SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di Perfezionamento in
Matematica per la Finanza

Ph.D. Thesis

Price formation and liquidity modeling in high frequency finance

Candidate:
Damian Eduardo Taranto

Supervisors:
Prof. Giacomo Bormetti
Prof. Fabrizio Lillo

September 2017

List of previously published work

1. Taranto, D. E., Bormetti, G., Lillo, F.

   The adaptive nature of liquidity taking in limit order books.

   Journal of Statistical Mechanics: Theory and Experiment, 2014(6), **P06002**, (2014).

2. Taranto, D. E., Bormetti, G., Bouchaud, J-Ph., Lillo, F., and Tóth, B.

   Linear models for the impact of order flow on prices. I. Propagators: Transient vs. History Dependent Impact.

   Arxiv preprint: arXiv:1602.02735, submitted to *Quantitative Finance* (2016).

3. Taranto, D. E., Bormetti, G., Bouchaud, J-Ph., Lillo, F., and Tóth, B.

   Linear models for the impact of order flow on prices. II. The Mixture Transition Distribution model for market impact and price dynamics.

   Arxiv preprint: arXiv:1604.07556, submitted and accepted by *Quantitative Finance* (2016).

Acknowledgements

I would like to express my special appreciation and thanks to my supervisors Professor Giacomo Bormetti and Professor Fabrizio Lillo, you have been a tremendous mentors for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I especially thank Professor Jean-Philippe Bouchaud and Dr. Bence Tóth from Capital Fund Management (Paris), with whom I had the pleasure to collaborate with during the course of the PhD.

I acknowledge the former and the current members of the Mathematical Finance group at Scuola Normale Superiore, Dario Alitab, Giulia Livieri, Davide Pirino, Daniele Regoli, Michael Schneider, Adam Majewski, and Professor Stefano Marmi, for all the interesting discussions we had and for the exceedingly long time we spent in front of the coffee vendor machines.

I would also thanks my formers colleagues Caterina Rizzi, Francesco Mazza, Alberto Biella, Stefano Valentini, Marco Ce', Riccardo Bosisio, and Ilirjan Aliaj for sharing joy and sorrow during our academic experience at Scuola Normale Superiore.

A special thanks to my family. Words cannot express how grateful I am to my mother Marta, my father Eduardo, and my brothers Maria and Matias, which are not only brothers but also my closest friends. I would also like to thank all of my friends who supported me during the days in Pisa, and incented me to strive towards my goal. Thanks to all, I will never forget your support.

Contents

Introduction

The basic purpose of a financial market is to provide a place where buyers and sellers meet and exchange securities. These market participants can either physically meet in a trading floor, or they can communicate through an electronic trading platform. Trades can be initiated by financial intermediaries, such as dealers or brokers, or arranged directly by investors of different intentions without the intervention of intermediaries. All of them contribute to the supply and demand of a specific asset. Whatever the setting, there are rules either explicit or implicit that govern the trading mechanisms and define the market structure. This organizational structure of trading determines traders' behaviour – what, when, where and how they can trade – and is one of the origins of market liquidity and price formation. The aim of the market microstructure is the study of the process and outcomes of exchanging assets, which follow explicit trading structures used for financial securities (O'Hara, 1995).

Madhavan (2000) considers that market microstructure is the framework within which the latent beliefs and intentions of investors are ultimately translated into prices and volumes. An important implication drawn from these definitions is that market microstructure is shaped by market structure and trading rules. The market microstructure studies derive from the specific market structure how different trading mechanisms affect the price formation process, and why prices generate particular time series. Hasbrouck (2007) identifies three main themes in the empirical microstructure analysis. First, these analyses try to understand the sources of value and reasons for trade, which emerge in a trading framework where a wide range of market participants with different information decide to trade. The second theme is the understanding of trade mechanisms used to accomplish trade, such as Limit Order Book, continuous and non-continuous auction trading. The last theme is the process of equilibrium price setting. At any given time, there may be many prices depending on the direction of trade (buying or selling), the trade quantity, the required speed for the trade, and the trader's latent intentions.

The previous themes can be explored by using the enormous quantity of accessible data of trading activity. In particular, it allows to study the crucial role of information in the price discovery process. This information can contain the past dynamics of the market (prices, order signs, etc.) or exogenous information (news). In this Thesis we consider a specific market mechanism of trading in which participants can place two types of orders: market orders, which generate an immediate

transaction, or limit orders, which are placed in a queue, the limit order book, waiting for counterparties that have intentions to trade the specific volume and price. The last important component of the limit order book dynamics are the cancellations: i.e. the removal of an existing limit order from the book. In general, limit orders increase liquidity, while market orders and cancellations decrease liquidity. Clearly, the interplay of market order executions, limit orders placements, and cancellations from all the market participants generated the time series of prices and returns. Part of market microstructure studies focus on finding the relationship between the characteristics of the previous processes and the statistical proprieties of the resulting price process.

An important empirical evidence, which has been shown by several studies, is the characteristic unpredictability of returns in financial markets. These findings led to the formulation of the Efficient Market Hypothesis, which is one of the most important results of twentieth century financial research. Samuelson (1965) and Fama (1970) state that a market is informationally efficient if all available information on the market determines the formation of prices, that is, prices incorporate instantaneously the information and expectations of all market participants. Consequently, there are no possibilities of further gains or arbitrages on the market using only this information. If a financial market consists of rational and fully informed traders, the price of each financial asset will equal the asset's fundamental value (Arthur et al., 1997).

The concept of efficient prices is strongly related with the concept of Market Impact, that is, the empirically verified correlation between the direction of an incoming order (to buy or to sell) and the subsequent price change. We can state that one of the mechanisms which transmit the private information to prices is exactly the market impact. For example, on average a buy order pushes the price up, whereas a sell order pushes the price down. In fact, the sign of the order, plus one for buy orders and minus one for sell orders, is positively correlated with the subsequent price change. Essentially, the market reacts to the revealed intention of the market participant, by adjusting the equilibrium price. A striking statistical property of the times series of order signs is that the order flow is correlated in time, which means that there exists a non vanishing predictability of the future order signs. This evidence has been deeply analysed in the literature, in particular the puzzling combination of efficient prices, correlated order flow, and market impact. If the order flow is strongly autocorrelated in time, then also market impacts should be autocorrelated and the returns would be predictable. Clearly, there must exist a mechanism that ensures efficient prices. Several studies in the literature were focused on finding this mechanism, with the idea that the assumption of permanent and fixed market impact should be relaxed. Bouchaud et al. (2004) proposed the propagator model which are able to reconcile the efficient market hypothesis and the correlated order flow. Within this framework, the price process at a given time consists in a linear superposition of past order signs, weighted by the so-called 'propagator' function. This function decreases in time, which means that the impact of trades are transient but of fixed values. In (Lillo and Farmer, 2004) the authors proposed an alternative model for the market impact of trades, by introducing the

asymmetric liquidity mechanism. This mechanism states that the price impact of a type of order (buy or sell) is inversely related to the probability of its occurrence. This means that if at a certain point in time it is more likely that the next trade is a buy rather than a sell, a buyer initiated trade will have a smaller impact than a seller initiated trade. There is therefore a compensation between probability of an event and its effect on the price. Bouchaud et al. (2009) have shown that in a particular case the two models are equivalent. In this Thesis we study empirically which microstructural mechanisms are responsible for this striking behaviour of financial markets, by intensively analysing the dynamics of the limit order book and its characteristic variables (volumes of market and limit orders, market depth of the order book, etc.). After this analysis, we propose an adaptive liquidity model for the order book dynamics which can reproduce the mechanisms found. The key ingredient of the modeling is a liquidity dynamics that adapts itself to the degree of predictability of the order flow, where the market participants vary the request of liquidity on their own past order flow.

The above linear framework leads to an interesting approximate description of the price dynamics, which are able to model market impact and market liquidity. However, the previous models rely in strong assumptions on the determinants of the price dynamics. In fact, they take into account only certain type of events, such as the execution of market orders. There exists a huge dynamics missed in this description, such as time inhomogeneous effects, non linearities (the so-called square root law, see Tóth et al. 2011), missing events, etc. In this Thesis we focus our attention not only on how the prices are affected by the execution of order signs, but also on the other way around: We measure which is the impact of past returns on the subsequent order flow. We show that the previous linear models are not able to reproduce an additional anti-correlation between past returns and future order signs and, consequently, the diffusive behaviour of prices at very short timescales. Our efforts are focused in the generalization of the previous models, in order to better reproduce these statistical proprieties of the price dynamics.

The generalization of the propagator models proposed in this Thesis introduces the simple idea that orders that trigger or not trigger a price change have different impacts on the markets. This is particular true if the price path at very short time scales consists in changes of discrete values. The discretization effect is strong when the price of the security is small with respect to the minimum price change allowed by the limit order book. Within the linear framework described in (Eisler et al., 2012b), we decouple the the reaction of market to the two events of price change and not price change by using two different propagator functions. Therefore, the price process is a superposition of the past order signs, weighted by two propagator functions, which are selected by two indicator functions for each types of events. This apparently minor modification has lead to an extended class of propagator models which describe with remarkable realism the intertwined high-frequency dynamics of prices and order flow.

