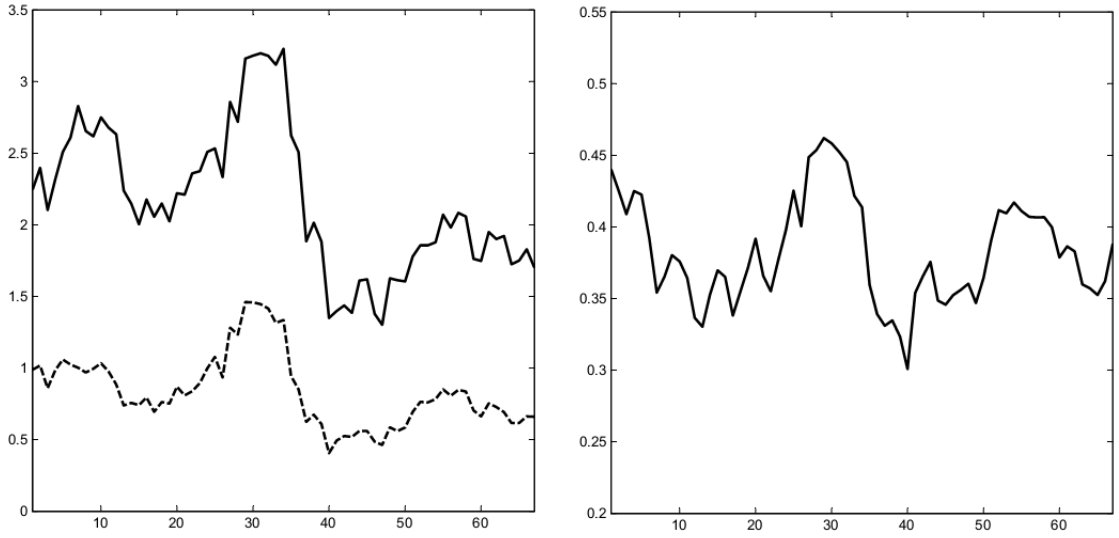
for each event, a random number distributed as $U(0,1)$ after the number that measures the second of each timestamp. A similar procedure was used by Bowsher (2007) and FS (2012). We re-estimate the Hawkes' model with data containing this randomized precision. Now, the number of observations is much higher because we are able to distinguish between same second events through the randomized precision. The increase in the number of observations allows us to estimate the model using small intra-day rolling windows of 20, 30 and 40 minutes. We verify that even though there is variation of the estimates calculated using the overlapping rolling windows, taking the average of the rolling window estimates gives roughly the same values as estimating the model using all observations at once. Comparing Table 6.2 with Table 6.1 we see that there is a great difference in the estimates obtained using the randomized timestamps. Now μ is much larger, n is approximately half of its value and α and β are more closer to unity than before. In fact, α and β are now roughly 80 times larger than in Table 6.1, this fact reduces the half-life of a shock to less than one second $t_{1/2} \sim 0.34$ seconds. The intra-day dynamics of the parameter n is relatively smooth with n varying around its mean value which is close to 0.4. The dynamics of μ, and of the the intensity $\mu/(1-n)$, display the well know U shaped intra-day pattern with higher intensity of trading at the opening and during the closing of the trading day. Comparing the estimates with those obtained by FS (2012) in the same period we now get values for μ that are slightly higher and values for n that are much lower than FS's (2012) estimates. Nevertheless, our results were estimated in a slightly different way than FS (2012) and our dataset is based on equity trades data and not on futures quotes data. The intensity of events is now 0.35 seconds^{-1} and is somewhat higher than in the previous case. The average duration of the data is $\tau_m = 2.7584$ seconds ($\sigma_\tau = 4.4324$ seconds). The value obtained from

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.2142	0.4015	0.8151	2.0301	0.3534	2.7584	8265
Standard Error	0.0038	0.0362	0.0302	0.0831	0.3558	4.4323	-
20 min.	0.2141	0.3788	0.8335	2.1671	0.3469	2.7584	8265
Standard Error	0.0166	0.0513	0.1372	0.3940	0.3473	4.4324	-
30 min.	0.2125	0.3816	0.8370	2.1626	0.3461	2.7584	8265
Standard Error	0.0135	0.0433	0.1126	0.3195	0.3493	4.4324	-
40 min.	0.2114	0.3825	0.8403	2.1705	0.3450	2.7584	8265
Standard Error	0.0116	0.0325	0.0982	0.2784	0.3498	4.4324	-

Table 6.2: Estimates Using Randomized TAQ Data on Trades

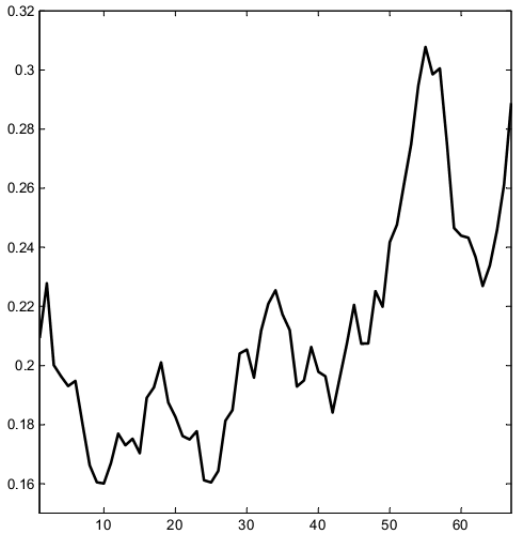
simulating data with parameters taken from the first line of Table 6.2 give 2.8439 seconds for the average duration and 4.1418 seconds for the standard deviation, values that are very close to those in the real data.

We have also calculated the fit of the model for the different parameters we estimated. We observe that now the model fits very well the lower region of the graph. Overall, only in the region that is approximately between 0.6 and 0.9 the line of fit touches the Kolmogorov-Smirnov bound. We associate the good performance of the model with random timestamps, when compared to the model estimated on data with one second precision, to the larger dispersion of durations. If we check some simple statistics of the residual process for the original data we get that the minimum value of the duration of the *residual* process is 0.0409 and the maximum is 10.2348. On the other hand, for data with random timestamps the minimum value of the duration of the residual process is of the order of 10^{-6} and the

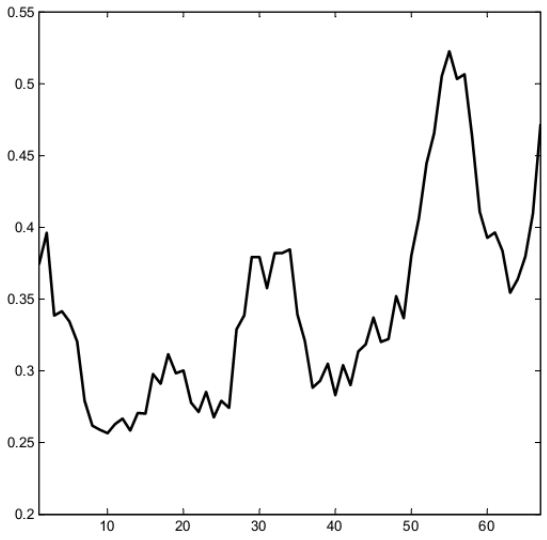


(a) α and β

(b) n



(c) μ



(d) Intensity

Figure 6.2: Estimation Using 40 minutes Rolling Window - TAQ Data on Trades

maximum value is 15.1764. The fit of the model presented here is based on the estimates of the first line of Table 6.2. Using the other parameters from Table 6.2 to assess the fit of the model gives very similar results (not presented here).

