Essays on High Frequency Econometrics

Thesis submitted
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Economics

by

Luca Mucciante
Department of Economics
Royal Holloway, University of London

December 2020

Contents

[1] This chapter is co-authored with Alessio Sancetta.

Declaration

I, Luca Mucciante, hereby declare that Chapters 1 and 3 of the thesis are my own research work. Chapter 2 comprises collaborative research with Alessio Sancetta: I contributed 50% of this work.

Name: Luca Mucciante

Signed:

Date:

Abstract

The thesis consists of four chapters.

Chapter 1 contains an empirical study of high frequency Bitcoin data. Bitcoin is the (first and) most important, in terms of market capitalization, cryptocurrency. Given its recent inception there is still a little literature about the statistical properties of this new type of financial asset. The chapter partially fills the gap analyzing a few basic stylized facts and some fundamental microstructural variables such as order flow and volume imbalance.

Chapter 2 introduces a statistical framework to model the intensity of the counting process representing the number of buy (or sell) order arrivals as additive functions of some covariates relative to the orders resting on the order book. The procedure allows to test whether those functions are increasing/convex and is suitable for high dimensional datasets. The methodology can be useful in order to flag “markets prone to be manipulated via high frequency spoofing algorithms”.

Chapter 3 extends to (a certain class of) counting processes the main results derived in Meinshausen (2013): the paper shows, for high dimensional linear regressions, that imposing a positivity constraint on the regression coefficients acts (under some circumstances) as a regularization technique comparable to the Lasso.

Chapter 4 suggests some possible extensions of the studies conducted in the previous chapters.

Acknowledgments

I would like to express my gratitude to Alessio Sancetta for his guidance throughout the PhD. I am also grateful to Graham Jones for kindly providing the Java script to collect Bitcoin data.

List of Figures

List of Tables

Introduction

Technological advancement has made it possible trading and recording financial data in real time. These High Frequency Data (HFD) are an exceptional source of information that allows to reconstruct the entire order flow and the underlying order book. Nevertheless, the statistical analysis of HFD poses two major challenges. First, HFD usually means very large dataset: the number of daily transactions can be greater than 100000. If on the one hand such large amount of data allows more precise statistical estimations, on the other hand, the computational cost of standard statistical techniques may become an issue, therefore it could be necessary to design a suitable statistical methodology. Second, unlike traditional low frequency financial time series, data are not equally spaced in time: for example, how can returns be computed? There are (at least) two ways to circumvent this difficulty and they, of course, can be applied to other financial variables (not only returns). The first possibility (clock time approach) is to fix a frequency then construct an equally time spaced grid: if for a given time in the grid does exist a record with the same timestamp then it is filled out in the obvious way otherwise the nearest predecessor is chosen. A complementary manner to cope with non equally time spaced time series is to perform an event time study: in this occasion the variable of interest is updated only when certain events have happened.

In Chapter 1 both strategies are applied to analyze high frequency Bitcoin data. Stylized facts are statistical qualitative properties common to a broad type of financial markets and instruments revealed by several years of empirical financial research. Bitcoin is a new type of financial asset, thus it is relevant to ascertain whether standard stylized facts apply. The principal outcome of the study is that returns, at short time scales, appear to be autocorrelated. This does not align with standard stylized facts, yet other studies, such as Zargar and Kumar (2019a), have reached the same conclusion.

The second chapter is dedicated to spoofing: an illegal trading strategy that makes profits misleading other market participants about the true imbalance between supply and demand in the order book. Typically, a spoofing algorithm places a relatively small buy order on the best bid, and almost contemporaneously it places a sequence of relatively large sell orders on the ask side of the book. This action provides a snapshot of the demand and supply schedule, where the market appears to be willing to sell. The reason is that there are considerably more orders to sell than to buy. This will often induce a trader (usually another algorithm) to place a sell order that crosses the bid-ask spread. In consequence the small limit order placed by the spoofing algorithm on the best bid will be filled. Once this happens, the spoofing algorithm will cancel all the large limit orders placed on the ask side of the book. The game then repeats reversing the role of the two sides of the book: a small resting order on the ask, and relatively large orders on the bid. The whole procedure lasts less than one second and as a result the manipulator gains the spread.

The core idea of the statistical methodology discussed in Chapter 2 is to model the intensity of the counting process representing the number of buy (or sell) arrivals as a function of some covariates, e.g., the volume imbalance, that the manipulator can affect in order to create a fictitious buying (or selling) pressure: if these functions turn out to be increasing and convex the market is considered "prone to be manipulated". Essentially, the statistical framework is a non-parametric one and allows to impose monotone and convex constraints in addition to the obvious non negative constraint (the intensity of a counting process is always non negative).

The empirical study in Chapter 2 reveals that crude oil futures could be a profitable market for a spoofer. Moreover, as a byproduct of the empirical analysis, we observe that the non-negativity constraint on the intensity leads to some form of regularization. Indeed, the intuition behind this phenomenon can be found in results by Meinshausen (2013) for signed constrained linear regression: the article shows that under appropriate circumstances, a non negative constraint is a regularization technique as powerful as the well known Lasso and is easier to implement because it does not require the specification of any tuning parameter. Chapter 3 extends to counting processes the main results of Meinshausen (2013), thus providing a theoretical justification of the empirical findings (of Chapter 2). In particular, two main technical assumptions (the same used in Meinshausen, 2013) are made: the *Compatibility Condition* (borrowed from the Lasso literature) and the *Positive Eigenvalue Condition* (introduced in Meinshausen, 2013).

The final chapter discusses possible generalizations and extensions of the results obtained in the first three chapters.

Chapter 1

The Information Content of Bitcoins Order Book and Trades

Abstract. Cryptocurrencies are a new type of financial asset. Bitcoin is the most important, in terms of market capitalization, cryptocurrency. Despite the increasing interest in this new market both from investors and regulators there is still a little literature about the empirical properties of the pair Bitcoin/USD. This paper aims to shed some light in this respect. Unlike most of the existing literature the study is based on high frequency Bitcoin data: some fundamental stylized facts are analyzed both in clock time (at different time scales) and in event time. Additionally a few characteristics of the volumes resting on the order book (ask/bid volume, volume imbalance, order flow) and of the market order arrivals are discussed.

1.1 Introduction

1.1.1 High Frequency Financial Econometrics

This paper analyzes high frequency Bitcoin data, therefore it belongs to the realm of high frequency empirical studies. In this framework the dynamics of the price process is described via a continuous time semimartingale. One of the most relevant characteristic (see next subsection) of the Bitcoin market is its large price movements: this raises a fundamental question in high frequency financial econometrics, i.e., whether jumps should be included in the price dynamics of the pair Bitcoin/USD (in other words, if the semimartingale representing the price process exhibits discontinuous trajectories).

Indeed, according to Scaillet et al. (2018) the largest fluctuations in prices are caused by the presence of jumps in the high frequency dynamics of the fundamental price of Bitcoin. More explicitly, according to that paper, the logarithm of the fundamental price at time t, say B_t, satisfies (except for a drift term) the following stochastic differential equation:

$$dB_t = \sigma dW_t + Y_t dJ_t \tag{1.1}$$

where W_t is a Brownian motion, J_t is a jump counting process, Y_t is the size of the jump (at time t) and σ is the diffusion parameter. Detecting the presence of jumps in the fundamental price of an asset is of paramount importance in high frequency financial econometrics, e.g., to design a ("jump robust") consistent estimator for the volatility. In order to detect jumps different testing methodologies have been proposed: for a comparison of the tests available in the literature see, e.g., Dumitru and Urga (2012) and Maneesoonthorn et al. (2020). Scaillet et al. (2018) adopt the test introduced in Lee and Mykland (2012).There are at least two additional difficulties (common to high frequency econometrics and in particular) when designing those tests: first, (the logarithm of) the fundamental price B is contaminated by the microstucture noise, therefore it is observed a different price $\tilde{B}$ rather than B. Second, variables can only be measured at discrete times, say $t_0 < t_1 < \ldots < t_N$. As a result the observed prices are $\tilde{B}_{t_0}, \ldots, \tilde{B}_{t_N}$ and for $i = 0, \ldots, N$

$$\tilde{B}_{t_i} = B_{t_i} + U_{t_i} \tag{1.2}$$

where U_{t_i} is a random variable modelling the effect of the market microstructure noise such as bid-ask spread, tick size, transaction costs, etc., see e.g., Black (1986) for a general discussion of noise in financial markets and Chapters 2 and 7 in Aït-Sahalia and Jacod (2014) for further technical details. According to Maneesoonthorn et al. (2020) the test introduced in Lee and Mykland (2012) (together with the one proposed in Aït-Sahalia et al., 2012) is the best choice when "microstructure noise alone is thought to be present". In addition, the test continues to perform "well when microstructure is absent, but only when the sampling frequency remains very high, the price jump size is large and volatility jumps are absent".

Another central issue in high frequency econometrics is that data are not equally spaced in time: there are (at least) two possible approaches to study them. One is to form an equally spaced time grid (i.e., fix a sampling frequency) and use the available data to fill it out (clock time analysis), a second possibility is to update the variable of interest only when an event that changes it occurs, see, e.g., Chapter 3 in Hautsch (2012). In particular the first approach raises the question: what is the optimal sampling frequency? To fix ideas consider the estimation of the quadratic variation of the price process. There are two opposite tendencies: from one hand increasing the sample frequency produces a more accurate estimator (given the larger sample size) for the quadratic variation, on the other hand at very high frequency the estimator is subject to the effect of the market microstructure noise therefore it is biased. Aït-Sahalia et al. (2005) conclude that it is optimal to sample "as often as possible provided one accounts for the presence of the noise when designing the estimator". In particular, Zhang et al. (2005) estimate the integrated volatility combining two estimators: the first uses all the data available, whereas the second is based on a 5 minute sampling. More generally, several estimators for the quadratic variation have been developed: a comprehensive comparison between them is contained in Liu et al. (2015). That study includes "jump-robust" estimators as well and concludes that, overall, when the 5-minute realized volatility is the benchmark estimator there is little evidence that more sophisticated estimators outperform it in terms of estimation accuracy. If the 5-minute realized volatility is no longer the benchmark and the best model is selected via the "model confidence set" (Hansen et al., 2011) the 5-minute realized volatility is outperformed by a few estimators. Among them there are: 1-minute sub-sampled realized volatility, 1- and 5-second realized kernels (Barndorff-Nielsen et al., 2008) and 1- and 5-second multiscale realized variance (Zhang, 2006).

1.1.2 The Bitcoin Market: Salient Features

Unlike traditional currencies cryptocurrencies are not issued by a central bank, instead they rely on a decentralized peer-to-peer network of users that transact digital tokens among them. All these transactions are validated via cryptographical protocols and are kept track of in a transparent database that is accessible to every user, the so called blockchain. In order to maintain such a public ledger an expensive computer network is needed: people who make available their computing power are rewarded through a proportional amount of cryptocurrencies, in this way new money is issued (mining).

The first and most important digital currency is, in terms of market capitalization, Bitcoin. It was introduced in 2008 by an anonymous programmer (or group of programmers) known as Satoshi Nakamoto. Since then it has experienced an increasing interest both from investors and regulators, yet there is still a little literature about the empirical properties of such market. This paper aims to shed some light in this respect.

From the existing literature two main abnormalities of Bitcoin returns emerge. First they have tails heavier than usual stocks or fiat currencies across several time scales ranging from 1 minute to 1 day (Begušić et al., 2018), this reflects the fact that since its inception the Bitcoin market has undergone several price bubbles culminating in just as many market crashes (Gerlach et al., 2018). Nevertheless, the decay of the tails is fast enough so that returns admit a finite second moment (Begušić et al., 2018). The second anomaly concerns the size of the volatility: it is six to seven times larger than the G10 fiat currencies (Osterrieder and Lorenz, 2017). A detailed study of Bitcoin volatility is contained in Shaw (2017), see also Lahmiri et al. (2018).

1.1.3 Related Literature

One of the main focus of this article is the analysis of stylized facts. Stylized facts are statistical qualitative properties common to a broad type of financial markets and instruments revealed by several years of empirical financial research (see Section 1.3). Many studies about the stylized facts of the Bitcoin market are low or medium frequency studies: Urquhart (2016), Bariviera et al. (2017), Caporale et al. (2018), Zhang et al. (2018a) and Zhang et al. (2018b), Aggarwal (2019) recover some basic stylized facts such as negative skewness, high kurtosis of returns (fat tails) and volatility clustering, yet they do not confirm the standard stylized fact according to which returns are not correlated.

More recently, several high frequency studies have appeared. Sensoy (2018), Zargar and Kumar (2019a), Zargar and Kumar (2019b), focus on the efficiency of the Bitcoin market: they use different techniques to prove its inefficiency that can be exploited by intraday algorithmic traders (Fisher et al., 2019). Nevertheless, Schnaubelt et al. (2019) do not find any significant autocorrelations of returns except for the first lag when the time scale is one minute long. Alvarez-Ramirez et al. (2018) conclude that the Bitcoin market at different time scales (day, hour, second) “exhibits periods of efficiency alternate with periods where the price dynamics are driven by anti-persistence”. In the high frequency literature, there is a general agreement about the volatility clustering phenomenon and the fact that the distribution of returns shows fat tail, whereas there are conflicting findings about the skewness of returns: this statement is valid not only for the Bitcoin market (cf., e.g., Eross et al., 2019, Zargar and Kumar, 2019a, Zargar and Kumar, 2019b, Schnaubelt et al., 2019) but it can be extended to the whole literature about stylized facts for high frequency data (cf. Section 1.3). Eross et al. (2019) find that returns (computed using a 5-min time grid) are negatively skewed in years 2015, 2016, 2017 and positively skewed in 2014. According to Zargar and Kumar (2019a), Zargar and Kumar (2019b) returns are positively skewed at the shortest time scale analyzed (15-min) while they are negatively skewed at any other time scale considered (30-min, 60-min, 120-min): their data sample ranges from 21st Jan 2013 to 8th Jan 2018. The empirical study in Schnaubelt et al. (2019), comprising data from 2 December 2017 to 12 October 2018, leads to the conclusion that “daily returns are slightly skewed to the left, minutely returns are slightly skewed to the right”.

Schnaubelt et al. (2019) is the only paper in the literature containing a (carefully) study of the order book of the Bitcoin market and reaches three main conclusions: the order book is relatively shallow with quick rising liquidity costs for larger volumes, many small trades occur, the limit orders distribution extends far beyond the current mid price.

The intraday patterns of trading activity are investigated in Wang et al. (2020a) and Eross et al. (2019): they analyze the Bitstamp exchange and find that the market is mainly driven by european and north american investors. More in details, “volume increases throughout the day and falls from around 2 pm until midnight, which is consistent with the intraday patterns found in currency markets” (Eross et al., 2019). The distribution of the intraday trading volume resembles a reversed V-shaped pattern (Wang et al., 2020a) or an inverted U-shaped pattern (Eross et al., 2019). In addition, Eross et al. (2019) find a positive correlation between volume and volatility and a negative one between returns and volatility.

Feng et al. (2018), Lennart (2020) and Wang et al. (2020b) discuss informed trading in the Bitcoin market. Feng et al. (2018) introduce a new indicator, tailored for cryptocurrencies, in order to detect information based trades: using it they “find evidence of informed trading in the Bitcoin market ahead of cryptocurrency-related negative Bitcoin market events, and ahead of large positive events”. Lennart (2020) confirms the presence of informed traders and concludes also that “abnormal trading volume negatively correlates with the degree of information asymmetry associated with transactions”. Wang et al. (2020b) reach different conclusions. In particular, they relate the autocorrelation of daily returns to the presence of uniformed traders, while informed traders are able to significantly reduce the volatility during “bull market” periods.

1.1.4 Contributions and Structure of the Paper

The contribution of the present analysis to the study of high frequency Bitcoin data is threefold. First, unlike the existing literature that adopts solely a clock time approach, we study the stylized facts listed above both in clock time (for five different time scales: 1 sec., 30 sec., 1 min., 5 min., 30 min.) and in event time: while negative skewness of returns, fat tails in the returns distribution and volatility clustering are strongly confirmed, returns appear to be autocorrelated at every time scale and for more than one lag. The same conclusions hold true when the event time approach is adopted. Second, it is one of the few studies, such as Schnaubelt et al. (2019), that deals with the order book of the Bitcoin market: the analysis includes variables directly linked to the orders resting on the order book such as bid-ask spread, volume imbalance, order flow. Third, it is the only study trying to quantify the probability of informed trades via the novel methodology introduced in Duarte et al. (2020) and analyzing the intraday trading volume using an approach tailored for high frequency econometrics (counting process). In particular, the traded volume is assumed to be proportional to the number of trade arrivals (given the short time scale this is an acceptable assumption) that is modelled via a counting process: the intensity of this counting process is estimated in a non parametric way, therefore in order to avoid overfitting the regularization technique introduced in Alaya et al. (2015) is adopted.

The rest of the paper is organized as follows. The next section presents the data and the methodology adopted to investigate them. Section 3 and Section 4 are dedicated to the study of some stylized facts and statistical properties of volumes and (best) prices resting on the order book. The intraday volume profile is studied in Section 5, while Section 6 investigates the probability of informed trading.

1.2 Data and Their Treatment

1.2.1 Data Description and Data Cleaning

The analysis of Bitcoin market is conducted using a data sample ranging from November 14, 2015 to December 31, 2016. The data were collected live via the internet from the Bitstamp exchange[1]. Each record comprises twenty-four fields: timestamp, sign of trade, trade volume, trade price (to USD) and the first five levels of quotes (to USD)/volumes size for both sides of the order book (in total twenty observations). From each working day (Monday to Friday) of this dataset we extract the records relative to the hours interval 8-18 (except in Section 1.5 where the whole time window 0-24 is considered): in fact, from a brief study of the traded volume, this period of the day seems the most active. Not all the collected data are reliable: days containing, in the time window 8-18, less than 10000 observations are deleted from the dataset. The final dataset comprises 276 days.

The Java script for collecting the data sometimes produced multiple records with the same timestamp. Trade prices having the same timestamp are reduced to a single trade price through a weighted average (using the traded volumes as weights). For the other variables (mid price, aggregate log volume and so on) we consider the latest value if multiple records have the same timestamp.

Buy and sell trades are classified using the tick rule: a trade is a buy (sell) order if the trade price is greater (lower) than the mid price. The case in which the trade price is equal to the mid price is very rare, in fact, this event concerns 140 cases out of 48293, so we have classified such trades simply as buy trade.

[1] The data have been collected live using a Java script written by Graham Jones. I am grateful to Alessio Sancetta for making the data available to me.

1.2.2 Variables of Interest and Analysis Methodology

Traditional low frequency financial econometrics is based on sampling the variables of interest at regular time intervals. High frequency financial data analysis deals with non equally time spaced time series, so as a first step we need to define the time variable, i.e., we have to specify when to update a given variable. The first possibility (clock time approach) is to fix a sampling frequency, in our study 1 second, 30 seconds, 1 minute, 5 minutes, 30 minutes. Then, for each frequency, an equally time spaced grid is constructed. If for a given time in the grid does not exist a record with the same timestamp then the nearest predecessor is chosen. This approach is carried out for the following variables: trade price returns, mid price returns, aggregate log volume, spread. This methodology reduces the impact of microstructure effects (overall at longer time scale), yet it ignores part of the information, i.e., the time duration between two consecutive updates. In order to model those time durations Engle and Russell (1998) introduced the so called Autoregressive Conditional Duration (ACD) model. Since that seminal paper numerous generalizations of the ACD model have appeared in the literature, see, e.g., Bhogal and Ramanathan (2019), Pacurar (2008) for a review and Chapters 11, 12 in Hautsch (2012). On the one hand the use of non equally time spaced data allows to gain more information, but on the other hand it makes more challenging the statistical analysis. Consider, for example, the estimation of the quadratic variation of the price process: in the general case, the adoption of endogenous and non deterministic sampling makes more difficult the study of the asymptotic distribution of the realized volatility, indeed, neither Gaussian approximations nor symmetry properties can be used (Fukasawa and Rosenbaum, 2012). Fukasawa (2010a) studies the asymptotic behaviour of the realized volatility when the sampling times are given by hitting times of a regular time grid while Fukasawa (2010b) and Li at al. (2014) derive central limit theorems in a more general setting (endogenous random sampling) under different technical assumptions. An alternative way to estimate the integrated volatility of a jump-diffusion process with stochastic volatility is introduced in Andersen et al. (2008): it is based on the theory of Brownian passage times and is robust to market microstructure noise. This latter approach (price duration estimators) has been recently revisited and improved, both in the parametric and nonparametric framework, in Hong et al. (2020):

the simulations carried out by the authors show that the non-parametric price duration estimators have the same accuracy as the best realized volatility type estimators, while the parametric price duration estimator significantly outperforms all realized volatility type estimators.

We perform an event time study as well: in this occasion we update the variable of interest only when certain events have happened. In the following we specify for which variables this approach is meaningful and what is/are the event/s that we take into account.

- *Trade price returns.* An event occurs if the current trade price is different from the previous trade price.
- *Mid price returns.* An event occurs if the current top ask (or bid) price is different from the previous ask (or bid) price.
- *Log Ask Volumes.* An event occurs for the i^{th} level aggregate log ask volume if one of the current ask volume up to the i^{th} level is different from the corresponding previous ask volume up to the i^{th} level.
- *Volume imbalance.* An event occurs, for the volume imbalance at level i, if the current i^{th} ask (or bid) volume, is different from the previous ask (or bid) volume at the same level.

Let us define the variables listed above. Trade price returns (or changes) R_t^{trade} and mid price returns (or changes) R_t^{mid} at time t are given by:

$$R_t^{trade} = P_t - P_{t-1}$$

$$R_t^{mid} = M_t - M_{t-1}$$

where: M_t, P_t are the mid price (the mid price at time t is given by the arithmetic average between top bid and top ask quote) and the trade price at time t respectively. Given $i = 1, \ldots, 5$ the i^{th} level aggregate log ask (or bid) volume at time t is defined as $log(X_t^{(1)} + \ldots + X_t^{(i)})$ if $X_t^{(j)}$ is the ask (or bid) volume size at level j and time t. The volume imbalance at time t and level k, $k = 1, \ldots, 5$, is given by

$$\frac{bidSize_t(k) - askSize_t(k)}{bidSize_t(k) + askSize_t(k)}$$

where $bidSize_t(k)$ ($askSize_t(k)$) is the volume bid (ask) size at level k. Finally for the order flow, we do not construct an artificial time grid nor consider an event time approach. We compute the five minute order flow. We do so computing the signed volume over non-overlapping five minutes time intervals, i.e., the order flow after t minutes $\mathcal{O}_t$ ($t \in \{5, 10, \ldots\}$) is given by $\mathcal{O}_t = \sum_{s \in Q_t} V_s$ where V_s is the signed volume at time s (positive in case of a buy trade, negative in the opposite case) and Q_t is the set of timestamps belonging to the time interval $(t, t-5]$.

Our study focuses on two kinds of statistical properties of the time series involved: linear dependence and their distributions, in particular we are interested in establishing their departure from normality. To assess linear dependence we use the Ljung-Box test with 10 lags, while normality is ascertained through the Jarque-Bera test.

1.3 Stylized Facts

More than fifty years of empirical financial research have revealed the existence of some statistical qualitative properties common to a broad type of financial markets and instruments, such properties are known under the name of stylized facts. For a general review of these stylized facts and the econometric techniques usually employed see Cont (2001) and Pagan (1996), inter alia. In the framework of high frequency econometrics Guillaume et al. (1997) focus on the analysis of intraday stylized facts relative to foreign exchange markets (cf. also Chapter 5 in Gençay et al., 2001), whereas Caporin et al. (2015) is dedicated to the study of precious metals. Chakraborti et al. (2011) discuss also some empirical facts relative to the limit order book. The following stylized facts are an excerpt from those listed in Cont (2001): they coincide with those listed in Caporin[2] et al. (2015) and appear also in Guillaume et al. (1997) and Chakraborti et al. (2011) except for the asymmetry of returns.

- **Absence of autocorrelations:** Returns are uncorrelated except for the first time lag at the highest frequencies.

[2] Actually Caporin et al. (2015) find out that the asymmetry can be positive or negative, whereas we assume that "asymmetry" stands for negative asymmetry. Notice also that in the FX markets "asymmetry" does not seem to be a stylized facts (cf., Guillaume et al., 1997 and the footnote on page 224 in Cont, 2001).

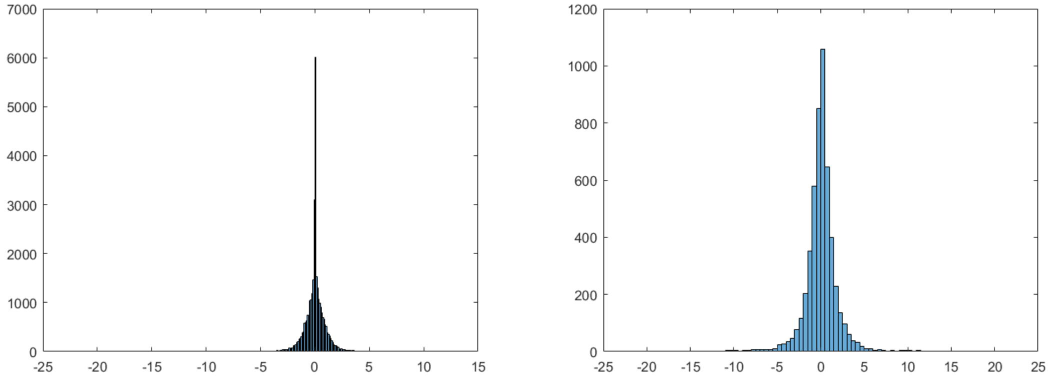


Figure 1.1: Histograms relative to trade price returns. Left: 30 sec. frequency. Right: 30 min. frequency. Even from an eye inspection the distributions do not resemble a normal one.

- **Fat tails:** The kurtosis of the unconditional distribution of returns is higher than a normal distribution, that is extreme events are more frequent than would be expected under a normal distribution.
- **Asymmetry:** The unconditional distribution of returns has negative skewness, that is extreme negative returns are more frequent than extreme positive returns: this is another evidence against normality assumption of the distribution of returns.
- **Aggregated normality:** Increasing the time scale over which returns are calculated, the distribution of returns resembles a normal one.
- **Volatility clustering:** Volatility of returns exhibits a positive autocorrelation, this means that a large absolute return tends to be followed by another large absolute return.

The rest of the section is dedicated to the analysis of the stylized facts listed above both for trade price returns and mid price returns.

1.3.1 Trade Price Returns

Figures 1.1 and Figure 1.2 represent Bitcoin returns. Table 1.1 summarizes the stylized facts discussed above. The skewness is negative at every time scale, the tails are heavier than the normal distribution and the Jarque-Bera test confirms the non normality. Figure 1.3 displays the autocorrelation function relative to two different frequencies (1 sec. and 30 min.), the bands represent the significance level at 5%. Roll's model may explain the first order dependence of returns. According to that model the observed price at time t, say P_t, is given by

$$P_t = P_t^* + \frac{s}{2} I_t$$

where P_t^* is the fundamental price at time t, s is the bid-ask spread and I_t are i.i.d. random variables indicating whether the transaction is buyer initiated or seller initiated, more explicitly

$$I_t = \begin{cases} 1 & \textit{with probability } 0.5 \ (\textit{buyer initiated}) \\ -1 & \textit{with probability } 0.5 \ (\textit{seller initiated}). \end{cases}$$

It can be shown that the correlation between two consecutive price changes is negative (cf. Chapter 3 in Campbell et al., 1997 for an intuitive explanation or Chapter 6 in de Jong and Rindi, 2009 for a more thoroughly discussion). Nevertheless Roll's model cannot justify any higher order linear correlation displayed by the time series: the autocorrelations are not strong but note that the p-value of Ljung-Box test (10 lags) is less than 1%. This phenomenon seems to be confirmed by other empirical studies (inter alia Zargar and Kumar, 2019a, Zargar and Kumar, 2019b, Sensoy, 2018, Caporale et al., 2018, Barivieria et al., 2017 and Urquhart, 2016) and it could be due to the prevalence of retail traders in the market (cf., e.g., Wang, 2020b). Volatility clustering is apparent from Figure 1.4. The persistence of the autocorrelations (of squared returns) strongly depends on the length of the time scale: from a detailed analysis emerges that if the time scale is 30 sec. then the autocorrelation is significant (5% level) even for more than 2000 lags, instead if the time scale is 30 min. the significant lags are about 200. Additional plots are contained in Appendix 1.8.

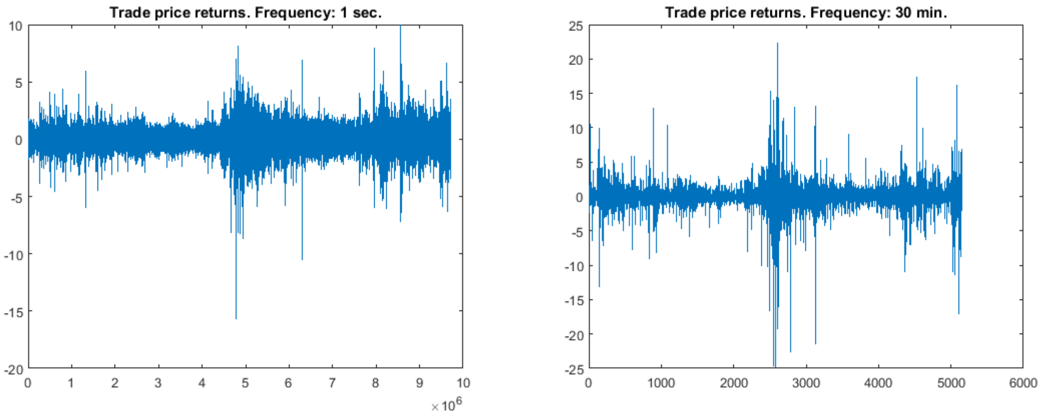


Figure 1.2: Trade price returns at different time scales. Increasing the time scale returns tend to be larger (in absolute value).

Time scale	Skewness	Kurtosis	p-val J-B test	p-val L-B test
1 sec.	-0.3714	368.2615	< 0.01	< 0.01
30 sec.	-0.2724	29.3286	< 0.01	< 0.01
1 min.	-0.2682	19.7957	< 0.01	< 0.01
5 min.	-0.6159	24.1563	< 0.01	< 0.01
30 min.	-0.9759	29.0751	< 0.01	< 0.01

Table 1.1: Skewness, kurtosis, p-value of Jarque-Bera test, p-value of Ljung-Box test relative to trade price returns. Returns are negatively skewed (especially at low frequency), display fat tails, appear to be autocorrelated and not normal at each time scale.

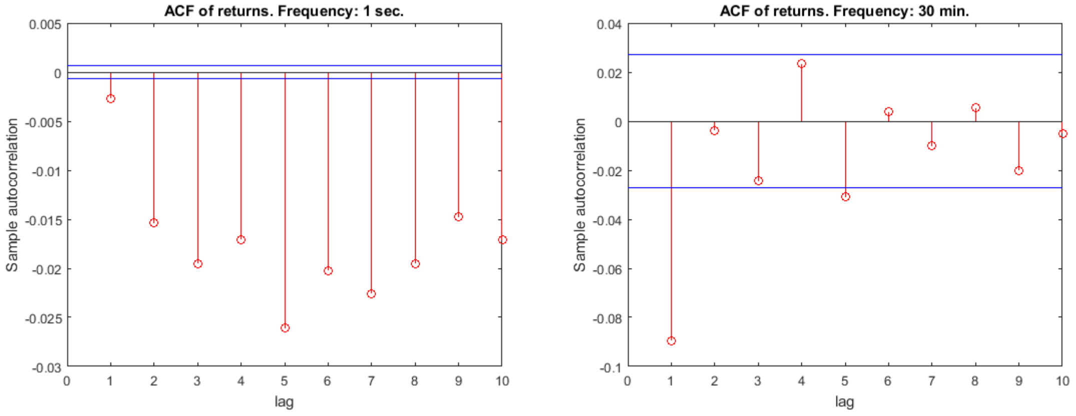


Figure 1.3: ACF of trade price returns at different time scales. Horizontal lines indicate the 95% confidence interval. Increasing the time scale the number of significant lags decreases because the effect of the microstructure noise vanishes.

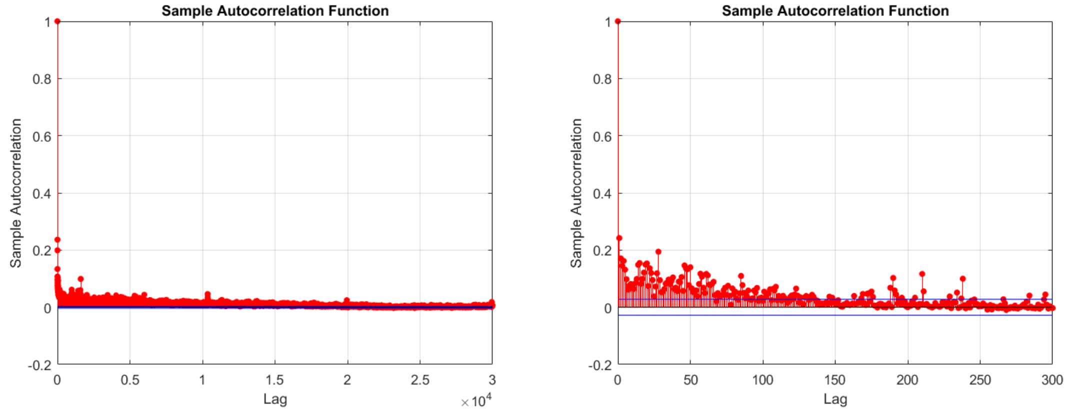


Figure 1.4: ACF of squared trade price returns at different time scales: 30 sec. (left) and 30 min. (right). Horizontal lines indicate the 95% confidence interval. Volatility clustering is apparent. In addition, increasing the time scale the number of significant lags decreases.

As discussed in Section 1.2.2 it is meaningful to consider trade returns in event time: the relative time series is shown in Figure 1.5, its distribution does not resemble a normal one (see the histogram in the same figure). In fact, although skewness is close to zero, kurtosis is large (about 11) and the p-value of the Jarque-Bera test is less than 1%. First order autocorrelation is evident. Higher autocorrelations are also significant at 5% level, but appear to be weak: in the best case, about five times weaker than the first order autocorrelation (the p-value of the Ljung-Box test is less than 1%). Squared returns are strongly autocorrelated: see Figure 1.5. We can conclude that the results in event time are similar to those in clock time.

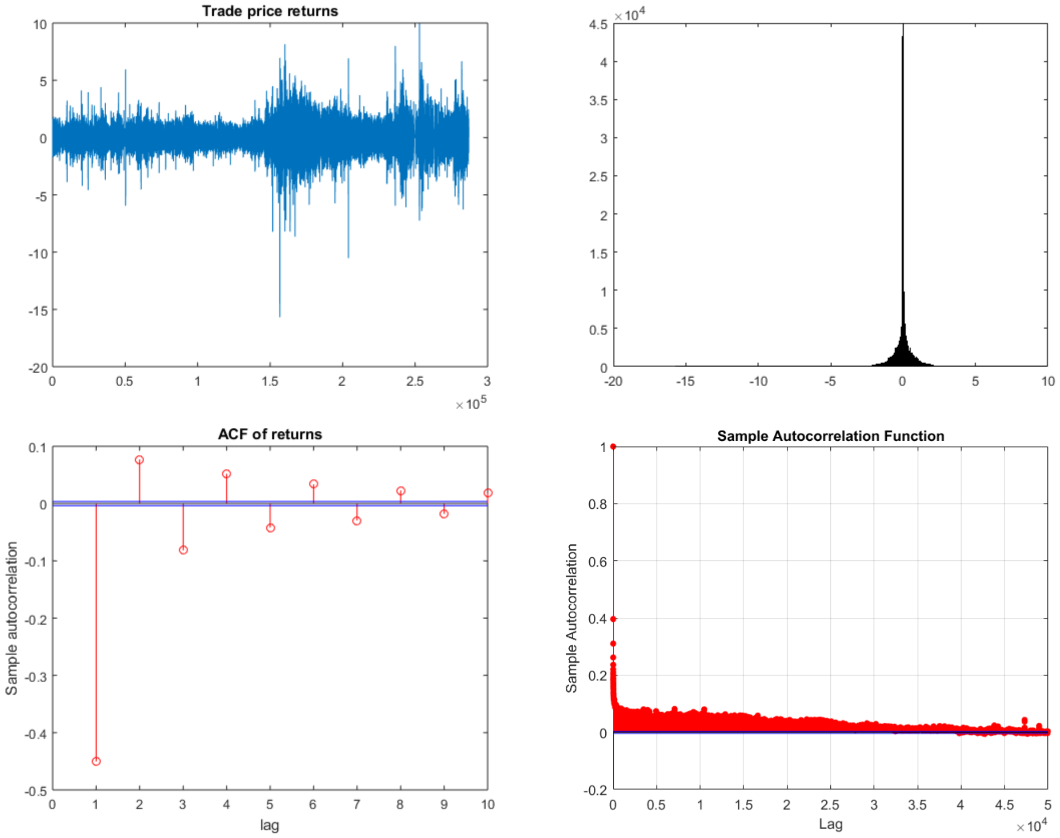


Figure 1.5: Trade returns in event time. From top left to bottom right: plot, histogram, ACF (of returns) and ACF of squared returns. Horizontal lines indicate the 95% confidence interval. Non normality and autocorrelation of returns are apparent as well as volatility clustering.