Nonetheless, the linear description achieved with two propagator functions and indicator functions of the market dynamics is still too rigid: these models are de-

signed to describe the evolution of the market with an exogenously specified order flow. This fact seriously limits the forecasting capabilities of linear impact models. In order to overcome this issue, we propose to model this price dynamics by using the Markov chains framework. Given the long range correlations observed in order flow, we consider high order Markov process. In order to cope with the large number of parameters, we consider a specific, yet general, class, namely Mixture Transition Distribution model proposed by Raftery (1985). This is an explicit stochastic model for the order flow, treated as an endogenous component of the dynamics, which is specially designed for variables which are inherently discrete. Moreover, we introduce a specific subclass of MTDg models, for which we prove that can be estimated via Generalized Method of Moments. The corresponding estimation procedure involves the numerical solution of an optimization problem, which is proved to be convex in the model parameters. Consequently, it solves the issue of the high dimensionality of the models needed in the case of time series of the price dynamics.

The Thesis is composed of six chapters. The first two chapters mainly contain a review of material found in literature. The last four, instead, present our original research and contributions. In the following we briefly summarize their content.

- **Chapter 1 – Market microstructure and price discovery** introduces some basic notions of financial markets and typical variables studied in market microstructure. It discusses about key characteristics of markets, such as market liquidity and the price formation process. It then focuses on the microstructure of financial markets, with the characteristics of the Limit Order Book and the trading mechanisms that allow transactions among market participants. Finally, it describes recent statistical models used to describe the order book dynamics.

- **Chapter 2 – Price movements: efficient prices and market impact** discusses the concept of Market Impact of trades and the statistical properties of the order flow. It defines formally the Efficient Market Hypothesis and the information set used to define it. Moreover, it explains the paradox that arises considering the strong autocorrelation of the order flow and the Efficient Market Hypothesis. It then describes two models for the Market Impact that resolve this contradiction. The Bouchaud et al. (2004) model resolves this problem with a fixed but transient market impact. The second one, the Lillo and Farmer (2004) model, proposes a permanent market impact, which is variable because it depends on the past history of order flow. The latter introduces the important mechanism of asymmetric liquidity which permits efficient prices.

- **Chapter 3 – The adaptive liquidity model** reports an extensive empirical analysis of high frequency dynamics in order to understand the microstructural mechanisms responsible for ensuring efficient prices. Then, it introduces a statistical model of the order book dynamics, where the liquidity adapts to the predictability of the order flow in order to obtain a diffusive price process.

- **Chapter 4 – Linear models for market impact** introduces a generalization of the propagator model, which assumes that different orders can have different impact on the market. It shows that if a linear model decouples the dynamics of events that trigger or not a price change, then it can better reproduce the total price dynamics and the diffusive propriety of the price process. This is evident in particular if the price changes are rare and have discrete values, such as in large tick stocks.

- **Chapter 5 – The Mixture Transition Distribution model** discusses about the application of high order Markov chains to the modeling of the price dynamics. In particular, it discusses an approximation of the full Markov chain, the MTD model. Despite the high order of the model, it can be estimated via Maximum Likelihood Estimation, or, in a subclass introduced in this Thesis, by Generalized Method of Moments. The optimization problem of this estimation method is proven to be convex, therefore it solves the issue of the high dimensionality of the model.

- **Chapter 6 – MTD for order flow and price impact** shows that the above framework of the MTD model can be applied to the study of time series of real trading activity, and it reports the empirical results of the estimation of the model.

Finally, we summarize briefly the findings of this Thesis in the Conclusions.

Chapter 1

Market microstructure and price discovery

The role of having a regulated market mechanism serves the security of the traders. There are some types of transactions (e.g. forward contracts, exotic options) that are made between two parties (individuals or institutions). In these cases the details of the transaction usually remain private (the so-called *over-the-counter* contracts). However, standardized contract types are offered by regulated markets that provide higher liquidity and lower counterparty risk. Market microstructure deals with the trading of these standardized financial assets. Their prices are determined in different ways on different markets. We study continuous trading double auction markets. This means that buyers enter competitive bids and sellers enter competitive offers (orders) simultaneously and in a continuous fashion.

In principle, this mechanism allows the matching of two different populations, called market participants: entrepreneurs that have some requirement of funding for an industrial project, and investors that have capitals to invest and share, on the long run, either the future profit or the risk of these industrial projects. So, financial markets can also play the role of insurance companies, with the availability of different contracts that offer the possibility of hedging, by securing against losses. This population, that protects itself against market risks or interest rate risks, is called *hedgers*. There are also two different broad categories of market participants. One is made by *broker-dealers* or *market makers* that provide liquidity to the markets by buying or selling some stock at a given time, i.e. they operate as intermediaries between others market participants and make trading more fluid, compensating any momentary lack of counterparties. The latter category is made by *speculators*, which have short or medium term positions on markets, and they make bets on the prices following different strategies and expectations. For example, if a price variation of a stock is considered excessive by speculators, they bet on a short term reversal. This is called contrarian strategy, while trend following strategies are used if the recent trend are believed to be persistent in the future.

In a well-functioning market, buyers and sellers can easily find each other, and

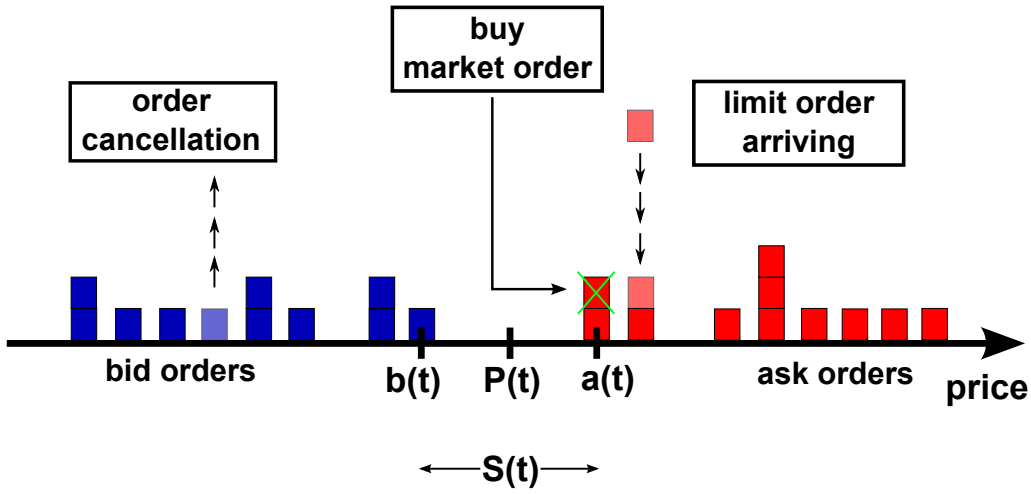

Figure 1.1: Original scheme of the double auction trading mechanism.

trade without significant adverse effect on prices. This is known as a liquid market where there are many standing orders which wait to match investors' orders of opposite intention. The price of the security being quoted at any given point in time will thus reflect the information held by market participants. This price formation process is known as price discovery. In a fair and manipulation-free market, prices obtained on the market are a reflection of genuine supply and demand. This concept refers to market integrity. The following section reviews each of these principal issues in market microstructure theory.

Trading activity involves transaction of financial products, such as stocks, bonds, commodities, and derivatives. The former, also called capital stock (or shares) of a business entity, represents the original capital paid into or invested in the business by its founders. Stock is different from the property and the assets of a business may fluctuate in quantity and value. In general, the terms "stocks" and "shares" are included under the term *equity*.

In a double auction market ("double" because the behaviour of the market is symmetric for buyers and sellers) the participants of the trading activity can either place *limit orders* or *market orders*. Traders can submit limit orders to buy or sell a certain amount of shares of a given stock at a price not worse than a given limit price. We call them "patient traders". Limit orders are not necessarily executed at the moment they are submitted. In this case they are stored in the queue of orders, the Limit Order Book which we analyse in details below. On the other hand, traders can put market orders, orders to buy or sell a certain amount of shares of a given stock at the best available price. We call them "impatient traders". Market orders are usually followed by an immediate transaction, matched to standing limit orders on the opposite side of the book according to the price and the arrival time. In Figure 1.1 we draw a scheme of this mechanism.

The third important constituent of market dynamics are cancellations: i.e. removal of an existing limit order from the book. In general, limit orders increase

liquidity, while market orders and cancellations decrease liquidity. Buy limit orders are called bids and sell limit orders are called asks (or offers). The activity of a market is therefore a succession of quotes (quoted bid and ask) and trades (transaction prices).