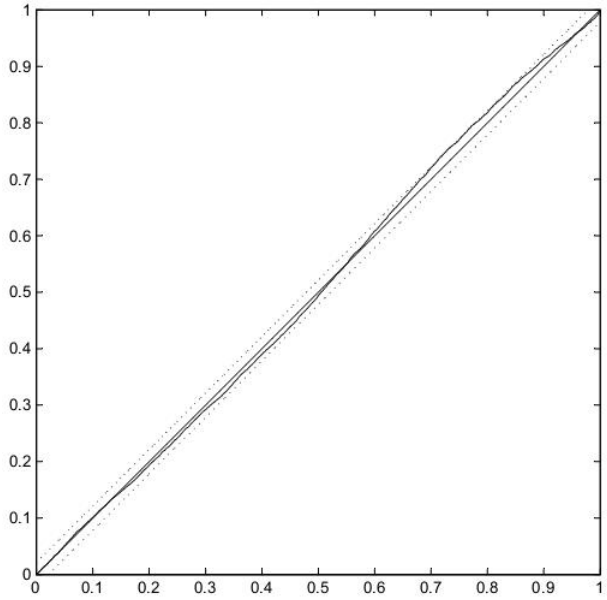


Figure 6.3: Fit of the Model Using Randomized TAQ Data

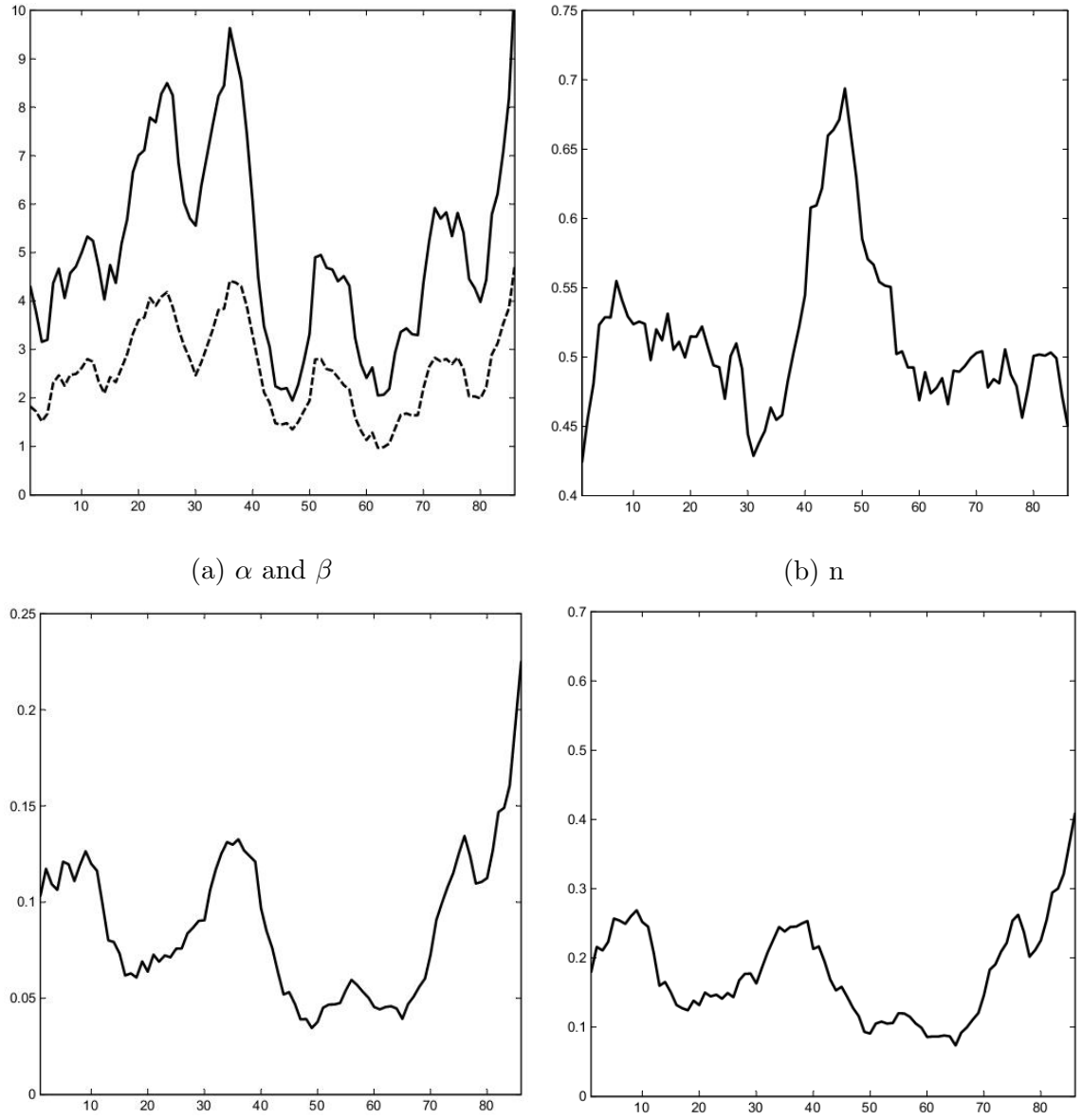
6.2 Results - Thomson Reuters Database

We now estimate the Hawkes' process using the Thomson Reuters database. Here we base our results on the exchange with highest liquidity, which is the London Stock Exchange (LSE) and analyze data on trades and quotes. To check how the model performs when the duration between events is larger we estimate the model using data from a venue with smaller trade intensity, in this case the NYSE Euronext Brussels. Due to the limited number of events in this exchange we conduct our analysis using data on quotes only. The estimations for both venues were based on data for one trading day (January, 2^{nd}, 2009) on the Vodafone (ticker - VOD) stock.

6.2.1 London Stock Exchange

Quotes Data. We begin estimating the model using data on quotes and considering four different estimation strategies. The results are presented in Table 6.3. Again there is little difference when one estimates the model using all available observations at once or when one considers a rolling window estimation. Note that the estimates for μ are small, similar to the values obtained with "raw" TAQ data. The values for n are much smaller than in that case but higher than in the random TAQ data case. Both α and β are relatively high. In fact, the half-life of a shock now is very low ($t_{1/2} \sim 0.15$ seconds). The average duration between trades is also much higher than in the TAQ data. Its value is now $\tau_m = 5.1724$ seconds ($\sigma_\tau = 13.6046$ seconds). Using simulated data obtained with parameters from the first line of Table 6.3 we get an average duration of 5.2947 seconds and a standard deviation of 9.2812 seconds. The intra-day dynamics of the parameters displays qualitatively the same behavior as in our previous estimation. The values for μ are subject to the intra-day seasonality, while n varies around its mean value. Figure 6.4 presents the results.

Using the estimates from the first row of Table 6.3 we calculate the fit of the model. Compared with the result using randomized timestamps the model fits the data relatively poorly. Two regions that comprise most of the goodness-of-fit plot cross the Kolmogorov-Smirnov bounds. Contrary to the case presented in Figure 6.1, where the line of fit crosses the KS bound below the 45° line, the line of fit here crosses the KS bound above the 45° line, indicating that some durations occur more frequently than the model can capture.



(a) α and β

(b) n

(c) μ

(d) Intensity

Figure 6.4: Estimation Using 40 Minutes Rolling Window - LSE Data on Trades

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.0931	0.5137	2.4308	4.7317	0.1848	5.1724	5798
Standard Error	0.0020	0.0107	0.0952	0.1978	0.4506	13.6046	-
20 min.	0.0917	0.5115	2.6111	5.2434	0.1793	5.1724	5798
Standard Error	0.0091	0.0567	0.4615	0.8981	0.4474	13.6046	-
30 min.	0.0901	0.5144	2.5759	5.1145	0.1792	5.1724	5798
Standard Error	0.0075	0.0466	0.1506	0.7346	0.4504	13.6046	-
40 min.	0.0883	0.5143	2.5676	5.0888	0.1776	5.1724	5798
Standard Error	0.0065	0.0408	0.3352	0.6584	0.4507	13.6046	-