1.3.2 Mid Price Returns

The analysis conducted in the previous subsection can be repeated using the mid prices instead of the trade prices. Skewness and kurtosis of returns appear significantly larger in absolute value (at any time scale) than the trade price case, and a fortiori the hypothesis of normality is always rejected (see Table 1.2). At short time scales the autocorrelations of mid price returns is more persistent than that of trade price returns. However, the persistency decreases at longer time scales (cf. Figure 1.8 and the graphs contained in Appendix 1.8) nevertheless the p-value of the Ljung-Box test is less than 1%. Volatility clustering is apparent, see Figure 1.9.

Time scale	Skewness	Kurtosis	p-val J-B	p-val L-B
1 sec.	-8.6849	1.5872e+03	< 0.01	< 0.01
30 sec.	-3.0086	127.9161	< 0.01	< 0.01
1 min.	-2.8356	104.0600	< 0.01	< 0.01
5 min.	-1.5861	55.7517	< 0.01	< 0.01
30 min.	-3.0660	76.2204	< 0.01	< 0.01

Table 1.2: Mid price returns: skewness, kurtosis, p-values of Jarque-Bera (p-val J-B) and Ljung-Box (p-val L-B) tests. Returns are negatively skewed, display fat tails, appear to be autocorrelated and not normal at each time scale.

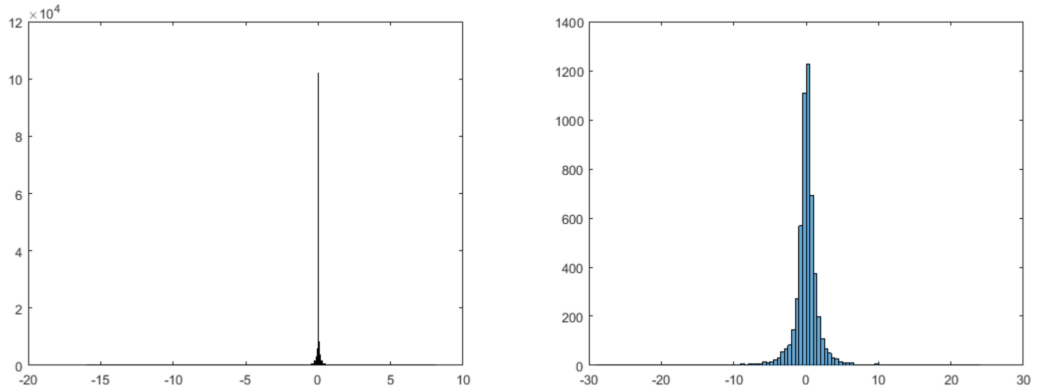


Figure 1.6: Mid price histograms. Left: 30 sec. frequency. Right: 30 min. frequency. Even from an eye inspection the distributions do not resemble a normal one.

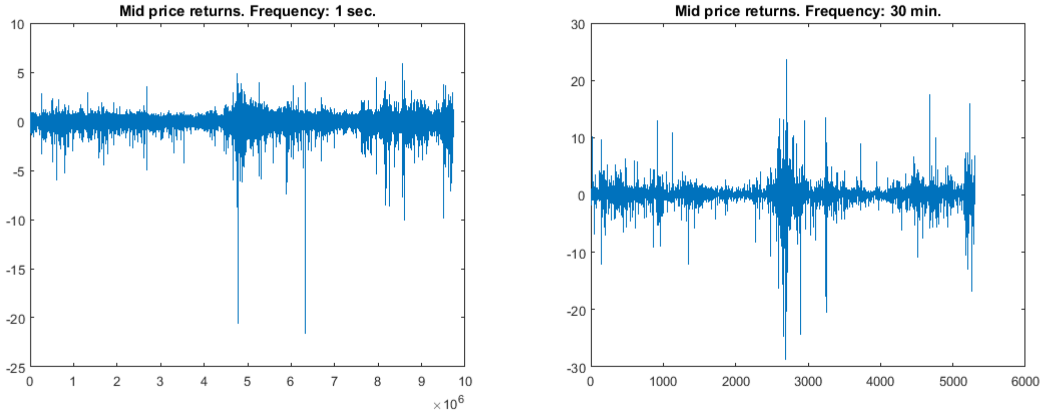


Figure 1.7: Mid price returns at different time scales. Increasing the time scale returns tend to be larger (in absolute value).

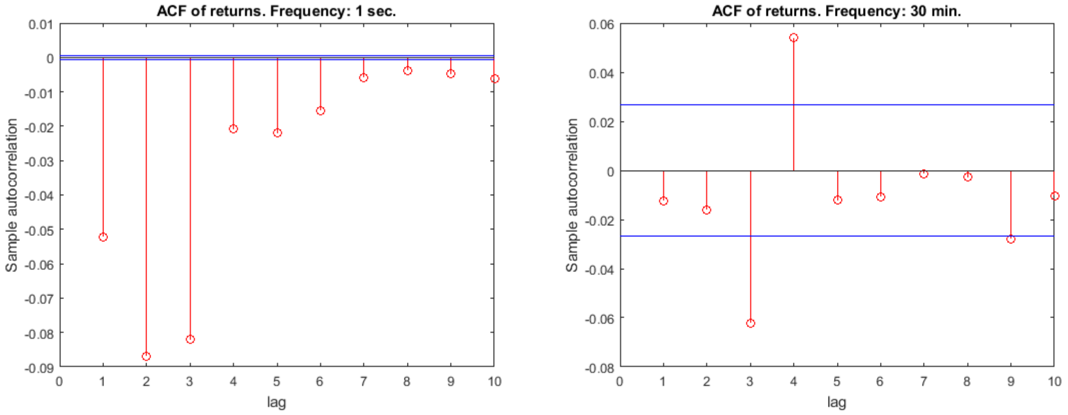


Figure 1.8: Mid price returns: ACF at different time scales. Horizontal lines indicate the 95% confidence interval. Increasing the time scale the number of significant lags decreases because the effect of the microstructure noise vanishes.

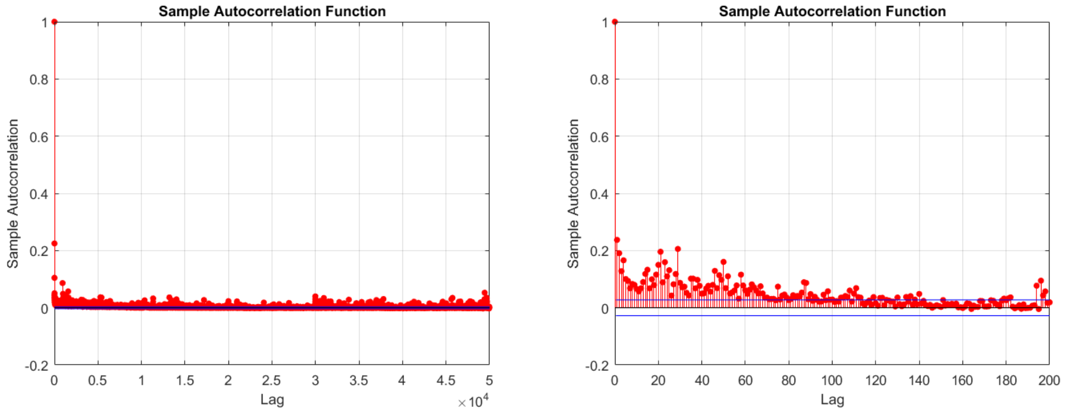


Figure 1.9: Mid price squared returns: ACF at different frequencies (left: 30 sec., right: 30 min.). Horizontal lines indicate the 95% confidence interval.Volatility clustering is apparent. In addition, increasing the time scale the number of significant lags decreases.

As for the trade prices we can carry out a statistical analysis similar to that above in event time (see Section 1.2.2). Figure 1.10 resumes our findings: non normality (skewness is negative and close to zero, large kurtosis and the p-value of Jarque-Bera test is less than 1%), p-value of Ljung Box test (relative to returns) is less than 1% (the first order autocorrelation is particularly significant, instead at higher lags the correlations appear very weak yet significant). Finally, once again the volatility clustering is strongly confirmed.

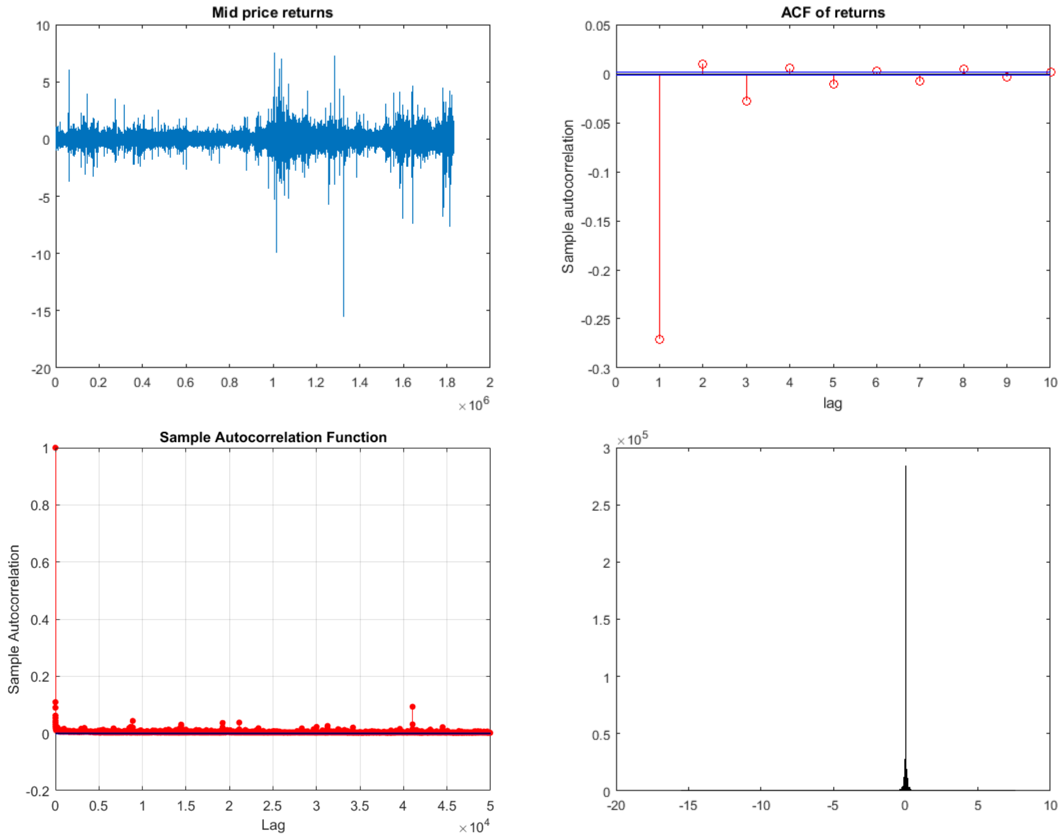


Figure 1.10: Mid price returns in event time. From top left: returns, ACF of returns, ACF of squared returns, histogram. Horizontal lines indicate the 95% confidence interval. Non normality and autocorrelation of returns are apparent as well as volatility clustering.

1.4 Additional Features of the Order Book

In this section we collect some statistical properties of the volumes and quotes resting on the order book.

1.4.1 Log Ask and Bid Sizes

As discussed in Section 1.2.2 log-volumes have been studied at different frequencies (30 sec., 1 min., 5 min., 30 min.) and aggregated over five levels. Aggregating volumes produces higher positive autocorrelations at any frequency and smoother histograms that are leptokurtic. Hence the Jarque Bera test strongly rejects the hypothesis of normality (all the p-values are less than 1%) independently on how many levels are aggregated and the time scale considered. Figure 1.11 , Figure 1.13 and Figure 1.12 are an excerpt of the analysis, further graphs can be found in Appendix 1.8. A similar study is possible considering event time instead of clock time. Again aggregating volumes over more levels strengthen the autocorrelations at every lag analyzed, additionally it has a regularizing effect on the distribution: aggregating five levels the shape of the distribution becomes smoother, nevertheless the Jarque Bera test has p-value less than 1%. The ACF and PACF plots in Figure 1.13 and 1.14 suggest an ARMA dynamics for the log volumes.

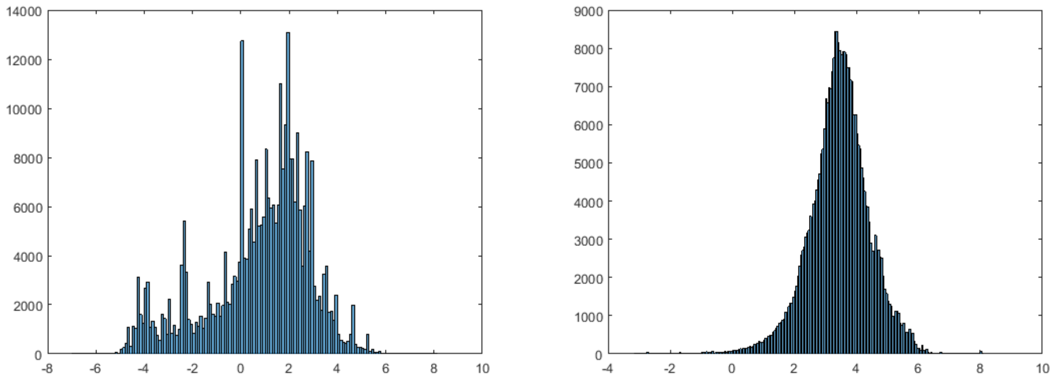


Figure 1.11: Log ask volume histograms at different time scales (and for different number of aggregated levels). Left: first level (30 sec.). Right: five levels aggregated (30 sec.). Aggregating volumes and lengthening time scales produces "smoother histograms".

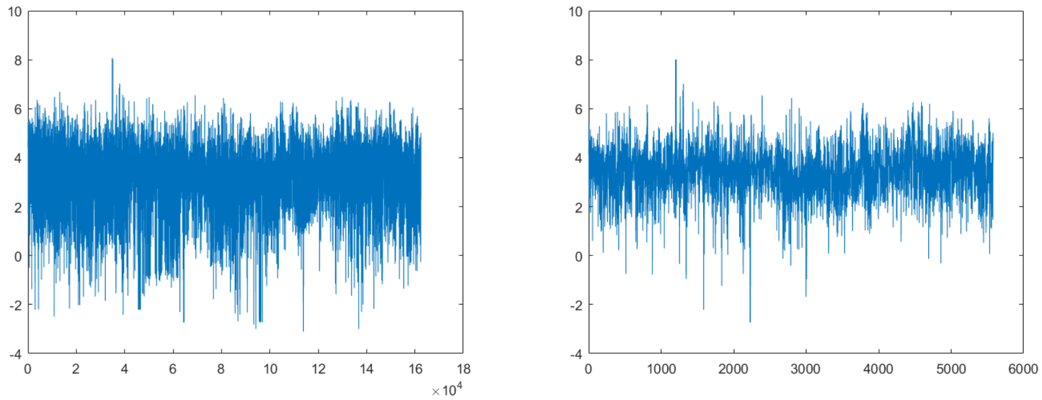
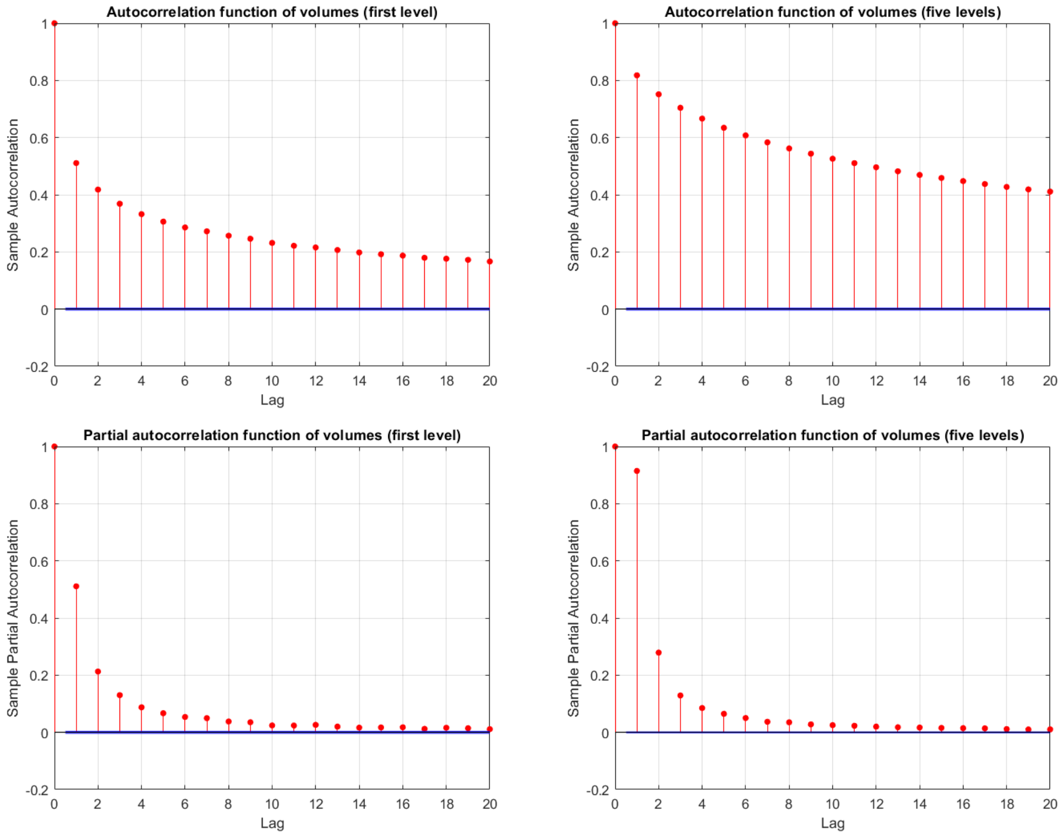


Figure 1.12: Five levels aggregate ask volumes at two different time scales: 60 sec. (left) and 30 min. (right). Volumes resting on the order book display a mean reverting behaviour.



Autocorrelation function of volumes (first level)

Autocorrelation function of volumes (five levels)

Partial autocorrelation function of volumes (first level)

Partial autocorrelation function of volumes (five levels)

Figure 1.13: (Bottom) Top: Log ask volume (P)ACF. Left: first level. Right: five levels aggregated. The time scale is 30 sec.. Horizontal lines indicate the 95% confidence interval. The plots show that the volumes resting on the order book are strongly serially correlated especially when different levels are aggregated.

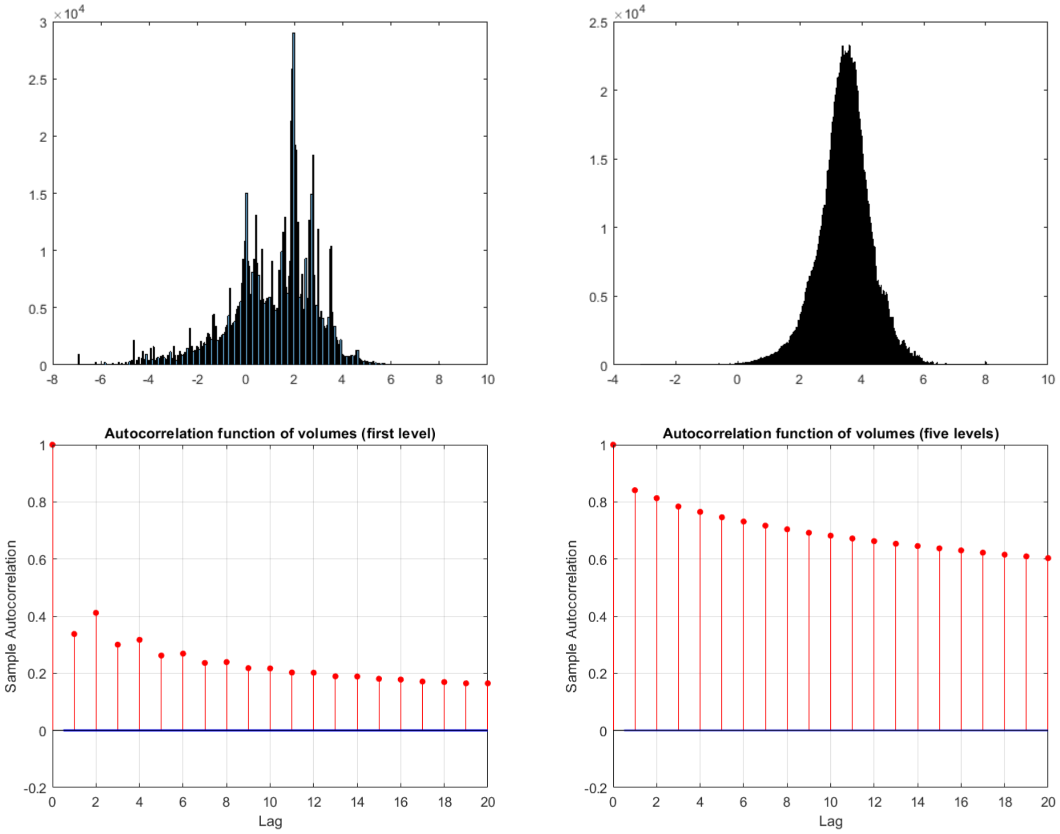


Figure 1.14: Log ask volumes in event time. From top left to bottom right: histogram relative to the first level, histogram relative to five levels aggregated, ACF relative to the first level, ACF relative to the five levels aggregated. Horizontal lines indicate the 95% confidence interval. The findings are similar to the clock time study: aggregating volumes and lengthening time scales produces higher positive autocorrelations and smoother histograms. In addition, the volumes resting on the order book appear to be strongly serially correlated especially when different levels are aggregated.

Repeating the same analysis above, both in clock time and in event time, for the other side of the order book (i.e., the bid side) produces results of identically nature, thus we omit them.

1.4.2 Order Flow

Order flow is an important microstructural variable in market movement forecasting (Cont et al., 2014). Figure 1.15 displays the distribution of the five-minute order flow and its ACF function: the distribution is peaked around zero, the ACF diagram reveals significant autocorrelatios for different lags. Figure 1.16 shows the correlation between order flow and trade price returns calculated using a 5 min. time scale: order flow is not linearly correlated with future returns.

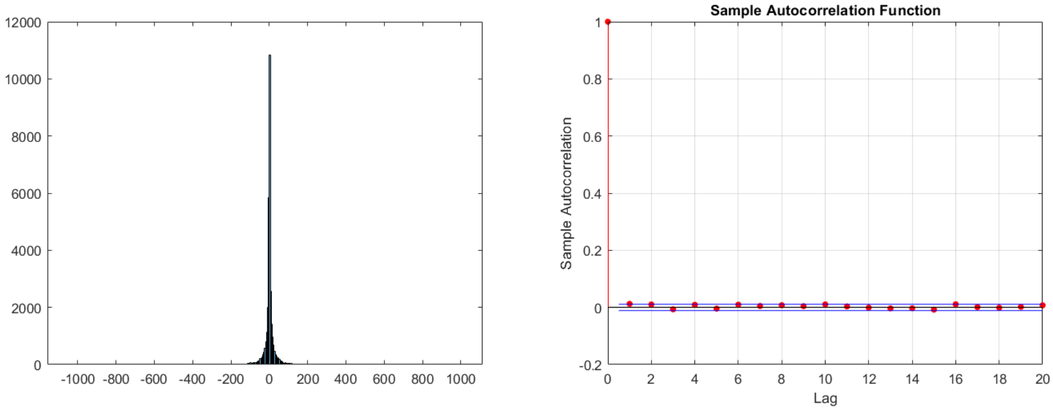


Figure 1.15: Order flow: histogram and ACF. Horizontal lines indicate the 95% confidence interval. The plots show that the distribution of the order flow is not normal and the order flow is not serially correlated.

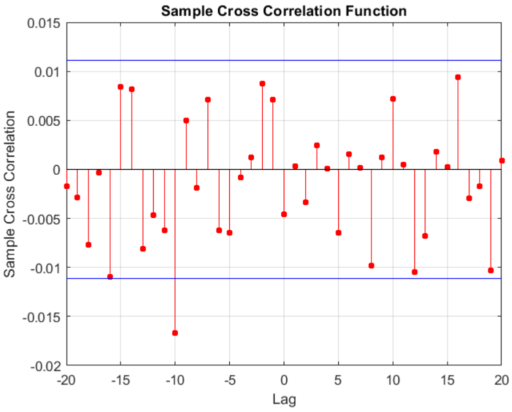


Figure 1.16: Cross correlation function between order flow and trade price returns calculated using a 5 min. time scale. Horizontal lines indicate the 95% confidence interval. Order flow does not seem to be useful in order to predict future returns.

1.4.3 Volume Imbalance

Volume imbalance is another fundamental variable in market movement forecasting as showed in Sancetta (2018) and Cartea et al. (2018). We study the volume imbalance relative to all the first five levels of the order book in event time (for a definition see Section 1.2.2): Figure 1.17 summarizes the results. Both the ACF diagram and the histogram distribution do not seem to be substantially affected by the level, so we report the findings concerning just two levels. The autocorrelations are positive and persistent and the distribution appears to be far different from a normal one, in particular there are peaks around the extreme values (plus and minus one).

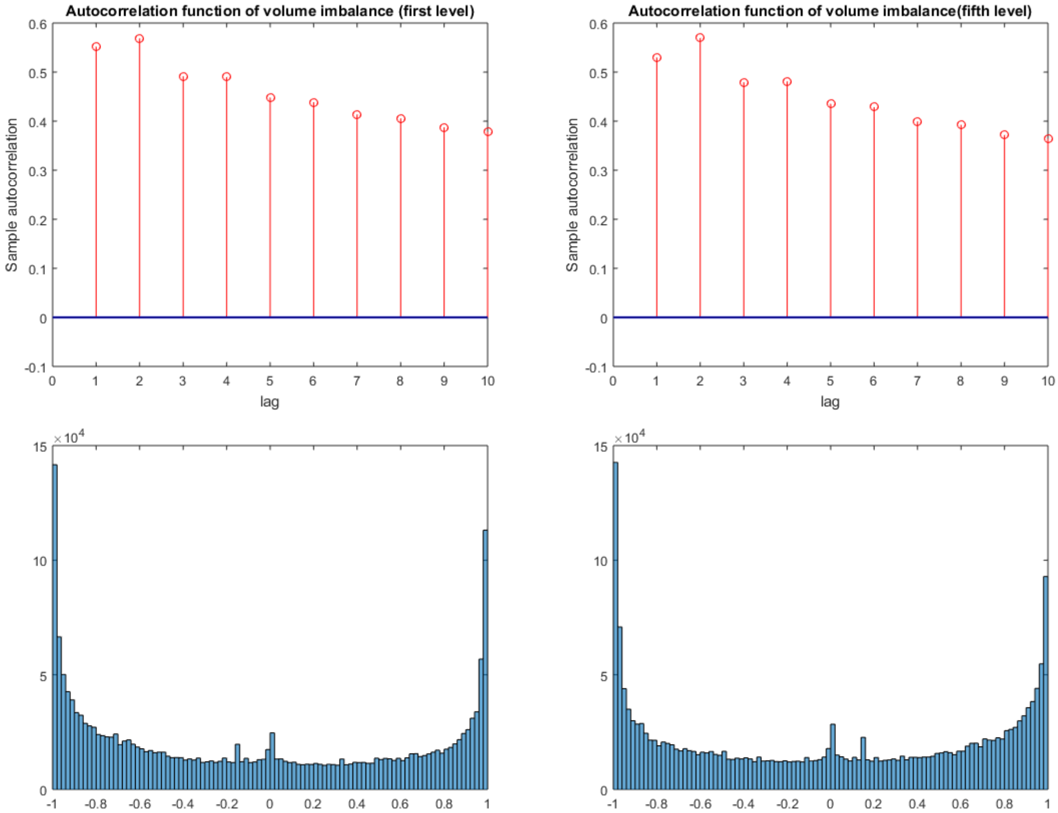


Figure 1.17: Volume imbalance. Top: ACFs relative to the first and fifth level, respectively. Horizontal lines indicate the 95% confidence interval. Bottom: histograms relative to the first and fifth level. The two top figures show that the volume imbalance is strongly serially correlated. In addition, the spikes, in the histograms, at -1 and +1 suggest that the distribution is a mixture of continuous and discrete random variables.

1.4.4 Spread

In this final subsection we address the study of the spread, i.e., the difference between best ask price and best bid price. Figure 1.18 summarizes the empirical findings showing the spread and its autocorrelation diagram. The Ljung-Box test has always p-value less than 1%, thus the hypothesis of no autocorrelation is rejected.

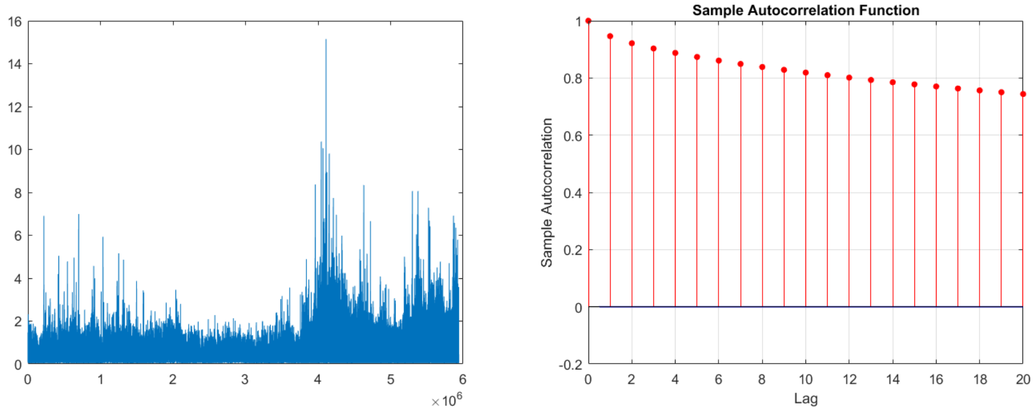


Figure 1.18: Spread and its ACF. Horizontal lines indicate the 95% confidence interval. The spread displays a mean reverting behaviour and it appears to be strongly serially correlated.

1.5 Trading Volume Estimation

The goal of this section is to model the intraday trading volume. For US equities the typical daily trading volume curve is U shaped with spikes corresponding to the market open and close: we do not observe this shape, yet jumps in the trading activity are clearly present. We want to capture these jumps but, at the same time, we aim at regularizing the trading volume curve. Each day is divided into $K = 288$ intervals, i.e., each interval spans a period of time five minutes long. Figure 1.19 represents the average number of trade arrivals during the 24 hours of the day over the 276 days.

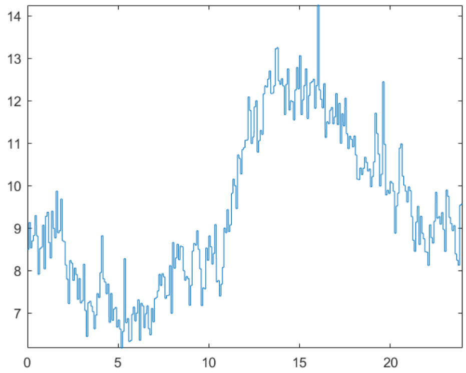


Figure 1.19: Average number of trade arrivals in the Bitcoin market during the 24 hours. Horizontal axis: hour of the day (GMT). Vertical axis: number of trades. The pattern does not resemble a U-shaped or M-shaped curve typically found in the literature: this is not completely unexpected given that the Bitcoin market, unlike traditional markets, is open 24/7.

Let v the $K-$dimensional vector having as i^{th} entry the average number of trade arrivals over the m days during the time interval $(5\times(i-1) \quad minutes, 5\times i \quad minutes]$. Then we look for a vector x that minimize the following function ($\|\cdot\|$ denotes the euclidean norm)

$$\frac{1}{2}\|x-v\|^2+\mu\sum_{i=1}^{K}|x_{i+1}-x_i| \tag{1.3}$$

i.e., a total variation penalty term is added to a quadratic type contrast function. The scope of the latter term is to get a vector "close" to v while the former term promotes the sparsity of the first differences of the coefficients (i.e., they tend to be locally constant): the role of the parameter $\mu>0$ is to control the trade off between fit and sparsity. This type of penalty term has been applied to the multiple change-point problem, cf., e.g., Alaya et al. (2015) and is a generalized version of the Lasso, the so called fused estimator (cf., e.g., Tibshirani and Taylor, 2011). Even if we do not impose a nonnegativity constraint we get nonnegative estimators for the vector v: this justifies the choice. In lieu of minimizing (1.3) we solve its dual formulation, cf. Equation (13) in Tibshirani and Taylor (2011). First, it is required to solve the following constrained quadratic problem

$$\max_{\|z\|_\infty\leq\mu}\left(\frac{1}{2}\|R'z\|^2-z'Rv\right) \tag{1.4}$$

where R' (z', respectively) is the transpose matrix (vector, respectively) of R (z, respectively) and

$$R_{ij}=\begin{cases}-1 & if \quad j=i, i=1,2,\ldots,K-1\\ 1 & if \quad j=i+1, i=1,2,\ldots,K-1\\ 0 & otherwise.\end{cases}$$

If $\hat{z}$ solves (1.4), then $\hat{x}$, i.e., the minimizer of (1.3), is given by

$$\hat{x}=v-R'\hat{z}.$$

We perform the optimization using different values of the parameter μ : to choose the "right" regularized model we adopt the Akaike Information Criteria

(AIC). AIC is given by:

$$AIC = \frac{2 \times (Number \quad of \quad jumps)}{K} + \log\left(\frac{\|x-v\|^2}{K}\right)$$

where the number of jumps is equal to cardinality of the set $\{i \geq 1 : |x_{i+1} - x_i| > 10^{-12}\}$[3]. Table 1.3 summarizes the results obtained for different values of μ: according to the AIC the best model has 102 jumps and corresponds to $\mu = 0.5$. Figure 1.20 represents the vector $\hat{x}$ relative to different values of the parameter μ : when μ increases the number of jumps decreases.

μ	AIC	Number of jumps
0.5	-1.557	102
1	-1.2883	68
1.5	-1.2936	48
2	-1.2993	42
2.5	-1.3006	41
3	-1.2856	41
5	-1.2328	37

Table 1.3: AIC and number of jumps relative to different values of the parameter μ. A jump occurs whenever $|x_i - x_{i-1}| > 10^{-12}$. As μ increase the number of jumps decreases and the AIC increases.

The top left plot in Figure 1.20 helps identify jumps in the trading activity: in particular the peak, reached at 16 GMT, is followed by a significant slow down. That time almost coincides with the closing time of the London Stock Exchange (16:30 GMT). More in details, from 10:30 GMT to 13 GMT there is a steeply increase in the trading activity, then during the period 13-16 GMT there is a stable and high trading activity: the peak is reached at 16 GMT, thereafter the traded volume gradually declines. The trading activity is concentrated during the hours in which the European and USA markets are opened: this may indicate that the Bitstamp exchange trading activity is mostly driven by European and USA traders.

[3] The results do not change if we define the number of jumps as the cardinality of the set $\{i \geq 1 : |x_{i+1} - x_i| > 10^{-5}\}$.