However, there are other types of markets with different rules for the participants. For example, the London Stock Exchange (LSE) consists of two markets. The on-book market (SETS) is a centralised order driven market governed by a continuous double auction, where each agent can publicly place bids or asks in the Limit Order Book. The off-book market (SEAQ) is a decentralized bilateral exchange, where trades are arranged privately. The New York Stock Exchange (NYSE) also consists of two markets. The downstairs market is a centralised quote driven market, and operates through a specialist system. Normal market participants are only allowed to put market orders. For each stock there is a specialist (or market maker). The specialist is given the monopoly to put limit orders for both the buy and sell side simultaneously and earn the difference between the sell and buy prices (and as a return regulates the market). The upstairs market is an informal, decentralized market, governed similarly to the off-book market on the LSE.

There exist markets which also display the broker identifier (ID) beside each quote. This is the indication of a transparent market, where brokers are able to identify the parties of other limit orders and the counterparties to trades after transactions have occurred. In Australia, ASX had displayed the full limit order book including broker identification number, until 28 November 2005 when its broker regime changed to be anonymous. World exchanges which have adopted an anonymous market include New Zealand, Paris, and Tokyo.

In order to analyse the dynamic of prices in financial markets, we need to define some key quantities. First of all, the time series of prices is indicated by $\{P_t\}$, where t runs over a defined time interval, for example one week, one day, or one month. The *return* of a stock between time t and $t+1$ is defined as

$$R_t \equiv \frac{P_{t+1} - P_t}{P_t}. \tag{1.1}$$

A useful and better quantity with wide usage is the logarithm of prices, called *log-prices*, $p_t \equiv \log P_t$, and the corresponding *log-returns* of stock prices

$$r_t \equiv \log \frac{P_{t+1}}{P_t} = \log(1 + R_t) \approx R_t. \tag{1.2}$$

where the last approximation is valid for short time scales, for which the returns are small.

In modern financial literature, it is common to study not the increment $P_{t+1} - P_t$ itself, but rather the return R_t or r_t. Indeed two stocks could have different absolute price changes, and therefore different absolute daily price changes, but similar daily returns. Let us describe in more details some key concepts of financial markets, such as the liquidity of markets and the price discovery process.

1.1 Market Liquidity

In market microstructure literature, liquidity is defined as "the willingness of some traders to take the opposite side of a trade that is initiated by someone else at low cost" (Harris, 1990). Thus, as indicated by Lee et al. (1993), market liquidity has two dimensions: the price dimension, represented by the bid-ask spread, and the quantity dimension, represented by the outstanding volume at the opposite side, or market depth. Therefore, the liquidity of the market is partially characterized by the bid-ask spread, which sets the cost of an instantaneous round trip of one share (a buy order instantaneously followed by a sell, or vice versa). A market is called liquid when this cost is small. A large body of microstructure research focuses on market liquidity based on a theory of asymmetric information, which assumes that one party has more or better information than the other in a transaction. Well-informed traders profit at the expense of less-informed traders. Less-informed traders therefore try to avoid well-informed traders (Harris, 2003). Lee et al. (1993) and Benveniste et al. (1992) argue that if specialists believe that there is a chance of informed trading, they will respond by increasing the bid-ask spread and or reduce the depth at the quoted prices. This also implies a negative relationship between the spread and depth. Kavajecz (1999) shows that specialists manage quoted depth to deal with risks associated with an information event. Specifically, Kavajecz finds that liquidity providers, both market-maker and limit order traders, reduce depths around earning announcements to decrease adverse selection costs.

By extending this reasoning to the case of the anonymous limit order book, Foucault et al. (2007) develop a theoretical model which enables them to conclude that anonymous quotes can lead to overall tighter bid-ask spreads. Their model explains that in a transparent market where broker identification codes are displayed, uninformed traders estimate the proportion of informed trades in the market before submitting orders. If they believe that the participation rate of informed traders is small, they will actively set the best quotes, as there is a relatively low chance that informed traders will pick off their limit orders. This leads to narrower spreads. Conversely, when informed traders' participation rate is high, wide spreads from uninformed traders are predicted. However, in an anonymous market, traders generally are unable to discriminate between informed and uninformed parties, and to pick off uninformed orders or free-ride informed orders. They will therefore place more aggressive limit orders, and not behave differently on informed and uninformed trades. This is consistent with the study by Garfinkel and Nimalendran (2003), who investigate the impact of insider trading on market-maker behaviour for anonymous NASDAQ and transparent NYSE. They find that NASDAQ dealers do not adjust to the presence of insider trading by raising effective spreads. The effective spreads of stocks traded in the anonymous NASDAQ dealer system are narrower than in transparent NYSE specialist system.

Moreover, O'Hara (1995), Foucault et al. (1997), Madhavan (2000) and Glosten and Milgrom (1985) summarize the three types of liquidity costs which fix the magnitude of the spread:

- *Order processing costs*, that include market makers' profit;
- *Adverse selection costs*: liquidity takers may have private information on the future price of a stock, in which case market makers suffer losses;
- *Inventory risk*, market makers may temporary take large long or short positions that carry risk. If market makers are sensitive to risk, they may add extra costs and increase the spread.

It has been argued that order processing costs explain large fraction of the spread. This reason might be true for quote driven markets, but it cannot be correct for highly liquid and competitive markets. Finally, in (Bouchaud et al., 2009) is argued that the main determinant of the spread is in fact the market impact and we define it in details in Chapter 2.

1.2 Price Discovery

Like liquidity, price discovery is another central function of financial markets, as stated in (O'Hara, 2003). While the former refers to the easiness of an asset to be traded, the latter refers to the ability of the market to find the efficient price (O'Hara, 2003). Price discovery has been defined as "the incorporation of new information into security prices" (Hasbrouck, 1995), and as "the process by which markets attempt to find equilibrium prices from new information" (Schreiber and Schwartz, 1986). In the Chapter 2 we will discuss briefly about the Efficient Market Hypothesis and the idea of efficient prices in financial markets.

1.2.1 Random walk models for asset price dynamics

The concept of price discovery drawn from previous definitions is that the prices for the same security in different markets should tend to converge in the long run but might deviate from one another in the short run. Each observable price of an asset in multiple markets can be conceived as an information-based common efficient price shared by all these markets, plus a transitory liquidity/noise trading shock such as bid-ask bounce and order imbalances on liquidity trades. Following Hasbrouck (1995, 1996), this concept can be expressed in a random walk model, when considering a security traded in two separate markets at potentially different prices $S_{1,t}$ and $S_{2,t}$:

$$\begin{aligned} P_t &= P_{t-1} + \eta_t \qquad \eta_t \sim \mathrm{IID}(0, \sigma^2), \\ S_t &= P_t + U_t, \qquad U_t = \begin{pmatrix} U_{1,t} \\ U_{2,t} \end{pmatrix}. \end{aligned} \tag{1.3}$$

Here, the common underlying implicit efficient price is P_t, which follows random walk; η_t reflects new information and represents increments of prices which are

Independent and Identically Distributed (IID) random variables; S_t is the observed security price; U_t shows the non-informational features, i.e. transitory liquidity shocks, and is assumed a zero-mean covariance stationary process. As shown in this equation, the process P_t is simply the sum of the disturbance term $\eta_{t-\ell}$, $\ell > 0$. A random walk is a stochastic process which changes by steps and any step does not depend on the previous history of the data. Hence, the returns are unpredictable.

Following the assumption that an implicit unobservable efficient price is common to all markets, Hasbrouck (1995) initiates the information share method as an explicit measure of relative contribution to price discovery by a particular market to the innovation in this common efficient price.

We can calculate with Equation (1.3) the conditional expectation and variance of the process at time t, given some initial value P_0 at time $t = 0$,

$$\begin{aligned} \mathbb{E}[P_t|P_0] &= P_0, \\ \mathbb{V}[P_t|P_0] &= \sigma^2 t. \end{aligned} \tag{1.4}$$

The most common distributional assumption for the increments η_t is normality. If the η_t's are IID $\mathcal{N}(0, \sigma^2)$, then 1.3 is equivalent to an arithmetic Brownian motion, sampled at regularly spaced unit intervals. The equations 1.4 yield the random walk model of Bachelier (1900) and Einstein (1905).

This assumption suffers from one important drawback: most financial assets exhibit limited liability, that is the largest loss an investor can realize is his total investment. This is clearly violated by normality. If the conditional distribution of P_t is normal, then there will always be a positive probability that $P_t < 0$ (Campbell et al., 1997).

To avoid violating limited liability, we can assert that the natural logarithm of prices $p_t = \log P_t$ follows a random walk with normally distributed increments. Hence

$$p_t = p_{t-1} + \eta_t \qquad \eta_t \sim \text{IID } \mathcal{N}(0, \sigma^2). \tag{1.5}$$

Using the previous definition of *log-returns*, we can state that log-returns are distributed as increments η,

$$r_t = p_t - p_{t-1} = \eta_t \qquad \eta_t \sim \text{IID } \mathcal{N}(0, \sigma^2), \tag{1.6}$$

This model defines the geometric Random Walk. The variance σ^2 of the returns distribution is usually called *volatility*.