Table 6.3: Estimates Using Thomson Reuters Data on Quotes - LSE

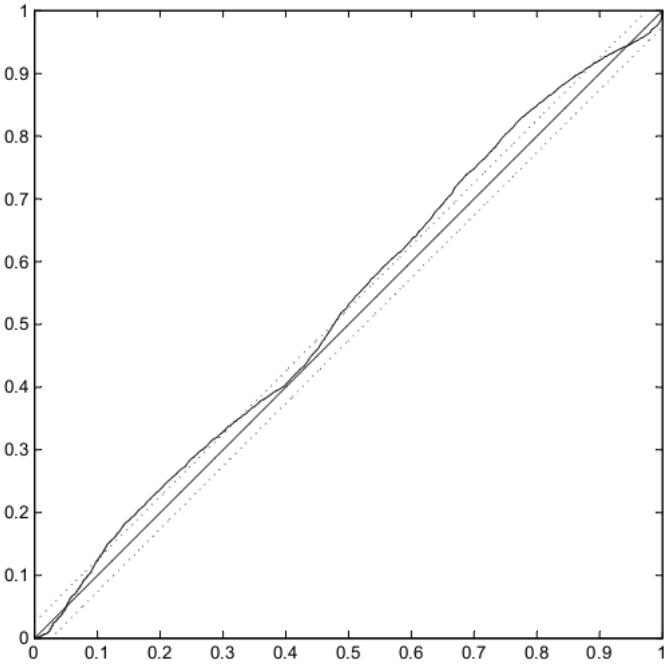


Figure 6.5: Fit of the Model Using Thomson Reuters Data on Quotes - LSE

We are now able to assess how the randomization of timestamps, performed before to

get parameters estimates using the TAQ database, impacts the parameters estimates of the Hawkes' process[2]. We do this by first rounding the timestamps of the Thomson Reuters database to its nearest integer (rounding the timestamps to the smallest and greatest integers leads to similar results) and then adding a random $U(0,1)$ component to the rounded timestamp (adding a $-0.5 + U(0,1)$ random component as in Bowsher (2007) also leads to similar results). We do not present the results using the rolling window procedure since the estimates are again very close to the estimates in Table 6.4 and the intra-day dynamics of the parameters are qualitatively the same as already discussed before.

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.0743	0.6118	0.9147	1.4949	0.1895	5.1725	5798
Standard Error	0.0019	0.0117	0.0309	0.0513	0.3452	13.5479	-

Table 6.4: Estimates Using Random Timestamps on Quotes - LSE

The parameters μ and n have values that are, respectively, a little bit higher and smaller than its values obtained in the estimation using the "true" data. But the parameters α and β have now values that are roughly 2.5 times smaller than its estimated values using the original timestamps. If one is interested in the parameters μ and n, introducing the random precision to the timestamps produces a relatively small bias in the estimates. Nevertheless, the values of α and β change a lot. It is also remarkable how the fit of the model changes when one considers a random component in the timestamps. The figure below shows that the model now provides a much better fit.

[2]I thank Andreas Rapp for giving me the idea to perform this check.

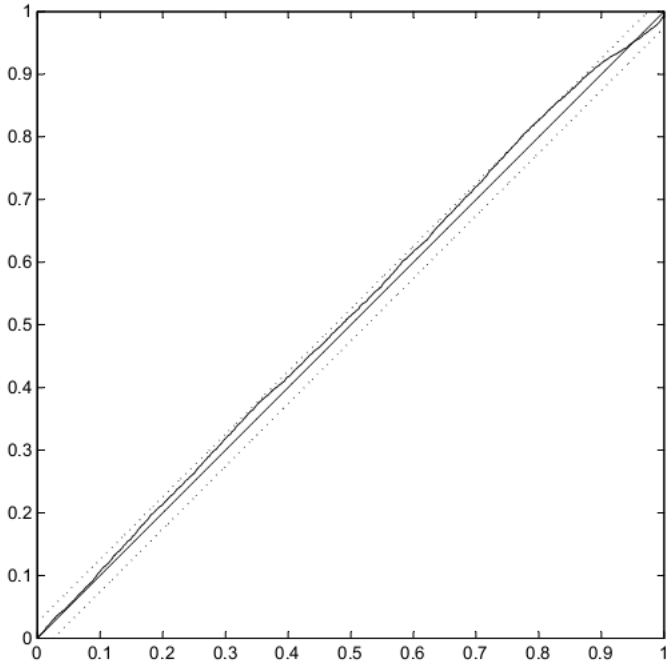


Figure 6.6: Fit of the Model Using Random Timestamps on Quotes - LSE

There are two remarkable facts when one compares estimates using the unchanged Thomson Reuters data with the randomized data: *i)* the parameters estimates for α and β are much smaller for the randomized data, and *ii)* the model provides a much better fit. The introduction of the random component into the timestamps smooths the data, providing a better fit. The question is: how to get reasonable estimates (and a good fit) without imposing such strong assumptions on the ordering of equal timestamps? A natural answer to that question is to round down the millisecond component of the timestamps of our data. We first round the millisecond component to get a timestamp with a centisecond (1/100 second) precision. This produces slightly (less than 5%) smaller estimates for α and β. We then round the centisecond component to get a timestamp with a decisecond component (1/10 second). Now we get values for α and β that are much smaller than the ones obtained before without changing μ and n that much. Actually, we get values for μ, α and β that are very

close to the values we obtained using randomized timestamps, but we get an estimate for n that is very close to the one obtained using the unmodified Thomson Reuters data. We present the results in Table 6.5.

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
Decisecond	0.0761	0.5214	0.8136	1.5605	0.1565	6.2258	4817
Standard Error	0.0020	0.0127	0.0352	0.0738	0.2420	14.7039	-

Table 6.5: Estimates Using Rounded Timestamps on Quotes - LSE

Nevertheless, the rounding of timestamps comes at a cost. As depicted by Figure 6.7, now the fit of the model is poor in the lower region of the goodness-of-fit plot.

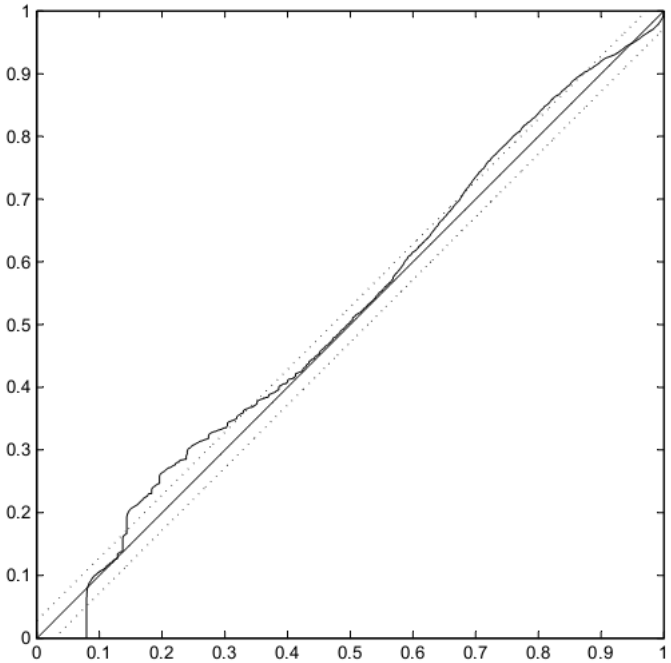


Figure 6.7: Fit of the Model Using Rounded Timestamps on Quotes - LSE

Another way to smooth the data without rounding the timestamps is by removing outliers. We do this by performing a Winsorisation of the data at the $95th$ percentile. If we perform the

Winsorisation using millisecond data we still get high estimates for α and β. Nevertheless, using millisecond data we get a very good fit of the model. We show here the estimates obtained when Winsorising the millisecond timestamps. The results for centi- and decisecond Winsorised timestamps are similar with the exception of α and β that become lower, like presented in Table 6.5. The goodness-of-fit is also presented in the case of millisecond timestamps. For centi- and decisecond timestamps the lower region of the goodness-of-fit plot crosses the KS bounds (not presented here).