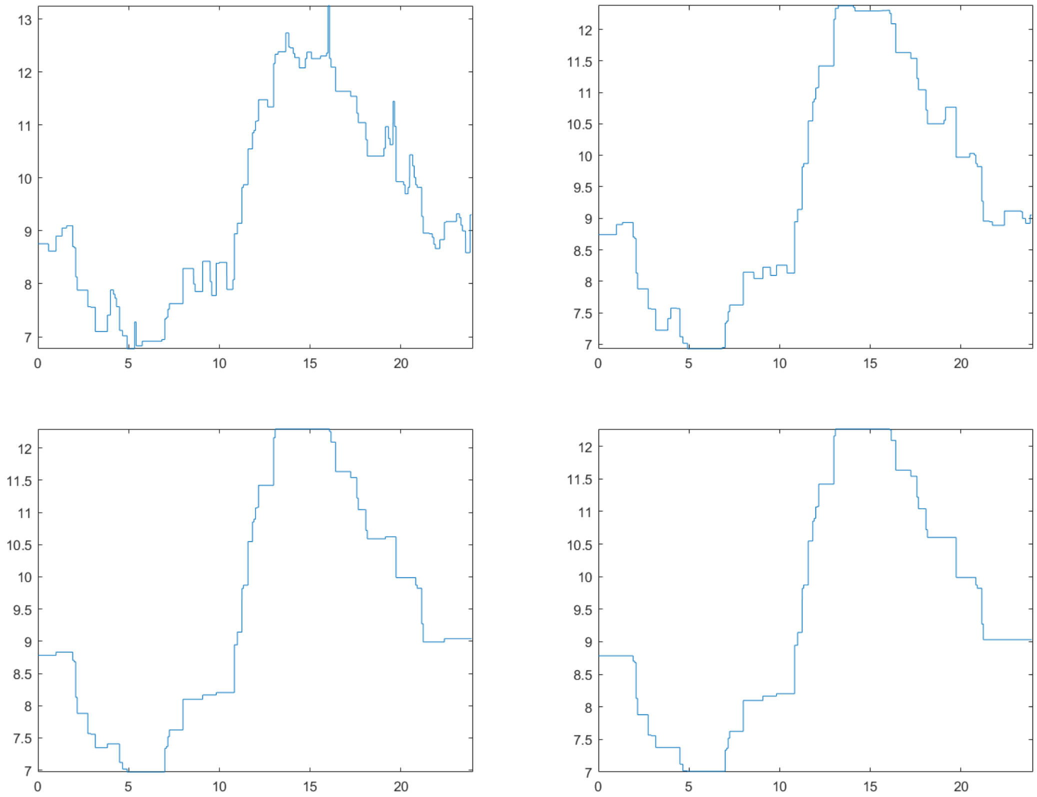


Figure 1.20: Estimated intensity relative to different values of the parameter μ. From top left to bottom right: $\mu = 0.5, 1, 1.5, 2$. Horizontal axis: hour of the day (GMT). Vertical axis: number of trades per unit of time. Increasing μ the graph is clearly smoother. The AIC criteria selects the top left plot.

1.6 Probability of Informed Trading

Market microstructure dynamics could be driven by informed trades. The PIN (Probability of Informed Trading) model was introduced in Easley et al. (1997) to compute, in illiquid markets, the probability that a trade is information based. As pointed out in Duarte and Young (2009) that model is unable to match some empirical findings, notably the positive correlation between buys and sells and their large variance. Duarte et al. (2020), in order to overcome those drawbacks, put forwards a few alternative models: we shall analyze the Bitcoin market via the GPIN (Generalized PIN) model. The GPIN allows to estimate the Conditional Probability of an Information Event at day t (CPIE(t)), i.e., the conditional probability of private-information arrival given the data observed on day t. Figure 1.22 contains the graph of the daily closing Bitcoin price and Figure 1.21 shows the filtered CPIE, i.e., the moving average of the CPIE based on a 20 days time window. An eye inspection of the two plots reveals that a surge in prices is followed by a decline in the CPIE. This could be due to a herd behavior: high prices attract uniformed (retail) traders that follow the upward trend.

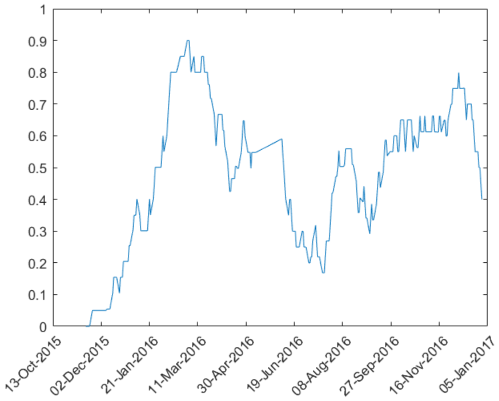


Figure 1.21: Filtered CPIE. The filtered CPIE is the moving average of the CPIE based on a 20-day time window. Horizontal axis: days. Vertical axis: filtered CPIE. Since mid March 2016 the probability that a trade is informed-type oscillates around 0.5.

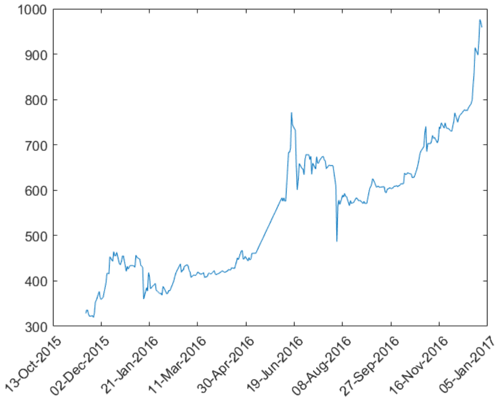


Figure 1.22: Bitcoin daily closing price. Horizontal axis: day. Vertical axis: closing price. The graphs suggests an exponential growth for the Bitcoin price.

1.7 Conclusions

The study confirmed the validity, for the Bitcoin market, of some well known stylized facts common to mature financial markets (negative skewness, fat tails and volatility clustering of returns) except for the efficient market hypothesis: returns appear to be autocorrelated at every time scale analyzed, i.e., 1 sec, 30 sec, 1 min, 5 min, 30 min. The same conclusion is reached by several other papers in the literature. The study of the information based trades via the Generalized Probability of Informed Trading (GPIN) model supports the idea that high prices attract retail traders that act as "noise traders": this could be one of the cause of the not perfect efficiency of the market. In addition, it is in line with the conclusion of Petukhina et al. (2019), according to them "the digital realm of cryptocurrencies has yet to be conquered by the machines and is still firmly in the hands of free-time-/holiday-traders or could be even driven by respective start-up's". The paper also analyzed some other variables related to the orders resting on the order book. They display some typical features of high frequency data such as strong serial correlation, long memory (ask and bid volumes, bid-ask spread) and distributions that are mixture of continuous and discrete random variables (volume imbalance). Finally, the trading volume profile does not resemble a classic pattern and suggests that the trading activity in the Bitstamp exchange is mainly driven by European and USA investors.

1.8 Appendix: Additional Graphs and Plots

In this section we collect further graphs and plots.

1.8.1 Trade Returns Plots

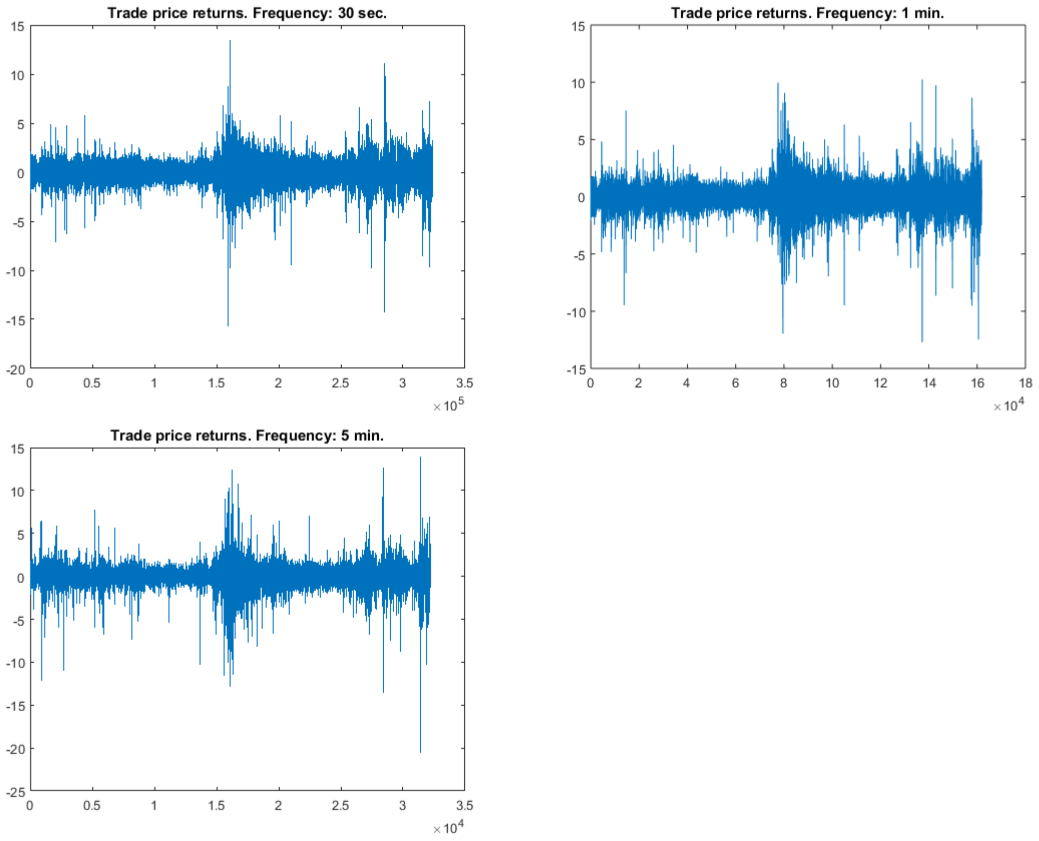


Figure 1.23: Trade price returns at different time scales. Increasing the time scale returns tend to be larger (in absolute value).

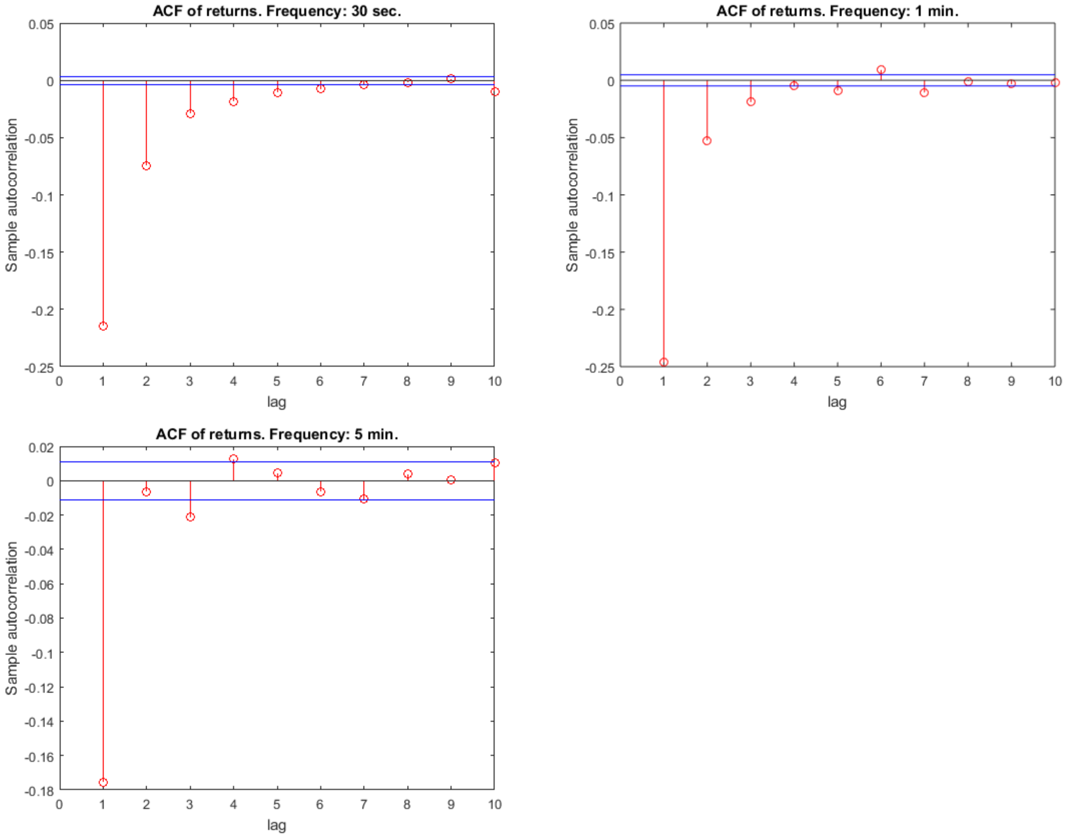
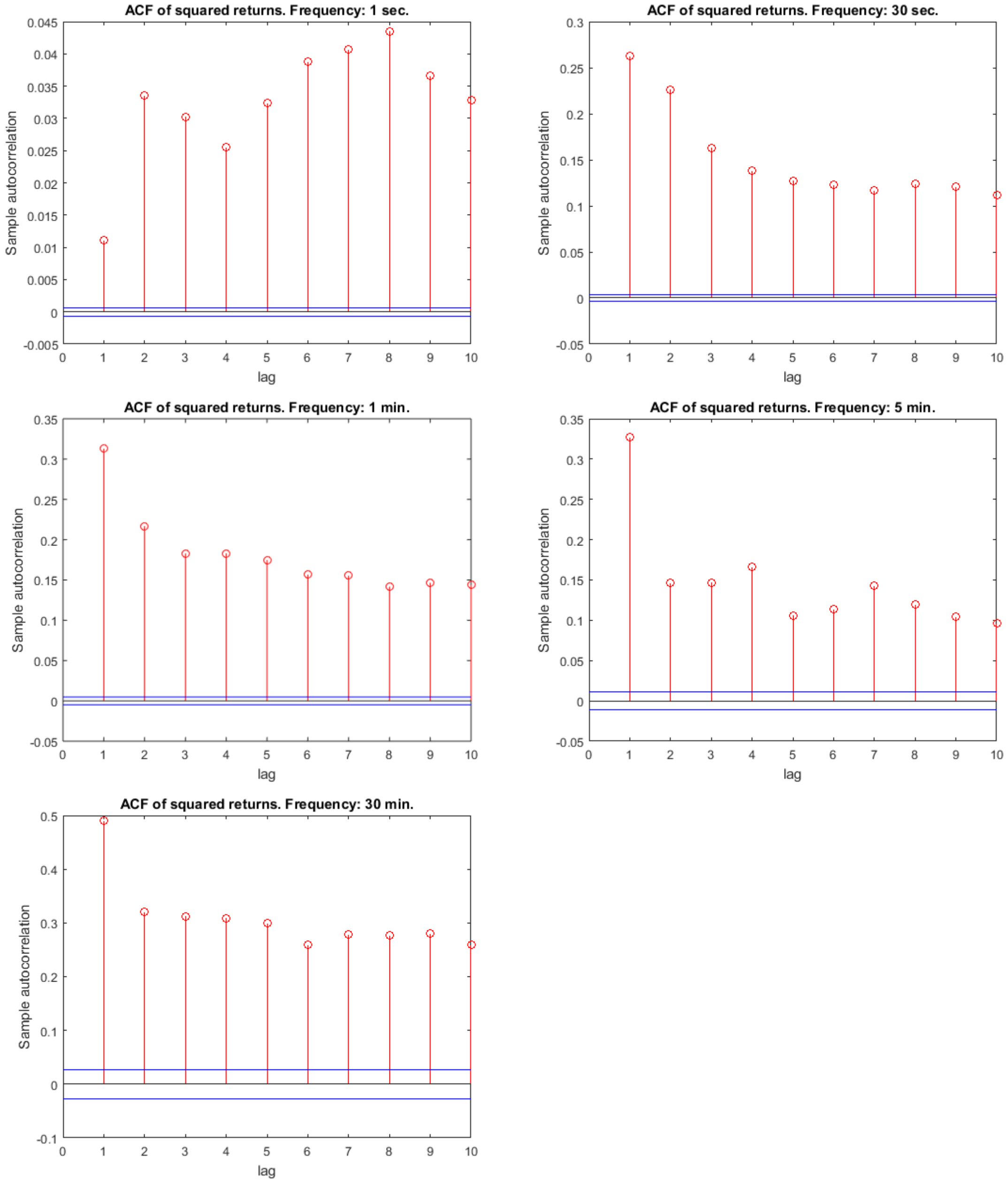


Figure 1.24: ACF of trade price returns at different time scales. Horizontal lines indicate the 95% confidence interval. Increasing the time scale the number of significant lags decreases.



ACF of squared returns. Frequency: 1 sec.

ACF of squared returns. Frequency: 30 sec.

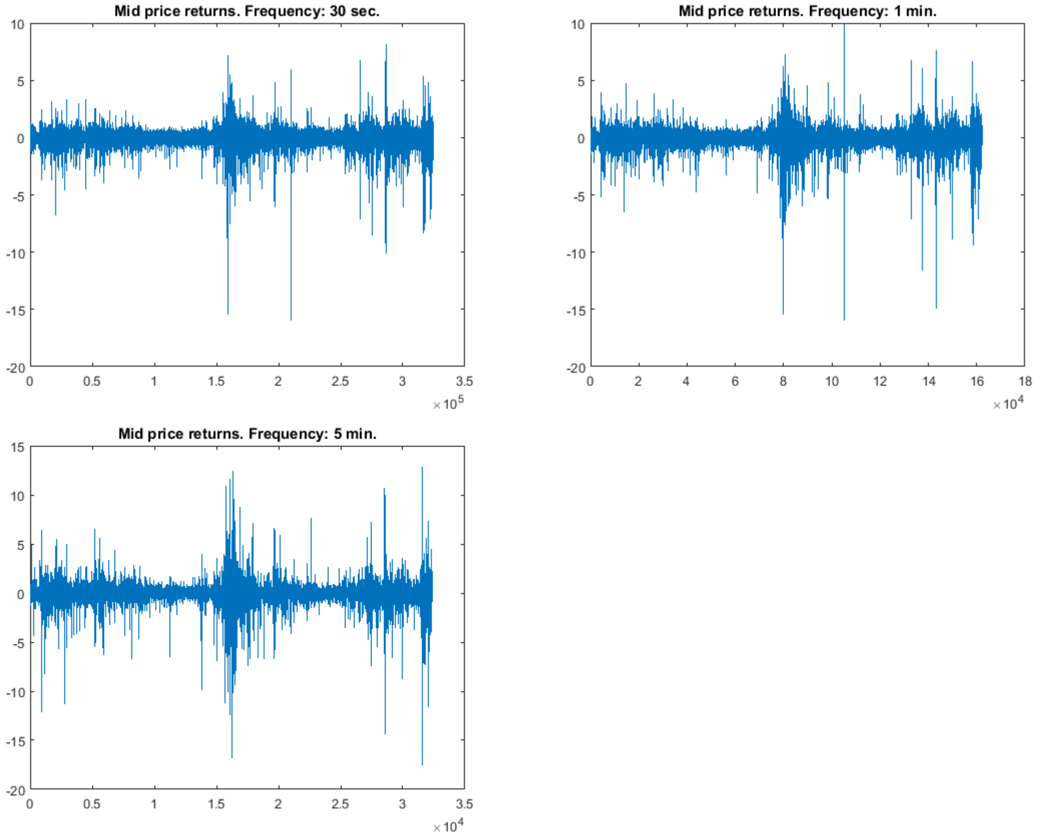
ACF of squared returns. Frequency: 1 min.

ACF of squared returns. Frequency: 5 min.

ACF of squared returns. Frequency: 30 min.

Figure 1.25: ACF of squared trade price returns at different time scales. Horizontal lines indicate the 95% confidence interval. Volatility clustering is apparent at every time scale.

1.8.2 Mid price returns plots



Mid price returns. Frequency: 30 sec.

Mid price returns. Frequency: 1 min.

Mid price returns. Frequency: 5 min.

Figure 1.26: Mid price returns at different time scales. Increasing the time scale returns tend to be larger (in absolute value).

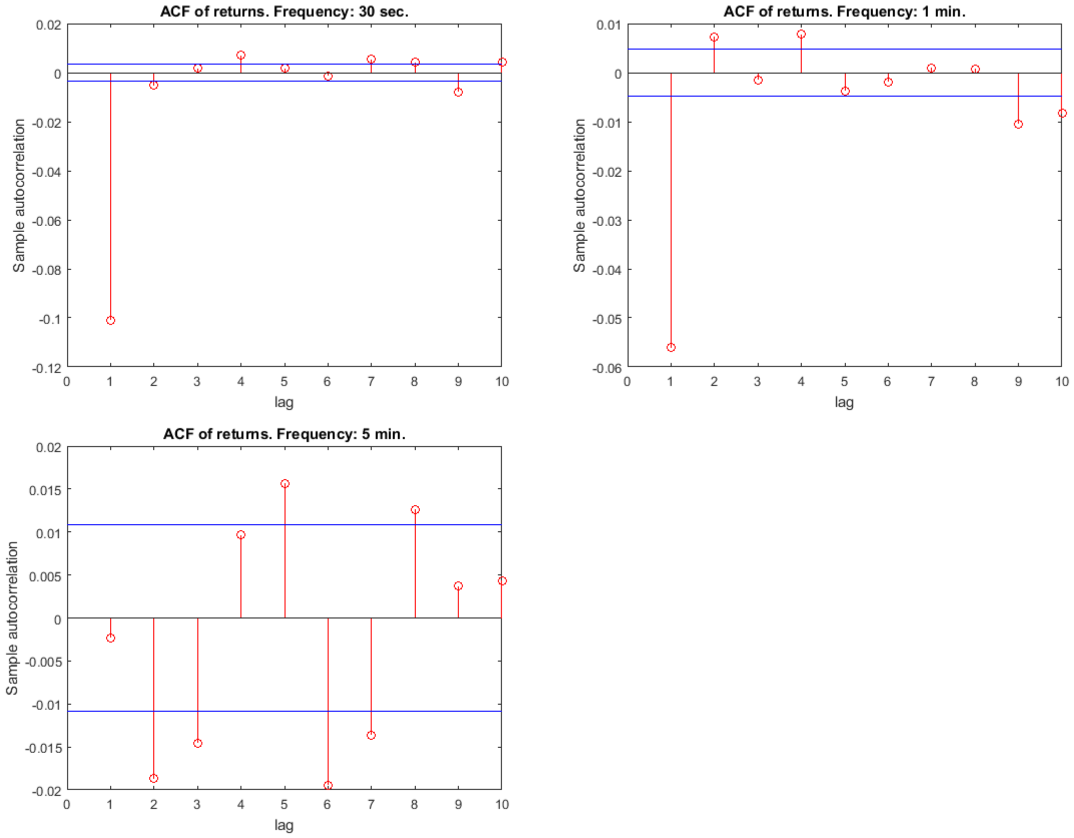


Figure 1.27: Mid price returns: ACF at different time scales. Horizontal lines indicate the 95% confidence interval. At every time scale returns appear to be autocorrelated.

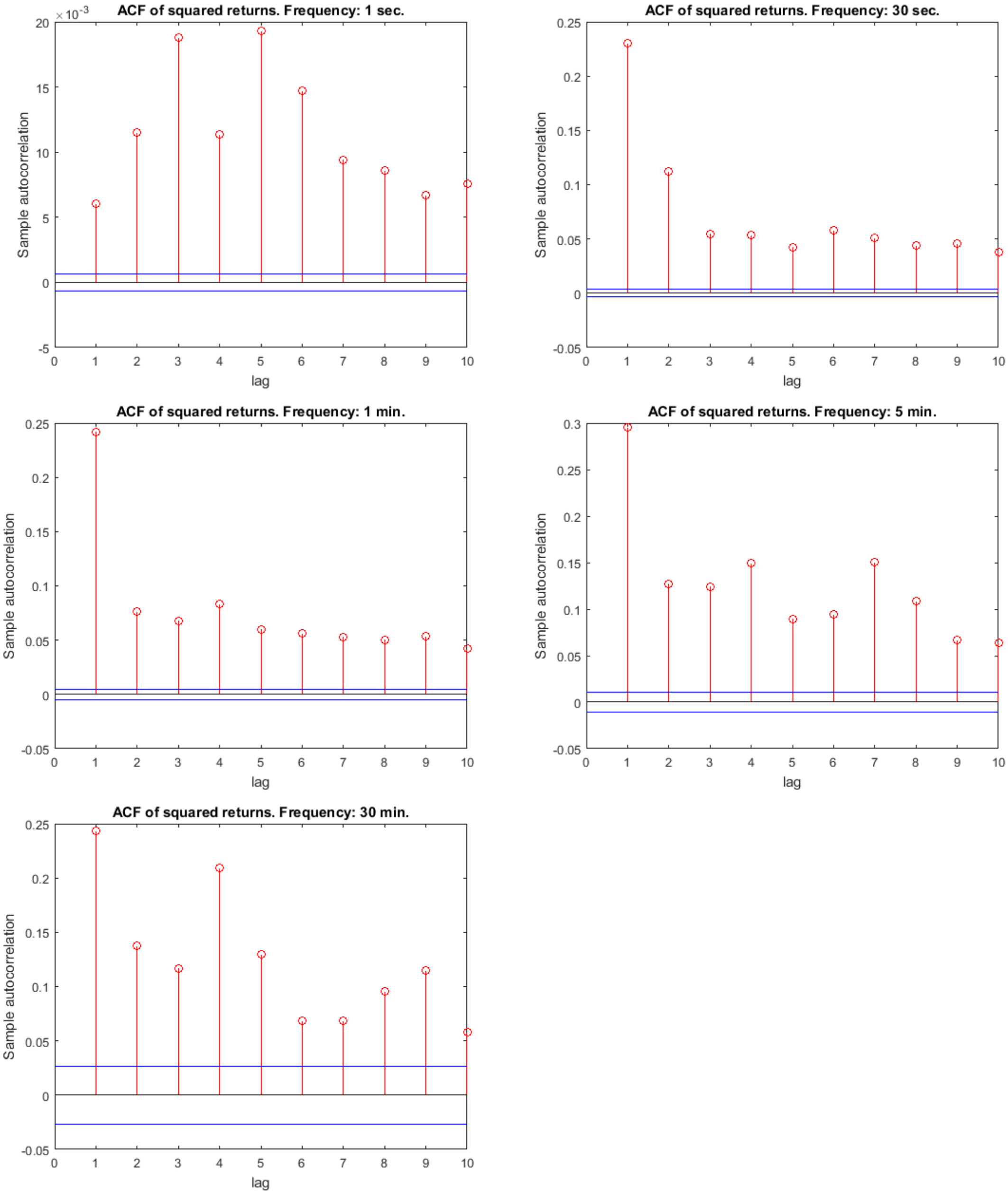


Figure 1.28: Mid price squared returns: ACF at different frequencies. Horizontal lines indicate the 95% confidence interval. Volatility clustering is apparent at every time scale.

1.8.3 Log Volumes Plots

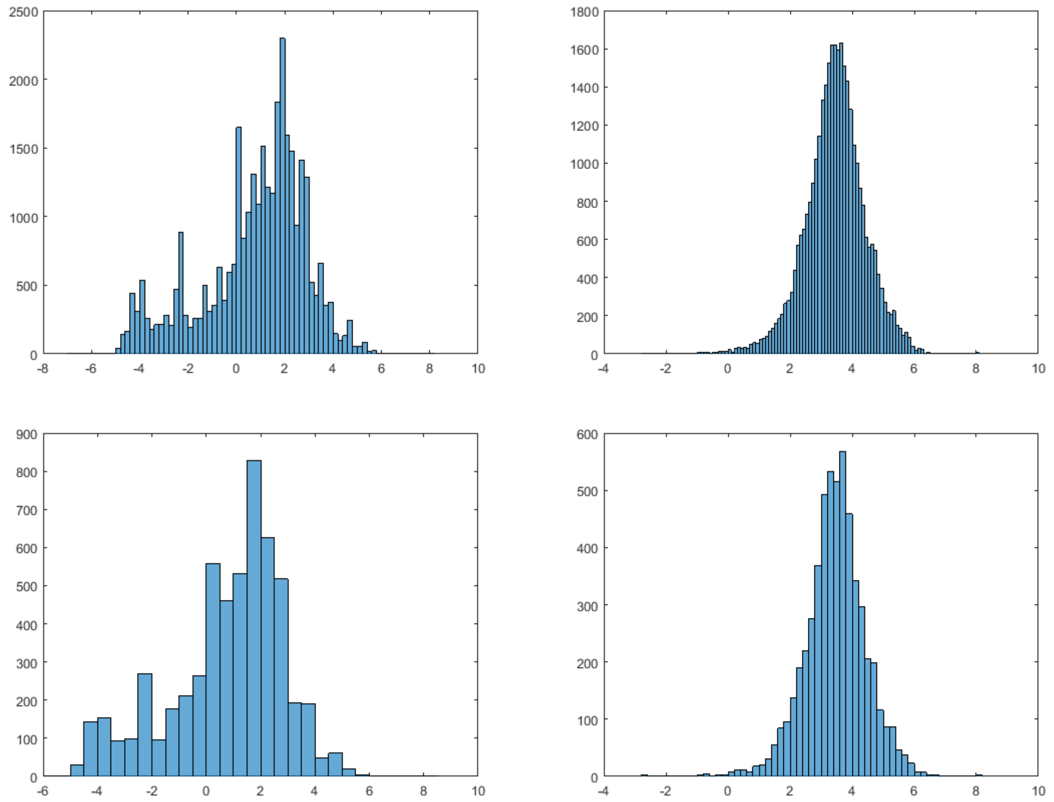


Figure 1.29: Log ask volume histograms at different time scales and number of aggregated levels. From top left to bottom right: first level (5 min.), five levels aggregated (5 min.), first level (30 min.), five levels aggregated (30 min.). Aggregating volumes and lengthening time scales produces “smoother histograms”.

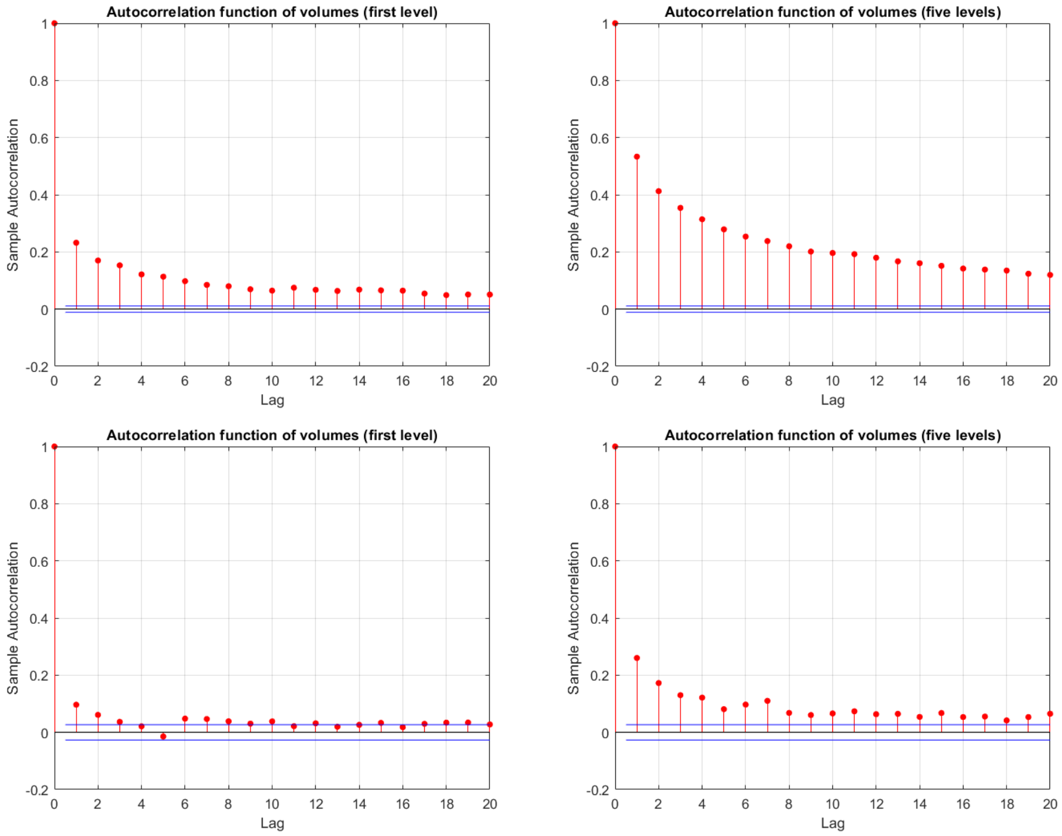


Figure 1.30: Log ask volume ACF for different time scales and number of aggregated levels. Top: 5 min.. Bottom: 30 min.. Horizontal lines indicate the 95% confidence interval. The volumes resting on the order book appear to be strongly serially correlated especially when more levels are aggregated.

Chapter 2

Estimation of an Order Book Based Intensity Model: How Prone is a Market to Manipulation?

Abstract. We model the intensity of trade arrivals to understand how electronic markets could be manipulated by high frequency algorithms. We relate trade arrivals to the impact of various order book events, such as buy and sell pressure, spread, etc. We adopt a stochastic intensity model. The intensity of trade arrivals is driven by a baseline intensity and additive functions of the covariates, which we call impact functions. If the impact functions for certain order book events satisfy conditions such as monotonicity, a spoofing algorithm could be successfully implemented. Such an algorithm would place fictitious resting orders to distort the view of demand and supply. The analysis requires the introduction of a statistical framework suitable for high frequency data with sample sizes in the order of possibly hundreds of million or billion data points.We apply our methodology to the study of the crude oil futures that trade on the Chicago Mercantile Exchange. We can conclude that, in our sample period, such futures contract could be manipulated under certain circumstances that we are able to identify.

2.1 Introduction [1]

Market manipulation is not uncommon in electronic markets. To fix ideas, consider the illegal practice of spoofing. Typically, a spoofing algorithm places a relatively small buy order on the best bid, and almost contemporaneously it places a sequence of relatively large sell orders on the ask side of the book. This action provides a snapshot of the demand and supply schedule, where the market appears to be willing to sell. The reason is that there are considerably more orders to sell than to buy. This will often induce a trader (usually another algorithm) to place a sell order that crosses the bid-ask spread. In consequence the small limit order placed by the spoofing algorithm on the best bid will be filled. Once this happens, the spoofing algorithm will cancel all the large limit orders placed on the ask side of the book. The game then repeats reversing the role of the two sides of the book: a small resting order on the ask, and relatively large orders on the bid. The final result is that the manipulator gains the spread. Examples of this practice can be found in the 3 July 2013 Final Notice given to Michael Coscia by the Financial Conduct Authority (URL:`https://www.fca.org.uk/publication/final-notices/coscia.pdf`).

High frequency trading strategies rely on order book features. In the above spoofing example, the volume imbalance between bid and ask quantities appears to be crucial. In fact, the literature has found that volume imbalances and other order book variables have an impact on price movements and trade arrivals at very short term horizons (Cont et al., 2014, Sancetta, 2018). MacKenzie (2017) reports anonymous interviews with ex algorithmic traders of the market maker Automated Trading Desk. These interviews confirm the importance of order book imbalances for price movement. Hence, a market manipulator can place fictitious orders (i.e., orders that will be soon canceled) with the purpose of modifying the view of demand and supply to mislead other traders.

A practical implementation of spoofing needs to account for other quantities such as bid-ask spread and quoted sizes (cf. page 2 point 9 in the Financial Conduct Authority document mentioned above). Despite these intricacies, recent cases in the press (Singh Sarao, Michael Coscia) have clearly shown that high frequency market manipulation is possible. The regulator, in order to

[1] This chapter is co-authored with Alessio Sancetta.

detect spoofing practices, may first have to find whether the market has some statistical characteristics that makes it particularly prone to be manipulated. Our model aims to capture one of this characteristic, i.e., the relationship between market orders and some (easily manipulatable) covariates.

For example, consider the case of manipulation with the purpose of filling a resting order on the ask (i.e., sell at the ask price). We would want the intensity of buy orders to increase relatively to sell orders. The market manipulator would act so to make this possible. To understand how this can be done, we need a model quantifying the relation between the number of trade arrivals and order book variables (e.g., volume imbalance and spread). The variables are the ones that an algorithm could distort with the aim to trade advantageously. Because of irregularly spaced time series data, it is natural to adopt a stochastic intensity counting processes model (Bauwens and Hautsch, 2009). Our intensity model is driven by a baseline process times additive functions of the covariates. We call these functions impact functions and one of the main goal of the paper is to estimate these functions.

2.1.1 Goals and Contribution

We analyze transaction and order book data for crude oil futures over a period of six months during liquid hours. We estimate the shape of the impact functions of market covariates of buy and sell trades. We find that the variables that have most impact on buy and sell trade arrivals are the ones that can be most easily manipulated by a trader. These variables are predominantly the quoted volume imbalances on the first few levels of the order book. The degree of top of book imbalances seem to dictate the urgency of aggressive orders by uninformed traders. This result suggests that spoofing does require a certain degree of risk as orders on top of book need to be manipulated. Attempting to manipulate the market by placing orders deeper in the book is not as effective. Orders deeper in the book are less likely to be filled, hence they are less risky when trying to manipulate the market.

Our empirical conclusions rely on an a methodology to model and estimate the intensity of trade arrivals as a function of market covariates. The estimation procedure can be used with large datasets. Our approach allows us to impose constraints on the impact functions using quadratic programming. Moreover, we observe that our non-negativity constraint on the intensity and the impact functions leads to some form of “shrinkage”. In particular, we estimate each impact function by a Bernstein polynomial and observe that most of the coefficients are zero due to the non-negativity constraint. This is what we mean by "shrinkage", and it does not require to explicitly employ a penalty on the coefficients. The intuition behind this phenomenon can be found in results in Chapter 3.