1.3 The Limit Order Book

We have mentioned previously that the Limit Order Book is the list of all buy and sell limit orders, with their corresponding price and volume, still active at a given

instant of time. In figure 1.2 we present an instant snapshot of the first levels of the limit order book of Google stock, as displayed by `www.island.com`.

As mentioned before, the trading activity in the limit order book consists in the interplay of different actions of the market participants. Real markets accept the execution of different types of limit and market orders, but in order to treat these differences we use a unified description of these types of orders. We simply classify orders based on whether an order results in an immediate transaction. In this case we call it an *effective market order* (or market order), if it leaves a limit order waiting in the book, we call it an *effective limit order* (or limit order). For example, crossing limit orders are limit orders that cross the opposing best price, resulting in a partial or total transaction. The fraction of the order that results in an immediate transaction is classified as market order, whereas the non-transacted part is counted as a limit order.

When an order is executed in the Limit order book, the corresponding price is printed as transaction price. If the requested volume is larger than the volume at the opposite best (that is, the best price available on the opposite side), a succession of trades at increasingly higher prices for buy orders or decreasingly lower prices for sell orders is triggered, until all the volume of the market order is executed. In general, limit orders increase liquidity, while market orders and cancellations decrease liquidity. Buy limit orders are called bids and sell limit orders are called asks (or offers). The activity of a market is therefore a succession of quotes (quoted bid and ask) and trades (transaction prices).

The smallest interval between two prices is fixed by market, and is called the *tick size*. For example, for British stocks, the tick size is typically 0.25 pence for stocks worth more than 300 pence. On EuroNext the tick size is 0.01 Euros for stocks worth less than 50 Euros and grows when the price increases. On NASDAQ the tick size is fixed to 0.01 Dollars. We define one *basis point* (bp) as a fraction of percent and it is equal to 0.01% (10^{-4}). The order of magnitude of the tick size is between 10^{-3} and 10^{-4} relative to the price of the stock or, in terms of basis points, between 10bp and 1bp.

We define the best ask price A_t as the lowest price among the sell limit orders in the book at time t. It is the lowest price at which someone requires to sell some shares. Symmetrically, the best bid price B_t is the highest price among the buy limit orders in the book. It is the highest price that someone offers to pay for buying some shares. Using these two quantities, we can define the *midpoint price* $P_t = (A_t + B_t)/2$ and another important quantity, the *bid-ask spread* $S_t = A_t - B_t$, that is the difference between the best sell price and the best offer price. It is a dynamic quantity, because market orders tend to deplete the order book and increase the spread, whereas limit orders tend to fill the gap and decrease the spread. We can also define the respective *logarithmic* quantities, such as the best ask log-price $a_t = \log A_t$, the best bid log-price $b_t = \log B_t$, and the bid-ask log-spread $s_t = a_t - b_t$.

The limit order book is also characterized by the bid/ask gap and the volume at best bid/ask. The former is the logarithmic price difference between the best quote

Figure 1.2: Part of the order book of Google stock. The first column gives the number of offered shares, the second column the corresponding prices with a tick size of 0.0001\$. Source: `www.island.com`

and the second best quote, defined as $g_t^B = b_t - b_t^{2nd}$ on the bid side and $g_t^A = a_t^{2nd} - a_t$ on the ask side, where the variables a_t^{2nd} and b_t^{2nd} are the second best ask and second best bid. The latter is the liquidity at best bid v_n^B or the liquidity at best ask v_n^A. These two quantities are strictly related to the instantaneous market liquidity of an asset. In fact, large outstanding volumes, or small gaps at the opposite side indicate high liquidity, whereas small volumes, or large gaps at the opposite side are proxies of low liquidity.

1.4 Order book dynamics

Prices in markets with continuous trading are formed in the limit order book, through fluctuations of liquidity on the book. Therefore, it is important to study the origin of these fluctuations and several studies in the past have discovered some regularities in the dynamic of limit order book, that bind order flow and liquidity.

1.4.1 Limit order flow and average shape of order book

Market orders and limit orders are characterized by the volume of execution, but the latters have a second key characteristic, the limit price. This is the price at which the market participants want to trade, and it is a measure of the their patience. In fact, if one considers the absolute value of the difference between the limit price and the best available price, Δ, a patient (impatient) trader places limit order with large (small) values of Δ, very far (close to) the spread.

Empirical studies have found a the power-law distribution of limit order price (Bouchaud et al., 2002), (Zovko and Farmer, 2002). Let us define $B_t - \Delta$ as the price of a new buy limit order and $A_t + \Delta$ as the price of a new sell limit order. In both cases the Δ is measured at the time when the limit order arrives. Asymptotically, *i.e.* for large values of Δ, the probability density of incoming limit order placed at distance Δ is symmetric between buy and sell limit order and is well described by a power-law relation, as

$$\rho(\Delta) \sim \frac{\Delta_0^\mu}{\Delta^{1+\mu}}, \tag{1.7}$$

where $\mu = 1.5$ was measured by Zovko and Farmer (2002) for stocks traded in London Stock Exchange and Bouchaud et al. (2002) have estimated a value of $\mu = 0.6$ for stocks traded in Paris Stock Exchange. In (Mike and Farmer, 2008), the distribution of limit order prices was fitted with a Student t-distribution, with 1.3 degrees of freedom, corresponding to $\mu = 1.3$. The relative price has a distribution that goes from 1 to 100 ticks and it means that market participants anticipate large price jumps that would lead to trading opportunities.

Order flow of market and limit orders and cancellation, and the interplay of the trading activity of different market participants, determine the instantaneous shape of limit order book. An important limit of this shape is the asymptotic profile of the

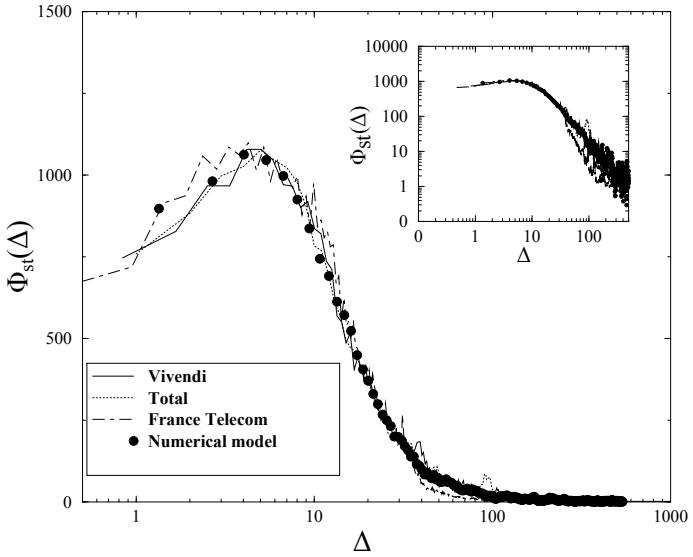

Figure 1.3: Average volume in the queue in function of the absolute value of the distance from best price, in a log-linear plot for three liquid French stocks. The shape is very similar for all three stocks studied. In the inset is showed log-log plot of the same data. From (Bouchaud et al., 2002).

limit order book. Essentially, if we average over time order book states at different time, we found that the asymptotic profile is almost symmetric between bid side and ask side and this fact is true for many different stocks. Surprisingly, the empirical maximum value of volume in the average shape was not found at the best price, but few ticks inside the book, as Figure 1.3 shows, even though the best price is the most likely place for a limit order to be placed. Bouchaud et al. (2002) have demonstrated that this evidence can be explained by the mechanical activity of the market orders, which has been reproduced by their proposed statistical model of the order book dynamics.

It is important to emphasize the big differences that exist between the average shape of limit order book and the instantaneous state of order book. In (Farmer et al., 2004) it is shown that for most stocks of LSE, when the ratio between tick size and price is small, there exists many unoccupied price levels that form gaps in the order book. The typical shape of the book is indeed extremely sparse. In particular, the gap between best price and second best price is responsible of price response to individual trades. This is true because the volume of market orders often matches the size at the opposite best (for buy or sell orders, the volume at the best ask or bid) and the probability that the volume of market orders exceeds the sum of the volume at best and at second best is very low.

1.4.2 Order book models: the zero-intelligence approach

Models of the dynamics of the limit order book are very complicated. These models should take into account the rational behaviour of each market participant. As-

suming limitations in human cognition, a stream of economists have increasingly explored models in which agents have bounded rationality, as alternative of the previous assumption of completely rational agents. In econophysics literature this concept has been stressed and to aim at modelling order book, the order flow has been explored and applied to different systems including market microstructure.