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
Winsorised	0.1351	0.4759	2.7991	5.8814	0.2449	5.1724	5798
Standard Error	0.0027	0.0102	0.1046	0.2245	0.4989	13.6046	-

Table 6.6: Estimates using Winsorised Timestamps on Quotes - LSE

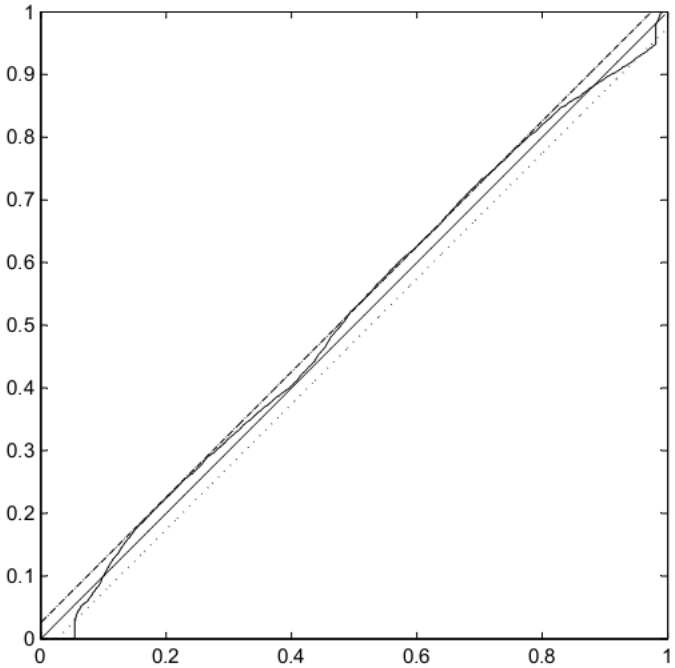


Figure 6.8: Fit of the Model Using Winsorised Timestamps on Quotes - LSE

We also estimate the model using data somewhat older than the data used in the main estimation presented before. Here we choose two trading days in January, 2006. Table 6.7 presents results for quotes. The incredibly small number of quotes for 2006 stems from the fact that, for this period, the quotes did not change as much as in 2009. So even if the initial dataset has a large number of quotes (which happens to be the case here), after considering only mid-price changes as events in the estimation, we end up with a very small number of quotes. Taking all mid-prices into the estimation does not change the estimates a lot (n is slightly smaller).

	μ	n	α	β	τ_m	Observations
2006-03-01	0.0021	0.6886	0.1608	0.2335	152.1555	196
Standard Error	0.0003	0.0622	0.0274	0.0383	546.5348	-
2006-06-01	0.0042	0.6011	0.2084	0.3467	94.7212	314
Standard Error	0.0004	0.0481	0.0341	0.0597	332.8917	-

Table 6.7: Estimates Using Thomson Reuters Data on Quotes - 2006

Trades Data. Switching to trades data seems to have a large impact on n and a relatively small impact on the other parameters. Here we have only 2070 observations (with mean duration of 14 seconds). The small number of observations makes it harder to estimate the model using the rolling window procedure. Even considering a window of 40 minutes, we have in some cases estimates that are based on less than 100 observations. Since the rolling windows estimation does not seem to bring new information, we choose not to perform such estimation here. Regarding the branching ratio, we observe in Table 6.8 that its value is now much lower than its value for quotes data. Since n is a measure of endogenous price

changes relative to all price changes, the conclusion is that the process for quotes has a higher endogeneity than the process for trades. The average duration is now much higher, as expected, with a value of $\tau_m = 14.4591$ seconds and a standard deviation $\sigma_\tau = 23.4307$ seconds. Simulated data produces an average duration of 14.4058 seconds and standard deviation of 18.3548 seconds.

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.0521	0.2395	1.2609	5.2643	0.0655	14.4591	2070
Standard Error	0.0014	0.0121	0.1076	0.4575	0.0970	23.4307	-

Table 6.8: Estimates Using Thomson Reuters Data on Trades - LSE

The interpretation of the branching ratio as the fraction of endogenously generate price changes among all price changes, as advocated by FS (2012), and the high values for n (specially for quotes and for the TAQ data) presented here, together with the increase in n in the last years as observed by FS (2012), makes it tempting to associate the higher proportion of endogenous price movements with the presence of High Frequency Traders. As documented in several works, High Frequency Traders started to trade a relatively small volume in the equity market around the 2000's to become the dominant force in equity trading nowadays. Much debate whether HFT is good or bad to market quality has been observed in both, the media and in the academia. How can we use the branching ratio as a measure of market quality? Is it good or bad to document a high (or increasing) branching ratio? While at a first sight a high proportion of endogenous market movements seems to be a bad thing (recall that in the efficient market hypothesis prices change only as a reaction to information), a high branching ratio could also be the reflection of a more thorough

price discovery process. The fact that n is much higher for quotes than for trades seems to corroborate this fact. Consider that an exogenous event happens and, as a result, increased activity in quotes of a given stock is registered. The more quotes are placed, as a result of the exogenous event, the higher the branching ratio will be. But this increased number of quotes might reflect simply a more thorough price discovery process among market participants that needs to place more quotes in order to agree into a "fair" price. Therefore, a high value for the branching ratio registered in quotes data is not necessarily an indication of poor market condition, as long as the branching ratio calculated from trades data remains relatively low, as it seems to be the case.

We have also calculated the fit of the model for trades. We see that the model fits the data relatively well.

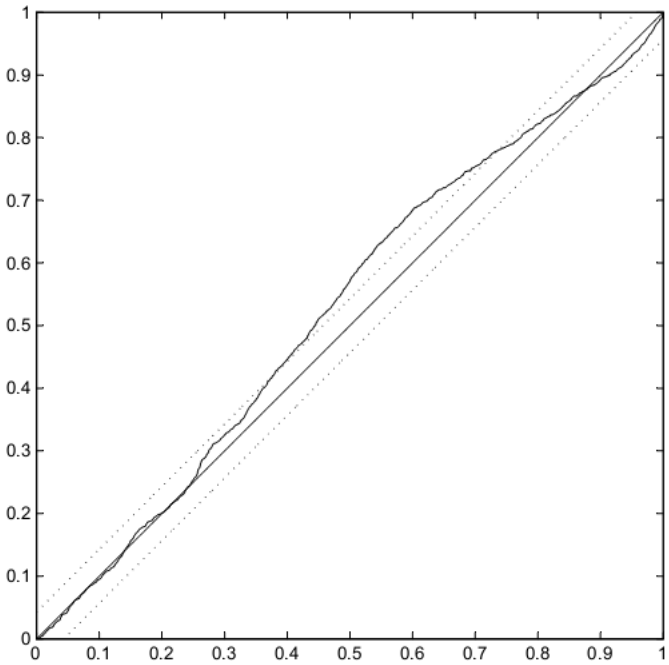


Figure 6.9: Fit of the Model Using Thomson Reuters Data on Trades - LSE

If we now perform the rounding of timestamps to the nearest second and then add a

random precision component to the rounded timestamps we get, qualitatively, the same results as obtained before for quotes. Both α and β are much smaller, but in a way that n stays relatively constant and is somewhat greater than using the original data. The value for μ hardly changes. The table below presents the results.

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
All obs.	0.0479	0.3007	0.3995	1.3285	0.0674	14.4593	2070
Standard Error	0.0014	0.0152	0.0321	0.1158	0.0750	23.3936	-

Table 6.9: Estimates Using Random Timestamps on Trades - LSE

The next table presents the results when the rounding of timestamps to a timestamp with a decisecond precision is performed, as well as results regarding the Winsorisation of millisecond timestamps. The situation is qualitatively the same as before. Rounding the data provides smaller estimates for α and β, while Winsorising the data produces a better fit (presented in Figure 6.10).