We represent the impact functions using Bernstein polynomials. We cast our estimation problem as a quadratic optimization problem with linear constraints. The estimation procedure allows us to deal with a large number of time series observations. For example in our empirical application we have approximately 46 million event updates. In order to recast our problem into a quadratic programming problem, we use a two step procedure, where in the first step the unknown baseline intensity is estimated.

Given the large sample size, to compare two competing estimators we can use the predictive sequential (prequential) log-likelihood (Dawid, 1984) or sample splitting (cf. Cox, 1975 and see also Section 2.2.2 for details). The derived test statistic is a martingale version of the Diebold-Mariano test statistic (Diebold and Mariano 1995), and it is asymptotically standard normal under the null of equal predictive performance.

As a byproduct, our use of Bernstein polynomials easily lends itself to estimation under linear restrictions to test monotonicity and convexity. As intuitively obvious, the use of constraints is useful when the estimation error dominates the approximation error. For example this is the case when considering recursive estimation with a rolling window to reduce bias. Then, the use of constraints reduces the higher estimation error resulting from the use of a smaller sample. This is substantiated by simulation results that are reported in the supplementary material.

Monotone and convex regression estimation and testing is a well established problem in statistics and often relies on splines (e.g., Ramsay, 1988, Meyer, 2008, Wang and Meyer, 2011, and references therein). This is an important problem in econometrics (e.g., Yatchew and Bos, 1997, Yatchew and Härdle, 2006). Unlike this literature, we do face some challenges due to the use of high frequency data and a continuous time model whose likelihood is not linear in the parameters. Our two step procedure allows us to impose constraints in a seamlessly way. Note that constraints are always needed to impose non-negativity of the intensity. Our methodology can be used with splines, but we found Bernstein polynomials easier to implement. Our simulations also report results using B-splines and we could not find any substantial improvement.

Finally, our focus is on predictive ability rather than consistency toward a true value. This is relevant if the goal is to manipulate the market in a profitable way. Nevertheless, we provide a concise heuristic explanation on why our procedure should be consistent under suitable regularity conditions (Section 2.4). We prefer an outline of the method of proof, as the tools to lend rigor to the arguments would not be new, but would require lengthy technical proofs.

2.1.2 Outline of the Paper

The paper is organized as follows. Section 2.2 introduces the stochastic intensity model for buy (sell) orders. There we describe the estimation methodology and the testing approach. Section 2.3 analyzes a high frequency dataset for crude oil futures. Section 2.4 provides a justification of our methodology via asymptotic arguments. Section 2.5 reports a set of numerical experiments to show that the asymptotic arguments hold in finite sample. The numerical experiments also compare the use of Bernstein polynomials to splines. Concluding remarks are in Section 2.6.

Supplementary material, in the form of an Appendix, collects further details about the testing procedure, the numerical experiments and the empirical study. Finally, the methodology discussed in the main part of the paper is applied to the Bitcoin market to assess its manipulability.

This paper also comes with companion code that can be downloaded from the following URL `https://github.com/asancetta/IntensityEstimation`. The code contains MATLAB functions to carry out the constrained estimation procedure proposed in this paper as well as an example of workflow to analyze datasets using our methodology.

2.2 The Model and its Estimation

Suppose that $(N(t))_{t\geq 0}$ is the number of trade arrivals. For the sake of definiteness, consider buy trades. Let $(X(t))_{t\geq 0}$ be a left-continuous stationary process representing K covariates. The counting process admits a stochastic intensity λ_0 such that

$$\lambda_0(t) = h_0(t)\, g_0(X(t)), \tag{2.1}$$

where h_0 is a baseline intensity (a stationary predictable process) and g_0 is a continuous additive function

$$g_0(x_{1,} \ldots, x_k) = \sum_{k=1}^{K} g_{0,k}(x_k). \tag{2.2}$$

We choose additivity as a compromise between flexibility and ease of interpretation of each covariate's impact. Notice that if $K > 1$ the model is identified up to a location shift. The intensity in (2.1) means that a.s.

$$\lim_{s \downarrow 0} \mathbb{E}(N(t+s) - N(t)|\mathcal{F}_t) = \lim_{s \downarrow 0} \Pr(N(t+s) - N(t) = 1|\mathcal{F}_t) = \lambda_0(t) \tag{2.3}$$

where we can assume that $\mathcal{F}_t$ is the σ-algebra generated by $(N(s), X(s))_{s \le t}$. Λ will denote the compensator of the counting process N, i.e., $\Lambda(t) = \Lambda((0,t])$ $= \int_0^t \lambda_0(s)\, ds$.

Hence, the counting process quantifies the magnitude of the trading activity, i.e., it counts the number of buy arrival orders instant by instant. It is influenced by a background noise (the baseline process h_0) and by a number of variables such as volume imbalance and spread. Exact definition of these variables shall be given in due course. At time t, these variables are represented by the K-dimensional variable $X(t)$, and the impact that the k^{th} variable has on the intensity is quantified by $g_{0,k}$. Next, we consider estimation of the model.

2.2.1 Estimation

By additivity of the model, there is no loss of generality in taking $K = 1$. This is to reduce the notational burden in favor of clarity of exposition.

We represent $g_0(x)$ in terms of a Bernstein polynomial of order J on $[0,1]$. This means that $g_0(x) = \sum_{j=0}^{J} a_j B_j(x)$, where $B_j(x) := \binom{J}{j} x^j (1-x)^{J-j}$ and the coefficients a_j are scalars. In consequence, the variables need to be mapped into $[0,1]$, see Section 2.7.5 in the Appendix for a discussion. In general, for arbitrary but continuous $g_0(x)$ on $[0,1]$, we have that

$$g_0(x) = \lim_{J\to\infty} \sum_{j=0}^{J} g_0\left(\frac{j}{J}\right) B_j(x), \tag{2.4}$$

where the equality holds under the uniform norm (Lorentz, 1986, Theorem 1.1.1).

The Loss Function

The sample size is large, hence particular attention has to be paid to computational aspects. To begin we replace the maximum likelihood estimator with a quadratic contrast function estimator. We follow a two step procedure. First h_0 is estimated (we shall discuss possible functional forms for h_0 in Section 2.2.1). Second we estimate the impact function g_0 using quadratic programming. We then iterate the procedure. In Section 2.4.2 we argue that this should minimize the full log-likelihood, under regularity conditions.

Suppose that h_T is a good estimator for h_0. To find an estimate of g_0, say g_T, we minimize the contrast function

$$R_T(a, h_T) := -2\int_0^T \frac{\sum_{j=0}^{J} a_j B_j(X(t))}{h_T(t)} dN(t) + \int_0^T \big(\sum_{j=0}^{J} a_j B_j(X(t))\big)^2 dt \tag{2.5}$$

with respect to (w.r.t.) $\{a_j \geq 0 : j = 0, 1, 2..., J\}$. Section 2.4 justifies this procedure. From (2.4), we can see that the constraint naturally ensures nonnegativity of the intensity.

Define the i^{th} jump time of N by $T_i := \inf\{s > 0 : N(s) \geq i\}$, with $T_0 = 0$. The covariates update at random event times possibly different from the T_i's. As a rule of thumb for order book covariates, the number of updates tends to be about 10 times more frequent than the number of trade updates. Let $\{t_j : j = 1, 2, ...\}$ be the times at which there is an update either in the counting process or the covariates. Note that $N(t) = N(T_i)$ for $t \in [T_i, T_{i+1})$ (right continuous) and $X(t) = X(t_{j-1})$ for $t \in (t_{j-1}, t_j]$ (left continuous). Suppose that we observe the process until time $T = T_n$ and that in this period there are m event updates, i.e., $0 = T_0 = t_0 < t_1 < t_2 < ... < t_m = T_n$.

Then, (2.5) becomes

$$-2\sum_{l=1}^{n}\frac{\sum_{j=0}^{J}a_jB_j\left(X\left(T_l\right)\right)}{h_T(T_l)}+\sum_{i=1}^{m}\left(\sum_{j=0}^{J}a_jB_j\left(X\left(t_{i-1}\right)\right)\right)^2\left(t_i-t_{i-1}\right).$$

The goal is to estimate the coefficients $a_0, a_1, ..., a_J$ subject to positivity constraints and possibly additional constraints.

From now on, let a be the vector of coefficients $(a_0, a_1, ..., a_J)'$, where the prime symbol $'$ stands for transpose. In matrix notation, the previous display becomes

$$-2a'\Phi'\Gamma + a'\Phi'\Sigma\Phi a. \tag{2.6}$$

Here, Φ has (i,j) entry $B_j\left(X\left(t_i\right)\right)$, Γ is a vector with i^{th} entry $1/h_T(T_l)$ if $t_i = T_l$ for some l (i.e., if t_i is a jump time of N) and zero otherwise; Σ is a diagonal matrix with $(i,i)^{th}$ entry $(t_i - t_{i-1})$.

When a new observation is collected, we only need to update $\Phi'\Gamma$ and $\Phi'\Sigma\Phi$, which are relatively low dimensional matrices ($(J+1)\times 1$ and $(J+1)\times(J+1)$, respectively). When we have $K>1$ covariates, the changes are conceptually trivial, and the dimensions of $\Phi'\Gamma$ and $\Phi'\Sigma\Phi$ become $K(J+1)\times 1$ and $K(J+1)\times K(J+1)$, respectively.

The Constraints

For expository reasons, we still consider $K=1$. We need to ensure that g_0 is positive if we do not want negative intensity. From (2.4) it is sufficient that the entries in the vector of coefficients a are positive. In certain circumstances, we may also wish $g_0\left(x\right)$ to be increasing. In this case, by the properties of Bernstein polynomials, we require the entries in a to satisfy $a_{j-1}\leq a_j$ for all j's, as a sufficient condition for monotonicity. Similarly, a convexity restriction can be imposed by requiring the second difference of a to be positive, i.e., $(a_{j+1}-a_j)-(a_j-a_{j-1})\geq 0$ for all j's. (See Section 2.7.1 in the Appendix, for details.) Mutatis mutandis, this is the same approach used for spline regression under shape constraints.

From the above remarks we deduce that we can solve the following minimization problem

$$\min_{a\in\mathbb{R}^{J+1}} -2a'\Phi'\Gamma + a'\Phi'\Sigma\Phi a \qquad s.t. \quad Ca\geq 0$$

for a matrix C suitably chosen to impose restrictions such as non-negativity, monotonicity and/or convexity.

The linear restrictions, to impose for example monotonicity, are sufficient but not necessary. They only become necessary in the limit as the order of polynomial $J \to \infty$. This latter set up is less feasible, as computational constraints require J relatively small. When using rolling window estimators with relatively small window sample, the noise level becomes high enough to make the distinction between necessary and sufficient practically irrelevant.

The Baseline Intensity and Feasible Two Step Estimation

The simplest case of baseline intensity is the standard exponential hazard function $h_0(t) = 1$. A more realistic choice is

$$h_0(t) = c_0 + \int_{(0,t)} e^{-\beta_0(t-s)} dN(s) = c_0 + \sum_{j \geq 0: T_j < t} e^{-\beta_0(t-T_j)} \tag{2.7}$$

where $\beta_0 > 1$ and $c_0 > 0$. This is proportional to a Hawkes process. A model similar to (2.1) with (2.7) as baseline intensity has been considered in Sancetta (2018). From Theorem 1 in Brémaud and Massoulié (1996) we can deduce that the process is stationary if $\mathbb{E}g_0(X(t))/\beta_0 < 1$. The Hawkes process is not always easy to estimate because of the possible presence of local maxima in the likelihood (Ogata and Akaike, 1982, for early mentions). In fact, amongst other reasons, alternative procedures to likelihood estimation have been proposed (e.g., Da Fonseca and Zaatour, 2014, Kirchner, 2017). However we shall use Hawkes process in Section 2.3 because of its relatively good fit to the data.

In Section 2.3 and in the simulations in the Appendix (Section 2.7.2), we consider the Weibull hazard function because of its flexibility and simplicity of estimation, as opposed to the Hawkes process. In this case $h_0(t) = \beta_0 (t - T_i)^{\beta_0 - 1}$ for $\beta_0 > 0$ when $t \in (T_i, T_{i+1}]$. Recall that T_i is the time of the i^{th} trade. When $g_0 = \gamma_0$, where γ_0 is a constant, this corresponds to durations being distributed as a Weibull random variable:

$$\Lambda((T_i, T_{i+1}]) = \int_{(T_i, T_{i+1}]} h_0(t)\gamma_0 dt = \gamma_0 (T_{i+1} - T_i)^{\beta_0}.$$

Then, by standard time change, $\Lambda((T_i, T_{i+1}])$ is an exponential random variable with mean one. In consequence, $1-\exp\left\{-\gamma_0 \left(T_{i+1}-T_i\right)^{\beta_0}\right\}$ is uniformly distributed and we deduce that the durations are distributed as Weibull random variables. We suggest to estimate the baseline intensity first supposing that $g_0 = \gamma_0$ and maximizing the log-likelihood. We then use the estimator for h_0 in the estimation of g_0 in (2.5). The procedure can then be iterated until convergence. The details are in Algorithm 2.1.

Algorithm 2.1 Intensity Estimation

Start with $g_T^{(0)}\left(X\left(t\right)\right) = \gamma$ an unknown constant. For each $v = 1, 2, ...,$ find the minimizer of $-\int_0^T \ln\left(h\left(t\right) g_T^{(v-1)}\left(X\left(t\right)\right)\right) dN\left(t\right) + \int_0^T h\left(t\right) g_T^{(v-1)}\left(X\left(t\right)\right) dt$ w.r.t. h and denote it by $h_T^{(v)}$. When $v = 1$ we shall also minimize w.r.t. $\gamma > 0$.
Minimize (2.6) w.r.t. $a \in \mathcal{A} \subseteq \left[0, \overline{a}\right]^K$ where $\overline{a}$ is some finite positive constant and $\mathcal{A}$ is defined by the linear constraints (see Section 2.2.1). Define the minimizer by $a_T^{(v)}$ so that $g_T^{(v)}\left(X\left(t\right)\right) = \sum_{j=0}^J a_{T,j}^{(v)} B_j\left(X\left(t\right)\right)$ is the estimator for $g_0\left(X\left(t\right)\right)$.
Stop when $h_T^{(v)}\left(t\right) g_T^{(v)}\left(X\left(t\right)\right)$ converges.

We may stop at $v = 1$ and still obtain reasonable results when h_0 and g_0 satisfy a certain orthogonality condition. In this case, we can estimate the baseline intensity using the durations only. In fact, the log-likelihood of N, at the true parameter $\lambda_0 = h_0 g_0$, is

$$L_T\left(\lambda_0\right) := \int_0^T \ln\left(h_0\left(t\right) g_0(X(t))\right) dN\left(t\right) - \int_0^T h_0\left(t\right) g_0\left(X(t)\right) dt \qquad (2.8)$$

(Ogata, 1978). If the following orthogonality condition holds

$$\frac{1}{T}\int_0^T h_0\left(t\right) g_0\left(X\left(t\right)\right) dt \simeq \frac{1}{T}\int_0^T h_0\left(t\right) dt \frac{1}{T}\int_0^T g_0\left(X\left(t\right)\right) dt$$

then,

$$\frac{L_T\left(\lambda_0\right)}{T} \simeq c_T + \frac{1}{T}\int_0^T \ln\left(h_0\left(t\right)\right) dN\left(t\right) - \frac{1}{T}\int_0^T g_0(X(t))dt \frac{1}{T}\int_0^T h_0\left(t\right) dt,$$

where $c_T := T^{-1}\int_0^T \ln\left(g_0\left(X(t)\right)\right) dN\left(t\right)$ does not depend on h_0. Hence, estimation of h_0 is approximately independent from estimation of g_0. In practice, this does not seem the case, and we require a few iterations.

2.2.2 Testing the Performance of Competing Models

Let $\lambda^{(1)}$ and $\lambda^{(2)}$ be two competing models for the intensity. Our aim is to assess which model is closest to the true one. For large datasets (millions of observations), it is natural to recast the inference problem within a predictive sequential (prequential) framework. We assess the value of the two models only on the basis of how the forecasts that they generate agree with the outcomes of the the point process N (e.g., Dawid, 1984, Dawid and Vovk, 1999). This is particularly suited for our purposes, as computational constraints can force us to rely on approximations.

In this framework at the beginning of each day i we re-estimate the model using data up to the $(i-1)^{th}$ day. The estimator for $\lambda^{(k)}$ $(k=1,2\ )$ using all the data until the $(i-1)^{th}$ day is denoted by $\hat{\lambda}^{(k)}_{i-1}$ and estimated using Algorithm 1, as described in Section 2.2.1. We use a hat instead of a subscript T to avoid notational oddities. This estimator is evaluated on data on the i^{th}-day. For two competing estimators of the intensity we form the prequential log likelihood for the i^{th} day:

$$L_i^{(k)} = \sum_{s=N_{i-1}+1}^{N_i} \left[\ln(\hat{\lambda}^{(k)}_{i-1}(T_s)) - \int_{T_{s-1}}^{T_s} \hat{\lambda}^{(k)}_{i-1}(t)dt \right]$$

where N_i is the number of jump times of the counting process until the i^{th}day. The prequential loglikelihood is given by $L_T^{(k)} = \sum_{i=1}^{I} L_i^{(k)}$ when we have a sample of I trading days so that $T = T_{N_I}$ (do not confuse $L_T^{(k)}$ with $L_i^{(k)}$). We suppose that on day 1 we have an estimator $\hat{\lambda}^{(k)}_0$ ($\hat{\lambda}^{(k)}_{i-1}$ with $i=1$) based on previous observations. The prequential loglikelihood ratio is $L_T^{(1)} - L_T^{(2)}$. Taking into account its asymptotic behaviour, we are able to design a test. Let q_α be the α quantile of the standard normal distribution, e.g., $q_{0.95} \simeq 1.64$. At the $(1-\alpha)\,\%$ significance level, reject model 2 in favor of model 1 if

$$\frac{L_T^{(1)} - L_T^{(2)}}{\sqrt{T\hat{\sigma}_T^2}} \geq q_\alpha \tag{2.9}$$

where

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{i=1}^{I} \sum_{s=N_{i-1}}^{N_i} \left[\ln\left(\hat{\lambda}^{(1)}_{i-1}(T_s) / \hat{\lambda}^{(2)}_{i-1}(T_s) \right) \right]^2. \tag{2.10}$$

To facilitate the graphical analysis of our results, in the empirical version we split the sample into two parts, the first is used to estimate the parameters, the second to test. Mutatis mutandis, this is a one period version of the methodology described above. Sample splitting has a long tradition in statistics and econometrics (inter alia, Cox, 1975, for an early reference, Yatchew, 1992, for testing restrictions in regression models). In this case, $\hat{\lambda}_0^{(k)}(\cdot)$ is estimated in the first part of the sample (the estimation sample), and it is used throughout in the validation sample $(0, T]$.

Suppose that for both competing models we compute

$$\frac{1}{N_{T_I}} \sum_{i=1}^{I} \sum_{s=N_{i-1}+1}^{N_i} \int_{T_{s-1}}^{T_s} \hat{\lambda}_{i-1}^{(k)}(t)dt = \bar{\lambda}^{(k)}. \tag{2.11}$$

Under stationarity and ergodicity, $\bar{\lambda}^{(k)}$ converges to a constant: in fact, the display in (2.11) is equal to the time average of the (predictable) intensity$\hat{\lambda}_{i-1}^{(k)}$ divided by the average counts N_{T_I}/T_I. By the definition of ergodicity and stationarity the numerator and denominator converge to a constant (Lemma 2, Ogata, 1978). We can scale our intensities by $\bar{\lambda}^{(k)}$ so that the first "out of sample" moment of $\int_{T_{s-1}}^{T_s} \hat{\lambda}_{i-1}^{(k)}(t)/\bar{\lambda}^{(k)}dt$ is one. This ensures that both models match the theoretical first moment of the true intensity measure $\Lambda\left((T_i, T_{i+1}]\right)$ (see Section 2.2.1). This standardization has the advantage of removing the effect of the first moment in the model comparison. We can think of this to be equivalent to removing the effect of the intercept in a regression context. Then, we can just define $L_i^{(k)} = \sum_{s=N_{i-1}+1}^{N_i} \ln(\hat{\lambda}_{i-1}^{(k)}(T_s)/\bar{\lambda}^{(k)})$ and use $\hat{\lambda}_{i-1}^{(k)}(T_s)/\bar{\lambda}^{(k)}$ instead of $\hat{\lambda}_{i-1}^{(k)}(T_s)$ in the calculation of the variance. We shall use this approach in the empirical results, as our main interest is not in the scaling of the intensity. However, this did not have a substantive impact in the results. In our simulations, we did not use this scaling or equivalently, we set $\bar{\lambda}^{(k)} = 1$.

2.3 Spoofing the Crude Oil Futures Market

We consider high frequency data on the crude oil front month futures traded on the CME (CME ticker CL). The sample period is from 01/May/2013 to 30/Sept/2013 each day from 13:30 to 18:00 GMT. The time interval for each day is based on liquidity considerations. We use a proprietary data set that comprises of all market trades and book updates. The data were collected by a proprietary trading group, in a server collocated in the Aurora data center in

Chicago. The messages were time stamped at the nanosecond resolution. The trades were accurately classified as buy or sell. Moreover, in busy times, when many trades are executed, CME might not send the resulting book update for some time as there is a limit in the size of each packet being sent through the network. For this reason, if a trade arrives and the book is not updated, we construct an imputed book. Again this operation is admissible (was carried out in live trading) and avoids any bias due to lack of synchronicity. Finally, we also subtract 400 microseconds from trade times in order to account for some delay on the side of CME when sending trade messages as opposed to order book messages. All these operations were chosen to match closely both trading and latency and were based on empirical analysis of network data. We do so to avoid the risk of asynchronicity and consequently spurious relations. To summarize, the data processing and variables construction is the same as in live trading to ensure that we do not "peep into the future".

A few days are missing in our data set. In total, we have complete data for 94 trading days. The total number of updates in the data set is in excess of 46 million. The number of trade events (both buy and sell) is about 3.4 million. For our best model we have 7 covariates and a Bernstein polynomial of order 8 (i.e., 9 basis functions for each covariate for a total of 63). This means a data matrix that has 46 million rows and 63 columns, i.e., almost 3 billion data entries. Our estimation procedure had no problems to deal with such problem in RAM, and was rather fast (in the order of minutes to parse data, a day at the time, and estimate a model).

We model the intensity for buy and sell trades separately using the model in (2.1). We consider two specifications for the baseline intensity h_0 in the first estimation step. In particular, we estimate a Weibull hazard function and a Hawkes process. We test which model is best suited to our data sample. We then produce a graphical plot of the impact functions for the best model. Heuristically, this allows us to see how the market could be manipulated.

2.3.1 The Model

We estimate (2.1) with Weibull and Hawkes baseline intensities (see Section 2.2.1). We consider g_0 modelled by a second and an eight order Bernstein polynomial (i.e., $J = 2, 8$ in (2.5)). We use Algorithm 2.1 for the estimation. The covariates are reported in Table 2.1. Details regarding the calculations of the variables are in Section 2.7.5 in the Appendix. Here, we just give an overview. We apply exponential moving average (EWMA) filters to some of the covariates. The EWMA of a variable $X(t_i)$ with smoothing parameter α is

$$EWMA(X(t_i)) = \alpha EWMA(X(t_{i-1})) + (1-\alpha) X(t_i) \tag{2.12}$$

where $EWMA(X(t_1)) = X(t_1)$. Here, t_1 is the time of the first update in the variable X at the start of each day. EWMA's are computed for each day. Note that the covariates update at discrete times that are different from the trade updates T_j's which have also been adjusted by 400 microseconds as mentioned above. We then sample the data at times that are the union of each covariate update and the times T_i. To ensure that the covariates are predictable, we make them left continuous by lagging them after sampling at times that are the union of all the observed updates.

All variables are mapped linearly into $[0,1]$, except for spread and durations that are first capped and then linearly mapped in $[0,1]$. The top of book volume imbalance $volImb$, is defined as

$$volImb = \frac{bidSize - askSize}{bidSize + askSize} \tag{2.13}$$

where $bidSize$ is the bid size (quantity) at the best bid, and similarly for $askSize$. This variable takes values in $[-1,1]$. We map it to $[0,1]$ by standard linear transformation: multiply by two and subtract one. The trade imbalance is computed from the EWMA of the signed traded volume every time there is a trade. We then divide it by the EWMA of the unsigned volumes. The EWMA's parameter is $\alpha = 0.98$ for both denominator and numerator. Durations are in seconds with nanosecond decimals, capped to one second. They are then passed to EWMA filters with parameter $\alpha = 0.98$ and 0.90. The spread is capped to 4 ticks and standardized by 4. Hence, the minimum value it can take (excluding choice prices) is 0.25. After the application of EWMA's filters, our additive model (2.1-2.2) has 7 covariates.

Table 2.1: Covariates used for estimation. The column “Smoothing” reports the smoothing parameter used if an EWMA had been applied to the original variable.

Variables	Short Name	Smoothing
Volume Imbalance Level 1	VolImb1	
Volume Imbalance Level 2	VolImb2	
Volume Imbalance Level 3	VolImb3	
Spread	Spread	
Trade Imbalance	TrdImb98	$\alpha = 0.98$
Durations	Dur98 Dur90	$\alpha = 0.98$ and 0.90

2.3.2 Comparison of the Models with Weibull and Hawkes Baseline Intensities

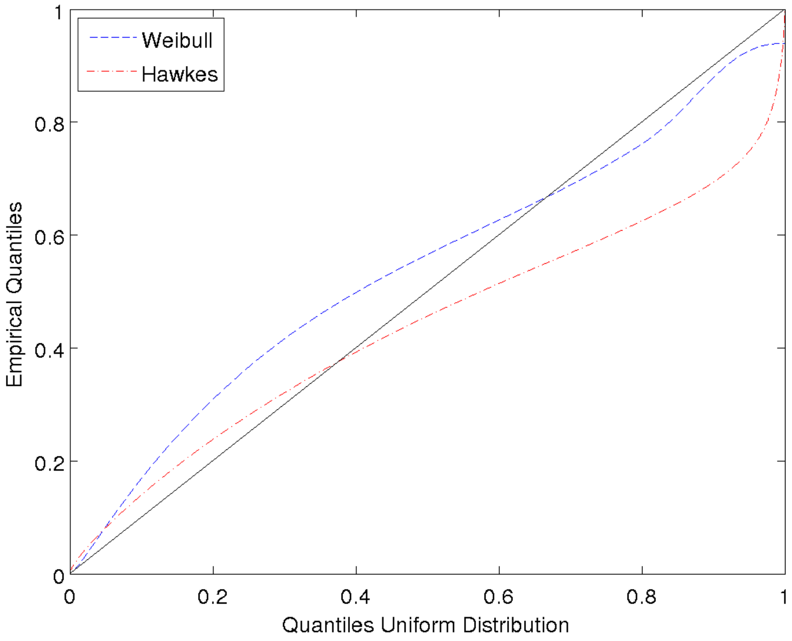
In this section, we have three main goals. First, we want to verify the extent of non-linearity in the impact functions. We test a low ($J = 2$) versus a high ($J = 8$) complexity model. Second, we want to verify whether a Weibull baseline intensity is an acceptable substitute to a Hawkes baseline intensity. In this respect, we are not just interested in whether we reject a Weibull in favour of a Hawkes intensity, but also to what extent the resulting estimators for g_0 may differ. A Weibull baseline intensity is easier and faster to estimate. Hence there needs to be a clear gain to justify the use of a Hawkes baseline intensity. Third, and most important, we want to understand the shape of the impact functions to see by visual inspection whether the crude oil futures market could be manipulated, at least during the sample period we consider.

We compute $\Lambda\left(\left(T_{i-1}, T_i\right]\right) = \int_{\left(T_{i-1}, T_i\right]} h_0\left(t\right) g_0\left(X\left(t\right)\right) dt$ for $i = 1, 2, ..., n$, where h_0 and g_0 are replaced by their estimators. As mentioned previously, here n is in the order of 3.4 million observations. If the estimate fits the data well, the data sequence

$$1 - \exp\left\{-\Lambda\left(\left(T_{i-1}, T_i\right]\right)\right\} \tag{2.14}$$

$i = 1, 2, ..., n$ forms a sequence of independent identically distributed (i.i.d.) uniform random variables in $[0, 1]$ (Brémaud, 1981, Ch.II, Theorem 16). Figure 2.1 shows the qq-plot of the estimated transformed data sequences that use the Weibull baseline intensity and the Hawkes baseline intensity with an

Figure 2.1: Plot of Uniform Transform of the Estimated Time Changed Data for Buy Trades for Hawkes and Weibull intensities. The closer are the quantiles of the sample to the theoretical uniform ones (the solid line) the better is the fit. The model with Weibull baseline intensity appears to be a better fit to the data.



eight order Bernstein polynomial estimator of g_0. The model with Weibull baseline intensity appears to be a better fit to the data.

We also analyzed the autocorrelation function for the same data sequence. Both sequences showed some autocorrelation, but the Hawkes model fared better than the Weibull baseline intensity, in this respect (see Section 2.7.6 in the Appendix for the actual plot).

Estimation of the Impact Functions

We estimate the models using a second order polynomial with Weibull (B2W) and Hawkes (B2H) baseline intensities. We compare to more complex models that use an eight order polynomial with Weibull (B8W) and Hawkes (B8H) baseline intensities. We have a total of 4 competing models. Their relative merits are assessed by the test procedure described in Section 2.2.2. To this end, the sample was split into two parts. The first 67% of the sample was used for the estimation of the four intensities. The last 33% was used to compute the test statistic. Results are reported in Table 2.2.

Table 2.2: Test of Model Performance. The test statistic (t-stat) is as in (2.9) after standardization of the intensity by (2.11). The columns identify the null hypothesis as described in Section 2.2.2. For example, B8W-B2W is the null that B8W and B2W perform the same, versus an alternative that B8W performs better than B2W. B8W (B8H, respectively) performs better than B2W (B2H, respectively) and B8H performs better than B8W.

	B8W-B2W		B8H-B2H		B8H-B8W	
	Buy	Sell	Buy	Sell	Buy	Sell
t-stat	563.98	589.55	5215.4	5139.5	10270	10108
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Given the size of the data set, an 8^{th} order Bernstein polynomial can be estimated with reasonable degree of accuracy (in terms of estimation error) and there is no need to impose a constraint beyond non-negativity for this case. When we impose non-negativity, we note that even though the total number of parameters to estimate is equal to 63, only approximately 30% are non zero. This is essentially irrespective of which baseline intensity we use. This is consistent with our remarks about sign constraint in Section 2.1.1. The test overwhelmingly favours B8H, the intensity with Hawkes baseline intensity and g_0 modelled by an eight order Bernstein polynomial. To gauge the impact of the variables, Figure 2.2 plots the estimated impact functions of $VolImb1$, $VolImb2$, $VolImb3$ and $Spread$ for B8H. We are interested in the functional form of these impact functions because the related covariates are the ones that can be distorted by a market manipulator in order to trigger a trade. Note that our transformation of the spread induces an artificial increase in the impact function at zero. However, the transformation of the spread results in $Spread$ taking values in $\{0.25, 0.5, 0.75, 1\}$ only. At first sight, it is surprising that a higher spread leads to higher intensity. However, a higher spread can result from a series of correlated aggressive trades that deplete liquidity on ones side of the book. This would be associated to a higher intensity. Moreover, a higher spread may induce a market participant with high urgency of trading to trade. This is because a high spread may reduce the probability to get filled sitting on the order book. In consequence, a participant with high urgency would just cross the book. Finally, note that liquidity providers keep the spread directly proportional to volatility. Periods of high volatility are periods with high trading activity.

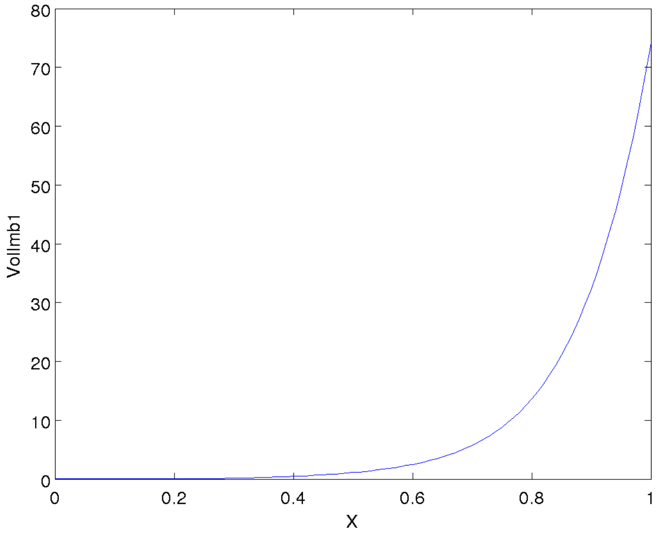
To make different models comparable, the vector of coefficients in g_T have been scaled to have unit Euclidean norm. In the Appendix we provide the complete set of results for all the impact functions using the Hawkes and Weibull intensities and we show that results appear visually similar despite the test results in Table 2.2.

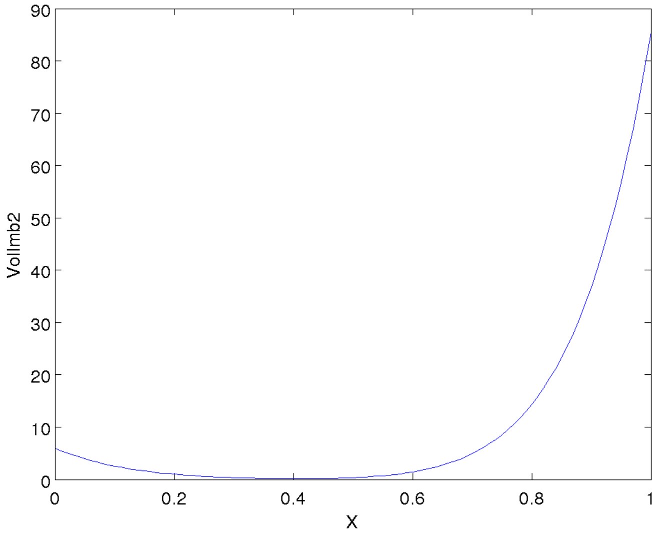
2.3.3 Implications for Market Manipulation

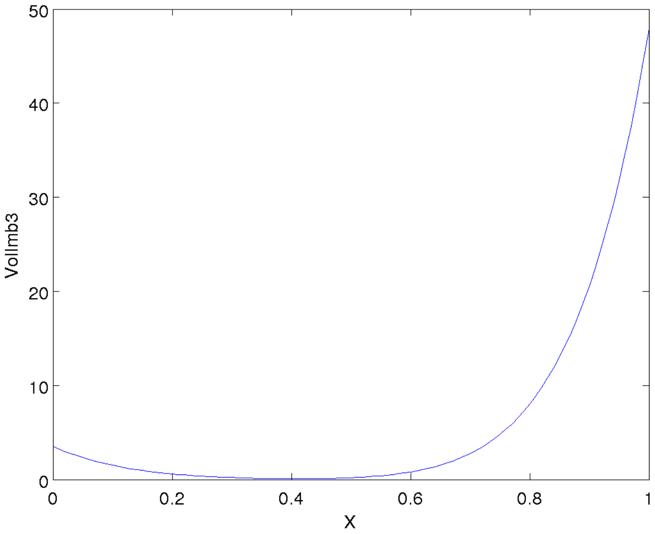
Heuristically, the empirical results suggest that under the right conditions, a trader could have manipulated the crude oil futures markets placing progressively large orders on the top of book. This could have been done as follows. Suppose that the objective is to sell at the ask rather than at the bid, in order to gain half of the spread. To do so, we need a relatively active period (small durations) where there is possibly a negative trade imbalance (recall TrdImb98 is mapped to $[0, 1]$ so TrdImb98 close to zero means a negative trade imbalance). Possibly the spread is greater than a tick; a tick is the minimum spread size which in the plot corresponds to 0.25. Moreover we want the volume quoted on the ask to be relatively thin. From Figure 2.2 (see also Figure 2.4 in the Appendix), we can see that in this case, the intensity of a buy order is relatively high. We can make the intensity even higher, by placing a relatively large order on the top of book bid and a small order on the top of book ask. This creates a positive volume imbalance. Given that the spread is wider than a tick, we can step inside the spread and place another relatively large order on the new top of book bid and/or a relatively small order on top of best ask. The latter ensures that we are top of book on the ask. In consequence of these actions, we have generated large VolImb1 and VolImb2 under particularly favorable conditions, i.e., a high intensity for buy trades. In consequence, our small order on the ask side should be filled despite the fact that we have also placed larger orders on the bid size.