The "zero-intelligence models" drop agent rationality almost completely. Although no one would dispute that agents in financial markets behave strategically, and that for some purposes taking this into account is essential, there are some problems where other factors might be more important. There are several models of this type in the literature, but we describe here the approach developed in Daniels et al. (2003); Smith et al. (2003). This approach has the merit that can be tested against real data, because it presents simple quantitative laws that relate one set of market properties to another, placing restrictions on the allowed values of variables.

This model assumes that two types of agents place orders randomly according to independent Poisson processes. Impatient agents place market orders randomly with a Poisson rate of μ shares per unit time. Patient agents, in contrast, place limit orders randomly in both price and time. Buy limit orders of log-price l_t are placed uniformly anywhere in the semi-infinite interval $-\infty < l_t < a_t$ and similarly sell limit orders are placed uniformly anywhere in $b_t < l_t < \infty$. Both buying and selling limit orders arrive with the same Poisson rate density ρ, which is measured in shares per unit price per unit time. Both limit and market orders are of constant size ν (measured in shares). Queued limit orders are cancelled according to a Poisson process, analogous to radioactive decay, with a fixed-rate υ per unit time. To keep the model as simple as possible, there are equal rates for buying and selling, and all these processes are independent, except for indirect coupling through the boundary conditions. When a new limit order arrives, it may change the previous boundary condition for the subsequent limit order placement. For example, a new buy limit order arrives inside the spread, immediately causing a variation of the best bid, that changes the boundary condition for a subsequent sell limit order. It is this feedback between order placement and price diffusion that makes this model interesting. However this model is rather simplistic, since it neglects all correlations between market and limit orders.

Within this framework, an approximate formula for the mean spread $\mathbb{E}[s]$ can be derived by dimensional considerations. The spread has dimensions of price and the unique combination of order flow rates with these dimensions is μ/ρ. Each order flow rate was defined above. While the dimensions indicate the scaling of the spread, they cannot determine multiplicative factors of order unity. A more intuitive argument can be made by noting that inside the spread, removal of limit orders due to cancellation is dominated by removal due to market orders. Thus the total limit order placement rate inside the spread, for either buy or sell limit orders ρs, must equal the order removal rate $\mu/2$, which implies that the spread is

$$\mathbb{E}[s] = \mu/2\rho. \tag{1.8}$$

In (Smith et al., 2003) it is found a more quantitative multiplicative factor for

the average spread,

$$\mathbb{E}[s] = \frac{\mu}{\rho} F\left(\frac{\upsilon}{\mu}\right), \tag{1.9}$$

where $F(u)$ is a monotonically increasing function that can be approximated as $F(u) \approx 0.28 + 1.86u^{3/4}$. The ratio u can be thought of as the ratio of removal by cancellation to removal by market orders and plays a crucial role in the order book dynamic. In the limit of $\upsilon \to 0$, we recover the result of Equation (1.8), *i.e.* $\mathbb{E}[s] = 0.28\mu/\rho$.

Predictions of Equation (1.9) can be tested against real data, as in (Farmer et al., 2005), by independent measure of order flow rates and unconditional mean spread.

Chapter 2

Price movements: efficient prices and market impact

Markets are places where buyers and sellers meet to exchange goods at given levels of prices. During this apparently simple process, the informations of market agents are incorporated in a single number, the price. One of the achievements of economics has been the formulation of simple and elegant equilibrium models that attempt to explain the end results of this process without going into the details of the mechanisms through which prices are actually set. In fact, there exist situations in which is very useful to understand in the details the process of trading and the strategic nature of its dynamic, details that these broad-brush equilibrium models do not provide. It is worth, indeed, to properly understand how prices change from a more microscopic point of view. First of all, we introduce briefly the idea of efficient prices, and the relation between the unpredictability of prices and the incorporation of information on prices. Then, we discuss about the relevant propriety of the order flow, which is highly correlated in time, and how to reconcile this characteristic in a market impact model with efficient prices.

2.1 The Efficient Market Hypothesis

The Efficient Market Hypothesis is one of the key concepts of the modern economics. It has been developed and formulated first by Samuelson (1965) and Fama (1970).

Market microstructure theory suggests that the evolution of asset prices depends on the nature of players in the market. An early model by Lintner (1969) on asset price formation shows that financial markets aggregate the beliefs of individual traders, and the market equilibrium price is a weighted average of these beliefs with the weights being determined by the investors' risk aversion. Grossman (1976) considers a rational expectations equilibrium model of a stock market in which there are two types of traders: informed and uninformed. Informed traders know the true value of traded asset, and take positions in the market based on their information.

Uninformed traders invest no resources in collecting information, but know that prices will reflect the information of the informed traders. Under this framework, when informed traders trade, the security price will reflect all of the information to all traders, and private information is transmitted from the informed to the uninformed. Kyle (1985) develops a model of speculative trading in which a monopolist insider trades sequentially in the asset market against uninformed noise traders, who trade randomly without information. Kyle's main result is that the insider trades slowly, so that his private information is incorporated into prices gradually.

Despite the development of various models of price formation, one feature in common is that they assume that market prices are based on private information. Uninformed investors act as price takers and price discovery occurs through trading with informed traders.

A market is called efficient (or more precisely informationally efficient) if all available information on the market is used to determine prices (Samuelson, 1965; Fama, 1970). More precisely, prices incorporate the information and expectations of all market participants. In a financial market consisting of rational and fully informed traders, the price of each financial asset will equal the asset's fundamental value (Arthur et al., 1997).

The Efficient Market Hypothesis has been one of the cornerstones of twentieth century financial research. The reason for the wide interest is due to the fact that it implies that arbitrage possibilities are not possible.

If financial markets are "informationally efficient", the best predictor for price P_t at time t, conditioned on the information set available at the same time, I_{t-1}, is the last price. Then, the Efficient Market Hypothesis can be written in the following way

$$\mathbb{E}[P_t|I_{t-1}] = P_{t-1}. \tag{2.1}$$

As seen above, market efficiency is defined subject to a certain information set. In practice, the classical taxonomy of information sets distinguishes between three forms of market efficiency (Fama, 1970).

- **Weak form of efficiency**: The information set I_t includes only historical prices. This means future price movements cannot be predicted by using past prices. Moreover, past data on stock prices are useless for predicting future stock price changes.
- **Semi-strong form of efficiency**: The information set I_t includes all information publicly available by all market participants. Therefore, only investors with additional inside information could have advantage on the market. Any price anomalies are quickly found out and the stock price adjusts.
- **Strong form of efficiency**: The information set I_t includes all information, private and public, known by any market participant. Therefore, no one can have advantage on the market in predicting prices since there is no data that

would provide any additional value to the investors. It also means that in case of a strong form efficient market, prices are only subject to change if new information arrives.

The weak form of efficiency and Equation (2.1) defines a *martingale*.

In the financial community there exists a long debate about the efficiency of real markets and numerous criticisms have been proposed against this hypothesis. First of all, it is difficult that the strong form of efficiency generally does hold for real markets. Nonetheless, it is believed that markets have a high level of efficiency. The principal theoretical argument against the Efficient Market Hypothesis has been given by Grossman and Stiglitz (1980). They have argued that efficient markets are in fact an impossibility, since in this kind of market the return for gathering information is zero, thus there is no reason for any agents to trade, causing the market to collapse eventually. Market efficiency is a very useful concept, but it is at best an approximation.

2.2 Market impact of trades

Dynamical models that incorporate market microstructure have triggered a widely interest in the economics literature, which calls into question the view that prices always remain in equilibrium and respond instantly and correctly to new information. The work reviewed in this chapter argues that trading is inherently an incremental process and that for this reason, prices often respond slowly to new information. The reviewed body of theory springs from the recent empirical discovery that changes in supply and demand constitute a long-memory process. This means that the supply and demand pressure of market agents vary very slowly, with a persistence that is observed on time scales of weeks or even months. The markets incorporate only with difficulty these movements in supply and demand, and react in order to keep the prices informationally efficient. This process involves adjustments in market liquidity, and its deep analysis requires the understanding of many properties of market microstructure, such as volatility, the bid-ask spread, and the market impact of individual incremental trades. By market impact we mean the average response of prices to trades and the quantification of such response is a crucial question for researchers and financial institution. Orders contain a variable amount of information about the hidden background of supply and demand. This affects how much prices move and therefore modulates the way in which information is incorporated into prices.

The debate on the origin of market impact is long in the economics literature and many interpretations were proposed in the past. Bouchaud et al. (2009) exposed the three main interpretations in literature. All these interpretations result in a positive correlation between trading volume and price impact:

1. *Traders make a successful short term prediction of price and trade accordingly.*

This leads to a correlation between orders and price movements, even if trades themselves do not impact prices at all. This interpretation is incoherent with real data because when "noise traders" initiate trades, with no information content, they must have no market impact.

2. *Trades are considered a signal about private information.* The arrival of new private information generates trades, other agents in the market may update their consideration, which changes prices. The result of this point of view is that all trades have to impact prices, because identities of traders in electronic markets are anonymous and it is hard to identify which ones are informed traders.