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
Decisecond	0.0497	0.2250	0.5093	2.2634	0.0627	15.4439	1938
Standard Error	0.0014	0.0133	0.0463	0.2198	0.0620	23.8990	-
Winsorised	0.0740	0.2222	1.3960	6.2819	0.0905	14.4591	2070
Standard Error	0.0019	0.0114	0.1119	0.4666	0.1090	23.4307	-

Table 6.10: Estimates Using Rounded and Winsorized Timestamps on Trades - LSE

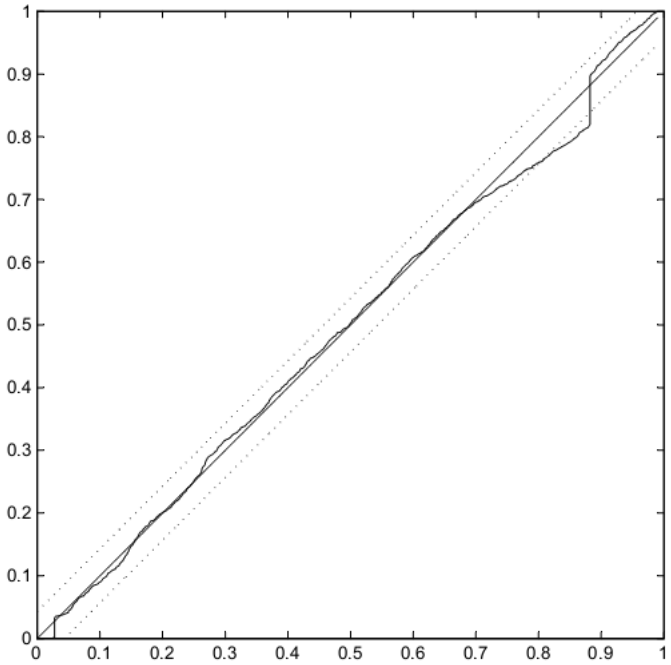


Figure 6.10: Fit of the Model Using Winsorised Thomson Reuters Data on Trades - LSE

As we did for quotes we also estimated the model using two trading days in the year of 2006. Table 6.11 presents results for trades. Again, the situation is similar to what was observed for quotes. The parameters α and β are extremely small and produce a branching ratio which is quite high. Recall that Figures 4.1 and 4.2 show that smaller values of α and β indicate less clustering, while a higher branching ratio indicates more clustering. The fact that α and β are much smaller than for 2009 data is indicating that the clustering of orders was less pronounced in 2006 even though the branching was higher (as can be inferred by Figures 4.1 and 4.2).

	μ	n	α	β	τ_m	Observations
2006-03-01	0.0224	0.6769	0.0056	0.0038	14.3932	2084
Standard Error	0.0051	0.0760	0.0009	0.0015	16.4448	-
2006-06-01	0.0119	0.8310	0.0048	0.0058	14.7324	2034
Standard Error	0.0030	0.0497	0.0008	0.0011	18.5308	-

Table 6.11: Estimates Using Thomson Reuters Data on Trades - 2006

6.2.2 BE - NYSE Euronext Brussels

We now estimate the model using quotes registered on the NYSE Euronext Brussles Exchange. The number of observations here is drastically decreased when compared to the LSE and therefore we estimate the model using only quotes data. Looking at the estimated parameters, the most striking difference is related to the parameters α and β. Both μ and n are close to the estimates for quotes using data from the LSE. Due to the low trade intensity (0.0235 events/second) the data presents a high average duration with $\tau_m = 42.4711$ seconds with a standard deviation of $\sigma_\tau = 63.164$ seconds. Simulated data produces average duration of 43.3455 seconds and standard deviation of 68.4741 seconds. Note that the results here are somehow related to the results using data from 2006, presented in Table 6.11. The branching ratio is high, but α and β are very low. As it was the case with the one-second timestamped TAQ data, the half-life of a shock is quite high with $t_{1/2} \sim 20$ seconds. Now, the model seems to fit the data very well. But given the small number of observations this could be induced by the lack of power of the Kolmogorov-Smirnov statistic. We do not present results of the estimation using timestamps rounded to centi- or decisecond because

	μ	n	α	β	$E(\lambda)$	τ_m	Observations
Quotes	0.0113	0.5191	0.0176	0.0340	0.0233	42.4711	707
Standard Error	0.0011	0.0460	0.0029	0.0065	0.0152	63.164	-

Table 6.12: Estimates Using Thomson Reuters Data on Quotes - NYSE Euronext Brussels

it does not influence the estimates in Table 6.12. Because the data displays less clustering (as confirmed by smaller values of α and β) than in the previous cases this is expected to be so. Winsorising the data also does not have a big impact in the estimates or the fit of the model.

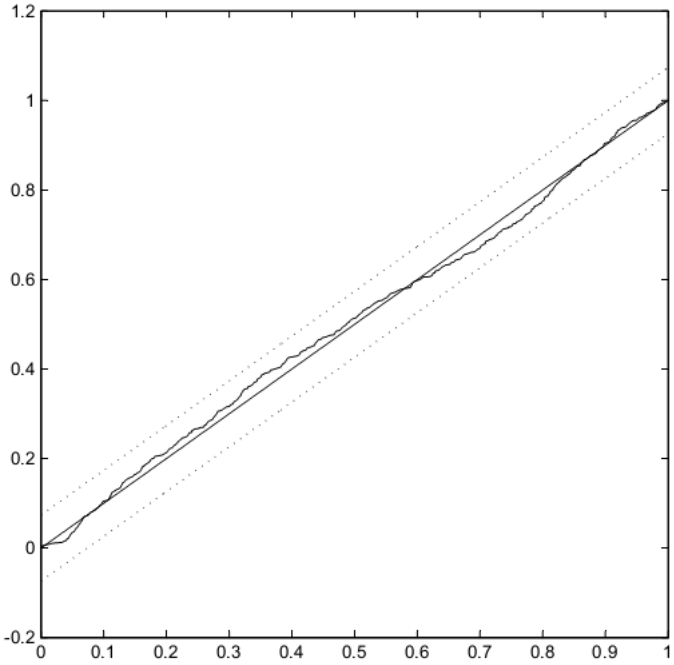


Figure 6.11: Fit of the Model Using Thomson Reuters Data on Quotes - NYSE Euronext Brussels

6.3 Hawkes' Process with a Weibull Response Function

As a robustness check we also fit the data using a Hawkes' process with a Weibull response function. The Weibull is a natural extension of the exponential kernel we used so far. Recall that the log-likelihood function of a Hawkes' process with an intensity $\lambda(t) = \mu + \sum_{u<t} g(t-u)$, is given by

$$\log L = -\Lambda(0, t_N) + \sum_{i=1}^{N} \ln \left[\mu + \sum_{t_j < t_i} g(t_i - t_j) \right], \tag{6.1}$$

where $\Lambda(0, t_N)$ is the compensator of the process and can be written as

$$\Lambda(0, t_N) = \sum_{j=1}^{N} \Lambda(t_{j-1}, t_j). \tag{6.2}$$

Now consider the p.d.f. of a Weibull (the case $\kappa = 1$ is equivalent to the exponential kernel)

$$g(t) = \left(\frac{\kappa}{\omega}\right) \left(\frac{t}{\omega}\right)^{\kappa - 1} \exp[-(t/\omega)^{\kappa}]. \tag{6.3}$$

We have to calculate the compensator

$$\Lambda(t_i, t_{i+1}) = \int_{t_i}^{t_{i+1}} \mu ds + \int_{t_i}^{t_{i+1}} \sum_{t_j < s} \left(\frac{\kappa}{\omega}\right) \left(\frac{s - t_j}{\omega}\right)^{\kappa - 1} \exp[-((s - t_j)/\omega)^{\kappa}] ds. \tag{6.4}$$