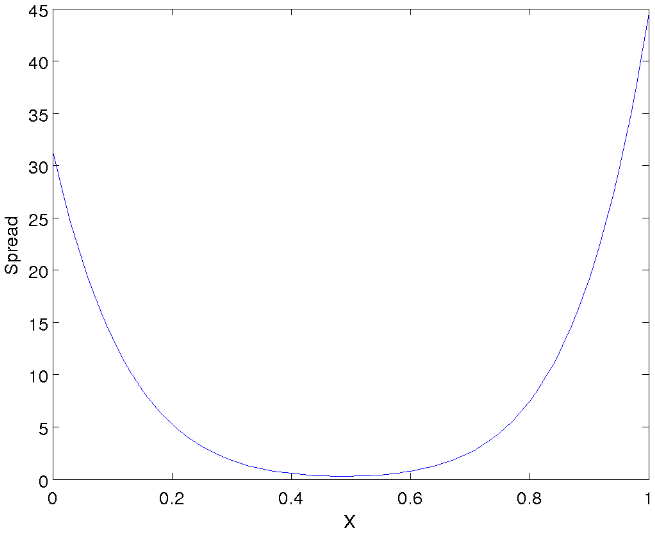
Such procedure can be easily implemented by a trading algorithm. Similarly, an algorithm can easily monitor trades that occur under these favorable circumstances and flag them as suspicious.

Figure 2.2: B8H Manipulable Impact Functions for Buy Trades. The estimated impact functions from B8H (the unconstrained 8^{th} order Bernstein estimator with Hawkes baseline intensity) are plotted for VolImb1, VolImb2, VolImb3, Spread. These variables are easy to manipulate. As expected, the impact functions, for the volume imbalance, are increasing in the interval $[0.5, 1]$ (this interval corresponds to a positive volume imbalance).






50
40
30
20
10
0
VolImb3
0
0.2
0.4
0.6
0.8
1
X


45
40
35
30
25
20
15
10
5
0
Spread
0
0.2
0.4
0.6
0.8
1
X

2.4 Justification for the Estimation and Testing Procedure

In this section we justify the inferential methodology discussed in the previous sections. The intensity $\lambda_0(t) = \lambda_0(t, \omega)$ is a continuous time stochastic process, i.e., a function of two variables $t \geq 0$ and $\omega \in \Omega$ where $(\Omega, \mathcal{B}, P)$ is a probability space and for each t, $\lambda_0(t, \cdot)$ is measurable on Ω. Similarly $h_0(t) = h_0(t, \omega)$ and $X(t) = X(t, \omega)$. The covariate process X and the baseline intensity are adapted stationary ergodic processes. All the quantities are supposed to be left continuous with right hand limits. For ease of notation, we may freely switch between $\lambda_0(t)$ and $h_0(t)\, g_0(X(t))$ and compactly write $\lambda_0 = h_0 g_0$. To make the notation simpler and more readable, we shall write $P h_0 g_0$ to mean $\int_\Omega h_0(0, \omega)\, g_0(X(0, \omega))\, dP(\omega)$ and similarly for other quantities. Both the set of positive functions and the set of monotonic/convex (or concave) functions form a convex set therefore we view g as an element in the closure of a convex subset $\mathcal{C}$ of the Hilbert space $L^2\left([0,1]^K\right)$. This framework is coherent with the hypotheses of the following classical result (see e.g., Corollary 3.23 in Brezis, 2011)

Theorem 2.1. *Let $(E, \langle \cdot, \cdot \rangle_E)$ be a Hilbert space and let $A \subset E$ be a nonempty, closed, convex subset of E. Let $\phi : A \to (-\infty, +\infty]$ be a convex lower semi-continuous function such that $\phi \not\equiv +\infty$ and*

$$\lim_{x \in A, \|x\|_E \to \infty} \phi(x) = \infty$$

if A is unbounded ($\|e\|_E^2 := \langle e, e \rangle$ for every elements $e \in E$). Then ϕ achieves its minimum on A, i.e., there exists some $x_0 \in A$ such that

$$\phi(x_0) = \inf_{x \in A} \phi(x).$$

Remark 2.1. Note that the closure of a convex set is a convex set.

We suppose that the true parameter g_0 belongs to the closure of $\mathcal{C}$ ($\lambda_0 = h_0 g_0$) so that the model for g_0 is not misspecified.

2.4.1 The Quadratic Risk Functional

We justify the procedure used to estimate g_0. Assume that the intensity is uniformly bounded away from zero. For any fixed h_0 we can define the following functional $g \mapsto P(g-g_0)^2$ on $\mathcal{C}$. A straightforward computation shows that this functional is strictly convex and its unique minimum is reached at $g = g_0$. Expanding the square and neglecting the term Pg_0^2 we get a functional having the same two properties (strict convexity and minimum at the same point), that is $g \mapsto R(g, h_0) := -2Pgg_0 + Pg^2$. In practice P is unknown, and we consider the sample counterpart

$$-\frac{2}{T}\int_0^T g(X(t))g_0(X(t))dt + \frac{1}{T}\int_0^T g^2(X(t))dt.$$

Given that $\lambda_0 = h_0 g_0$ is the compensator of dN, we have that the limit of the above display is equal to (cf. Equation (3.4) in Ogata, 1978)

$$\begin{aligned}
&\lim_{T\to\infty}\left(-\frac{2}{T}\int_0^T \frac{g(X(t))}{\lambda_0(t)} g_0(X(t))dN(t) + \frac{1}{T}\int_0^T g^2\left(X\left(t\right)\right)dt\right)\\
=&\lim_{T\to\infty}\left(-\frac{2}{T}\int_0^T \frac{g\left(X(t)\right)}{h_0(t)}dN\left(t\right) + \frac{1}{T}\int_0^T g^2\left(X\left(t\right)\right)dt\right)
\end{aligned}$$

almost surely. By ergodicity (Ogata, 1978, Lemma 2) the above display is equal to $R(g, h_0)$ almost surely. Hence we have that

$$-\frac{2}{T}\int_0^T \frac{g(X(t))}{h_0\left(X(t)\right)}dN(t) + \frac{1}{T}\int_0^T g^2(X(t))dt \to R(g, h_0) \tag{2.15}$$

almost surely. This is the motivation for estimating g_0, minimizing the objective function (2.5). Next we provide some heuristic justification for the procedure when h_0 needs to be estimated.

2.4.2 The Two-Step Procedure

We heuristically show that, asymptotically, the first order optimality condition w.r.t. the parameter g is the same regardless of whether we adopt the log-likelihood or the quadratic objective function defined in (2.15). This means that the minimization procedure described in Algorithm 1 should provide asymptotically consistent results under regularity conditions.

First we show that the negative log-likelihood ratio $-\left[L_T\left(\lambda\right)-L_T\left(\lambda_0\right)\right]$ is asymptotically quadratic for $\lambda=hg$. Recall that $L_T\left(\lambda\right)$ is as in (2.8), and λ_0 as in (2.1). The maximizer of $L_T\left(\lambda\right)$ is the same as the minimizer of the negative log-likelihood ratio $-\left[L_T\left(\lambda\right)-L_T\left(\lambda_0\right)\right]$, this is why we focus on the latter. Note that

$$\begin{aligned}
-\left[L_T(\lambda)-L_T\left(\lambda_0\right)\right] &= -\frac{1}{T}\int_0^T \ln\left(\frac{\lambda(t)}{\lambda_0(t)}\right)dN(t)+\frac{1}{T}\int_0^T\left(\lambda(t)-\lambda_0(t)\right)dt \\
&\simeq -\frac{1}{T}\int_0^T \ln\left(\frac{\lambda(t)}{\lambda_0(t)}\right)\lambda_0(t)dt+\frac{1}{T}\int_0^T\frac{\left(\lambda(t)-\lambda_0(t)\right)}{\lambda_0(t)}\lambda_0(t)dt.
\end{aligned}$$

By the approximation $\ln\left(1+x\right)\simeq x-x^2/2$ for $x>-1$, applied to $x=(\lambda(t)/\lambda_0(t))-1$, the above display is approximately equal to

$$\frac{1}{2T}\int_0^T\frac{\left(\lambda(t)-\lambda_0(t)\right)^2}{\lambda_0^2(t)}\lambda_0(t)dt\simeq\frac{1}{2}P\left[\frac{\left(\lambda-\lambda_0\right)^2}{\lambda_0}\right] \tag{2.16}$$

where the r.h.s. follows by ergodicity of the point process. We explicitly write $\lambda=hg$. The first order condition for a minimum of (2.16) w.r.t. $g\in\mathcal{C}$, satisfies the variational inequality

$$P\left[\left(hg-\lambda_0\right)\frac{h}{\lambda_0}v\right]\geq 0$$

for any $v=v\left(X\left(0,\omega\right)\right)$ such that $v+g_0\in\mathcal{C}$. Denote by $R_T(g,h)$ the term obtained from the l.h.s. of (2.15) replacing h_0 with h. By ergodicity,

$$\begin{aligned}
R_T\left(g,h\right) &\simeq -\frac{2}{T}\int_0^T\frac{g(X(t))}{h(t)}\lambda_0(t)dt+\frac{1}{T}\int_0^T g^2(X(t))dt \\
&\simeq P\left(-2\frac{g}{h}\lambda_0+g^2\right).
\end{aligned}$$

The first order condition for a minimum w.r.t. $g\in\mathcal{C}$ satisfies the variational inequality

$$P\left[-\frac{1}{h}\lambda_0 s+gs\right]\geq 0$$

for any $s=s\left(X\left(0,\omega\right)\right)$ such that $s+g_0\in\mathcal{C}$. The above display can be rewritten as

$$P\left[\left(hg-\lambda_0\right)\frac{s}{h}\right]\geq 0.$$

If g_0 is inside $\mathcal{C}$ (not on the boundary) and eventually $h \to h_0$, the variational inequalities become equalities and hold not just for $v + g_0, s + g_0 \in \mathcal{C}$ but for arbitrary continuous bounded functions v and s including the case $\frac{h}{\lambda_0} v = \frac{s}{h}$. Suppose that $h(t)$ is a continuous function of the time from the last jump, i.e., $t - T_i$ when $t \in (T_i, T_{i+1}]$. Then, we have that, asymptotically, whether we use the log-likelihood or the contrast function R_T, the first order conditions imply that g satisfies $P[(hg - \lambda_0)\rho] = 0$ for any $\rho = \rho(\omega, X(0, \omega))$ continuous in both arguments (e.g., set $\rho = s/h$ or $\rho = hv/(h_0 g_0)$).

To conclude the heuristic justification of the estimation procedure it suffices to rely on standard results about the convergence of the Gauss-Siedel method, e.g., Theorem 1 in Mammen et al. (1999).

2.4.3 Comparing Two Intensity Estimators

For ease of notation, we consider the sample split procedure. The result also applies to the intensity that is recursively estimated. All that we need is a measurable intensity that is bounded away from zero and infinity. Define the predictable part of the log-likelihood $L_T^{(k)}$ as

$$H_T^{(k)} := \int_0^T \ln \hat{\lambda}_0^{(j)}(t)\, d\Lambda(t) - \int_0^T \hat{\lambda}_0^{(j)}(t)\, dt.$$

Here $\hat{\lambda}_0^{(k)}$ is estimated on a sample up to time 0 (see the end of Section 2.2.2). Recall that $\Lambda(t)$ is the compensator of $N(t)$. Define the predictable part of the log-likelihood ratio as

$$\epsilon_T := H_T^{(1)} - H_T^{(2)}. \tag{2.17}$$

Under the null hypothesis, we suppose that $\epsilon_T = o_p\left(\sqrt{T}\right)$. Loosely speaking, the intensities $\hat{\lambda}_0^{(1)}$ and $\hat{\lambda}_0^{(2)}$ give similar predictions asymptotically, i.e., the predictable part of the log-likelihood ratio diverges at a rate slower than $\sqrt{T}$. The following can be used to justify (2.9).

Theorem 2.2. *Suppose that the $\hat{\lambda}_0^{(j)}$s are bounded away from zero and infinity. If $\hat{\sigma}_T^2$ in (2.10) converges in probability to a strictly positive constant, and $\epsilon_T = o_p\left(\sqrt{T}\right)$, then,*

$$\frac{L_T^{(1)} - L_T^{(2)}}{\sqrt{T\hat{\sigma}_T^2}} \to Z \tag{2.18}$$

in distribution where Z is a standard normal random variable.

Proof. Using the definition of ϵ_T we have that

$$\frac{L_T^{(1)} - L_T^{(2)}}{\sqrt{T\hat{\sigma}_T^2}} = \frac{\int_0^T (\ln(\hat{\lambda}_0^{(1)}) - \ln(\hat{\lambda}_0^{(2)}))dM(t)}{\sqrt{T\hat{\sigma}_T^2}} + \frac{\epsilon_T}{\sqrt{T\hat{\sigma}_T^2}}$$

where M is the martingale given by $M(t) = N(t) - \Lambda(t)$. Given that the intensities $\hat{\lambda}_0^{(k)}$ are bounded away from zero and infinity by hypothesis, the same holds true for their logarithm. Hence, we can follow the proof of Proposition 1 in Sancetta (2018) to bound the first term on the right hand side of the display. Given that $\epsilon_T = o_p\left(\sqrt{T}\right)$, we apply Slutsky Theorem to deduce the convergence in distribution of the left hand side of the above display. □

2.5 Numerical Experiments

In the previous section we have provided a heuristic justification of the statistical methodology introduced in Section 2.2: in the present section we use simulations to further validate it. We compare the true known parameters to their estimators and the estimated level of significance of the test to its theoretical one.

We have multiple goals in mind. 1. We verify that our estimator approaches the true value as the sample size increases. 2. We show under what circumstances the use of an additional constraint such as monotonicity can improve the estimation relative to the unconstrained estimator. 3. We compare our estimation results using Bernstein polynomials to the more classical spline smoothing estimator. 4. We estimate the size and power of the test procedure in Section 2.2.2.

Model simulation. We consider four different models for the intensity λ_0, namely $\lambda_0(t) = g_0(X(t))h_0(t)$, where $g_0(x) = 1, x+0.1, x^2+0.1, -x^3+x+0.5$; h_0 is proportional to the Weibull hazard function and in particular we set β_0 as defined in Section 2.2.1 equal to one, i.e., $h_0 = 1$. However, in the estimation β_0 is unknown and need to be estimated.

To simplify the framework, we consider a one dimensional covariate $X\left(t\right)$ such that $X\left(t\right) = X\left(T_{i-1}\right)$ for $t \in (T_{i-1}, T_i]$ where T_i it the time of the i^{th} jump of the counting process N. The sample period is $[-S, T]$ where $S = T_{-(n-1)}$ and $T = T_n$. We split the sample into estimation sample $[-S, 0]$ and test sample $(0, T]$ and consider $n = 100, 1000, 10000$. The $X\left(T_{i-1}\right)$'s are independent identically distributed uniform random variables in $[0, 1]$. We shall refer to the number of jump points n as the sample size.

Recall that $\Lambda((T_i, T_{i+1}])$ is an exponential random variable with mean one. Given the assumptions made above, we have that

$$\Lambda((T_i, T_{i+1}]) = \int_{T_i}^{T_{i+1}} g_0(X(t))h_0(t)dt = g_0(X(T_i))(T_{i+1} - T_i).$$

Then, the $(i+1)^{th}$ duration can be simulated from a exponential distribution with parameter $g_0(X\left(T_i\right))$. We use 1000 simulations, and for each simulation, the simulated sample is $\{(T_i, X\left(T_{i-1}\right)) : i = -\left(n-1\right), -\left(n-2\right), ..., n\}$.

Model estimation and test. We use sample split and estimate the model on the first $n/2$ observations. We do so as we focus on assessing the testing procedure. We use the procedure in Algorithm 1 in Section 2.2.1 to estimate the model from $\left\{\, (T_i, X\left(T_{i-1}\right)) : i = -\left(n-1\right), -(n-2), \ldots 0\right\}$. The number of iterations is five. We correctly suppose a Weibull baseline intensity. We use a Bernstein basis of order $J = 5, 10, 20$ for estimation of g_0 in the second step. For comparison we also estimate g_0 in the second step using B-splines of degree three with 4, 8, 16 knot points.

The unconstrained estimator is denoted by $\hat{g}^{(uncon)}$ and estimated on $[-S, 0]$ (we avoid a subscript to avoid notational oddities). Note that we always impose the non-negativity constraint even for this estimator. We denote by $\hat{g}^{(con)}$ the estimator based on convex increasing constraints estimated on $[-S, 0]$. The simulated models for g_0 do satisfy this constraint except for the case $g_0\left(x\right) = -x^3 + x + 0.5$. Hence, $\hat{g}^{(con)}$ will be biased in this case. In this case the null that the true g_0 satisfies the constraints is false. This will allow us to evaluate the power of the test.

On the second half of the sample, we carry out a test for the null $H_0 : g_0$ is convex and increasing, against an unconstrained alternative. We use Theorem 2.2 to construct the t-statistic and the critical values.

2.5.1 Test and Model Fit

To assess the sample fit, we approximate the integrated mean square error. To evaluate the finite sample performance of the test procedure in Section 2.2.2 we compute the frequency of rejections of the null. In particular, we compute the following quantities which we shall refer to throughout this section.

T1: the frequency of rejections of the the null hypothesis according to the rule (2.9) with $q_{0.95}$, where model 1 is the unconstrained and model 2 is the constrained: reject the null if the unconstrained model performs better. We use a 5% level of significance, hence $q_{0.95}$.

Recall the definition of ϵ_T in (2.17). Here, we want to test that $\epsilon_T/T \leq 0$ under the null. In fact, if the constrained model performs better than the unconstrained $\epsilon_T/T < 0$. If we use Theorem 2.2, this makes the test undersized: the probability of rejecting the null is lower than the nominal level. Hence, in a simulation, we cannot use T1 to assess a Type I error, unless the unconstrained and the constrained model are equally good. This is unlikely and it would defeat the point of imposing a constraint in order to improve fit. To ensure that $\epsilon_T = o\left(\sqrt{T}\right)$ as in Theorem 2.2, we compute an alternative to T1.

T2: the frequency of rejections of the the null hypothesis according to the rule

$$\frac{\left[L_T^{(uncon)} - L_T^{(con)}\right] - \hat{\epsilon}_T}{\sqrt{T\hat{\sigma}_T^2}} \geq q_{0.95},$$

where $\hat{\epsilon}_T$ is the Monte Carlo estimate of (2.17). This is as T1 once we subtract $\hat{\epsilon}_T$ from the loglikelihood ratio. Hence, this allows us to study the the size of the test under the null that both models are equally good.

In summary, T1 is useful to compute the power of the test, i.e., the probability of rejecting the null when it is false. On the other hand to verify whether the normal approximation is acceptable when $\epsilon_T = 0$, we focus on T2. We also need to verify that when the constraint holds ($\epsilon_T/T < 0$), T1 is undersized.

MSECon: the Monte Carlo approximation of

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=-(n-1)}^{0}\left(g_0(X(T_{i-1})) - \hat{g}^{(con)}(X(T_{i-1}))\right)^2\right]$$

which is the mean square error of the constrained estimator.

MSEUncon: same as MSECon but using $\hat{g}^{(uncon)}$ instead of $\hat{g}^{(con)}$.

StdCon: the standard deviation of the errors of the constrained estimators. Note that we use 1000 simulations, so the standard errors are obtained dividing by (approx.) 31.6.

StdUncon: the standard deviation of the errors of the unconstrained estimators.

Tables 2.3 and 2.4 report an excerpt for our largest $n = 10000$, which is still very small given the sample sizes we deal with in empirical work. These results are for the higher dimensional models using Bernstein and splines basis. The two cases are comparable: the dimension of the Bernstein basis is 21 whereas the dimension of the vector space generated by the B-spline is 19.

Table 2.3: Bernstein Basis. Twentieth degree, n=10000. All the functions, except the fourth, satisfy the constraints. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	1e-03	0.047	2.144e-04	0.0013	2.9831e-04	0.0024
$x + 0.1$	1e-10	0.038	1.6899e-04	6.1285e-04	3.6829e-04	0.0016
$x^2 + 0.1$	0.003	0.047	2.0363e-04	4.5414e-04	5.1982e-04	0.0015
$-x^3 + x + 0.5$	1	0.055	0.0131	6.454e-04	0.0136	8.9886e-04

Table 2.4: B-Spline. Third degree, sixteen knot points, n=10000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	1e-10	0.044	2.4886e-04	0.002	5.453e-04	0.0032
$x + 0.1$	1e-10	0.053	2.1047e-04	9.0067e-04	6.8835e-04	0.0021
$x^2 + 0.1$	1e-03	0.047	2.6141e-04	6.3981e-04	8.635e-04	0.019
$-x^3 + x + 0.5$	1	0.043	0.0131	0.001	0.0137	0.0014

The results show that imposing the constraint, when this is true, does improve the fit. However, as expected, the ratio MSEUncon/MSECon does decrease as we increase the sample size. This means that the marginal benefit is decreasing, but still positive. Depending on the instrument, the sample size $n = 10000$ could be equivalent to a day of trading. Hence, if we were to use a one day rolling window estimator, imposing the constraint could be advantageous. On the other hand, when carrying out our empirical estimation, where the sample size was five months, we found out that imposing a constraint is not advantageous. In this case, the sample size was in the order of millions of observations. In the empirical study, an additive eight order Bernstein polynomial with 7 covariates corresponds to the estimation of 63 linear coefficients. With millions of observations these can be estimated with high degree of precision .

From Tables 2.3 and 2.4 we see that (when the true model is the constrained one) the constrained model performs better, which implies that $\epsilon_T/T < 0$. In consequence, looking at T1, we observe that the test is undersized. To assess whether the normal approximation is good we verify that T2 has the right size, which appears to be the case. When the unconstrained model performs better, we also observe that the test has power going to one, i.e., T1 has probability going to one.

The results in Tables 2.3 and 2.4 suggest that the Bernstein and spline estimators have similar performance. In fact, the corresponding errors, their standard deviations, T1 and T2 tend to have similar order of magnitude in both cases (additional comparison are contained in Section 2.7.2: they further validate the similarity between Bernstein and spline estimators). Nevertheless, estimation via Bernstein polynomials is simpler to implement. Hence, these results support our choice of Bernstein polynomials in a high frequency context.

The full set of results from our simulations are reported in Section 2.7.2. From these, we conclude that differences between Bernstein polynomials and splines are marginal. Of course, we could have used a spline basis instead of Bernstein polynomials throughout the paper with no additional conceptual difficulty.

2.6 Concluding Remarks

This paper investigates the relationship between the intensity of trade arrivals and covariates that an algorithmic trader can manipulate. Our empirical study provide a relatively detailed picture of the relationships between order arrivals and order book quantities such as order book volume imbalances. The empirical results suggest that the crude oil futures could be manipulated using a practice called spoofing.

The statistical analysis is conducted via a counting process whose intensity is the product of a baseline intensity and an additive function of order book related covariates. In the analysis of high frequency data sets, particular attention has to be paid towards computational aspects because the sample size can be very large (in the order of hundreds of million observations). In particular, in our case, a direct optimization of the maximum log-likelihood of the counting process considered, would be computationally unfeasible. Therefore, we propose a statistical methodology suitable for large datasets. The estimation approach uses a two step procedure and solved a quadratic programming problem under linear constraints in the second step. The nature of the constraints exploits the properties of Bernstein polynomials that we use in the definition of our model.

We also report results from simulation experiments. These simulations show that using a Bernstein bases is comparable to the more classical approach that relies on splines, though the former is simpler to implement. The simulations confirm that imposing the non-negativity constraint has an implicit shrinkage effect. In our empirical analysis we found that about 70% of the coefficient were zero out of 63 in the case of the eight order Bernstein polynomial with 7 additive covariates. The simulations also show that for relatively small sample sizes (e.g., thousands of jump observations, corresponding say to a day of trading) the imposition of additional correct constraints such as monotonicity can be beneficial. If we were to consider interactions between variables rather than a purely additive model, the number of coefficients, even for an eight order polynomial would grow substantially. This would make the use of functional constraints useful even for smaller sample sizes (e.g., weeks of trading).

Finally, the paper comes with companion code that can be used for estimation of the model: it can be found at `https://github.com/asancetta/IntensityEstimation`.

2.7 Appendix

2.7.1 Functional Restrictions via Bernstein Polynomials

We still consider the case $K = 1$ for expository simplicity. Given a function g defined on $[0, 1]$ its Bernstein polynomial of order J (a positive integer) is the polynomial

$$\sum_{j=0}^{J} g\left(\frac{j}{J}\right)\binom{J}{j} x^j (1-x)^{J-j}.$$

In our case the function g is represented by the impact function g_0. We approximate g_0 using a Bernstein polynomial P_J

$$P_J(x) = \sum_{j=0}^{J} a_j \binom{J}{j} x^j (1-x)^{J-j} \tag{2.19}$$

where the a_j's are the coefficients to be estimated. Some functional constraints are simple to implement via Bernstein polynomials: they result in linear constraints on the coefficients a_j.

1. *Non-negativity.* It is clear that each summand of (2.19) is equal or greater than zero (for $x \in [0, 1]$) if $a_j \geq 0$ for each j. This implies $P_J(x) \geq 0$ for all $x \in [0, 1]$.

2. *Monotonicity.* By Equation 1.4(1) in Lorentz (1986) we obtain

   $$\frac{dP_J(x)}{dx} = J \sum_{j=0}^{J-1} (a_{j+1} - a_j) \binom{J-1}{j} x^j (1-x)^{J-1-j} \tag{2.20}$$

   From equation (2.20) to obtain an increasing function, it is sufficient but not necessary to impose the restriction $a_{j+1} \geq a_j$ for all j's.

3. *Convexity.* Equation 1.4(2) in Lorentz (1986) says that

   $$\frac{d^2 P_J(x)}{dx^2} = J(J-1) \sum_{j=0}^{J-2} (a_{j+2} - 2a_{j+1} + a_j) \binom{J-2}{j} x^j (1-x)^{J-2-j}$$

   The above display implies that $a_{j+2} - 2a_{j+1} + a_j \geq 0$ for all j's is sufficient to ensure convexity of P_J.

2.7.2 Additional Numerical Results

Bernstein Results

We report the additional results from the numerical experiments obtained using Bernstein polynomials. We can draw two conclusions. First, increasing the sample size the results get closer to the expected ones independently of the degree of the polynomial. Second, increasing solely the degree of the polynomial does not lead to better results, for example, the errors relative to the estimators in Table 2.7 are smaller and more precise than those in Table 2.10.

Table 2.5: Bernstein Basis. Fifth degree, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.01	0.042	0.023	0.0597	0.0242	0.0954
$x + 0.1$	0.034	0.048	0.0182	0.0288	0.0304	0.0561
$x^2 + 0.1$	0.036	0.052	0.0143	0.0187	0.029	0.0413
$-x^3 + x + 0.5$	0.079	0.06	0.0249	0.0323	0.0326	0.0438

Table 2.6: Bernstein Basis. Fifth degree, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect, in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error and its standard deviation) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.009	0.052	0.0019	0.0057	0.0016	0.0087
$x + 0.1$	0.018	0.055	0.0014	0.0027	0.0019	0.0053
$x^2 + 0.1$	0.012	0.042	0.0013	0.0019	0.0025	0.0043
$-x^3 + x + 0.5$	0.661	0.058	0.0141	0.003	0.0158	0.0038

Table 2.7: Bernstein Basis. Fifth degree, n=10000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.008	0.054	1.7886e-04	6.0831e-04	1.3892e-04	9.1482e-04
$x + 0.1$	0.013	0.043	1.2996e-04	2.9221e-04	1.7084e-04	5.9315e-04
$x^2 + 0.1$	0.028	0.055	1.4506e-04	1.3702e-04	2.8116e-04	4.4202e-04
$-x^3 + x + 0.5$	1	0.049	0.0131	3.0959e-04	0.0137	3.6772e-04

Table 2.8: Bernstein Basis. Tenth degree, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.04	0.047	0.0314	0.099	0.0533	0.1936
$x + 0.1$	0.021	0.056	0.0272	0.05	0.0635	0.1244
$x^2 + 0.1$	0.011	0.049	0.0215	0.0329	0.0567	0.0933
$-x^3 + x + 0.5$	0.04	0.059	0.0266	0.0485	0.0382	0.0816

Table 2.9: Bernstein Basis. Tenth degree, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect, in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error and its standard deviation) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.007	0.047	0.002	0.0081	0.0022	0.0134
$x + 0.1$	0.008	0.047	0.0016	0.0038	0.0028	0.0088
$x^2 + 0.1$	0.009	0.046	0.0017	0.003	0.004	0.0086
$-x^3 + x + 0.5$	0.578	0.065	0.014	0.0041	0.0156	0.0054

Table 2.10: Bernstein Basis. Tenth degree, n=10000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.07	0.063	1.9856e-04	9.2454e-04	1.9172e-04	0.0015
$x + 0.1$	0.006	0.052	1.4666e-04	4.2783e-04	2.4697e-04	0.001
$x^2 + 0.1$	0.005	0.046	1.8283e-04	3.2006e-04	4.2241e-04	9.5181e-04
$-x^3 + x + 0.5$	1	0.063	0.0131	4.6612e-04	0.0136	6.1307e-04

Table 2.11: Bernstein Basis. Twentieth degree, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.002	0.058	0.0377	0.1681	0.0937	0.4477
$x + 0.1$	0.007	0.059	0.0335	0.0749	0.1114	0.2333
$x^2 + 0.1$	0.003	0.046	0.0341	0.0605	0.1288	0.2335
$-x^3 + x + 0.5$	0.015	0.05	0.0276	0.0769	0.0457	0.1614

Table 2.12: Bernstein Basis. Twentieth degree, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect, in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error and its standard deviation) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	1e-03	0.056	0.0023	0.012	0.0035	0.0221
$x + 0.1$	0.004	0.046	0.002	0.0057	0.0048	0.0149
$x^2 + 0.1$	0.005	0.063	0.002	0.0042	0.0058	0.0136
$-x^3 + x + 0.5$	0.451	0.051	0.014	0.0058	0.0157	0.0081

Spline Results

We report the additional results from the numerical experiments obtained using spline bases. We can draw two conclusions. First, increasing the sample size the results get closer to the expected ones independently of the dimension of the linear space generated by the B-spline. Second, increasing solely the degree of the polynomial and the knot points of the B-spline does not lead to better results as Table 2.15 and Table 2.18 show.

Table 2.13: B-Spline. Third degree, four knot points, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.006	0.057	0.0331	0.1079	0.0665	0.2403
$x + 0.1$	0.013	0.054	0.0279	0.484	0.0776	0.1308
$x^2 + 0.1$	0.016	0.046	0.0267	0.039	0.0826	0.122
$-x^3 + x + 0.5$	0.043	0.048	0.0264	0.0493	0.0389	0.0878

Table 2.14: B-Spline. Third degree, four knot points, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect, in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error and its standard deviation) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.008	0.068	0.0022	0.0076	0.0029	0.0117
$x + 0.1$	0.016	0.057	0.0018	0.0036	0.0035	0.0074
$x^2 + 0.1$	0.009	0.053	0.0018	0.0026	0.0045	0.0069
$-x^3 + x + 0.5$	0.63	0.061	0.0141	0.0038	0.0157	0.0049

Table 2.15: B-Spline. Third degree, four knot points, n=10000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.005	0.058	2.0733e-04	7.1834e-04	2.5127e-04	0.0011
$x + 0.1$	0.01	0.054	1.6066e-04	3.4213e-04	3.122e-04	7.2144e-04
$x^2 + 0.1$	0.022	0.048	1.8928e-04	2.5057e-04	4.3684e-04	6.3876e-04
$-x^3 + x + 0.5$	1	0.05	0.0131	3.6278e-04	0.0136	4.4564e-04

Table 2.16: B-Spline. Third degree, eight knot points, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size ofof the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.003	0.054	0.0565	0.2305	0.2164	0.4054
$x + 0.1$	0.004	0.051	0.0537	0.1071	0.2416	0.4054
$x^2 + 0.1$	0.002	0.045	0.0655	0.1025	0.3493	0.5014
$-x^3 + x + 0.5$	0.008	0.052	0.0388	0.1122	0.1265	0.3531

Table 2.17: B-Spline. Third degree, eight knot points, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect, in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error and its standard deviation) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.003	0.051	0.0025	0.012	0.0046	0.0208
$x + 0.1$	0.003	0.048	0.0022	0.0058	0.0057	0.0139
$x^2 + 0.1$	0.004	0.048	0.0023	0.0042	0.0073	0.0129
$-x^3 + x + 0.5$	0.488	0.057	0.0141	0.0061	0.0158	0.0085

Table 2.18: B-Spline. Third degree, eight knot points, n=10000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases, whereas, in the fourth case the unconstrained estimator performs better (in terms of mean square error and its standard deviation). In addition, we expect T2 (T1, respectively) to be close to 0.05 (0.95, respectively) for the first three function (the fourth function, respectively).

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.005	0.06	2.2823e-04	0.0011	3.8331e-04	0.0018
$x + 0.1$	0.002	0.04	1.8449e-04	5.2338e-04	4.5687e-04	0.0012
$x^2 + 0.1$	0.006	0.042	2.2186e-04	3.8424e-04	5.9165	0.0011
$-x^3 + x + 0.5$	1	0.054	0.0131	5.7877e-04	0.0136	7.573e-04

Table 2.19: B-Spline. Third degree, sixteen knot points, n=100. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. Contrary to what we expect, in the fourth case the constrained estimator performs better than the unconstrained one and T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	0.002	0.06	0.1514	0.6742	1.0666	3.1967
$x + 0.1$	0.003	0.059	0.1797	0.4708	1.2568	2.661
$x^2 + 0.1$	0.002	0.035	0.2354	0.4011	1.8995	2.8712
$-x^3 + x + 0.5$	0.002	0.046	0.0992	0.3813	0.7054	2.0683

Table 2.20: B-Spline. Third degree, sixteen knot points, n=1000. As expected, the mean square error and its standard deviation is smaller for the constrained estimator in the first three cases and T2 is close to 0.05. In addition, as we expect in the fourth case the unconstrained estimator performs better than the constrained one (in terms of mean square error) but T1 is not close to 0.95. This is due to the fact that the size of the sample n is not large enough.

g_0	T1	T2	MSECon	MSEUncon	StdCon	StdUncon
1	1e-10	0.04	0.0028	0.0213	0.0076	0.0388
$x + 0.1$	1e-10	0.059	0.0028	0.0104	0.011	0.0278
$x^2 + 0.1$	1e-10	0.045	0.0026	0.0069	0.0106	0.0231
$-x^3 + x + 0.5$	0.18	0.04	0.0141	0.011	0.0158	0.0167

2.7.3 Shrinkage Effect of Non-negativity Constraint

In order to support the second point discussed at the beginning of this section we estimate the intensity function $\lambda_0(t) = h_0(t)g_0(X(t))$, where h_0 and g_0 are as previously defined. We us a tenth degree Bernstein polynomial and 1000 simulations. The simulations are conducted as in Section 2.5.1, but we only impose the non negativity of the estimator for g_0. For each simulation, we compute the fraction of zero coefficients of the Bernstein polynomial estimator. Table 2.21 reports the mean and standard deviation from the simulations. The shrinkage effect is apparent, and decreases with the sample, as expected. Remarkably, no penalty is used in the estimation.

Table 2.21: Frequency of Estimated Coefficients Equal to Zero. The frequency is computed from counting the estimated coefficients equal to zeros and dividing by 11. A tenth degree Bernstein polynomial requires estimation of 11 coefficients. The table reports the mean and the standard deviation over 1000 simulations. Increasing the sample size n the shrinkage effect decreases. The results are uniform respect to g_0.

n	g_0	Mean	Std
100	1	0.40	0.88
100	$x + 0.1$	0.40	0.88
100	$x^2 + 0.1$	0.41	0.86
100	$-x^3 + x + 0.5$	0.39	0.87
1000	1	0.26	0.76
1000	$x + 0.1$	0.25	0.79
1000	$x^2 + 0.1$	0.27	0.78
1000	$-x^3 + x + 0.5$	0.25	0.76
1000	1	0.17	0.65
10000	$x + 0.1$	0.16	0.64
10000	$x^2 + 0.1$	0.18	0.64
10000	$-x^3 + x + 0.5$	0.16	0.65

2.7.4 The Role of the Number of Iterations in the Algorithm

We estimate the model using Algorithm 2.1 with one and five iterations. Moreover, to ensure that the estimator does actually fare better than the best constant intensity, we compute the Monte Carlo approximation of

$$\mathbb{E}\left[\inf_{\gamma>0}\left(\frac{1}{n}\sum_{i=-(n-1)}^{0}\Big(g_0(X_i)-\gamma\Big)^2\right)\right]. \tag{2.21}$$

This is the mean square error for the best constant intensity. Table 2.22 reports the results. Recall that $g_0\left(x\right) = -x^3 + x + 0.5$ does not satisfy the monotonicity and convexity constraint imposed on $\hat{g}^{(con)}$.