3. *Market impact is due to a pure statistical mechanism.* We have discussed in Section 1.4.2 about models of order book where order flow is a purely random process. When an extra buy appears in the market, with other conditions kept constant, the price rises up. These fluctuations in supply and demand can be completely random, with no information content, and the effect of market impact remains the same. In this case market impact is a completely statistical phenomenon.

In the first two interpretations, orders do not impact prices themselves, but it is right to say that orders "forecast" prices. If agents know the future price and trade accordingly, other market participants may change their valuations and their quotes. In this case the incorporation of information in prices is a function of the observed order flows. In the last interpretation, information revelation is an ambiguous concept, trades themselves move prices, even if the fraction of informed traders is zero.

2.3 Long memory of the order flow

From a mechanical point of view, price formation process is the outcome of the flow of orders arriving in the market and the response of prices to individual orders. The statistical proprieties of the order flow is crucial in order to understand this process, which can be studied by considering the time series of signs of orders. Specifically, consider the symbolic time series $\{\epsilon_t\}$ obtained in event time by replacing buy market orders with $+1$ and sell orders with -1, irrespective of the volume of the order size in order to avoid problems created by large fluctuations in order size. We will see that the statistical properties are almost the same for the order flow or market order, limit orders and cancellations.

Figure 2.1 shows the sample autocorrelation functions of order signs $C(\ell) = \mathbb{E}[\epsilon_t \epsilon_{t+\ell}]$ for different stocks traded on Paris Bourse in double logarithmic scale. We note that the decay of the functions is very slow in time, in fact it is still above the statistical noise level even after 10^4 transactions, which for this stock corresponds to roughly 10 days. This evidence indicates that if one observes a buy market order

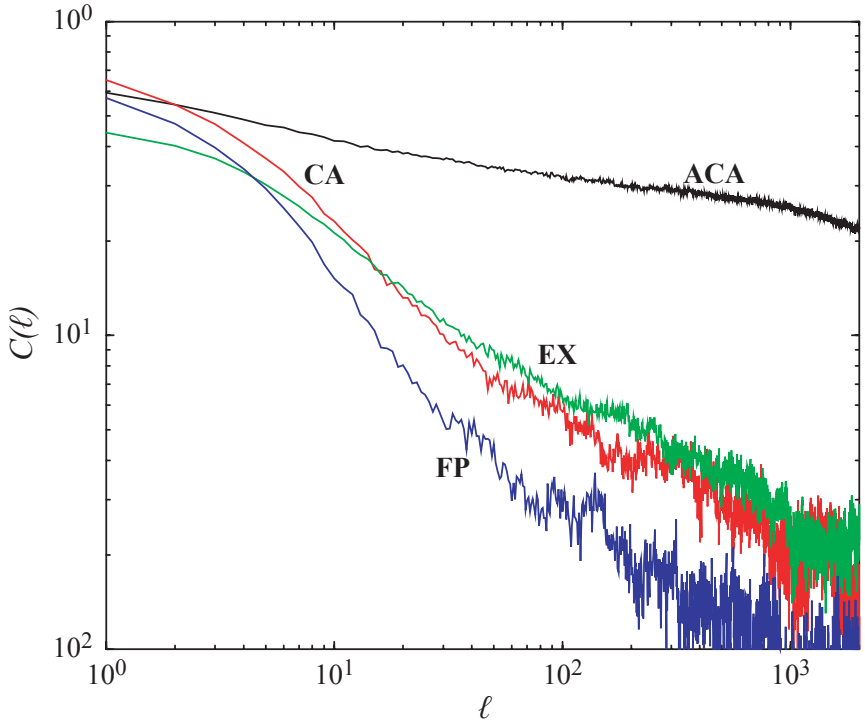

Figure 2.1: Plot of the sign correlation $C(\ell)$, for a selection of four stocks (label with ACA, CA, EX and FP) traded on Paris Bourse. In log-log scale the curves are approximately linear and show the long-range nature of the correlation of signs. In fact, they can be approximate by a power-law relation. From (Bouchaud et al., 2006).

now, based on this information alone, there is an high probability that a market order of the same signs is executed two weeks from now.

The sample autocorrelation function shown in the figure is roughly linear in double logarithmic scale for large lags ℓ, which means that it can be fit by a power-law relation $C(\ell) \sim \ell^{-\gamma}$, where $0 < \gamma < 1$. The values of the exponent γ is an important indicator of the persistence of the underlying process, which in the literature is indicated as a long-memory process where the smaller is γ, the longer is the memory. Long memory can also be discussed in terms of the Hurst exponent H, which is simply related to γ. For a long-memory process, $H = 1-\gamma/2$ or $\gamma = 2-2H$. Short-memory processes have $H = 1/2$, and the autocorrelation function decays faster than $1/\ell$. A positively correlated long-memory process is characterized by a Hurst exponent in the interval $(0.5, 1)$. Lillo and Farmer (2004) have found that for all the stocks under analysis, the order flow of market orders shows the same characteristic of positive persistence of difference degree. The value of H observed in the London Stock Exchange was generally about $H \approx 0.7$, which corresponds to $\gamma = 0.6$. Bouchaud et al. (2004, 2006) measured a larger interval of γ values in the Paris Stock Exchange, ranging from 0.2 to 0.7.

We have discussed about the statistical proprieties of the order flow in financial markets, but is not clear the origin of this persistence. As stated in (Bouchaud et al.,

2009), the evidence of positive correlations of the order flow suggest two possible classes of explanations. The first class considers that this is a propriety of the order flow of each investors, in an independent way with respect to the other investors. The second one states that the interplay between different market agents through strategies of herding leads to a correlated order flow. In the literature it has been shown that the first explanation is the prevalent, despite the empirical evidence also of the second strategic behaviour of investors.

In a pioneer work, Lillo et al. (2005) have developed a model which generate a correlated order flow. It is based on the strategy of a large investor, which decides to buy an amount of shares of a company, but she submits to the market small volume orders and trades incrementally over a long period of time, with the aim of keeping her intentions as secret as possible. These *metaorders* are split in small pieces and, by definition, every small order executed has the same sign. The delay in time of the execution of the whole metaorder generates long-range temporal correlation in the time series of ϵ_t, and they provided a relationship between the distribution of sizes of the large orders and the statistical characteristic of the correlated order flow. We will describe it in more details in Chapter 3.

Empirical evidences of the previous mechanism are difficult to find because it is not easy to collect data of the trading activity of individual investors. There exists indirect way to recover partial information about the identity of participants by identifying the broker or the member of the exchange who places orders on the market, which is characterized by the so-called *membership code*. In many stock markets, such as the LSE, the Spanish Stock Exchange, the Australian Stock Exchange, and the NYSE, it is possible to obtain data containing this information. It is important to stress that knowing the membership code is not the same as knowing the individual participant, since the member may either trade on its own account or act as a broker for other trades, or do both at once.

Tóth et al. (2015) show the conditional autocorrelation function of market order signs with the same membership code and different brokerage. The first curves have a clear power-law behaviour, whereas the second curves show that, after a few lags, the correlation of order signs from different brokers are slightly negative. Under the assumption that most investors use only a few brokers to execute a given hidden order, these evidences strongly support the hypothesis that the long-memory of order signs is due to the presence of investors which place small pieces of hidden orders on the market of the same sign, in a large period of time.

2.4 Theory of market impact

From the perspective of an investors the key question about market impact is its dependence from the volume traded and the temporal evolution of the immediate impact of their trades. For many purposes it is useful to separate the dependence on volume from the dependence on time, by making the hypothesis that the impact

function can be written as a product of two functions. These functions depend independently from the volume and the time.

We define some useful variables to study market impact: p_n as the logarithm of the mid-price of the stock just before the n-th transaction executed on the market at time t_n, the logarithmic return is $r_n = p_{n+1} - p_n$, the traded volume v_n and the sign ϵ_n of n-th transaction. If the trade is initiated by a buyer, the sign is $+1$, whereas if it is initiated by a seller, it is -1.

2.4.1 Volume dependence: empirical results and theory

The dependence of the price shift of an individual trade from the executed trade has been intensively studied and measured in several studies. The common characteristic found in all these works is the concaveness of the empirical function on the volume, which increases rapidly for small v_n and more slowly for larger v_n.

The detailed functional form, however, varies from market to market and even period to period. Lillo et al. (2003) found that the average price impact of trades for a set of 1000 NYSE stocks behaves as a concave function,

$$\mathbb{E}[r_n|v_n] = \frac{\epsilon_n v_n^{\psi}}{\lambda}, \tag{2.2}$$

where the exponent $\psi(v_n)$ is approximately 0.5 for small volumes and 0.2 for large volumes. The liquidity parameter λ varies for different stocks; in particular, there is a clear dependence on market capitalization M that is well approximated by the functional form $\lambda \sim M^{\delta}$, with $\delta \approx 0.4$. Potters and Bouchaud (2003) have found that for orders traded at the Paris Bourse and NASDAQ the impact function can be fit very well by a logarithm function of the volume,

$$\mathbb{E}[r_n|v_n] \propto \epsilon_n \log v_n. \tag{2.3}$$

Thus all the studies find strongly concave functions but report variations in functional form that depend on the market and possibly other factors as well.