Before we tackle the integral above let's make a comparison with the standard case of a Hawkes' process with an exponential kernel. In this case the integrals we have to solve are

$$\Lambda(t_i, t_{i+1}) = \int_{t_i}^{t_{i+1}} \mu ds + \int_{t_i}^{t_{i+1}} \sum_{t_j < s} \alpha \exp[-\beta(s - t_j)] ds, \tag{6.5}$$

which are very similar (in fact, the first integral is identical) to the case of a Weibull kernel. To make this similarity even more clear note that we can make a change of variables in the

second integral of Equation (6.4). Let $\zeta = -((s-t_j)/\omega)^\kappa$, then $d\zeta = -ds(\kappa/\omega)((s-t_j)/\omega)^{\kappa-1}$ and Equation (6.4) can be rewritten as

$$\Lambda(t_i, t_{i+1}) = \int_{t_i}^{t_{i+1}} \mu ds + \int_a^b \sum_{t_j<s} e^\zeta d\zeta, \tag{6.6}$$

with $a = -((t_{i+1} - t_j)/\omega)^\kappa$, $b = -((t_i - t_j)/\omega)^\kappa$ and $\zeta = \zeta_j$. Then, after some algebra, the log-likelihood function for a Weibull kernel can be written as

$$\log L = -\mu t_N + \xi \sum_{i=1}^{N} \left\{ \exp\left[(t_N - t_i)/\omega\right]^\kappa - 1 \right\} + \sum_{i=1}^{N} \log\left[\mu + \xi U(i)\right], \tag{6.7}$$

where

$$U(i) = \sum_{t_j<t_i} \frac{\kappa}{\omega} \left(\frac{t_i - t_j}{\omega} \right)^{\kappa-1} \exp\left[-\frac{t_i - t_j}{\omega} \right]^\kappa \tag{6.8}$$

and the parameter ξ was inserted to control for the "strength" of coupling of the kernel function. If we take $\kappa = 1$ we get

$$\log L = -\mu t_N + \xi \sum_{i=1}^{N} \left\{ \exp\left[(t_N - t_i)/\omega\right] - 1 \right\} + \sum_{i=1}^{N} \log\left[\mu + \xi U_1(i)\right], \tag{6.9}$$

with

$$U_1(i) = \sum_{t_j<t_i} \frac{1}{\omega} \exp\left[-\frac{t_i - t_j}{\omega} \right]. \tag{6.10}$$

If we now compare the last two equations with the log-likelihood of a Hawkes' process with an exponential kernel, given by Equation (2.22), we construct the following mapping: $\xi = \alpha/\beta$ and $\omega = 1/\beta$. It follows that $\xi/\omega = \alpha$ and that the branching ratio is given by the parameter ξ. This last fact can also be seen from the definition of the branching ratio

$$n = \int_0^\infty \xi \left(\frac{\kappa}{\omega}\right) \left(\frac{t}{\omega}\right)^{\kappa-1} \exp(-t/\omega)^\kappa dt. \tag{6.11}$$

Solving this integral we get that $n = \xi$.

Estimates Using a Weibull Response Function

Using the log-likelihood function derived in the last section we estimate a Hawkes' process with a Weibull kernel using Thomson Reuters data for the LSE (trades only) and for the NYSE Euronext Brussels (quotes only). The results obtained are similar to the results obtained when the estimation was performed using an exponential kernel. For the LSE we get $\hat{\eta} = \left(\hat{\mu} = 0.0525, \hat{\xi} = 0.2333, \hat{\omega} = 0.1747, \hat{\kappa} = 1.0632\right)$ the standard errors are $\hat{\sigma}_{\hat{\eta}} = (0.0013, 0.0112, 0.0126, 0.0633)$. Recall that $\hat{\xi} = \hat{n} = 0.2333$, $\hat{\beta} = 1/\hat{\omega} = 5.7241$ and $\hat{\alpha} = \hat{\xi}/\hat{\omega} = 1.3354$. Comparing these values with the values presented in Table 6.10 we see that both models provide similar estimates (as it should be already clear by the fact that $\hat{\kappa} \sim 1$). Using data from NYSE Euronext Brussels we get $\hat{\eta} = \left(\hat{\mu} = 0.0107, \hat{\xi} = 0.5426, \hat{\omega} = 34.1487, \hat{\kappa} = 0.9141\right)$ the standard errors are $\hat{\sigma}_{\hat{\eta}} = (0.0008, 0.0339, 5.5349, 0.0862)$. Again, the value of κ is very close to one. Comparing with Table 6.12 we get very similar estimates (now $\hat{\alpha} = 0.0159$ and $\hat{\beta} = 0.0293$).

6.4 Relation with High Frequency Trading

FS (2012) presented evidence of an increase in the branching ratio over time. In their work, it was documented a branching ratio of approximately 0.2 at the beginning of their sample period in 1998, and a branching ratio of almost 0.8 at the end of their sample period in 2010. FS (2012) notes that the increase observed in the branching ratio coincides with the increase in HFT activity. Here we try to provide some more insight into this issue. Recall that we have previously calculated the parameters of the Hawkes' process using small intraday rolling windows of 20, 30 and 40 minutes. The results are presented in Table 6.3. If we

have a measure of intra-day HFT activity, we may compare the intra-day dynamics of the branching ratio as presented on panel (b) of Figure 6.4 with HFT activity to check if there is correlation between the two[3].

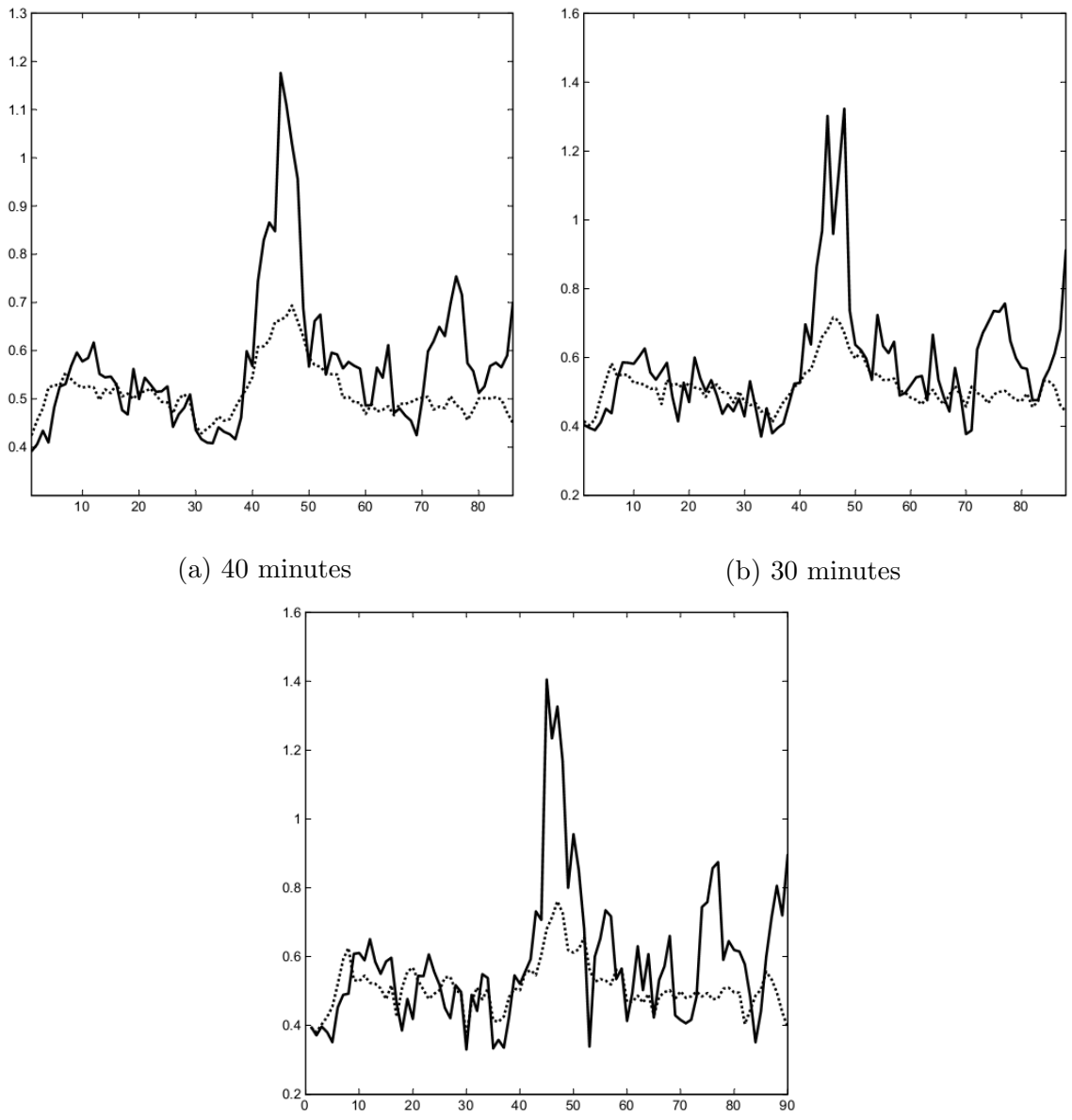
It is not easy to find a proxy for HFT activity when one has available only data on size, price and timing of trades/quotes. Thus, we decide to use a very "crude" measure of electronic trading activity, namely, we calculate the ratio of quotes per trades in a given time interval. Since many of the HFT strategies rely on placing quotes and quickly cancelling or modifying these quotes, the higher the number of quotes per trade the higher will be the activity of HFT [4].