Table 2.22: Error Ratios. MSECon and MSEUncon relative to (2.21) for one and five iterations of Algorithm 2.1. A number below one indicates an improvement over (2.21). The order of the polynomial is 10. Increasing the number of iterations the fit improves especially when g_0 is increasing and convex.

n	g_0	Iterations: 1		Iterations: 5	
		MSECon	MSEUnc	MSECon	MSEUnc
100	$x + 0.1$	0.52	0.87	0.30	0.57
100	$x^2 + 0.1$	0.57	0.77	0.25	0.39
100	$-x^3 + x + 0.5$	1.87	3.57	1.87	3.45
1000	$x + 0.1$	0.23	0.27	0.02	0.05
1000	$x^2 + 0.1$	0.29	0.32	0.02	0.03
1000	$-x^3 + x + 0.5$	1.02	0.31	1.02	0.29
10000	$x + 0.1$	0.21	0.21	0.00	0.01
10000	$x^2 + 0.1$	0.27	0.27	0.00	0.00
10000	$-x^3 + x + 0.5$	0.96	0.04	0.96	0.03

The results confirm that as the number of iterations increase, the fit does improve and we do fare better than the best possible constant estimator.

2.7.5 Additional Details for Section 2.3

The variable TrdImb98 is computed as follows. Let

$$TrdImb98\left(t_i\right) := \begin{cases} \frac{EWMA(signedTradedVolume(t_i))}{EWMA(tradedVolume(t_i))} & \text{if } t_i \text{ is a trade update} \\ TrdImb98\left(t_{i-1}\right) & \text{otherwise} \end{cases}$$

where the EWMA's are as in (2.12) with parameter $\alpha = 0.98$. Both signed traded volumes and traded volumes are updated only when a trade is reported. The EWMA is computed and updated only at these event times. When using trade variables as covariates, we do not adjust their timestamp by 400 microseconds in order to ensure that they can only be used once received, as in live trading. Note that if t_i is not an update for the trade imbalance, we just report the last available value of the trade imbalance. A similar approach is applied to the durations.

The duration variables are in nanosecond resolution with nanoseconds as decimals. Hence to map durations in $[0, 1]$ we cap them to one second maximum.

We compute the spread in ticks and cap it to 4 ticks. We also force the spread to take the minimum value of one tick. This is because there are no tradable choice prices, i.e., spread equal to zero is not a tradable event. We then map this spread into $[0, 1]$ dividing it by 4. In consequence, the transformed spread variable only takes values in $\{0.25, .5, 0.75, 1\}$.

2.7.6 Plots

In this subsection we collect the plots of the ACF of Buy Trades Model, the impact functions of non manipulable covariates (relative both to the Weibull and Hawkes baseline intensities) and the impact functions of manipulable covariates relative to the Weibull baseline intensity.

Figure 2.3: Autocorrelation Function (ACF) of (2.14) for Buy Trades Model. The top panel is the ACF for the model with Weibull baseline intensity. The bottom panel is the ACF for the model with Hawkes baseline intensity. For both models, the estimator for g_0 is based on a second order Bernstein polynomial. The model with Hawkes baseline intensity fits the data better than the model with Weibull baseline intensity: even if for both models the first 20 lags are significant, the autocorrelations of the model with Hawkes baseline intensity appear to be much smaller.

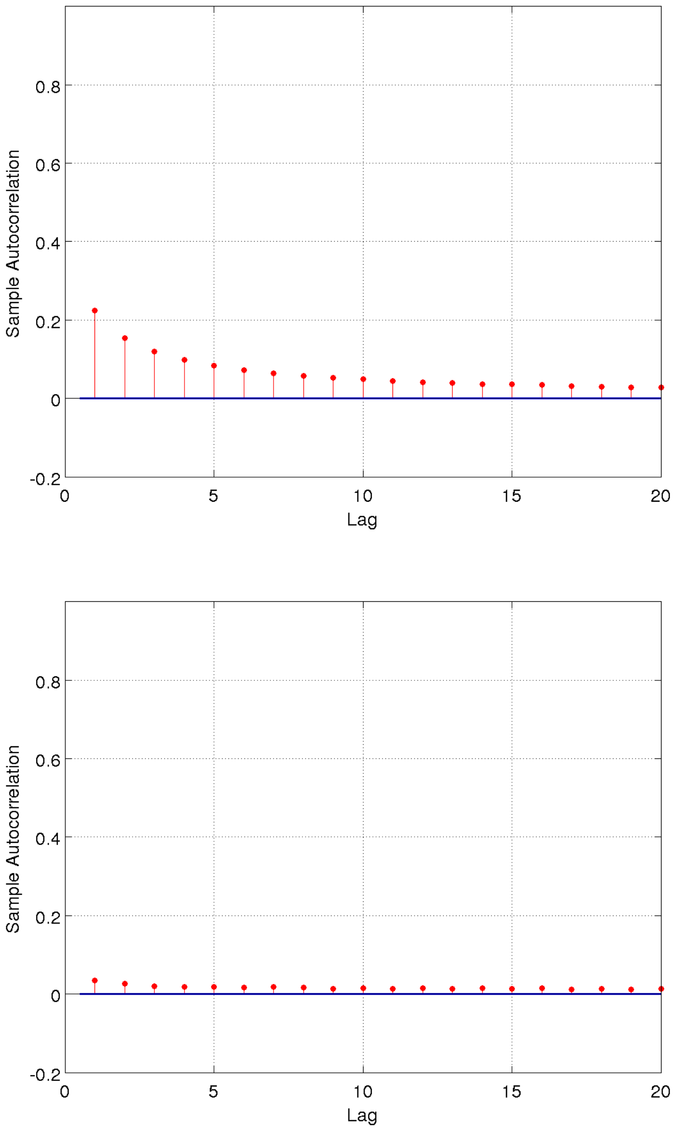
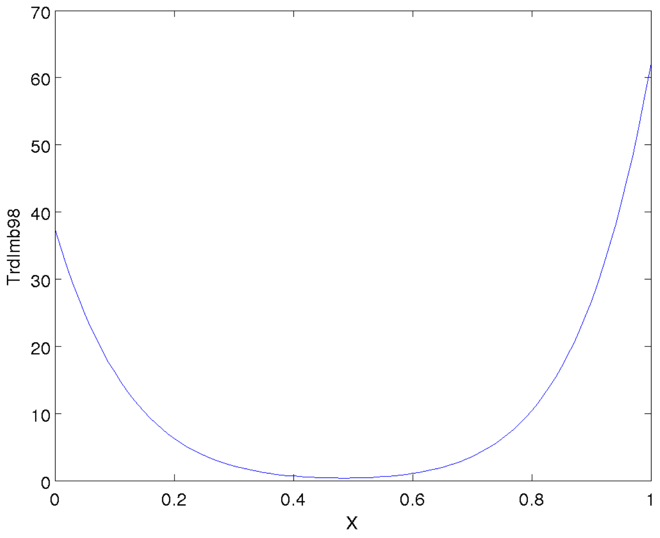
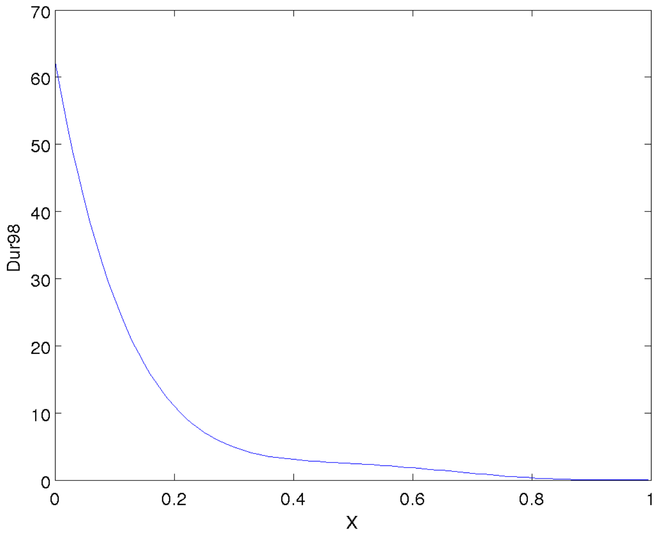


Figure 2.4: B8H Non-Manipulable Impact Functions for Buy Trades. The estimated impact functions from B8H (the unconstrained 8^{th} order Bernstein estimator with Hawkes baseline intensity) are plotted for TrdImb98, Dur98, Dur90. These variables are not easy to manipulate: in fact, the corresponding impact functions are not increasing.





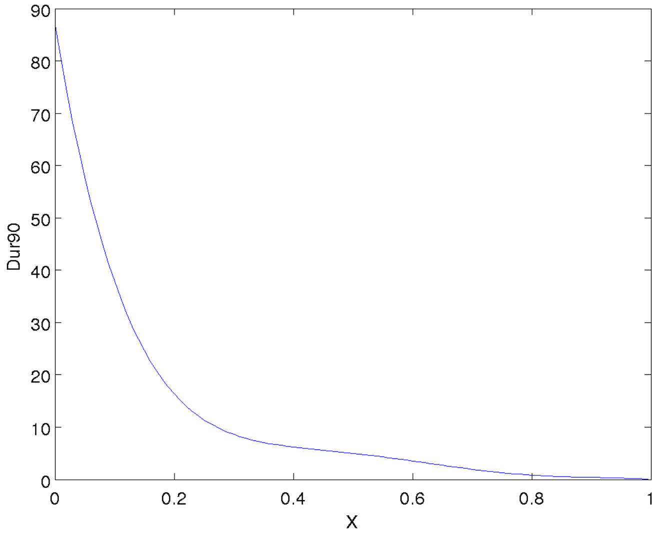
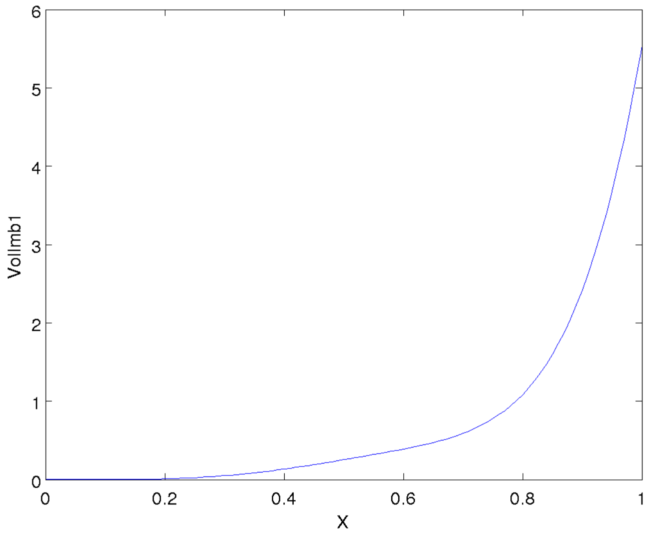
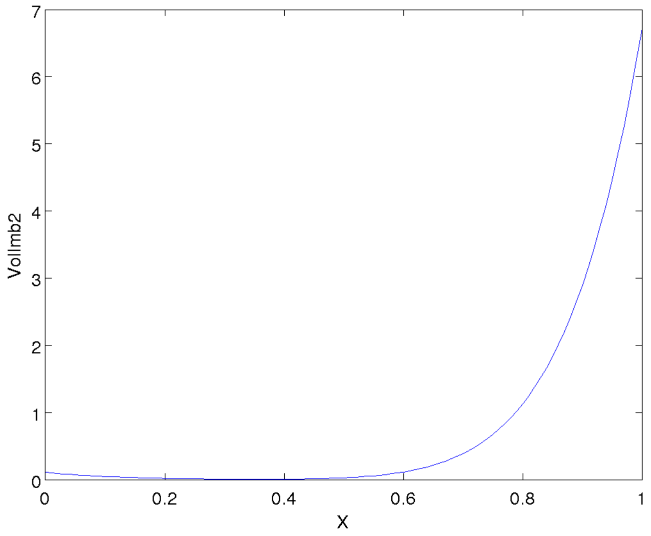
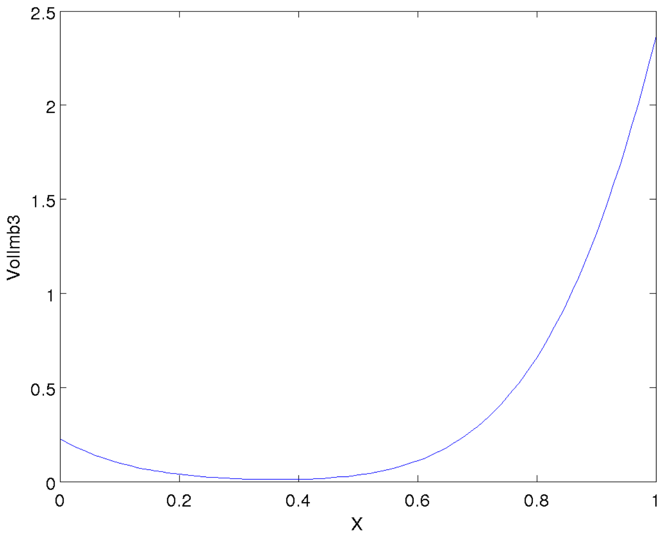

Dur90
X
0
10
20
30
40
50
60
70
80
90
0
0.2
0.4
0.6
0.8
1

Figure 2.5: B8W Manipulable Impact Functions for Buy Trades. The estimated impact functions from B8W (the unconstrained 8^{th} order Bernstein estimator with Weibull baseline intensity) are plotted for VolImb1, VolImb2, VolImb3, Spread. These variables are relatively easy to manipulate. As expected, the impact functions, for the volume imbalance, are increasing in the interval $[0.5, 1]$ (this interval corresponds to a positive volume imbalance).






2.5
2
1.5
1
0.5
0
VolImb3
0
0.2
0.4
0.6
0.8
1
X

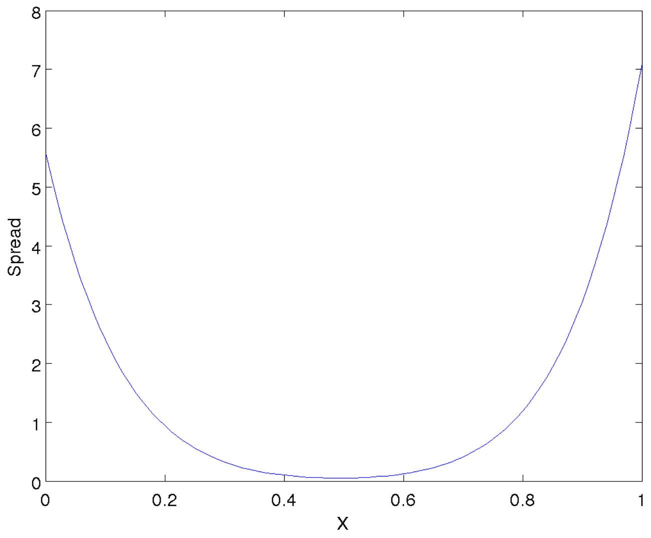
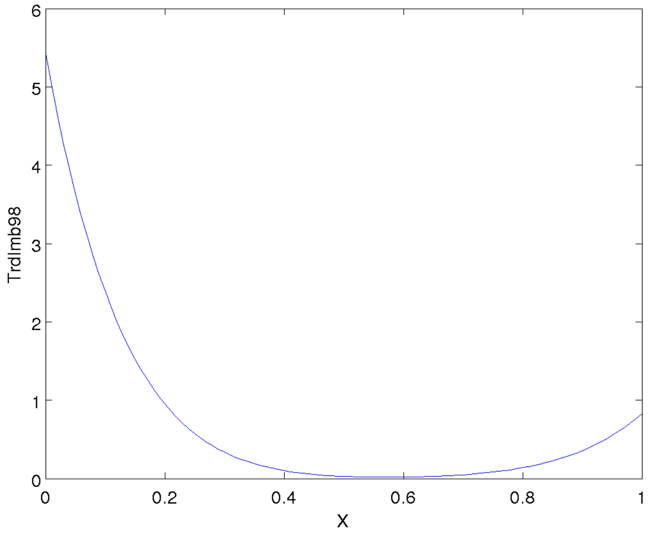
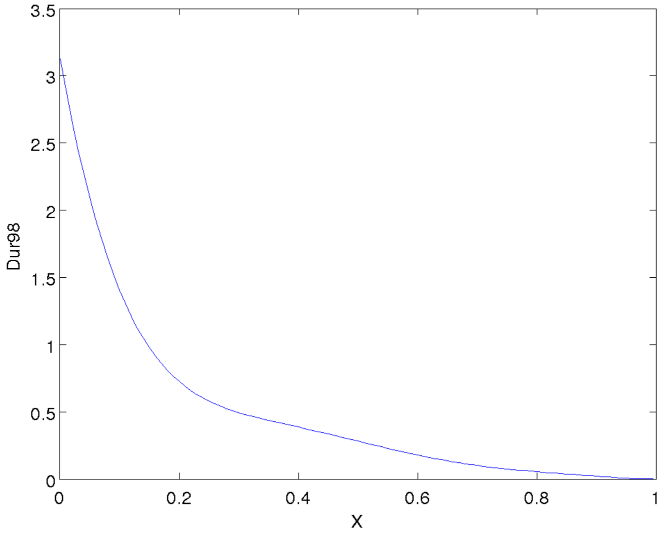
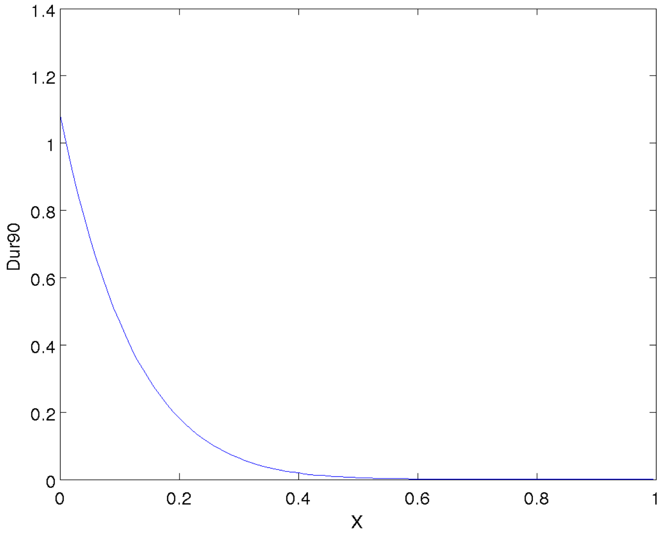

8
7
6
5
4
3
2
1
0
Spread
0
0.2
0.4
0.6
0.8
1
X

Figure 2.6: B8W Non-Manipulable Impact Functions for Buy Trades. The estimated impact functions from B8W (the unconstrained 8^{th} order Bernstein estimator with Weibull baseline intensity) are plotted for TrdImb98, Dur98, Dur90. These variables are not easy to manipulate: in fact, the corresponding impact functions are not increasing.







2.7.7 Spoofing the Bitcoin Market

As of now the Bitcoin market is a non regulated market. This could be one of the main reason to attract manipulators. In this section we assess the possibility to manipulate the Bitcoin market via the methodology introduced in the main body of this paper.

As for the empirical study relative to crude oil futures in order to estimate the intensity, only the first 67% of the sample is used while the remaining 33% of the sample is used to assess which model best fits the data. Figure 2.7 represents the impact functions obtained imposing no constraint (except the non negativity), whereas Figure 2.8 shows the impact functions obtained imposing (in addition to a non negativity constraint) a convex and increasing (volume imbalance variables) or decreasing constraint (spread variable).

Variable	Unconstrained	Model 1	Model 2	Model 3	Model 4
VI (First Level)	-	C and I	C and I	C and I	C and I
VI (Second Level)	-	-	C and I	C and I	C and I
VI(Third Level)	-	-	-	C and I	C and I
Spread	-	-	-	-	C and D

Table 2.23: Competing models: constraints imposed. All the models impose a non negativity constraint in addition to those indicated in the table. Legend: VI=Volume Imbalance, C=Convex, I=Increasing, D=Decreasing.

It is also possible to constraint some variables and not all the four covariates: Table 2.23 lists the competing models. To choose the one that best fit the data the model confidence set is used (Hansen et al., 2011) as reviewed in Algorithm 2.2. In the end the algorithm comes up with a (set of) model(s) that, asymptotically, contain the "best" model with probability at least $1-\alpha$.

Algorithm 2.2 Model confidence set

Step 0. Set $\mathcal{M} = \mathcal{M}^0 :=$ *{Unconstrained model, Model 1, Model 2, Model 3, Model 4}* and a level of confidence α, e.g., we set $\alpha = 0.05$.
Step 1. Test, via the rule $\delta_{\mathcal{M}}$, the null hypothesis $\mathcal{H}_{0,\mathcal{M}}$, i.e., all the models in $\mathcal{M}$ are "equally good", at level α.
Step 2. If $\mathcal{H}_{0,\mathcal{M}}$ is accepted then stop: every model in $\mathcal{M}$ is "equally good" and $\mathcal{M}$ is the model confidence set (at level α). Otherwise, if $\mathcal{H}_{0,\mathcal{M}}$ is rejected at level α, use an elimination rule $e_{\mathcal{M}}$ to eliminate the "worst" model(s) $\mathcal{E}_{\mathcal{M}}$, from the set $\mathcal{M}$ and repeat the procedure from *Step* 1 setting $\mathcal{M} := \mathcal{M} \setminus \mathcal{E}_{\mathcal{M}}$.

To implement it we need to detail the test $\delta_{\mathcal{M}}$ and the elimination rule $e_{\mathcal{M}}$. Let us start discussing $\delta_{\mathcal{M}}$. To begin we specify the loss function that allows to measure the "goodness" of a given model. We adopt the opposite of the loglikelihood function, e.g., if $\hat{\lambda}^U$ is the unconstrained estimator of the intensity its loss L_U is given by

$$L_U := -\int_0^T \log\left(\hat{\lambda}^U(s)\right) dN(s) + \int_0^T \hat{\lambda}^U(s)\, ds. \tag{2.22}$$

Following Hansen et al. (2011) in order to check whether a set of models are all "equally good" we design a multiple test procedure. For example the initial test comprises five hypotheses, i.e., in order to accept the null hypothesis $\mathcal{H}_{0,\mathcal{M}^0}$ that the five models perform equally good we compare the loss of the first model with the average loss of the five models, the loss of the second model with the average loss of the five models and so on. The test statistics used to make these comparisons are given by the standardized difference of the losses relative to the two competing models, e.g., if the two competing models are the unconstrained and the "averaged five models" we have the following test statistics

$$\frac{L_U - \bar{L}}{\sqrt{\int_0^T \left(\log\left(\frac{\hat{\lambda}^U(s)}{\left(\hat{\lambda}^U(s)\hat{\lambda}^{(1)}(s)\hat{\lambda}^{(2)}(s)\hat{\lambda}^{(3)}(s)\hat{\lambda}^{(4)}(s)\right)^{\frac{1}{5}}}\right)\right)^2 dN(s)}} \tag{2.23}$$

where $\hat{\lambda}^{(1)}, \ldots, \hat{\lambda}^{(4)}$ are the estimators of the intensity relative to Model 1, ..., Model 4, respectively and $\bar{L} := \frac{1}{5}\left(\sum_{i=1}^4 L_i + L_U\right)$, L_i is defined as in Equation

(2.22) replacing $\hat{\lambda}^U$ with $\hat{\lambda}^{(i)}$. The above display is asymptotically distributed as a standard normal variable (Sancetta, 2018). As said above being the test a multiple test the Holm procedure is adopted, see pages 350-351 in Lehmann and Romano (2005).

The elimination rule $e_{\mathcal{M}}$ is quite straightforward: eliminate from $\mathcal{M}$ the model(s) $\mathcal{E}_{\mathcal{M}}$ having the worst t-statistics, i.e., the highest value of (2.23).

In our case the procedure sequentially eliminates the following model: Model 4, Model 3, Model 2, Model 1. Thus the surviving model, i.e., the "best" model according to our loss function, is the Unconstrained Model with (asymptotic) probability 95%. This shows that the impact of these variables is highly nonlinear. It is also surprising that a higher spread leads to higher intensity. However, a higher spread can result from a series of correlated aggressive trades that deplete liquidity on ones side of the book. This would be associated to a higher intensity.

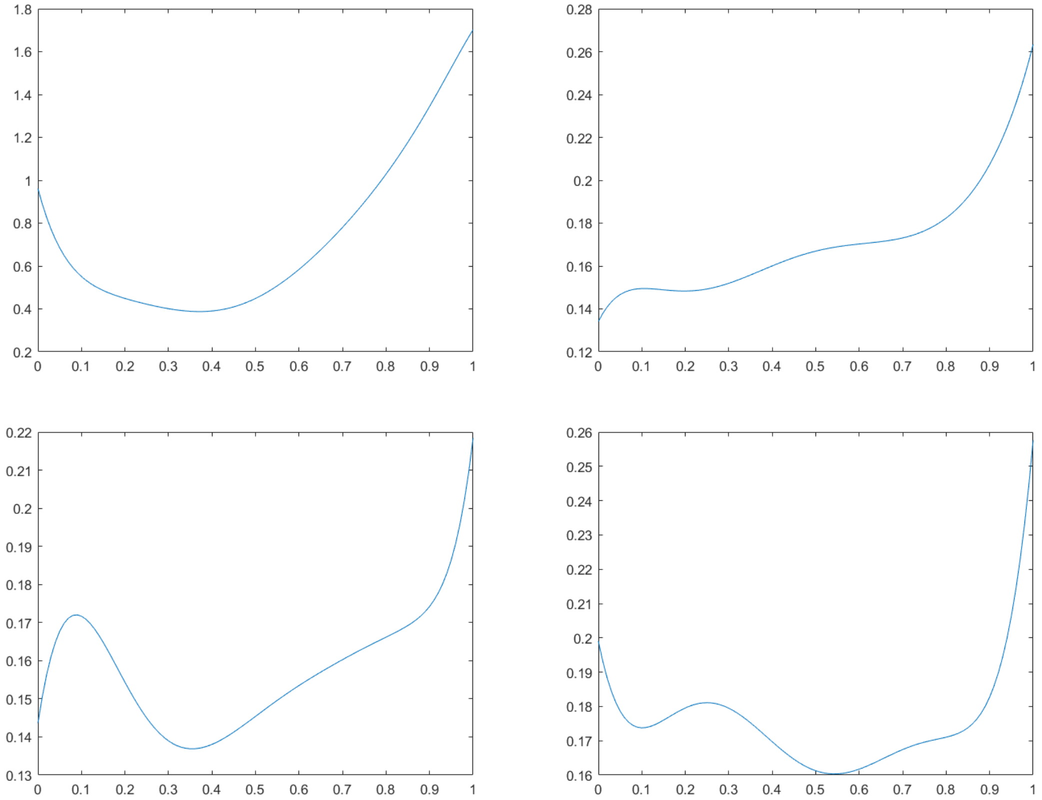


Figure 2.7: From top left to bottom right. Unconstrained impact functions relative to: spread, first level volume imbalance, second level volume imbalance, third level volume imbalance. In each case the Bernstein polynomial has degree 8. The estimators have been obtained using 67% of the data. The functions appear to be increasing in the interval $[0.5, 1]$ (that corresponds to a positive volume imbalance) even without imposing a monotonicity constraint.

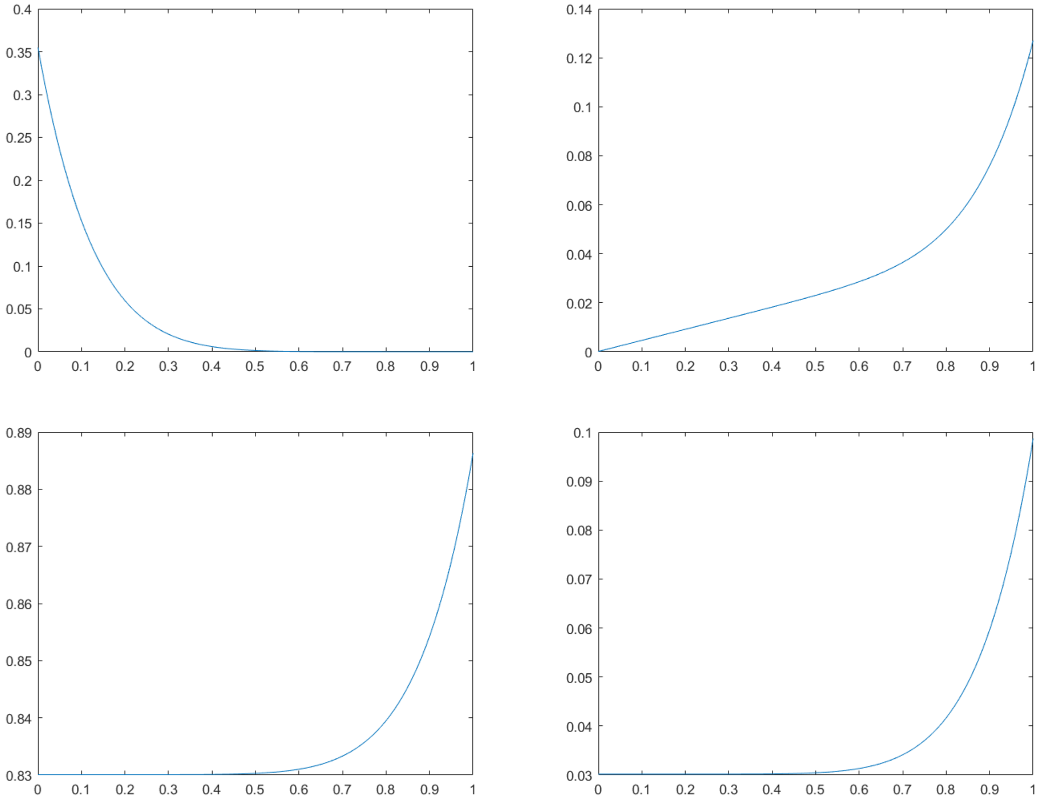


Figure 2.8: From top left to bottom right. Constrained impact functions relative to: spread, first level volume imbalance, second level volume imbalance, third level volume imbalance. All the function are constrained to be convex and non negative. In addition, the spread is constrained to be decreasing while the volume imbalances are constrained to be increasing. In each case the Bernstein polynomial has degree 8. The estimators have been obtained using 67% of the data.

Chapter 3

Estimation of a High Dimensional Counting Process Without Penalty

Abstract. Regularization techniques play a central role in high dimensional statistics. It has recently been shown that, under certain circumstances, a sign constraint is a regularization technique as effective as the more traditional Lasso approach: these results are derived in the framework of Gaussian regressions. The empirical study and the simulations conducted in Chapter 2 suggest similar results for (a certain class of) counting processes. This paper aims to provide a theoretical justification of those empirical findings.

3.1 Introduction

The analysis of high dimensional data has made it necessary the introduction of new statistical techniques. Among them regularization techniques have played a central role. In fact, in high dimensional statistics the number of predictors can be larger than the sample size. To fix ideas, consider the linear regression

$$Y = X\beta^* + \epsilon \tag{3.1}$$

such that Y is a $n \times 1$ real vector, X a $n \times p$ dimensional real valued matrix, ϵ are i.i.d. centered Gaussian random variables and $p > n$. This last assumption prevent us from using the standard OLS, instead different alternatives have been proposed over the last decades, such as ridge regression, non negative

Garrote, Lasso (and its variants). Every regularization technique relies on a sparsity assumption: β^*, i.e., the true parameter, is supported on the set $S := \{\beta_j^* \neq 0\}$ such that $|S| = s < n$ ($|S|$ denotes the cardinality of the set S). One of the most successful regularization techniques was introduced in Tibshirani (1996), i.e., the Lasso. It consists in finding $\hat{\beta}_{Lasso}$ (its uniqueness is discussed in Tibshirani, 2013) such that

$$||Y - X\hat{\beta}_{Lasso}||^2 + \gamma||\hat{\beta}_{Lasso}||_1$$

is minimized, where: $||\cdot||$ is the usual euclidean norm, $||\cdot||_1$ is the ℓ^1 norm and $\gamma > 0$ is a tuning parameter to be appropriately chosen. The role of the parameter γ is to control the trade-off between the fit of the estimator and its sparsity. Under the so called *Restricted Eigenvalue Condition* it can be shown (cf. Bickel et al., 2009) an error bound of the type $s\log(p)/n$.

Meinshausen (2013) showed that if it is known, a priori, that all the entries of the parameter β^* in (3.1) are non negative and a *Positivity Eigenvalue Condition* is fulfilled then the Non Negative Least Square (NNLS) performs as good as the Lasso (the paper proves an error bound of the type $s^2\log(p)/n$), yet it is simpler to implement because it does not require to specify any tuning parameter. Practically speaking, the NNLS estimator $\hat{\beta}$ solves the following convex minimization problem

$$\min_{\beta\geq 0} ||Y - X\beta||^2.$$

Slawsky and Hein (2013) is a more comprehensive study and obtain similar results for several norms relying on different technical assumptions (notice that in Slawsky and Hein, 2013, the *Positivity Eigenvalue Condition* is called *Self-Regularizing Property*). The empirical study and simulations carried out in Chapter 2 suggest a similar result for counting processes. Counting processes are continuous time stochastic processes with nondecreasing, càdlàg trajectories taking values in the set of non negative integers. They are constant between two consecutive events and jump one unit at each event time. The use of counting processes (and more in general of point processes) in high frequency financial modelling was pioneered by the Nobel laureate Robert Engle (Engle, 2000, notice that in that paper the meaning of the word "high frequency" is not the same as the usual usage nowadays. Instead "ultra high frequency" is what corresponds to the intraday data). Since then, they have acquired an increasing popularity in the literature, see e.g., Bacry et al. (2015) for a survey focusing on Hawkes processes, Bauwens and Hautsch (2009) and Hautsch (2012) for an extensive treatment of the econometric applications of point processes. A possible way to characterize a counting process is via its intensity: intuitively speaking the intensity of a counting process at time t is the instantaneous rate of occurrence of events conditional to the past history, i.e., the probability that the process will increase (of one unit) during the time interval $(t, t + dt]$ conditional to the past history and divided by dt. It can also be thought as the expected number of events during the time interval $(t, t + dt]$ conditional to the past history and divided by dt.

Essentially, the model considered in Chapter 2 is a (generalization of a) counting process N whose intensity process has the following form

$$\lambda^*(t) = X(t)'b^* \tag{3.2}$$

where X is a time dependent (column) vector of covariates and X' is its transpose. In that chapter N counts the number of buy or sell trade arrivals and the covariates are some relevant microstructural variables, e.g., volume imbalance, spread. The intensity is a positive stochastic process and assuming that the covariates are non negative processes it is natural to look for an estimator of the parameter b^* constrained to be nonnegative: the scope of the present paper is to theoretically justify the regularization property of that constraint. More in detail, if the true intensity is given by (3.2) then the estimator, say $\hat{b}$, introduced in Chapter 2 (see Section 2.4.1) reads

$$\arg\min_{b\geq 0}\left(-2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt\right).$$

We are able to derive the following fundamental estimates:

$$\left\|\hat{b} - b^*\right\|_1 = O_P\left(\sqrt{\frac{s^5 \log(K)}{T}}\right)$$

and

$$\frac{1}{T}\left(\hat{b} - b^*\right)' \int_0^T X(t) X(t)' dt \left(\hat{b} - b^*\right) = O_P\left(\frac{s^4 \log(K)}{T}\right)$$

for $T \to \infty$. These bounds are not entirely dissimilar to those obtained in the Gaussian regression problem (Meinshausen, 2013). The main difference between the existing literature and the present framework is that, being the error term a "generalized poissonian martingale", the proofs rely on exponential inequalities for counting processes (cf. Chapter 2 in Nishiyama, 1998) rather than to classical exponential inequalities for gaussian random variables.

The paper is structured as follows. The next section contains the description of the model together with the technical assumptions necessary to state the two main results: the first is about the consistency of the estimator of b^* and its rate of convergence, while the second one concerns the convergence of the prediction error (Section 3.4 is dedicated to their proofs). The subsequent section is dedicated to the discussion of the results of the simulations conducted.

3.2 Assumptions and Results

3.2.1 General assumptions

Let $K > 1$ an integer and $T > 0$ an arbitrary time horizon. We consider the following model. $N(t)$ is a counting process[1] with jump times $0 = T_0 < T_1 < \ldots < T_n = T$. We assume that the compensator[2] of $N(t)$ is an absolute continuous function: $\lambda^*(t)$ denotes its derivative, the so called intensity (of the counting process N). The intensity is given by (3.2)

$$\lambda^*(t) = X(t)'b^*$$

where the column vector $b^* \in \mathbb{R}^K$ is the (true) parameter to be estimated which entries are non negative, i.e., $b^* \geq 0$ (here and in the sequel such types of inequalities have to be understood elementwise). To simplify the results we also assume that $\|b^*\|_\infty \leq 1$ ($\|\cdot\|_p$ is the usual ℓ^p norm for $p \in [0, \infty]$) though this is not a necessary condition. $M(t) := N(t) - \int_0^t \lambda^*(s)\, ds$ is the martingale corresponding to the counting process. We make the following

Assumption 3.1. (Model Assumption) X is a $K-$dimensional (column vector) adapted, ergodic, càglàd stochastic process taking values in $[0,1]^K$.