Different explanations have been proposed in literature in order to explain the concavity of the impact function. They are related to the different informativeness of trades of small and large size, or the instantaneous shape of the limit order book. Farmer et al. (2004) suggested that the underlying mechanism which explains very well the concavity of the impact function is the "selective liquidity taking". In a way, the agents condition the size of their transaction on the available liquidity in the market: They trade large volumes when the outstanding liquidity is high and making small transactions when it is low. For example, for Astrazeneca they measured that approximately 87% of the market orders creating an immediate price change have a volume equal to the volume at the opposite best. Moreover, approximately 97% of the market orders creating an immediate price change have a volume that is either equal to the opposite best or larger than this value but smaller than the sum of

volume at the second best opposite price. The concavity of the impact function is the statistical effect of this strategic behaviour of market agents.

2.4.2 Temporal evolution: a naive model

In the previous section we have discussed about the price shift triggered by an order of a given volume v_n. We focus in particular on the immediate impact, that is the price change occurred after a trade is completed. This immediate impact of individual transactions, defined as $\mathbb{E}[r_n|\epsilon_n, v_n]$, is non zero and can be written as $\mathbb{E}[r_n|\epsilon_n, v_n] = \epsilon_n f(v_n)$, where f is a function that grows with v_n. It is important to understand the temporal evolution of this immediate impact, that is if the impact can be composed by transient or permanent components, and if its magnitude is fixed or time dependent.

We need to define some quantities to study the temporal behaviour of the price impact. We define a series of discrete times t_n, indexed by $n \in \mathbb{N}$, called *trade time*. The price p_n is the log-mid-price of the stock at, or right before, time t_n, and the series of traded volumes v_n of the individual transactions. The measure of correlation between an order sign at time t_n and the future midpoint change at time $t_{n+\ell}$ is called *response function* $\mathcal{R}(\ell)$, with *return variance* $\mathcal{V}(\ell)$

$$\begin{aligned} \mathcal{R}(\ell) &\equiv \mathbb{E}[(p_{n+\ell} - p_n) \cdot \epsilon_n], \\ \mathcal{V}(\ell) &\equiv \mathbb{E}[(p_{n+\ell} - p_n)^2]. \end{aligned} \tag{2.4}$$

The simplest model for the impact is the usual random walk model, where the price at a given time is the sum of the past steps and price changes. That is, the price impact of individual trades is permanent in time and affects forever the price. In more mathematical terms, the price at a given time t_n is

$$p_n = \sum_{k<n} \left[\epsilon_k f(v_k) + \eta_k\right], \tag{2.5}$$

and the mid-price change just before the n-th transaction and just before the $(n+1)$-th transaction is

$$r_n = p_{n+1} - p_n = \epsilon_n f(v_n) + \eta_n, \tag{2.6}$$

where η_n are IID random variables added to account for price changes not due to trading itself, such as the arrival of news that updates instantaneously the quotes. We assume that η_n are independent on the order flow and we set $\mathbb{E}[\eta] = 0$ and $\mathbb{E}[\eta^2] = \Sigma^2$.

The model predictions for the response function and the return variance of Equations (2.4) are

$$\begin{aligned} \mathcal{R}(\ell) &= \mathbb{E}[f], \\ \mathcal{V}(\ell) &= \left(\mathbb{E}[f^2] + \Sigma^2\right) \ell, \end{aligned} \tag{2.7}$$

where $\mathcal{R}(\ell)$ is a constant price impact, that is independent on lag time ℓ, and $\mathcal{V}(\ell)$ denote a pure price diffusion.

Recalling that η_n are IID random variables and $\mathbb{E}[\epsilon_n] = 0$, we write the autocovariance of price returns within the model of Equation (2.6):

$$\mathbb{E}[r_n r_{n+\ell}] - \mathbb{E}[r_n]^2 = \mathbb{E}[\epsilon_n \epsilon_{n+\ell} f(v_n) f(v_{n+\ell})]. \tag{2.8}$$

We now assume (as in Bouchaud et al. 2006) that the last expression can be factorized:

$$\mathbb{E}[r_n r_{n+\ell}] = \mathbb{E}[\epsilon_n \epsilon_{n+\ell}] \cdot \mathbb{E}[f(v_n) f(v_{n+\ell})]. \tag{2.9}$$

We recall that the empirical autocorrelation function of order signs behaves as a power-law, $C(\ell) \sim c_0 \ell^{-\gamma}$. In addition, we consider the impact function $f(v_n)$ as an independent random variable, with finite variance, *i.e.* $\mathbb{E}[f(v_n) f(v_{n+\ell})] = \mathbb{E}[f(v_n)]^2$. Finally, we found that

$$\mathbb{E}[r_n r_{n+\ell}] \simeq {c'}_0 \ell^{-\gamma}. \tag{2.10}$$

which means that price returns are strongly autocorrelated in time. Therefore, the previous simple model of the random walk is incompatible with the empirical evidence of the positively correlated order flow, because it would violate market efficiency with price returns easily predictable. In other words one of the assumptions of the random walk model must be relaxed. Among the various possibilities we will relax either the assumption that price impact is permanent or the assumption that price impact is independent of the order flow. As we will see, these two possibilities are related one to each other, but for the sake of clarity we present them in two different subsections.

2.4.3 Transient but fixed impact: the propagator model

We have seen that an impact model of fixed and permanent is incompatible with the efficient market hypothesis if the correlations of order signs is taken into account. Bouchaud et al. (2004) generalized the previous model of Equation (2.5) in a trade superposition model, where at a given trade time t_n the price is written as the sum over all impact of past trades propagated up to time t_n,

$$p_n = \sum_{k<n} \left[\mathcal{G}(n, k, v_k) \epsilon_k + \eta_k \right], \tag{2.11}$$

where $\mathcal{G}(n, k, v_k)$ is the "bare" impact function (or propagator) at time t_n of a single trade executed at time t_k, with traded volume v_k and order sign ϵ_k. As in the previous model, η_k are IID random variables added to account for exogenous sources of price changes, such as the arrival of news. This bare impact function is quite general, in fact the authors made some assumptions on the form and the dependence on time and volume: first of all, they decoupled the contribution of the volume and time by factorizing the function as in (Bouchaud et al., 2004), (Daniels et al., 2003) and (Potters and Bouchaud, 2003); secondly, the impact function is

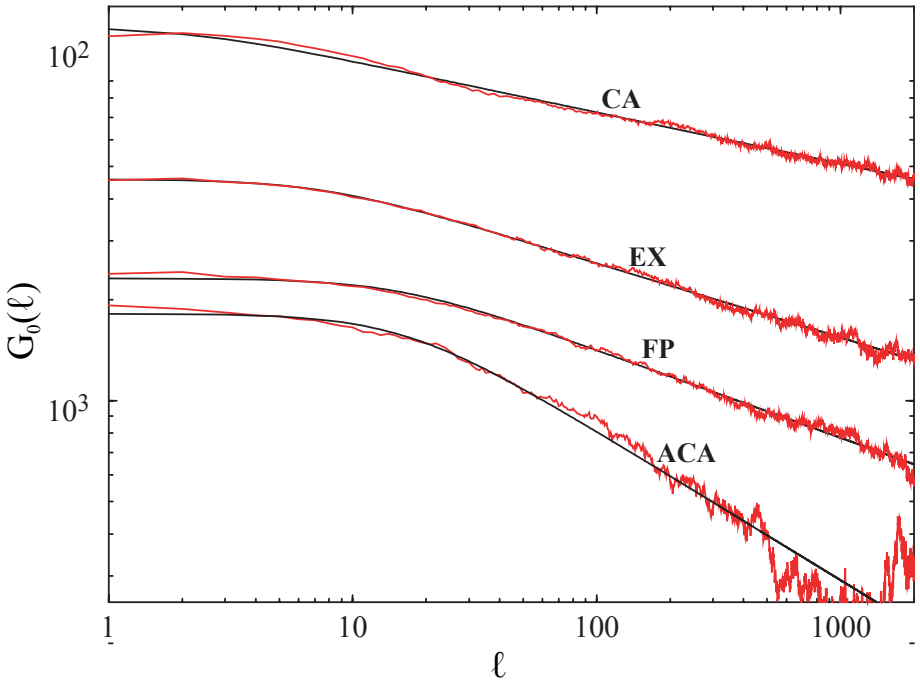

Figure 2.2: Comparison for four different stocks (ACA, CA, EX, FP) between the empirical propagator, extracted from $\mathcal{R}(\ell)$ and $C(\ell)$, and the fit using the improved relation $\Gamma_0/(\ell_0^2+\ell^2)^{\beta/2}$. From (Bouchaud et al., 2006).