Using intra-day intervals of the same length of those used in Table 6.3 and Figure 6.4 we calculate the quotes per trade ratio and compare these values against the branching ratio. Figure 6.12 shows the results, where the quotes per trade ratio is the solid line and its value was divided by 5 to give it a scale comparable to the scale of n.

Table 6.13 presents the estimates of a simple regression where n is taken as the independent variable and the quotes per trade (QPT) ratio, together with a constant, are use as dependent variables. The fact that, in the first specification, the QPT ratio has a positive coefficient indicates the positive relationship between HFT and market endogeneity. We also test a similar specification where the trades per quote ratio is used as independent variable. Obviously, we get a negative estimate for the coefficient of the TPQ ratio. Note that in this specification both the value of the TPQ ratio and the value of n are limited between zero

[3]Note that we do not intend to check whether there is a causal relationship between HFT activity and the branching ratio in this work.

[4]See, among others, `http://www.nanex.net/research/MsgRates/EquityMessageRates.html`



(a) 40 minutes

(b) 30 minutes

(c) 20 minutes

Figure 6.12: Branching Ratio (dotted line) and Quotes per Trade Ratio (solid line)

and one. Our last specification consists in taking logs of both variables. In all cases standard errors were calculated using the Newey-West estimator that is consistent to autocorrelation in the error terms[5]. Naturally, the evidence of a link between HFT and market endogeneity presented here is a relatively weak evidence, given all the data limitations and the simple specification of the regressions in which the results are based.

[5]We used 8, 6 and 4 lags for the time invervals of 40, 30 and 20 minutes, respectively.

	(I)			(II)			(III)			
	a_1	b_1	R^2	a_2	b_2	R^2	a_3	b_3	R^2	Observations
20 min.	0.3779	0.0456	0.4279	0.6579	-0.3907	0.3356	-0.9531	0.2665	0.3648	90
Standard Error	0.0265	0.0104	-	0.0512	0.1205	-	0.0668	0.0689	-	-
30 min.	0.3715	0.0492	0.5234	0.6907	-0.4087	0.4575	-0.9800	0.300	0.4853	88
Standard Error	0.0204	0.0083	-	0.0589	0.1416	-	0.0614	0.0665	-	-
40 min.	0.3437	0.0595	0.6333	0.7179	-0.5555	0.5497	-1.0273	0.3477	0.5865	86
Standard Error	0.0166	0.0071	-	0.0597	0.1441	-	0.0547	0.0617	-	-

This table shows the results of three different specifications of a regression of n on HFT activity. The first specification is $n_t = a_1 + b_1 QPT_t + \epsilon_t$, the second specification is $n_t = a_2 + b_2 QPT_t^{-1} + \epsilon_t$ and the third specification is $\ln n_t = a_3 + b_3 \ln QPT_t + \epsilon_t$. The standard errors, robust to heteroscedasticity and autocorrelation, are estimated using the Newey-West approach.

Table 6.13: Regression of n on HFT activity

Chapter 7

Conclusion and Future Research

In this work we fitted a special type of point process, known as Hawkes' process (Hawkes, 1971), to model duration of trades and quotes arrival in the equity market. The Hawkes' process is a very popular point process that has been widely applied in several different fields ranging from seismology to finance. The model is also relatively simple to interpret and estimate. The ability of generating endogenous clustering, controlled by the parameters α and β when the model has an exponential kernel, makes the Hawkes' process a good choice to model durations of price and mid-price changes in financial markets.

We have estimated a Hawkes' process on high-frequency data using two different databases. First, we modelled the duration of mid-price changes using the TAQ database. The fact that timestamps in the TAQ database are rounded to the nearest second poses a challenge in the estimation of the model since there is a large proportion of events that occur at the same second (roughly 30% of all mid-price changes occurred with equal timestamps in our TAQ dataset). Also, the estimation of the model using durations bounded by one second gives a poor model fit. A common solution to this problem is to add a uniform random number as an artificial precision component that disentangles events within the same second (Bowsher, 2007; FS, 2012). We implemented this procedure and observed that the model fits very well the randomized TAQ data. Nevertheless, we are not satisfied with the randomization of timestamps since this is equivalent to make very strong assumptions about the ordering of events.

The best robustness check one can perform to assess the impact of the randomization of timestamps in the estimates and in the fit of the Hawkes' process is use data that distinguishes between events at the same second. This check is conducted by constructing a randomized dataset from a "real" dataset that has a precision higher than one second. This randomized dataset is obtained by simply rounding to the nearest second, timestamps with high precision and adding a randomized precision component to the rounded data. We perform this check using data on European equities provided by Thomson Reuters. The Thomson Reuters data has timestamps with a millisecond component precision that distinguishes between same-second events and thus allows us to check directly how the randomization of timestamps affects the overall performance of the model.

Using data on the London Stock Exchange (LSE) for trades and quotes we showed that the randomization of timestamps introduces a relatively small bias in the estimates of μ (the underlying trading intensity) and n (the branching ratio, that represents the proportion of endogenously generated events to all events) and a large bias in the estimates of α and β, the parameters that control the decaying of the intensity function of the model. The randomization procedure also introduces some sort of smoothing in the data that leads to a better fit of the model when compared with the "crude" millisecond precision data. This fact shows that one should be cautious when deriving conclusions from the model estimated using randomized timestamps.

Inspired by FS (2012), who used the branching ratio derived from the Hawkes' process as a tool to predict flash-crashes, we would like to propose a tighter connection between the value of n and High Frequency Traders. As documented by FS (2012) the branching ratio has increased a lot in recent years. FS (2012) documents a branching around 0.3 till the

early 2000's, a value close to 0.5 in 2002 with a branching ratio above 0.6 after 2004. Even though the results obtained by FS (2012) used data on the E-mini S&P 500 contract and their estimation made use of the aforementioned randomization procedure, the leap observed in the branching ratio from 2000 to 2002 is a remarkable fact. Note that the rise in the value of the branching ratio coincides with the rise in HFT activity.

The finance literature has started to pay some attention to the effects of HFT in equity markets. Several papers like Brogaard (2010), Hasbrouck and Saar (2010), Zhang (2010), Hendershott, Jones and Menkveld (2011) and Hendershott and Riordan (2011) analyze how HFT is impacting several measures of market quality like liquidity, volatility and price discovery. The conclusions of many of these studies seem to associate HFT with increased liquidity, decreased volatility and more efficient price discovery. Nevertheless, some other studies like Zhang (2010) conclude the opposite. Also in the media there is a very fierce debate whether HFT are making markets more efficient, in the sense that prices are becoming more informative.

We tried to provide a first connection between HFT and market endogeneity, comparing the intra-day dynamics of the branching ratio with a "crude" measure of HFT activity defined as the ratio of quotes per trades. Even though it seems that there is some positive correlation between the branching ratio and HFT activity, a more formal treatment of the subject, pointing towards a causal link between the two, would necessarily involve the expansion of our dataset and a more robust approach to analyze the relationship between HFT activity and price endogeneity.

If we interpret the branching ratio, as derived from the Hawkes' process, as a measure of market quality, in the sense that a low branching ratio reflects a healthier market where

price changes are driven by exogenous events and a high branching ratio reflects a less healthier market where price changes are driven by "positive feedback mechanisms" and herding behavior, as proposed in FS (2012), then a more robust version of the regression we used here to exploit the relation between HFT activity and the branching ratio n could give a more satisfying answer to this question. This regression would necessarily incorporate, as additional variables, factors other than the HFT activity that are likely to impact the branching ratio.

Bibliography

BACRY, E.; DELATTRE, S.; HOFFMANN, M.; MUZY, J. F. (2011). Modelling microstructure noise with mutually exciting point porcesses, Proceedings of the ICASSP.