[1] Here and in the sequel we tacitly assume that we have defined an underlying stochastic basis $\left(\Omega, P, (\mathcal{F}_t)_{0 \leq t \leq T}\right)$ whose filtration $\mathcal{F}_t$ satisfies the so called "usual assumptions".

[2] The compensator of a general counting process $N(t)$ is an increasing predictable stochastic process, say $\Lambda(t)$, such that $N(t) - \Lambda(t)$ is a martingale. Its existence is guaranteed by Doob-Meyer theorem.

3.2.2 Regularity assumptions

We assume that

Assumption 3.2. (Eigenvalues Assumption) Let $\underline{\sigma}_T$ the smallest eigenvalue and $\bar{\sigma}_T$ the largest eigenvalue of the matrix $\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, dt$. Then

$$\frac{1}{\underline{\sigma}_T} = O_P(1) \qquad \bar{\sigma}_T = O_P(1) \tag{3.3}$$

as $T \to \infty$.

We introduce the two main technical assumptions of the papers: the *Compatibility Condition* and the *Positive Eigenvalue Condition* (Meinshausen, 2013). To this end we introduce the following notation. Let S the set of non-zero entries of b^*, i.e., $S = \{i : b_i^* > 0\}$, $N := S^c = \{i : b_i^* = 0\}$ and $s = |S|$. Note that N denotes both the counting process and a set of indexes. For a generic vector $a \in \mathbb{R}^K$ we denote by $a_S \in \mathbb{R}^K$ ($a_{0S} \in \mathbb{R}^s$, respectively) the vector having the same entries of a except in the set S^c: in this set a_S (a_{0S}, respectively) is equal to zero (not defined because it has dimension s). Similarly, we can define a_N and a_{0N}. X_S is the $s-$dimensional subvector of X obtained by removing from X all the entries with index not belonging to the set S.

- *Compatibility Condition.* Let $L, \phi > 0$ two constants and S the index set introduced above. We say that the $(L, S)-$Compatibility Condition holds with ϕ if $\phi^2_{comp}(L, S) \geq \phi$, where

  $$\phi^2_{comp}(L,S) := \min\left\{ s\frac{b'\int_0^T X(t)\, X(t)'\, dtb}{T||b||_1^2} : b \in \mathcal{R}(L,S) \right\} \tag{3.4}$$

  and $\mathcal{R}(L,S) := \{b : \|b_N\|_1 \leq L\,\|b_S\|_1\}$. Note that $\phi^2_{comp} \leq s$. In fact, if $\hat{\Sigma} := \int_0^T X(t)\, X(t)'\, dt/T$ then

  $$\frac{b'\int_0^T X(t)\, X(t)'\, dtb}{T||b||_1^2} = \frac{1}{||b||_1^2}\sum_{i,j=1}^K \hat{\Sigma}_{ij} b_i b_j \leq \frac{1}{||b||_1^2}\sum_{i=1}^K |b_i| \sum_{j=1}^K |b_j| = 1$$

  because $\hat{\Sigma}_{ij} \leq 1$. The *Compatibility Condition* appears in the Lasso literature and is the weakest assumption that guarantees its success, cf. van de Geer and Bühlmann (2009).

- *Positive Eigenvalue Condition.* Let $\nu > 0$ a constant. We say that the

Positive Eigenvalue Condition holds with ν if $\phi_{pos}^2 \geq \nu$, where

$$\phi_{pos}^2 := \min\left\{\frac{b'\int_0^T X(t)\,X(t)'\,dtb}{T||b||_1^2} : \min_k b_k \geq 0\right\}. \tag{3.5}$$

Reasoning as in the point above we get $\phi_{pos}^2 \leq 1$. This condition has been introduced in Meinshausen (2013) and Slawsky and Hein (2013) (in the latter paper the *Positivity Eigenvalue Condition* is called *Self-Regularizing Property*).

We shall discuss a couple of examples to clarify the concepts of *Compatibility Condition* and *Positive Eigenvalue Condition* (cf. van de Geer and Bühlmann, 2009, Meinshausen, 2013, Slawsky and Hein, 2013 for more advanced examples).

Example 3.1. Assume that the least eigenvalue $\underline{\sigma}$ of the matrix $\frac{1}{T}\int_0^T X(t)$ $X(t)'\,dt$ is strictly positive, i.e., $\underline{\sigma} > 0$, and z is an eigenvector corresponding to $\underline{\sigma}$ that belongs to $\mathcal{R}(L,S)$ for some $L > 0$ and S. If $b \in \mathcal{R}(L,S)$ then

$$\|b\|_1 = \|b_S\|_1 + \|b_N\|_1 \leq (1+L)\|b_S\|_1 \leq (1+L)\sqrt{s}\,\|b_S\|.$$

In consequence

$$\begin{aligned}\frac{s}{T}\frac{b'\int_0^T X(t)\,X(t)'\,dtb}{||b||_1^2} &\geq \frac{s}{T}\frac{b'\int_0^T X(t)\,X(t)'\,dtb}{(1+L)^2 s\|b_S\|^2}\\ &= \frac{1}{T}\frac{\|b\|^2}{\|b_S\|^2}\frac{b'\int_0^T X(t)\,X(t)'\,dtb}{\|b\|^2}\frac{1}{(1+L)^2}\\ &\geq \frac{1}{T}\frac{z'\int_0^T X(t)\,X(t)'\,dtz}{\|z\|^2}\frac{1}{(1+L)^2}\\ &= \frac{\underline{\sigma}}{(1+L)^2} > 0.\end{aligned}$$

The above display implies that the $(L,S)-$*Compatibility Condition* holds with (every strictly positive real number less or equal to) $\frac{\underline{\sigma}}{(1+L)^2}$. Conversely, if $\underline{\sigma} = 0$ and $||z_N||_1 \leq L||z_S||_1$ ($z \geq 0$, respectively) then the $(L,S)-$Compatibility Condition (the Positive Eigenvalue Condition, respectively) cannot be fulfilled.

Example 3.2. (Meinshausen, 2013) If every entry of $\hat{\Sigma} := \frac{1}{T}\int_0^T X(t)\,X(t)'\,dt$ satisfies $\Sigma_{ij} \geq \nu > 0$, then the Positive Eigenvalue Condition holds true. In fact,

$$\frac{b'\hat{\Sigma}b}{||b||_1^2} = \sum_{i,j=1}^{K} \frac{\hat{\Sigma}_{ij}b_ib_j}{||b||_1^2} \geq \nu \frac{(\sum_{i=1}^{K} b_i)^2}{||b||_1^2} = \nu > 0$$

for every $b \geq 0$.

We are now able to make the following

Assumption 3.3. The Positive Eigenvalue Condition holds with $\nu > 0$.

Assumption 3.4. The (L, S) −Compatibility Condition holds with $\phi > 0$ for $L = \frac{3}{\nu}$ and with $\phi_\infty > 0$ for $L = 0$.

Note that, from Assumption 4, as ν gets closer to zero, i.e., the positive eigenvalue condition is easier to be met, the compatibility condition has to hold in a larger set, therefore it is more unlikely to be verified.

3.2.3 Additional notation

The symbols $\lesssim$ and $\gtrsim$ will be used to indicate inequality up to an absolute constant. The acronyms lhs and rhs stand for left hand side and right hand side, respectively.

We denote with $\hat{b}$ a solution of the following convex minimization problem

$$\min_{b\geq 0} \left(-2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt \right). \tag{3.6}$$

Note that a priori the solution of that problem can be not unique, nevertheless if we require that it is sparse enough such a result holds (Bruckstein et al., 2008).

The vector b^{oracle} is a solution of the following minimization problem

$$\min_{b\geq 0} \left(-2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt \right) \qquad s.t. \quad b_N = 0. \tag{3.7}$$

3.2.4 Main results

Now we can state the two main theorems of the paper. The first theorem is a consistency result for the estimator $\hat{b}$

Theorem 3.1. *If Assumption 3.1, Assumption 3.2, Assumption 3.3, Assumption 3.4 hold true then*

$$\left\| \hat{b} - b^* \right\|_1 = O_P\left(\sqrt{\frac{s^5 \log(K)}{T}} \right)$$

for $T \to \infty$.

The second result is an estimation of the prediction error.

Theorem 3.2. *If Assumption 3.1, Assumption 3.2, Assumption 3.3, Assumption 3.4 hold true then*

$$\frac{1}{T}\left(\hat{b} - b^*\right)' \int_0^T X(t) X(t)' \, dt \left(\hat{b} - b^*\right) = O_P\left(\frac{s^4 \log(K)}{T}\right)$$

for $T \to \infty$.

Their proofs are in Section 3.4.

3.3 Numerical Examples

This section presents the simulations conducted in order to validate the theoretical results discussed above. In particular it is enlightened the role played by the parameters k, s, n. We consider the estimation of two different types of intensity functions. In the first case the covariates are constant functions between two consecutive jumps (linear design) while in the second case they are linear combinations of indicator functions (localized basis).

Let us denote, as usual, by $\hat{b}$ and b^* the estimated parameter and the true parameter, respectively. The goodness of fit of the estimators is measured via four statistics (in parentheses the acronyms displayed in the subsequent tables).

Relative Mean Square Error (MSE): it is the Monte Carlo approximation of the norm l^2 of the relative error, i.e., $\left\|b^* - \hat{b}\right\| / \|b^*\|$.

Norm one of the relative error (Norm1): it is the Monte Carlo approximation of the norm l^1 of the relative error, i.e., $\left\|b^* - \hat{b}\right\|_1 / \|b^*\|_1$.

Norm zero of the relative error (Norm0): it is the Monte Carlo approximation of the norm l^0 of the relative error, i.e., $\left\|b^* - \hat{b}\right\|_0 / \|b^*\|_0$ ($\|\cdot\|_0$ denotes the l^0 norm).

Missing active features (Type1): it is the number of the estimated coefficients that are set to zero instead of being strictly positive. To have meaningful results a generic entry of the vector $\hat{b}$, say $\hat{b}_k$, is considered to be equal to zero if $\hat{b}_k < 10^{-4}$.

False Discovery (Type2): it is the number of the estimated coefficients that are strictly positive instead of being zero. To have meaningful results a generic entry of the vector $\hat{b}$, say $\hat{b}_k$, is considered to be equal to zero if $\hat{b}_k < 10^{-4}$.

To get the interarrival times $(T_i - T_{i-1})_{i=1}^n$ of the counting process N the classic time change theorem (cf. Brémaud, 1981, Chapter II, Theorem 16) turns out to be crucial. It assures that the left side of the following display

$$\int_{T_{i-1}}^{T_i} \lambda(t)\, dt = \int_{T_{i-1}}^{T_i} X(t)' \, b^* dt \tag{3.8}$$

are i.i.d. exponential random variable with unitary parameter.

The results mentioned below are obtained running 500 simulations.

3.3.1 Linear Design

The true model is given by $\lambda^*(t) = b_0^* + \sum_{i=1}^K X_k(t)\, b_k^*$ where $b_k^* = 1$ if $k \leq s$ and 0 otherwise and $b_0^* = 0.001$. The number of active variables s will be set to 1 or 11. The covariates X_k are assumed to be constant between two consecutive jumps of the counting process N so that the jump times can be easily obtained via (3.8) once we know the values of the covariates at each jump time: they are generated as follows. An $n-$dimensional random sample vector $\{Z_j\}_{j=1}^n$ is generated from a $K-$multivariate normal distribution with zero mean and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$ (Toeplitz design) or $\Sigma = I + \rho(1_K 1_K' - I)$ (equicorrelated design) where I is the $K-$dimensional identity matrix, 1_K is the $K-$dimensional column vector having all entries equal to one and ρ a parameter to be fixed. The matrix $X_k(T_i)$ is given by $X_k(T_i) = \Phi\left(Z_i^k\right)$ where Φ is the cdf of a standard normal random variable and Z_i^k is the $k^{th}-$component of Z_i. Note that, consequently, $X_k(T_i)$ are uniform random variable in $[0,1]$.Tables 3.1, 3.2 and 3.3 show the results for $n = 100$ and different values of K, s, ρ. The vector b^* is estimated via (3.6). As the ratio K/n or s/n increases the errors (MSE, Norm1, Norm2, Type1, Type2) get worse as we expect (the sample size n is fixed). In the uncorrelated case $(\rho = 0)$ the results are, in general, better than the equicorrelated case with $\rho = 0.9$ or the Toeplitz design if the active covariate is just one, i.e., $s = 1$. The reason is intuitively clear: when the covariates are uncorrelated, it is not difficult to "distinguish" them and in particular the only active covariate is easily spotted. When the correlation (among the covariates) and the number of active covariates increase within a cluster (as in the Toeplitz case) the results improve. However, an increase in correlation among all the variables makes the variable selection problem harder.

(K, s, n)	MSE	Norm1	Norm0	Type1	Type2
$(10, 1, 100)$	0.0227	0.1674	2.6860	0.9960	1.6860
	(0.0013)	(0.0051)	(0.0425)	(0.0028)	(0.0425)
$(100, 1, 100)$	0.0436	0.2733	4.5180	1.0000	3.5180
	(0.0029)	(0.0086)	(0.0718)	(0)	(0.0718)
$(100, 11, 100)$	2.3729	1.8203	2.0000	8.4540	10.0000
	(0.0331)	(0.0108)	(0.0109)	(0.0568)	(0.1088)
$(1000, 1, 100)$	0.0631	0.3802	6.3740	1.0000	5.3740
	(0.0031)	(0.0098)	(0.1012)	(0)	(0.1012)
$(1000, 11, 100)$	2.7148	2.2271	2.6568	10.4080	16.5680
	(0.0298)	(0.0083)	(0.0298)	(0.0135)	(0.1345)

Table 3.1: Equicorrelated matrix design with $\rho = 0$. Estimated standard errors in parentheses. As the ratio K/n or s/n increases the errors (MSE, Norm1, Norm2, Type1, Type2) get worse.

(K, s, n)	MSE	Norm1	Norm0	Type1	Type2
$(10, 1, 100)$	0.3170	0.7186	2.9300	0.7140	1.9300
	(0.0172)	(0.0236)	(0.0571)	(0.0216)	(0.0571)
$(100, 1, 100)$	0.7423	1.4001	5.6700	1.0700	4.6700
	(0.0212)	(0.0257)	(0.0870)	(0.0246)	(0.0870)
$(100, 11, 100)$	3.9346	1.9290	1.5350	9.7380	5.3500
	(0.0653)	(0.0084)	(0.0078)	(0.0432)	(0.0785)
$(1000, 1, 100)$	0.9981	1.7627	7.8420	1.4400	6.8420
	(0.0183)	(0.0215)	(0.1067)	(0.0249)	(0.1067)
$(1000, 11, 100)$	3.8268	2.0865	1.7934	10.5920	7.9340
	(0.0579)	(0.0071)	(0.0100)	(0.0268)	(0.0996)

Table 3.2: Equicorrelated matrix design with $\rho = 0.9$. Estimated standard errors in parentheses. As the ratio K/n or s/n increases the errors (MSE, Norm1, Norm2, Type1, Type2) get worse as we expect. This set up makes the predictions harder because the covariates are confounded.

(K,s,n)	MSE	Norm1	Norm0	Type1	Type2
$(10,1,100)$	0.1011	0.3498	2.4360	0.7620	1.4360
	(0.0085)	(0.0142)	(0.0450)	(0.0191)	(0.0450)
$(100,1,100)$	0.1090	0.4134	4.1740	1.0000	3.1740
	(0.0075)	(0.0140)	(0.0691)	(0)	(0.0691)
$(100,11,100)$	2.5101	1.4001	1.3312	6.9840	3.3120
	(0.0545)	(0.0096)	(0.0074)	(0.0513)	(0.0742)
$(1000,1,100)$	0.1266	0.4870	6.1760	1.0060	5.1760
	(0.0093)	(0.0154)	(0.0980)	(0.0035)	(0.0980)
$(1000,11,100)$	2.4541	1.5048	1.6732	7.3580	6.7320
	(0.0584)	(0.0100)	(0.0120)	(0.0529)	(0.1202)

Table 3.3: Toeplitz matrix design with $\rho = 0.9$. Estimated standard errors in parentheses. As the ratio K/n or s/n increases the errors (MSE, Norm1, Norm2, Type1, Type2) get worse as we expect. In this set up when the active covariates are more than one, i.e., $s = 11$, the predictions can be even better than the uncorrelated case.

3.3.2 Localized basis

The scenario discussed in this subsection is inspired by the multiple change point problem for counting processes (Alaya et al., 2015). In econometric, that framework can be used to detect spikes in the intraday volume curve relative to, e.g., futures contracts (see also Section 1.5). Let $T = K = 100$ and $s = 10$. In this setup the true intensity is given by

$$\lambda^* (t) = 1 + 10 \sum_{i=1}^{T} 1_{(i-1,i]} (t) \, b_i^*$$

where $b_i^* = 1$ if $i \leq s$ and $b_i^* = 0$ otherwise. The covariates are $\left\{1_{(k-1,k]} (t)\right\}_{k=1}^{K-1}$ and $\left\{1_{[0,T]} (t)\right\}$. Notice that in order to include the intercept, i.e., the covariate $\left\{1_{[0,T]} (t)\right\}$, and avoid multicollinearity the covariate $\left\{1_{(K-1,T} (t)\right\}$ has not been included. This means that jumps occurring during the time interval $(K-1, K]$ will not be detected.

Unlike the previous scenario now we consider m trading days, i.e., we have m independent counting processes N^j, $j = 1, \ldots, m$ (they have the same intensity λ^*). For each of them the relative jump times are, recursively, generated according to (3.8), i.e., $T_0 = 0$ by definition and

$$T_{i+1} = \begin{cases} T_i - \frac{\ln(U_{i+1})}{11} & if \ \{T_i \leq s\} \bigcap \mathcal{U}_i \\ 11T_i - 10s - \ln(U_{i+1}) & if \ \{T_i \leq s\} \bigcap \mathcal{U}_i \\ T_i - \ln(U_{i+1}) & if \ \{T_i > s\} \end{cases}$$

for $i \geq 0$, where U_i are i.i.d. uniform random variables in the interval $[0, 1]$ and $\mathcal{U}_i := \{-\ln(U_{i+1}) \leq 11(s - T_i)\}$. The number of jumps, in a given day, is not fixed in advance, nevertheless the length of each day is T. To get the estimator $\hat{b}$ a slight generalization of the function (3.6) is minimized, that is

$$m \int_0^T \lambda(t)^2 \, dt - 2 \sum_{i=1}^{m} \int_0^T \lambda(t) \, dN^i(t)$$

where $\lambda(t) = \sum_{k=1}^{K-1} 1_{[k-1,k)}(t) \, b_k + 1_{[0,T]}(t) \, b_K$. Table 3.4 summarizes the results: as m increases the MSE, Norm1, Norm0 and Type2 decrease and at the same time they become more precise. Nevertheless, Type1, i.e., the number of missing active features increases: this may be due to the low ratio signal to noise. In other words, missing the detection of an active feature does not degrade the error substantially.

m	MSE	Norm1	Norm0	Type1	Type2
100	0.0878	0.3586	5.0255	4.9192	49.1912
	(0.0010)	(0.0015)	(0.0278)	(0.2446)	(1.9790)
1000	0.0834	0.3228	3.8187	6.3793	37.375
	(0.001)	(0.0018)	(0.0181)	(0.2428)	(1.9644)
10000	0.0824	0.3188	3.7318	6.48475	36.524
	(0.001)	(0.0019)	(0.0125)	(0.242)	(1.9579)

Table 3.4: Localized basis scenario. Estimated standard errors in parentheses. As m increases MSE, Norm1, Norm0 and Type2 decrease.

3.4 Proofs

We start proving two lemmas that will be useful in the sequel. The first one assures that $\int_0^T X_S(t) X_S(t)' \, dt$ is an invertible matrix (see the proof of Lemma 2 in Meinshausen, 2012).

Lemma 3.1. *Let assume that the* $(0, S)-$*Compatibility condition holds with* $\phi_\infty > 0$. *Then the matrix* $\int_0^T X_S(t) X_S(t)' \, dt$ *admits inverse.*

Proof. From the compatibility condition we can write

$$b' \int_0^T X(t) X(t)' dtb \geq \phi_\infty \frac{T}{s} \|b\|_1^2$$

for every $b \in \mathbb{R}^K$ such that $b_N = 0$. This is equivalent to say that

$$b_{0S}' \int_0^T X_S(t) X_S(t)' dtb_{0S} \geq \phi_\infty \frac{T}{s} \|b_{0S}\|_1^2$$

or for every $b_{0S} \in \mathbb{R}^s \setminus \{0_{\mathbb{R}^s}\}$ ($0_{\mathbb{R}^s}$ denotes the null vector of the space $\mathbb{R}^s$)

$$\frac{b_{0S}' \int_0^T X_S(t) X_S(t)' dtb_{0S}}{\|b_{0S}\|^2} \geq \frac{\phi_\infty \frac{T}{s} \|b_{0S}\|_1^2}{\|b_{0S}\|^2} \geq \frac{\phi_\infty T \|b_{0S}\|_1}{s \|b_{0S}\|_\infty} \geq \frac{\phi_\infty T}{s} > 0$$

(because $\|b_{0S}\|^2 \leq \|b_{0S}\|_\infty \|b_{0S}\|_1$) i.e., zero cannot be an eigenvalue of the matrix $\int_0^T X_S(t) \; X_S(t)' dt$. □

The second lemma gives a bound for $\max_{1 \leq i \leq K} \left| \int_0^T X_i(t) \, dM(t) \right|$ and is crucial in order to prove the main theorems. Its proof is based on the following two lemmas (cf. Lemma 2.1.1 and Lemma 2.1.2 in Nishiyama, 1998, Corollary 3.3(a) in Nishiyama, 1997 and also page 693 in Nishiyama, 2000 for a more general statement).

Lemma 3.2. *Let Z an $\mathbb{R}-$valued, locally square integrable martingale such that $Z_0 = 0$ and that $|\Delta Z_t| \leq a$ [3] for (every t and) a constant $a \geq 0$, and τ a bounded stopping time. Then, it holds that for every $\Gamma > 0$*

$$P\left(\sup_{t \in [0,\tau]} |Z_t| > \epsilon, \langle Z, Z \rangle_\tau \leq \Gamma \right) \leq 2 \exp\left(-\frac{\epsilon^2}{2(a\epsilon + \Gamma)} \right) \qquad \forall \epsilon > 0$$

where $\langle Z, Z \rangle$ is the predictable quadratic variation of the process Z, i.e., the compensator of the quadratic variation of Z.

Lemma 3.3. *Let $N \in \mathbb{N}$ and let $Z_1, \ldots, Z_N$ be arbitrary $\mathbb{R}-$valued random variables. Assume that for a measurable set B and some constants $a \geq 0$ and $\Gamma > 0$*

$$P(|Z_i| > \epsilon, B) \leq 2 \exp\left(-\frac{\epsilon^2}{2(a\epsilon + \Gamma)} \right) \qquad \forall \epsilon > 0, \forall i = 1, \ldots, N.$$

[3] $\Delta Z_t := Z_t - Z_{t-}$, where $Z_{t-} := \lim_{s<t, s \to t} Z_s$.

Then, it holds that

$$E\left(\max_{1\leq i\leq N}|Z_i|\,1_B\right)\lesssim a\log(1+N)+\sqrt{\Gamma\log(1+N)}\,.$$

Now we can prove the announced result.

Lemma 3.4. *If* $\frac{\log(1+K)}{T}\to 0$ *as* $T\to\infty$ *then it holds true*

$$\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right|=O_P\left(\sqrt{sT\log(1+K)}\right)\tag{3.9}$$

as $T\to\infty$.

Proof. Thanks to Markov's inequality boundedness in probability is implied by boundedness in norm L^1. We prove that the lhs of (3.9) is bounded in norm L^1 : the proof is a consequence of Lemma 3.3.

Let $B:=\left\{\max_{1\leq i\leq K}\left|\int_0^T X_i^2(t)\,\lambda^*(t)\,dt\right|\leq\Gamma\right\}$ with $\Gamma>0$ a constant that we will fix soon. Note that $B=\bigcup_{i=1}^K\left\{\left|\int_0^T X_i^2(t)\,\lambda^*(t)\,dt\right|\leq\Gamma\right\}$, i.e., B is finite union of measurable sets (in fact, $X_i^2(t)\,\lambda(t)$ is predictable and, a fortiori, progressive measurable), thus B itself is a measurable set. To apply Lemma 3.3 we need to check the validity of the following display

$$P\left(\left\{\left|\int_0^T X_i(t)\,dM(t)\right|>\epsilon\right\}\bigcap B\right)\ \leq\ 2\exp\left(-\frac{\epsilon^2}{2(a\epsilon+\Gamma)}\right)$$

for every $i=1,\ldots,K$, $\epsilon>0$ and appropriate constants $a\geq 0$ and $\Gamma>0$. We claim that the above display holds true with $a=1$ and $\Gamma=Ts$. To prove the claim we rely on Lemma 3.2 with $\tau=T$. Let check that the hypotheses of the lemma are satisfied. From the assumptions we have made $\int_0^t X_i(s)\,dM(s)$ is a locally square-integrable martingale (X_i is a bounded predictable process). In addition,

$$\int_{t-}^t X_i(s)\,dM(s)\leq\int_{t-}^t X_i(s)\,dN(s)\leq N(t)-N(t-)\leq 1\tag{3.10}$$

(where $N(t-):=\lim_{s<t,s\to t}N(s)$) for every $i=1,\ldots,K$, because X_i takes values in $[0,1]$ and also ($\|b^*\|_\infty\leq 1$ by assumption)

$$\max_{1\leq i\leq K}\left|\int_0^T X_i^2(t)\,\lambda^*(t)\,dt\right|\leq\int_0^T\lambda^*(t)\,dt\leq Ts\,\|b^*\|_\infty\leq Ts.\tag{3.11}$$

The predictable quadratic variation of $\int_0^t X_i(s)\,dM(s)$, i.e., the compensator of the quadratic variation process, is given by $\int_0^t X_i^2(s)\,\lambda^*(s)\,ds$ so that, taking into account (3.10) and (3.11), the hypotheses of Lemma 3.2 are met and we have

$$P\left(\left\{\left|\int_0^T X_i(t)\,dM(t)\right| > \epsilon\right\}\bigcap B\right) \;\leq\; 2\exp\left(-\frac{\epsilon^2}{2(\epsilon + Ts)}\right).$$

The above display allows us to apply Lemma 3.3 obtaining

$$\mathbb{E}\left(\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right| 1_B\right) \lesssim \log(1+K) + \sqrt{Ts\log(1+K)} \quad (3.12)$$

where 1_B is the indicator function of the event B. Now, the assumption $\frac{\log(1+K)}{T} \to 0$ as $T \to \infty$ implies the first inequality of the following display

$$\log(1+K) < T \leq Ts \quad (3.13)$$

for T large enough or, equivalently

$$\sqrt{\log(1+K)} < \sqrt{Ts} \quad (3.14)$$

for T large enough. Multiplying both members of (3.14) for $\sqrt{\log(1+K)}$ we obtain that the inequality (3.12) can be rewritten as

$$\mathbb{E}\left(\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right| 1_B\right) \;\lesssim\; \sqrt{Ts\log(1+K)}. \quad (3.15)$$

To conclude the proof notice that the event B occurs anyway, i.e., $P(B) = 1$, so that

$$\mathbb{E}\left(\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right| 1_B\right) = \mathbb{E}\left(\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right|\right).$$

□

Next we prove that b^{oracle} is a “good approximation” of b^*.

Proposition 3.1. *Let assume that the $(0,S)-$Compatibility Condition holds with $\phi_\infty > 0$. Then*

$$||b^* - b^{oracle}||_1 \to 0 \quad (3.16)$$

in probability as $\sqrt{s^3}/\sqrt{T} \to 0$ for $T \to \infty$.

Proof. Let X_S the matrix obtained selecting the columns of the matrix X having support in S. Define $\hat{b}_{OLS}$ as

$$\hat{b}_{OLS} := \arg\min_{b\in\mathbb{R}^s} \left\{ -2\int_0^T X_S(t)' b dN(t) + \int_0^T \left(X_S(t)' b\right)^2 dt \right\}$$

i.e., $\hat{b}_{OLS} := \left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \left(\int_0^T X_S(t) dN(t)\right)$ (note that Lemma 3.1 guarantees that $\hat{b}_{OLS}$ is well defined). Let $\hat{\lambda}_{OLS} := X_S(t)' \hat{b}_{OLS}$, then $\lambda^{oracle} := X(t)' b^{oracle}$ minimizes the following functional

$$\lambda \to \left|\hat{\lambda}_{OLS} - \lambda\right|_2^2 := \int_0^T \left(\hat{\lambda}_{OLS}(t) - \lambda(t)\right)^2 dt \tag{3.17}$$

among the functions $\lambda = X(t)' b$,where $b \geq 0$ and $b_N = 0$. In fact, we know, by definition, that b^{oracle} minimizes $-2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt$ s.t. $b_N = 0$ and $b \geq 0$. This function coincides with (3.17) except for the two terms $\int_0^T \hat{\lambda}_{OLS}(t)^2 dt$, $-2\int_0^T \hat{\lambda}_{OLS}(t)\lambda(t) dt$. Nevertheless, the former term is a constant while the latter is equal to

$$\begin{aligned} -2\int_0^T \hat{\lambda}_{OLS}(t)\lambda(t) dt &= -2\int_0^T b_S' X_S(t) X_S(t)' \hat{b}_{OLS} dt - \\ &\quad 2\int_0^T b_N' X_N(t) X_S(t)' \hat{b}_{OLS} dt \\ &= -2\int_0^T b_S' X_S(t) X_S(t)' \hat{b}_{OLS} dt \\ &= -2\int_0^T b_S' X_S(t) X_S(t)' dt \\ &\quad \left[\left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \int_0^T X_S(t) dN(t)\right] \\ &= -2\int_0^T X_S(t)' b_S dN(t) \end{aligned}$$

where in the first equality we have used the identity $\lambda(t) = X(t)' b = X_S(t) b_S + X_N(t) b_N = X_S(t) b_S$. The above display proves our claim. Thanks to this property, being λ^* a feasible vector, we have

$$\left|\hat{\lambda}_{OLS} - \lambda^{oracle}\right|_2^2 \leq \left|\hat{\lambda}_{OLS} - \lambda^*\right|_2^2. \tag{3.18}$$

By the triangle inequality we deduce that $\left|\lambda^{oracle} - \lambda^*\right|_2 \leq \left|\lambda^{oracle} - \hat{\lambda}_{OLS}\right|_2 +$

$\left|\hat{\lambda}_{OLS} - \lambda^*\right|_2$ and using (3.18) we get

$$\left|\lambda^{oracle} - \lambda^*\right|_2 \leq 2\left|\hat{\lambda}_{OLS} - \lambda^*\right|_2. \tag{3.19}$$

We want to find a bound for the rhs of (3.19). Using the definition of $\hat{b}_{OLS}$ we obtain

$$\begin{aligned}\hat{b}_{OLS} &= \left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \int_0^T X_S(t)\, dN(t) \\ &= \left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \int_0^T X_S(t) \left(dM(t) + X_S(t)' b^*_{0S} dt\right) \\ &= \left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \int_0^T X_S(t)\, dM(t) + b^*_{0S}\end{aligned}$$

where b^*_{0S} is the “unsparsified” population parameter obtained by deleting the zero entries in b^* so that $\lambda^*(t) = X_S(t)' b^*_{0S}$. In consequence

$$\begin{aligned}\left|\hat{\lambda}_{OLS} - \lambda^*\right|_2^2 &= \left|X_S'\left(\hat{b}_{OLS} - b^*_{0S}\right)\right|_2^2 \\ &= \left(\int_0^T X_S(t)' dM(t)\right)\left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1} \\ &\quad \left(\int_0^T X_S(t)\, dM(t)\right). \end{aligned} \tag{3.20}$$

Using the fact that the trace of a scalar is the scalar itself and the property $Trace(ABC) = Trace(BCA)$ for arbitrary matrices A, B, C, the rhs of the above display can be rewritten as

$$Trace\left(\left(\int_0^T X_S(t) X_S(t)' dt\right)^{-1}\left(\int_0^T X_S(t)\, dM(t)\right)\left(\int_0^T X_S(t)' dM(t)\right)\right). \tag{3.21}$$

Denote $\left(\frac{1}{T}\int_0^T X_S(t) X_S(t)' dt\right)^{-1}$ by A_T and $\frac{1}{T}\int_0^T X_S(t)\, dM(t) \int_0^T X_S(t)'$ $dM(t)$ by B_T. Then the Cauchy-Schwartz inequality applied to the above display yields

$$Trace(A_T B_T) \leq \sqrt{Trace(A_T^2)}\sqrt{Trace(B_T^2)}. \tag{3.22}$$

Now we want to find a bound in probability for (3.22). We start considering the term $Trace(A_T^2)$. The trace of a matrix is the sum of its distinct eigenvalues, thus

$$Trace\left(A_T^2\right) \leq \frac{s}{\underline{\sigma}_T^2} \tag{3.23}$$

where $\underline{\sigma}_T$ is the smallest eigenvalue of the matrix $\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, dt$ and using (3.3)

$$\sqrt{Trace\left(A_T^2\right)} = O_P\left(\sqrt{s}\right) \tag{3.24}$$

as $T \to \infty$. Next we bound the second term on the rhs of (3.22), i.e., $\sqrt{Trace\left(B_T^2\right)}$. By assumption X is an ergodic process so that, using the ergodicity and the isometry property for counting martingales (Brémaud, 1981, Ch. III, Theorem 13) we get

$$\begin{aligned}
\lim_{T\to\infty}\left(\frac{1}{T}\int_0^T X_S(t)\, dM(t) \int_0^T X_S(t)'\, dM(t)\right) &= \\
\lim_{T\to\infty}\mathbb{E}\left(\frac{1}{T}\int_0^T X_S(t)\, dM(t) \int_0^T X_S(t)'\, dM(t)\right) &= \\
\lim_{T\to\infty}\mathbb{E}\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, \lambda^*(t)\, dt\right) &= \\
\lim_{T\to\infty}\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, \lambda^*(t)\, dt\right)
\end{aligned}$$

almost surely. Thanks to the above display and the continuous mapping theorem

$$Trace\left(B_T^2\right) \to Trace\left(\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, \lambda^*(t)\, dt\right)^2\right)$$

almost surely for $T \to \infty$ and a fortiori

$$Trace\left(B_T^2\right) = O_P\left(Trace\left(\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, \lambda^*(t)\, dt\right)^2\right)\right) \tag{3.25}$$

as $T \to \infty$. The trace of a matrix is the sum of its distinct eigenvalues, in consequence

$$\begin{aligned}
Trace\left(\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, \lambda^*(t)\, dt\right)^2\right) &\leq \\
Trace\left(\left(\frac{1}{T}\int_0^T X_S(t)\, X_S(t)'\, dt\right)^2 s^2 \left\|b^*\right\|_\infty^2\right) &\leq \\
\bar{\sigma}_T^2 s^3
\end{aligned}$$

where $\bar{\sigma}_T$ is the largest eigenvalue of the matrix $\frac{1}{T}\int_0^T X_S(t) X_S(t)' \, dt$. Taking into account (3.3) and the above display Equation (3.25) becomes

$$Trace\left(B_T^2\right) = O_P\left(s^3\right). \tag{3.26}$$

Finally putting together (3.24) and (3.26) into (3.22) we obtain

$$Trace\left(A_T B_T\right) = O_P\left(s^2\right).$$

as $T \to \infty$. Substituting the above display into Equation (3.20) we get

$$\left|\hat{\lambda}_{OLS} - \lambda^*\right|_2^2 \;=\; O_P\left(s^2\right) \tag{3.27}$$

as $T \to \infty$. Since $b_N^{oracle} - b_N^* = 0$, using the compatibility condition with $\phi_{comp}^2(0, S) \geq \phi_\infty$, we get

$$\begin{aligned} \left|\lambda^{oracle} - \lambda^*\right|_2^2 &= \left(b^{oracle} - b^*\right)' \int_0^T X(t) X(t)' \, dt \left(b^{oracle} - b^*\right) \\ &\geq \frac{T}{s}\phi_\infty \left\|b^{oracle} - b^*\right\|_1^2. \end{aligned} \tag{3.28}$$

Finally, putting together (3.19), (3.27) and (3.28) we conclude

$$\left\|b^{oracle} - b^*\right\|_1 = O_P\left(\sqrt{\frac{s^3}{T}}\right)$$

for $T \to \infty$. □

The next step is to prove that $\hat{b}$ is “close” to b^{oracle}.