BERMAN, M. (1983) Comment on "Likelihood Analysis of Point Processes and Its Applications to Seismological Data," by Y. Ogata, Bulletin of the International Statistical Institute, 50, Book 3, 412-418.

BOWSHER, C. (2007) Modelling security market events in continuous time: intensity based, multivariate point process models, Journal of Econometrics, 141(2).

BROGAARD, J. A. High Frequency Trading and its Impact on Market Quality, Working Paper, Northwestern University, 2010.

CRANE, R.; SORNETTE, D. (2008) Robust dynamic classes revealed by measuring the response function of a social system, Proceedings of the National Academy of Sciences of the United States of America, 105.

DALEY, D. J.; VERE-JONES, D. (2003) An introduction to the theory of point processes. Volume I. Springer: Heidelberg.

DUFOUR, A.; ENGLE, R. F. (2000) Time and the price impact of a trade. The Journal of Finance, 55.

EASLEY, D.; O'HARA, M. (1992) Time and the process of security price adjustment, The Journal of Finance, 42.

ENGLE, R. F.; RUSSEL, J. R. (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data, Econometrica, 66, (5).

EPPS, T.W. (1979) Comovements in Stock Prices in the Very Short Run, Journal of the American Statistical Association, 74.

FILIMONOV, V.; SORNETTE, D. (2012) Quantifying reflexivity in Financial markets: towards a prediction of flash crashes, `arXiv:1201.3572v1.pdf`.

HASBROUCK, J. (1999) Trading fast and slow: security market events in real time. Working Paper, Stern School of Business. New York University.

HASBROUCK, J; SAAR, G. (2010) Low-latency trading. Working Paper, Stern School of Business. New York University.

HAWKES, G. A. (1971) Point spectra of some mutually exciting point processes, Journal of the Royal Statistical Society, 33(2).

HENDERSHOTT, T.; JONES, C. M.; MENKVELD, A. (2011) Does algorithmic trading improve liquidity?, The Journal of Finance.

HENDERSHOTT, T; RIORDAN, R. (2011) High frequency trading and price discovery, Working Paper, UC Berkeley.

HEWLETT, P. (2006) Clustering of order arrivals, price impact and trade path optimisation, Workshop on Financial Modelling with Jump Processes, Ecole Polytechnique.

KIRILENKO, A.; KYLE, A. S.; SAMADI, M.; TUZUN, T. (2011) The Flash Crash: The Impact of High Frequency Trading on an Eletronic Market, Working Paper.

LARGE, J. (2007) Measuring the resiliency of an electronic limit order book, Journal of Financial Markets, 10.

LO, A. W.; WANG, J. (2001) Stock market trading volume, Working Paper, MIT, Sloan School of Management.

OGATA, Y. (1978) The asymptotic behaviour of maximum likelihood estimators for stationary point processes, Annals of the Institute of Statistical Mathematics, 30(1).

OGATA, Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association, 83.

OZAKI, T. (1979) Maximum likelihood estimation of Hawkes self-exciting point processes, Annals of the Institute of Statistical Mathematics, 31(1).

RUSSEL, J. R. (1999) Econometric modeling of multivariate irregularly-spaced high-frequency data, Working Paper, University of Chicago.

SECURITIES AND EXCHANGE COMISSION. (2010) Concept release on market microstructure, Washington.

SHEK, H. (2011) Modelling high frequency order market dynamics using self-excited point processes, Working Paper, Stanford University.

SORNETTE, D.; DESCHTRES, F.; GILBERT, T.; and AGEON, Y. (2004) Endogenous versus exogenous shocks in complex networks: an empirical test using book sale ranking, Physical Review Letters, 93.

THOMPSON, W. A. (1988) Point process models with applications to safety and reliability, Chapman and Hall: New York.

ZHANG, F. (2010) The effect of high-frequency trading on stock volatility and price discovery, Working paper, Yale University.

Appendix A

Derivation of the Log-Likelihood Function of a Hawkes' Process

The log-likelihood function of a Hawkes' process with an intensity $\lambda(t) = \mu + \sum_{u<t} g(t-u)$ is given by

$$\log L = -\Lambda(0, t_N) + \sum_{i=1}^{N} \ln \left[\mu + \sum_{t_j < t_i} g(t_i - t_j) \right], \tag{A.1}$$

where $\Lambda(0, t_N)$ is the compensator of the process and can be written as

$$\Lambda(0, t_N) = \sum_{j=1}^{N} \Lambda(t_{j-1}, t_j). \tag{A.2}$$

Now consider the exponential response function

$$g(t) = \alpha e^{-\beta(t)}. \tag{A.3}$$

We have to calculate the compensator

$$\Lambda(t_{i-1}, t_i) = \int_{t_{i-1}}^{t_i} \mu ds + \int_{t_{i-1}}^{t_i} \sum_{t_j < s} \alpha e^{-\beta(s-t_j)} ds. \tag{A.4}$$

This integral was already solved and leads to (see equations (2.27) - (2.29))

$$\Lambda\left(t_{i-1}, t_i\right) = \mu(t_i - t_{i-1}) - \sum_{k=1}^{i-1} \frac{\alpha}{\beta} \left[e^{-\beta(t_i - t_k)} - e^{-\beta(t_{i-1} - t_k)} \right]. \tag{A.5}$$

Using Equation (A.2) we have

$$\Lambda(0, t_N) = \sum_{i=1}^{N} \mu(t_i - t_{i-1}) - \sum_{i=1}^{N} \sum_{k=1}^{i-1} \frac{\alpha}{\beta} \left[e^{-\beta(t_i - t_k)} - e^{-\beta(t_{i-1} - t_k)} \right]. \tag{A.6}$$

It is easy to see that $\sum_{i=1}^{N} \mu(t_i - t_{i-1}) = \mu t_N$, since μ is a constant and the summation over $t_i's$ leaves only the last term (assuming $t_0 = 0$): $(t_1 - t_0) + (t_2 - t_1) + ... + (t_{N-1} - t_{N-2}) + (t_N - t_{N-1}) = t_N$. The situation is similar to the summation over the exponentials

$$\sum_{i=1}^{N} \sum_{k=1}^{i-1} \left[e^{-\beta(t_i - t_k)} - e^{-\beta(t_{i-1} - t_k)} \right]. \tag{A.7}$$

Consider the $i = 2$ term

$$e^{-\beta(t_2 - t_1)} - e^{-\beta(t_1 - t_1)}, \tag{A.8}$$

for $i = 3$ we have

$$\left[e^{-\beta(t_3 - t_1)} - e^{-\beta(t_2 - t_1)} \right] + \left[e^{-\beta(t_3 - t_2)} - e^{-\beta(t_2 - t_2)} \right]. \tag{A.9}$$

Note that adjacent terms will cancel out leaving the terms that are equal to one (i.e. $e^{-\beta(t_j - t_j)}$ terms). For $i = N$ we finally have

$$\left[e^{-\beta(t_N - t_1)} - e^{-\beta(t_{N-1} - t_1)} \right] + \left[e^{-\beta(t_N - t_2)} - e^{-\beta(t_{N-1} - t_2)} \right] + ... + \left[e^{-\beta(t_N - t_{N-1})} - e^{-\beta(t_{N-1} - t_{N-1})} \right], \tag{A.10}$$

and terms of the form $e^{-\beta(t_N - t_i)}$ will remain. Therefore we can write $\Lambda(0, t_N) = \sum_{j=1}^{N} \Lambda(t_{j-1}, t_j)$ as

$$\mu t_N - \sum_{i=1}^{N} \frac{\alpha}{\beta} \left[e^{-\beta(t_N - t_i)} - 1 \right]. \tag{A.11}$$

Using Equation (A.1) we get the log-likelihood function for a Hawkes' process with exponential kernel

$$\log L = -\mu t_N + \sum_{i=1}^{N} \frac{\alpha}{\beta} \left[e^{-\beta(t_N - t_i)} - 1 \right] + \sum_{i=1}^{N} \ln \left[\mu + \sum_{t_j < t_i} \alpha e^{-\beta(t_i - t_j)} \right]. \tag{A.12}$$