Proposition 3.2. *Assume the Positive Eigenvalue Condition holds with* $\nu > 0$ *and the* $\left(\frac{3}{\nu}, S\right)$ *−Compatibility Condition holds with* $\phi > 0$. *Then*

$$||\hat{b} - b^{oracle}||_1 \to 0 \tag{3.29}$$

in probability as long as $\sqrt{s^5 \log(K)}/\sqrt{T} \to 0$ *(for* $T \to \infty$*).*

Proof. By definition of $\hat{b}$ the vector $\delta b := \hat{b} - b^{oracle}$ solves the following minimization problem

$$\begin{cases} \min_w \left(-2\int_0^T X(t)' \left(b^{oracle}+w\right) dN(t) + \int_0^T \left(X(t)'\left(b^{oracle}+w\right)\right)^2 dt\right) \\ s.t. \quad w + b^{oracle} \geq 0. \end{cases}$$

Being the null vector a feasible solution of the above problem it holds

$$\left(-2\int_0^T X(t)' \left(b^{oracle}+\delta b\right) dN(t) + \int_0^T \left(X(t)'\left(b^{oracle}+\delta b\right)\right)^2 dt\right) \quad \leq$$

$$\left(-2\int_0^T X(t)' b^{oracle} dN(t) + \int_0^T \left(X(t)' b^{oracle}\right)^2 dt\right)$$

and expanding the square

$$-2\int_0^T X(t)' \left(b^{oracle}+\delta b\right) dN(t) + \int_0^T \Bigg(\left(X(t)' b^{oracle}\right)^2 + \left(X(t)'\delta b\right)^2 +$$

$$2X(t)' b^{oracle} X(t)' \delta b\Bigg) dt \leq$$

$$\left(-2\int_0^T X(t)' b^{oracle} dN(t) + \int_0^T \left(X(t)' b^{oracle}\right)^2 dt\right)$$

i.e., simplifying

$$-2\int_0^T X(t)' \delta b dN(t) + \int_0^T \left(\left(X(t)'\delta b\right)^2 + 2X(t)' b^{oracle} X(t)'\delta b\right) dt \leq 0$$

or, equivalently, adding and subtracting $2\int_0^T X(t)'\delta b X(t)' b^* dt$ and rearranging the terms

$$\begin{aligned} \int_0^T \left(X(t)'\delta b\right)^2 dt \quad &\leq \quad 2\int_0^T X(t)'\delta b dM(t) \\ &\quad +2\delta b' \int_0^T X(t) X(t)' \left(b^* - b^{oracle}\right) dt. \end{aligned} \qquad (3.30)$$

We start analyzing the rhs of (3.30). For the first term, using Lemma 3.4, we have

$$\begin{aligned}\int_0^T X(t)'\,\delta b dM(t) &= \sum_{i=1}^{K}\int_0^T X_i(t)\,dM(t)\,dt\delta b_i \\ &\leq \max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\,dt\right|\|\delta b\|_1 \\ &= O_P\left(\sqrt{sT\log(1+K)}\right)\|\delta b\|_1 \qquad (3.31)\end{aligned}$$

for $T \to \infty$, while the second term can be bounded as follows

$$\begin{aligned}\int_0^T \delta b' X(t)\,X(t)'\left(b^* - b^{oracle}\right)dt &= T\sum_{i,j=1}^{K}\delta b_i\hat{\Sigma}_{ij}\left(b_j^* - b_j^{oracle}\right) \\ &\leq T\sum_{i,j=1}^{K}|\delta b_i|\left|b_j^* - b_j^{oracle}\right| \\ &= T\|\delta b\|_1\left\|b^* - b^{oracle}\right\|_1 \qquad (3.32)\end{aligned}$$

because $T\hat{\Sigma}_{ij} = \int_0^T X_i(t)\,X_j(t)\,dt \leq T$.

Now we want to find a lower bound for the lhs of (3.30). Set $M^c := \{k : \delta b_k \geq 0\}$, thus by definition $M^c \supseteq N$ and $M \subseteq S$ (notice that M is a set of indexes and not the martingale process). We distinguish two different scenarios.

Case I: In the first case it holds true $||\delta b_{M^c}||_1 \geq \frac{3}{\nu}||\delta b_M||_1$. Recalling that $T\hat{\Sigma} = \int_0^T X(t)\,X(t)'\,dt$, we have

$$\begin{aligned}\delta b'\hat{\Sigma}\delta b &= (\delta b_M + \delta b_{M^c})'\,\hat{\Sigma}\,(\delta b_M + \delta b_{M^c}) \\ &= \delta b_M'\hat{\Sigma}\delta b_M + \delta b_{M^c}'\hat{\Sigma}\delta b_{M^c} + 2\delta b_M'\hat{\Sigma}\delta b_{M^c} \\ &= \sum_{i,j\in M}\delta b_i\hat{\Sigma}_{ij}\delta b_j + \delta b_{M^c}'\hat{\Sigma}\delta b_{M^c} + 2\sum_{i\in M, j\in M^c}\delta b_i\hat{\Sigma}_{ij}\delta b_j \\ &\geq \delta b_{M^c}'\hat{\Sigma}\delta b_{M^c} - 2\|\delta b_M\|_1\|\delta b_{M^c}\|_1 \qquad (3.33)\end{aligned}$$

where the last inequality follows from the fact that $\hat{\Sigma}_{ij} \leq 1$ and $\delta b_M \leq 0$. More in details, $\sum_{i,j\in M}\delta b_i\hat{\Sigma}_{ij}\delta b_j \geq 0$ and

$$\sum_{i\in M, j\in M^c}\delta b_i\hat{\Sigma}_{ij}\delta b_j \geq -\left|\sum_{i\in M, j\in M^c}\delta b_i\hat{\Sigma}_{ij}\delta b_j\right| \geq -\sum_{i\in M, j\in M^c}|\delta b_i||\delta b_j|.$$

Being $\delta b_{M^c} \geq 0$ the *Positive Eigenvalue Condition* can be applied and using

$||\delta b_{M^c}||_1 \geq \frac{3}{\nu}||\delta b_M||_1$ the inequality (3.33) becomes

$$\begin{aligned}\delta b'\hat{\Sigma}\delta b &\geq \nu \left\|\delta b_{M^c}\right\|_1^2 - 2\frac{\nu}{3}\left\|\delta b_{M^c}\right\|_1^2 \\ &= \frac{\nu}{3}\left\|\delta b_{M^c}\right\|_1^2 \\ &= \frac{\nu}{3\left(1+\frac{\nu}{3}\right)^2}\left(\left(1+\frac{\nu}{3}\right)\left\|\delta b_{M^c}\right\|_1\right)^2.\end{aligned}$$

Thanks to the inequality $||\delta b_{M^c}||_1 \geq \frac{3}{\nu}||\delta b_M||_1$ we get

$$\begin{aligned}\delta b'\hat{\Sigma}\delta b &\geq \frac{\nu}{3\left(1+\frac{\nu}{3}\right)^2}\left(\left(1+\frac{\nu}{3}\right)\left\|\delta b_{M^c}\right\|_1\right)^2 \\ &\gtrsim \left(\left\|\delta b_{M^c}\right\|_1 + \left\|\delta b_M\right\|_1\right)^2 = \left\|\delta b\right\|_1^2.\end{aligned}$$

Using the above display, equations (3.30), (3.31), (3.32) and Proposition 3.1 we can conclude that $||\delta b||_1 = O_P\left(\sqrt{\frac{s^3\log(1+K)}{T}}\right)$ for $T \to \infty$.

Case II: Otherwise it happens that $||\delta b_M||_1 > \frac{\nu}{3}||\delta b_{M^c}||_1$, but $N \subseteq M^c$ (that implies $S \supseteq M$) thus

$$||\delta b_N||_1 \leq ||\delta b_{M^c}||_1 \leq \frac{3}{\nu}||\delta b_M||_1 \leq \frac{3}{\nu}||\delta b_S||_1.$$

This allows us to apply the *Compatibility Condition,* in fact, $\delta b \in \mathcal{R}(\frac{3}{\nu}, S)$, and conclude $\delta b'\hat{\Sigma}\delta b \geq (\phi/s)||\delta b||_1^2$. Using again (3.30), (3.31) and (3.32) we have $||\delta b||_1 = O_P\left(\sqrt{\frac{s^5\log(1+K)}{T}}\right)$ for $T \to \infty$.

So in both cases we get the claimed result: this completes the proof. □

Putting together the results of the above propositions we get

Proof. (of Theorem 3.1). By triangle inequality we obtain

$$\left\|\hat{b} - b^*\right\|_1 \leq \left\|\hat{b} - b^{oracle}\right\|_1 + \left\|b^{oracle} - b^*\right\|_1$$

and using Proposition 3.1 and Proposition 3.2

$$\begin{aligned}\left\|\hat{b} - b^*\right\|_1 &= O_P\left(\sqrt{\frac{s^3}{T}}\right) + O_P\left(\sqrt{\frac{s^5\log(1+K)}{T}}\right) \\ &= O_P\left(\sqrt{\frac{s^5\log(1+K)}{T}}\right)\end{aligned}$$

for $T \to \infty$. □

Finally we derive a bound for the prediction error, i.e., Theorem 3.2.

Proof. (of Theorem 3.2). Putting (3.31), (3.32) in (3.30) we get ($\delta b := \hat{b} - b^{oracle}$)

$$\begin{aligned}\delta b' \int_0^T X(t) X'(t)\, dt \delta b &= O_P\left(\sqrt{sT \log(1+K)} \left\|\delta b\right\|_1\right) + \\ & \quad O_P\left(T \left\|\delta b\right\|_1 \left\|b^* - b^{oracle}\right\|_1\right)\end{aligned}$$

for $T \to \infty$ and using the error bound in Proposition 3.1 and Proposition 3.2

$$\begin{aligned}\delta b' \int_0^T X(t) X'(t)\, dt \delta b &= O_P\Bigg(\sqrt{sT \log(1+K)} \times \\ & \quad \sqrt{\frac{s^5 \log(1+K)}{T}}\Bigg) + \\ & \quad O_P\left(T \sqrt{\frac{s^5 \log(1+K)}{T}} \sqrt{\frac{s^3}{T}}\right)\end{aligned}$$

as $T \to \infty$. The above display implies (for $T \to \infty$)

$$\frac{1}{T} \delta b' \int_0^T X(t) X'(t)\, dt \delta b = O_P\left(\frac{s^4 \log(1+K)}{T}\right).$$

□

3.5 Conclusion

We extend to a particular class of counting processes the main findings of Meinshausen (2013) that are tailored for Gaussian regressions. The intensity of those counting process is a function of high dimensional variables (in other words the number of the covariates can be larger than the sample size) and is estimated minimizing a quadratic contrast functional. The natural non-negativity constraint for the intensity estimator acts as a regularization technique and together with the *Positive Eigenvalue Condition* and the *Compatibility Condition* allow us to prove a rate of convergence, both for the estimation error and the prediction error, not entirely dissimilar to the rate provided by Lasso theory with no need to include a penalty term and tune additional parameters. Further, the simulation study conducted confirms the shrinkage property of the non negative constraint under different scenarios.

Chapter 4

Conclusions and Further Developments

4.1 Conclusions

The previous three chapters discuss a few central themes in high frequency financial econometrics: empirical study of non equally time spaced time series, counting processes (to model financial variables such as traded volumes, buy/sell trade arrivals) and their estimation in scenarios commonly encountered in high dimensional statistics, i.e., when the sample size is very large (computational feasibility of the estimation procedure) and when the number of observation is less than the number of features (regularization techniques).

Chapter 1 is dedicated to the empirical side of the subject. It contains a statistical analysis of high frequency Bitcoin data: the study supports the idea that the Bitcoin market is not a completely mature market yet. The main finding that leads to that conclusion is the autocorrelation of returns even for relatively long time scale (30 minutes): indeed, many papers in the literature sustain that also daily returns are autocorrelated (for a review cf. Kyriazis, 2019). Nevertheless, it seems that the market, over the years, has become more efficient (Drożdż et al., 2018, Kyriazis, 2019): this could be due to the increasing interest in the Bitcoin market from institutional investors. In addition to some other statistical proprieties of returns the chapter contains a short analysis of the order book: the only article in the literature that thoroughly studies the order book of the Bitcoin market is Schnaubelt et al. (2019). Finally, a novel way to study intraday volumes is presented: volumes are assumed proportional to the number of trade arrivals that are modeled via a counting process (the intensity of this counting process is smoothed using the fused Lasso). The approach can be useful for intraday trading volume prediction.

High frequency finance has also introduced new market manipulation strategies: high frequency spoofing is one of them. According to Cartea et al. (2020) “The literature on spoofing is scant”. The second chapter aims to shed some lights on this particular type of market manipulation investigating the fundamental mechanism of the strategy: trigger a buy (sell) market order placing limit buy (sell) orders that create an upward (downward) price trend. This mechanism is investigated via a counting process: computational issues prevents its estimation maximizing the log-likelihood. Therefore, a novel statistical methodology is introduced. The major shortcoming of the proposed methodology is that it is not rigorously justified (however, see next section), nevertheless heuristic arguments and simulations to support it are presented.

Finally, Chapter 3 explores the “high dimensional” set up. Over the last decades the “high dimensional” paradigm in statistics, i.e., when the number of the observed features is larger than the size of the sample, has acquired a central role in statistics: what allows these models to be estimated is the so called sparsity, i.e., just few of the numerous features really matter. These relevant features are usually selected adding a penalty term to the contrast function together with a parameter (to be calibrated) controlling the trade off between fit and sparsity: the framework presented in Chapter 3, i.e., a counting process whose intensity is driven by many covariates, does not need the inclusions of that term and simply relies on a non negativity constraint. In other words, there is no need to choose any tuning parameter. The result is quite interesting given that point processes are widely used in high frequency econometrics and the non negativity constraint is a natural constraint for the intensity of a generic counting process.

4.2 Further Developments

In the sequel we list some possible extensions of the study conducted in:
Chapter 1.

1. The statistical analysis can be broaden in (at least) three ways:

   - Produce a quantitative analysis for some variables already analyzed in a qualitative manner. For example, in the mainstream literature returns and volumes at the top of the order book fit a power law distribution (Bouchaud et al., 2002, Potters and Bouchaud, 2003): it can be checked whether this is the case for Bitcoins as well. Some financial quantities, such as bid-ask spread, display a long memory behaviour: are they better described via fractional models (or more sophisticated models, cf., Groß-KlußMann and Hautsch, 2013) rather than ARMA-type models? In addition, the distribution of the volume imbalance resembles a mixture of continuous and discrete random variables. It could be interesting to investigate which distribution is a good fit for the volume imbalance.

   - The order book can be studied in more detail, in particular its shape (Potters and Bouchaud, 2003) and other features (of utmost importance from a practical point of view) such as the resiliency of the order book (Large, 2007). Moreover, a study of the virtual price impact (Maslov and Mills, 2001) and of the liquidity (chapter 9 in Hautsch, 2012) can be added.

   - According to part of financial econometrics literature the paths of volatility are rougher than those of a Brownian motion (Gatheral et al., 2018, Fukasawa et al., 2019): is that the case also for the Bitcoin market? This could be interesting given the fact that recently Bitcoin options have recently been issued and a correct model of the volatility is crucial in order to price them. Indeed, there is a recent paper about the topic (Takaishi, 2019).

2. According to Scaillet et al. (2018) the dynamics of the Bitcoin price includes a jump term, it could be interesting to investigate whether there are jumps in the volatility process as well (Jacod and Todorov, 2010).

3. If there were available high frequency data for other cryptocurrencies (e.g., Litecoin) it could be interesting to:

   - Test whether co-jumps occur (e.g., chapter 14 in Aït-Sahalia and Jacod, 2014) and if they could lead to some profitable trading strategy.

- Study multivariate stylized facts (Breymann et al. 2003).

Chapter 2.

1. The theoretical justification given of the statistical methodology introduced in the chapter is purely heuristic. Ideally, it should be transformed into a rigorous one. This point is partially addressed in Mucciante and Sancetta (2020) when the baseline intensity is that of a Hawkes process and the functions $g_{0,k}$ in (2.2) are linear (notice also that in Mucciante and Sancetta, 2020 a one-hot encoding approach is adopted rather than Bernstein polynomials or splines).

2. In the current set up buy orders and sell orders are treated separately, i.e., they are modeled via two different counting processes. A more realistic model should incorporate the dependence between them.

3. It could be interesting to extend the purely additive model (2.2) to the case in which there are interactions between the covariates.

4. Cartea et al. (2020) study the optimal execution for a spoofer. In this framework, the dynamics of the order book is crucial, therefore a more realistic picture of it could be added to the basic one discussed in that paper. In fact, e.g., if the order book depth is allowed to be stochastic the optimal execution is qualitatively different from the "standard scenarios" discussed in the literature: see, e.g., Fruth et al. (2019) and Ackermann et al. (2020).

Chapter 3.

1. While the "consistency" of the non negative least square estimator has been proved it could be interesting to derive (in a high dimensional set up) its asymptotic distribution, so that tests of significance and confidence intervals can be constructed. Notice, for example, that in the Lasso framework this is not trivial at all. In fact, the asymptotic distribution of the Lasso estimator is not uniform therefore it cannot be used to derive tests and the problem is approached via other strategies, see, e.g., van de Geer et al. (2014).

2. Other possible extensions of the non negative OLS could be the following (separately or combining them):

   - Replace the quadratic loss function with a more general convex loss function. Indeed, there is a recent work (Koike and Tanoue, 2019) about this possibility (it could also be extended to the "counting process framework").
   - Consider a high dimensional Functional Data Analysis set up (Roche, 2019).
   - Consider high dimensional time series type models (Basu and Michailidis, 2015).

Bibliography

[1] Ackermann, J., K. Thomas and M. Urusov (2020) Càdlàg semimartingale strategies for optimal trade execution in stochastic order book models. Preprint available at `https://arxiv.org/abs/2006.05863`.

[2] Aggarwal, D. (2019) Do Bitcoins Follow a Random Walk Model? Research in Economics 73, 15-22.

[3] Aït-Sahalia, Y. and J. Jacod (2014) High-Frequency Financial Econometrics. Princeton University Press.

[4] Aït-Sahalia, Y., J. Jacod and J. Li (2012). Testing for Jumps in Noisy High Frequency Data. Journal of Econometrics 168, 207-222.

[5] Aït-Sahalia, Y., P.A. Mykland and L. Zhang (2005) How Often to Sample a Continuous-Time Process in the Presence of market Microstructure Noise. Review of Financial Studies 18, 351–416.

[6] Alaya, M.Z., S. Gaïffas and A. Guilloux (2015) Learning the Intensity of Time Events with Change-Points. IEEE Transactions on Information Theory 61, 5148-5171.

[7] Alvarez-Ramirez, J., E. Rodriguez and C. Ibarra-Valdez (2018) Long-Range Correlations and Asymmetry in the Bitcoin Market. Physica A: Statistical Mechanics and its Applications 492, 948-955.

[8] Andersen, T.G., D. Dobrislav and E. Schaumburg (2008) Duration-Based Volatility Estimation. Working paper.

[9] Bacry, E., I. Mastromatteo and J.F. Muzy (2015) Hawkes Processes in Finance. Market Microstructure and Liquidity 1, 1550005.

[10] Bariviera, A.F., M.S. Basgall, W. Hasperué and M. Naiouf (2017) Some Stylized Facts of the Bitcoin Market. Physica A: Statistical Mechanics and its Applications 484, 82-90.

[11] Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde and N. Shephard (2008) Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise. Econometrica 76, 1481–1536.

[12] Basu, S. and G. Michailidis (2015) Regularized Estimation of Sparse High-Dimensional Time Series Models. Annals of Statistics 43, 1535-1567.

[13] Bauwens, L. and N. Hautsch (2009) Modelling Financial High Frequency Data Using Point Processes. In T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), Handbook of Financial Time Series, 953-982. New York: Springer.

[14] Begušić, S., Z. Kostanjčar, H.E. Stanley and B. Podobnik (2018) Scaling Properties of Extreme Price Fluctuations in Bitcoin Markets. Physica A 510, 400–406.

[15] Bickel, P.J., R. Ya'acov and A.B. Tsybakov (2009) Simultaneous Analysis of Lasso and Dantzig Selector. The Annals of Statistics 37, 1705-1732.

[16] Bhogal, S.K. and T.V. Ramanathan (2019) Conditional Duration Models for High-Frequency Data: A Review on Recent Developments. Journal of Economic Surveys 33, 252-273.

[17] Black, F. (1986) Noise. Journal of Finance 41, 529–543.

[18] Bouchaud, J.P., M. Mézard and M. Potters (2002) Statistical Properties of Stock Order Books: Empirical Results and Models. Quantitative Finance 2, 251-256.

[19] Brémaud, P. (1981) Point Processes and Queues: Martingale Dynamics. New York: Springer.

[20] Brémaud, P. and L. Massoulié (1996) Stability of Nonlinear Hawkes Processes. Annals of Probability 24, 1563-1588.

[21] Breymann, W., A. Dias and P. Embrechts (2003) Dependence Structures for Multivariate High-Frequency Data in Finance. Quantitative Finance 3, 1-14.

[22] Brezis, H. (2011) Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer.

[23] Bruckstein, A., M. Elad and M. Zibulevsky (2008) On the Uniqueness of Nonnegative Sparse Solutions to Underdetermined Systems of Equations. IEEE Transactions on Information Theory 54, 4813–4820.

[24] Campbell, J.Y., A.W. Lo and A.C. MacKinlay (1997) The Econometrics of Financial Markets. Princeton University Press.

[25] Caporale, G.M., L. Gil-Alana and A. Plastun (2018) Persistence in the Cryptocurrency Market. Research in International Business and Finance 46, 141-148.

[26] Caporin, M., A. Ranaldo and G.G. Velo (2015) Precious Metals Under the Microscope: a High-Frequency Analysis. Quantitative Finance 15, 743-759.

[27] Cartea, A., R. Donnelly and S. Jaimungal (2018) Enhancing Trading Strategies with Order Book Signals. Applied Mathematical Finance 25, 1-35.

[28] Cartea, A., S. Jaimungal and Y. Wang (2020) Spoofing and Price Manipulation in Order-Driven Markets. Applied Mathematical Finance 27, 67-98.

[29] Chakraborti, A., I.M. Toke, M. Patriarca and F. Abergel (2011) Econophysics Review: I. Empirical facts. Quantitative Finance 11, 991-1012.

[30] Cont, R. (2001) Empirical Proprieties of Asset Returns: Stylized Facts and Statistical Issues. Quantitative Finance 1, 223-236.

[31] Cont, R., A. Kukanov and S. Stoikov (2014) The Price Impact of Order Book Events. Journal of Financial Econometrics 12, 47-88.

[32] Cox, D.R. (1975) A Note on Data-Splitting for the Evaluation of Significance Levels. Biometrika 62, 441-444.

[33] Da Fonseca, J. and R. Zaatour (2014) Hawkes Process: Fast Calibration, Application to Trade Clustering, and Diffusive Limit. Journal of Futures Markets 34, 548-579.

[34] Dawid, A.P. (1984) Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach. Journal of the Royal Statistical Society A 147, 278-292.

[35] Dawid, A.P. and V. Vovk (1999) Prequential Probability: Principles and Properties. Bernoulli 5, 125-162.

[36] de Jong, F. and B. Rindi (2009) The Microstructure of Financial Markets. Cambridge University Press.

[37] Diebold, F.X. and R.S. Mariano (1995) Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253-263.

[38] Drożdż, S., R. Gębarowski, L. Minati, P. Oświęcimka and M. Wątorek (2018) Bitcoin Market Route to Maturity? Evidence From Return Fluctuations, Temporal Correlations and Multiscaling Effects. Chaos 28, 071101.

[39] Duarte, J., E. Hu and L. Young (2020) A Comparison of Some Structural Models of Private Information Arrival. Journal of Financial Economics 135, 795-815.

[40] Duarte, J. and L. Young (2009) Why is PIN Priced? Journal of Financial Economics 91, 119-138.

[41] Dumitru, A.M. and G. Urga (2012) Identifying Jumps in Financial Assets: A Comparison Between Nonparametric Jump Tests. Journal of Business and Economic Statistics 30, 242-255.

[42] Easley, D., N.M. Kiefer, M. O'Hara and J.B. Paperman (1996) Liquidity, Information, and Infrequently Traded Stocks. The Journal of Finance 51, 1405-1436.

[43] Engle, R.F. (2000) The Econometrics of Ultra High Frequency Data. Econometrica 68, 1–22.

[44] Engle, R.F. and J.R. Russell (1998) Autoregressive Conditional Duration: a New Model for Irregularly Spaced Transaction Data. Econometrica 66, 1127-1162.

[45] Eross, A., F. McGroarty, A. Urquhart and S. Wolfe (2019) The Intraday Dynamics of Bitcoin. Research in International Business and Finance 49, 71-81.

[46] Feng, W., Y. Wang and Z. Zhang (2018) Informed Trading in the Bitcoin Market. Finance Research Letters 26, 63-70.

[47] Fischer, T.G., C. Krauss and A. Deinert (2019) Statistical Arbitrage in Cryptocurrency Markets. Journal of Risk and Financial Management, 12, 31.

[48] Fruth, A., T. Schoneborn, and M. Urusov (2019) Optimal Trade Execution in Order Books with Stochastic Liquidity. Mathematical Finance, 29, 507-541.

[49] Fukasawa, M. (2010a) Central Limit Theorem for the Realized Volatility Based on Tick Time Sampling. Finance and Stochastics 14 , 209–233.

[50] Fukasawa, M. (2010b) Realized Volatility with Stochastic Sampling. Stochastic Processes and Their Applications 120, 829–552.

[51] Fukasawa, M. and M. Rosenbaum (2012) Central Limit Theorems for Realized Volatility Under Hitting Times of an Irregular Grid. Stochastic Processes and Their Applications 122, 3901-3920.

[52] Fukasawa, M., T. Takabatake and R. Westphal (2019) Is Volatility Rough? Preprint available at `https://arxiv.org/abs/1905.04852`.

[53] Gatheral, J., T. Jaisson and M. Rosenbaum (2018) Volatility is Rough. Quantitative Finance 18, 933-949.

[54] Gençay, R. M. Dacorogna, U.A. Muller, O. Pictet and R. Olsen (2001) An Introduction to High-Frequency Finance. Academic Press.

[55] Gerlach, J.C., G. Demos and D. Sornette (2018) Dissection of Bitcoin's Multiscale Bubble History from January 2012 to February 2018. Royal Society Open Science 6, 180643.

[56] Groß-KlußMann, A. and N. Hautsch (2013) Predicting Bid–Ask Spreads Using Long-Memory Autoregressive Conditional Poisson Models. Journal of Forecasting 32, 724-742.

[57] Guillaume, D.M., M.M. Dacorogna, R.R. Davé, U.A. Müller, R.B. Olsen and O.V. Pictet (1997) From the Bird's Eye to the Microscope: A Survey of New Stylized Facts of the Intra-Daily Foreign Exchange Markets. Finance and Stochastics 1, 95–129 .

[58] Koike, Y. and Y. Tanoue (2019) Oracle Inequalities for Sign Constrained Generalized Linear Models. Econometrics and Statistics 11, 145-157.

[59] Hansen, P.R., A. Lunde and J.M. Nason (2011) The Model Confidence Set. Econometrica 79, 453-497.

[60] Hautsch, N. (2012) Econometrics of Financial High-Frequency Data. Springer.

[61] Hong, S.Y., I. Nolte, S.J. Taylor and X. Zhao (2020) Volatility Estimation and Forecasts Using Price Durations. Submitted to Journal of Financial Econometrics.

[62] Jacod, J. and V. Todorov (2010) Do Price and Volatility Jump Together? Annals of Applied Probability 20, 1425-1469.

[63] Kirchner, M. (2017) An Estimation Procedure for the Hawkes Process. Quantitative Finance 17, 571-595.

[64] Kyriazis, N.A. (2019) A Survey on Efficiency and Profitable Trading Opportunities in Cryptocurrency Markets. Journal of Risk and Financial Management, 12, 67.

[65] Large, J. (2007) Measuring the Resiliency of an Electronic Limit Order Book. Journal of Financial Markets Volume 10, 1-25.

[66] Lahmiri, S., S. Bekiros and A. Salvi (2018) Long-Range Memory, Distributional Variation and Randomness of Bitcoin Volatility. Chaos, Solitons & Fractals 107, 43-48.

[67] Lee, S.S. and P.A. Mykland (2012) Jumps in Equilibrium Prices and Market Microstructure Noise. Journal of Econometrics 168, 396–406.

[68] Lehmann, E.L. and J.P. Romano (2005) Testing Statistical Hypotheses. New York: Springer.

[69] Lennart, A. (2020) Bitcoin Transactions, Information Asymmetry and Trading Volume. Quantitative Finance and Economics 4, 365-381.

[70] Li, Y., P.A. Mykland, E. Renault, L. Zhang, and X. Zheng (2014) Realized Volatility When Sampling Times are Possibly Endogenous. Econometric Theory 30, 580–605.

[71] Liu, L.Y., A.J. Patton and K. Sheppard (2015) Does Anything Beat 5-minute RV? A Comparison of Realized Measures Across Multiple Asset Classes. Journal of Econometrics 187, 293-311.

[72] Lorentz, G.G. (1986) Bernstein Polynomials. New York: Chelsea Publishing Company.

[73] MacKenzie, D. (2017) A Material Political Economy: Automated Trading Desk and Price Prediction in High - Frequency Trading. Social Studies of Science 47, 172-194 .

[74] Mammen, E., O. Linton and J. Nielsen (1999) The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions. The Annals of Statistics 27, 1443-1490.

[75] Maneesoonthorn, W., G.M. Martin and C.S. Forbes (2020) High-Frequency Jump Tests: Which Test Should We Use? Accepted for publication in Journal of Econometrics.

[76] Maslov S., M. Mills (2001) Price Fluctuations From the Order Book Perspective- Empirical Facts and a Simple Model. Physica A: Statistical Mechanics and its Applications 299, 234-246.

[77] Meinshausen, N. (2012) Sign-Constrained Least Squares Estimation for High-Dimensional Regression. Available at `https://arxiv.org/pdf/1202.0889.pdf`.

[78] Meinshausen, N. (2013) Sign-Constrained Least Squares Estimation for High-Dimensional Regression. Electronic Journal of Statistics 7, 1607-1631.

[79] Meyer, M.C. (2008) Inference Using Shape-Restricted Regression Splines. The Annals of Applied Statistics 2, 1013-1033.

[80] Mucciante, L. and A. Sancetta (2020) An Order Book Dependent Hawkes Process for Large Datasets. Submitted to Journal of Financial Econometrics.

[81] Nishiyama, Y. (1997) Some Central Limit Theorem for $\ell^{\infty}-$valued Semimartingale and Their Application. Probability Theory Related Fields 108, 459-494.

[82] Nishiyama, Y. (1998) Entropy Methods for Martingales. Phd Thesis, University of Utrecht. Available at `http://www.f.waseda.jp/nishiyama/`.

[83] Nishiyama, Y. (2000) Convergence of Some Classes of Martingales with Jumps. The Annals of Probability 28, 685-712.

[84] Ogata, Y. (1978) The Asymptotic Behaviour of the Maximum Likelihood Estimator for Stationary Point Processes. Annals of the Institute of Statistical Mathematics 30, 243-261.

[85] Ogata, Y. and H. Akaike (1982) On Linear Intensity Models for Mixed Doubly Stochastic Poisson and Self-Exciting Point Processes. Journal of the Royal Statistical Society B 44, 102-107.

[86] Osterrieder, J. and J. Lorenz (2017) A Statistical Risk Assessment of Bitcoin and its Extreme Tail Behaviour. Annals of Financial Economics 12, 1750003.

[87] Pagan, A. (1996) The Econometrics of Financial Markets. Journal of Empirical Finance 3, 15-102.

[88] Pacurar, M. (2008) Autoregressive Conditional Duration Models in Finance: A Survey of the Theoretical and Empirical Literature. Journal of Economic Surveys 22, 711-751.

[89] Petukhina, A.A., R.C.G. Reule and W.K. Hardle (2020) Rise of the Machines? Intraday High-Frequency Trading Patterns of Cryptocurrencies. The European Journal of Finance, DOI: 10.1080/1351847X.2020.1789684.

[90] Potters, M. and J.P. Bouchaud (2003) More Statistical Properties of Order Books and Price Impact. Physica A: Statistical Mechanics and its Applications 324, 133-140.

[91] Ramsay, J.O. (1988) Monotone Regression Splines in Action. Statistical Science 3, 425-441.

[92] Roche, A. (2019) Variable Selection and Estimation in Multivariate Functional Linear Regression via the LASSO. Preprint available at `https://arxiv.org/abs/1903.12414`.

[93] Sancetta, A. (2018) Estimation for the Prediction of Point Processes with Many Covariates. Econometric Theory 34, 598–627.

[94] Scaillet O., A. Treccani and C. Trevisan (2018) High-Frequency Jump Analysis of the Bitcoin Market. Journal of Financial Econometrics, 1–24.

[95] Schnaubelt, M., J. Rende and C. Krauss (2019) Testing Stylized Facts of Bitcoin Limit Order Books. Journal of Risk and Financial Management, 12, 25.

[96] Sensoy, A. (2019) The Inefficiency of Bitcoin Revisited: A High-Frequency Analysis with Alternative Currencies. Finance Research Letters 28, 68-73.

[97] Shaw, C. (2018) Conditional Heteroskedasticity in Crypto-Asset Returns. Journal of Statistics: Advances in Theory and Applications 20, 15-65.

[98] Slawsky, M. and M. Hein (2013) Non-negative Least Squares for High-Dimensional Linear Models: Consistency and Sparse Recovery without Regularization. Electronic Journal of Statistics 7, 3004-3056.

[99] Takaishi, T. (2019) Rough Volatility of Bitcoin. Preprint available at `https://arxiv.org/abs/1904.12346`.

[100] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B 58, 267-288.

[101] Tibshirani, R.J. (2013) The Lasso Problem and Uniqueness. Electronic Journal of Statistics 7, 1456–1490.

[102] Tibshirani, R.J. and J. Taylor (2011) The Solution Path of the Generalized Lasso. The Annals of Statistics 39, 1335–1371.

[103] Urquhart, A. (2016) The Inefficiency of Bitcoin. Economics Letters 148, 80-82.

[104] van de Geer, S.A. and P. Bühlmann (2009) On the Conditions Used to Prove Oracle Results for the Lasso. Electronic Journal of Statistics 3, 1360–1392.

[105] van de Geer, S., P. Bühlmann, Y. Ritov and R. Dezeure (2014) On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. Annals of Statistics 42, 1166-1202.

[106] Wang, C.J. and M.C. Meyer (2011) Testing the Monotonicity or Convexity of a Function Using Regression Splines. The Canadian Journal of Statistics 39, 89-107.

[107] Wang, J.N., H.C. Liu and Y.T. Hsu (2020a) Time-of-day Periodicities of Trading Volume and Volatility in Bitcoin Exchange: Does the Stock Market Matter? Finance Research Letters 34, 101243.

[108] Wang, J.N., H.C. Liu, S. Zhang and Y.T. Hsu (2020b) How Does the Informed Trading Impact Bitcoin Returns and Volatility? Applied Economics, DOI: 10.1080/00036846.2020.1814944.

[109] Yatchew, A.J. (1992) Nonparametric Regression Tests Based on Least Squares. Econometric Theory 8, 435-451.

[110] Yatchew, A.J. and L. Bos (1997) Nonparametric Regression and Testing in Economic Models. Journal of Quantitative Economics 13, 81-131.

[111] Yatchew, A.J. and W.K. Härdle (2006) Nonparametric State Price Density Estimation Using Constrained Least Squares and the Bootstrap. Journal of Econometrics 133, 579-599.

[112] Zargar, F.N. and D. Kumar (2019a) Informational Inefficiency of Bitcoin: a Study Based on High-Frequency Data. Research in International Business and Finance 47, 344-353.

[113] Zargar, F.N. and D. Kumar (2019b) Long Range Dependence in the Bitcoin Market: a Study Based on High-Frequency Data. Physica A: Statistical Mechanics and its Applications 515, 625-640.

[114] Zhang, L. (2006) Efficient Estimation of Stochastic Volatility Using Noisy Observations: a Multi-Scale Approach. Bernoulli 12, 1019–1043.

[115] Zhang, L., P.A. Mykland and Y. Aït-Sahalia (2005) A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data. Journal of the American Statistical Association 100, 1394–1411.

[116] Zhang, W., P. Wang, X. Li and D. Shen (2018a) Some Stylised Facts of the Cryptocurrency Market. Applied Economics 50:55, 5950-5965.

[117] Zhang, W., P. Wang, X. Li and D. Shen (2018b) The Inefficiency of Cryptocurrency and its Cross-Correlation with Dow Jones Industrial Average. Physica A: Statistical Mechanics and its Applications 510, 658-670.

[118] Zhang, Y., S. Chan, J. Chu and S. Nadarajah (2019) Stylised Facts for High Frequency Cryptocurrency Data. Physica A: Statistical Mechanics and its Applications 513, 598-612.