fixed and deterministic; finally, the function is exogenous in time, which means that the impact depends on the time difference and not on the time itself. The latter is a strong assumption as is explained by Lillo and Farmer (2004), and we discuss about its relaxation in the Chapeter 4. The result of these assumptions is that the propagator $\mathcal{G}$ can be written in this simple form

$$\mathcal{G}(n,k,v_k) \approx f(v_k)\cdot G(n-k). \tag{2.12}$$

By using the previous expression, we can write the current price as a superposition of order signs, weighted by the propagator, and external shocks (news)

$$p_n = \sum_{k<n} \left[G(n-k)f(v_k)\epsilon_k + \eta_k\right], \tag{2.13}$$

and using this representation, the price change between time t_n and time $t_{n+\ell}$ is

$$\begin{aligned} p_{n+\ell} - p_n = &\sum_{0\leqslant k<\ell} \left[G(\ell-k)f(v_{n+k})\epsilon_{n+k} + \eta_{n+k}\right] \\ &+ \sum_{k>0} \left[G(\ell+k) - G(k)\right] f(v_{n-k})\epsilon_{n-k}. \end{aligned} \tag{2.14}$$

Let us focus only on the temporal evolution of price impact, by factorizing out the dependence from the volume as made by Bouchaud et al. (2004, 2006). We can

compute the response function $\mathcal{R}(\ell)$ and the return variance $\mathcal{V}(\ell)$ of Equations (2.4). The first is

$$\mathcal{R}(\ell) = \sum_{0 \leqslant k < \ell} G_0(\ell - k) C(k) + \sum_{k>0} \left[G_0(\ell + k) - G_0(k) \right] C(k), \tag{2.15}$$

which, in the case of permanent impact of each trade, $G_0(\ell) = G_0$, and $C(\ell) \sim c_0 \ell^{-\gamma}$ with $\gamma < 1$, leads to an average market impact that grows like $\ell^{1-\gamma}$. The solution proposed in this model is that the bare propagator decay with time, in such a way to reduce the amplification effect due to trade correlations. To quantify the decay of the propagator, we analyse the return variance

$$\begin{aligned} \mathcal{V}(\ell) = & \sum_{0 \leqslant k < \ell} G^2(\ell - k) + \sum_{k>0} \left[G(\ell + k) - G(k) \right]^2 \\ & + 2\Delta(\ell) + \Sigma^2 \ell, \end{aligned} \tag{2.16}$$

where $\Delta(\ell)$ is the correlation-induced contribution

$$\begin{aligned} \Delta(\ell) = & \sum_{0 \leqslant j < k < \ell} G(\ell - j) G(\ell - k) C(k - j) \\ & + \sum_{0 < j < k} \left[G(\ell + j) - G(j) \right] \left[G(\ell + k) - G(k) \right] C(k - j) \\ & + \sum_{0 \leqslant j < k} \sum_{k>0} G(\ell - j) \left[G(\ell + k) - G(k) \right] C(k + j). \end{aligned} \tag{2.17}$$

Assume that $G_0(\ell)$ itself decays at large ℓ as a power law, $\Gamma_0 \ell^{-\beta}$. When $\beta, \gamma < 1$, the asymptotic analysis of $\Delta(\ell)$ yields:

$$\Delta(\ell) \propto \ell^{2 - 2\beta - \gamma} \tag{2.18}$$

If the single trade impact does not decay ($\beta = 0$), we recover the above superdiffusive result. This regime indicates that the diffusion coefficient of the price process is no longer constant, but increases with the time lag at witch the returns are measured. If the impact decays faster, superdiffusion is reduced, until

$$\beta = \beta_c = \frac{1 - \gamma}{2}, \tag{2.19}$$

for which $\Delta(\ell)$ grows exactly linearly with ℓ and contributes to the long-term value of the volatility. However, as soon as β exceeds β_c, $\Delta(\ell)$ grows sublinearly with ℓ, and impact only enhances the high-frequency value of the volatility compared to its long-term value Σ^2, dominated by "news". We therefore reach the conclusion that the long-range correlation in order flow does not induce long-term correlations nor anticorrelations in the price returns if and only if the impact of single trades is transient ($\beta > 0$) but itself nonsummable ($\beta < 1$). The convolution of this semipermanent impact with the slow decay of trade correlations gives only a finite contribution to the long-term volatility.

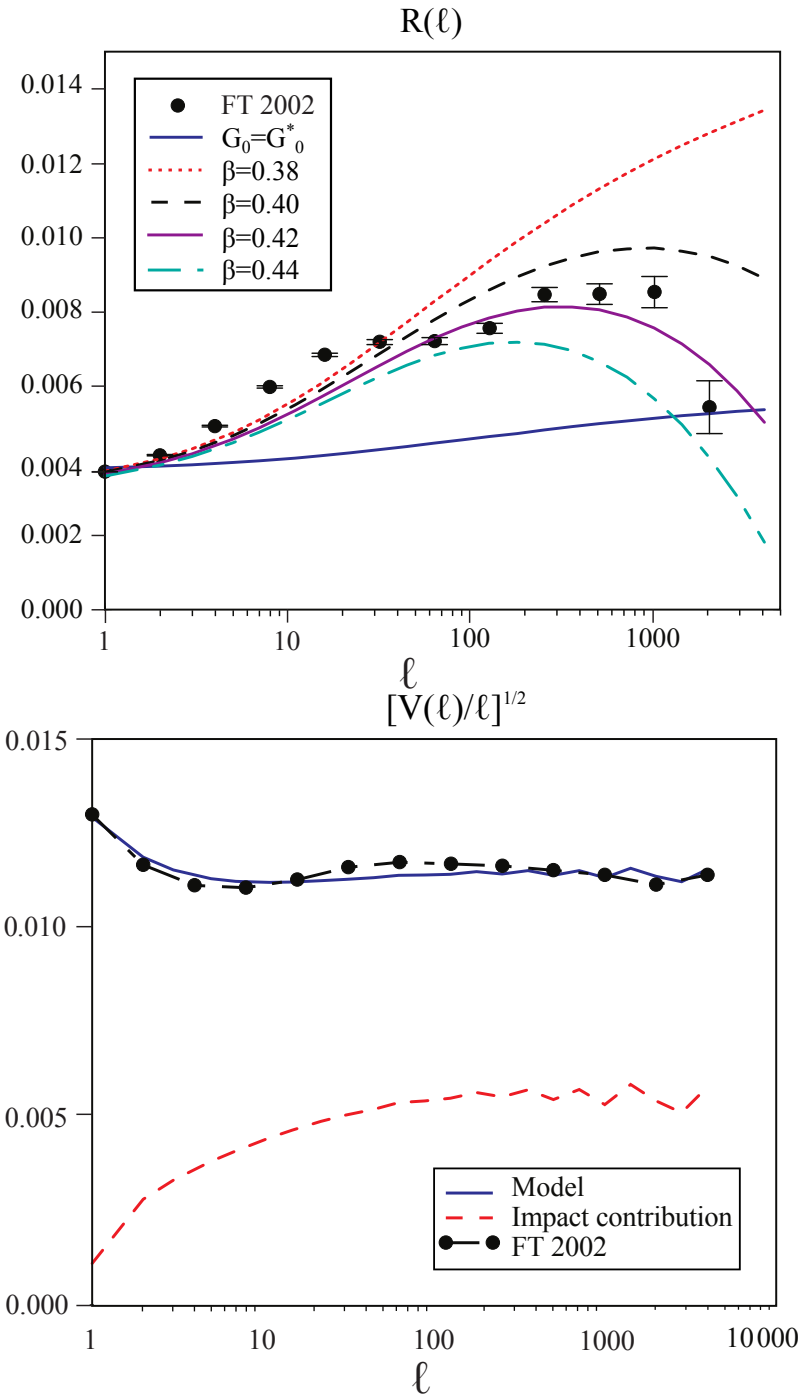

Figure 2.3: (Top) Theoretical response function $\mathcal{R}(\ell)$ of Equation (2.15) for different values of β close to $\beta_c = 0.38$ and for the France Telecom stock traded in 2002. The empirical shape of the function is reproduced using $\beta = 0.42$. (Bottom) Diffusion constant divided by ℓ of Equation (2.16), with the values of γ, β, C_0 determined from $\mathcal{R}(\ell)$. One extra parameter was used, $\Sigma = 10^{-4}$. The lower graph is the "impact contribution" to $\mathcal{V}(\ell)$, given for $\Sigma = 0$. The "oscillations" at long times are a numerical artefact. From (Bouchaud et al., 2004).

Figure 2.2 shows the fit of the empirical propagator, obtained from $\mathcal{R}(\ell)$ and $C(\ell)$ using real data, with the relation $G_0(\ell) = \Gamma_0/(\ell_0^2 + \ell^2)^{\beta/2}$, which matches quite well the initial behaviour of empirical $G_0(\ell)$. Moreover, Figure 2.3 shows the empirical market impact function for France Telecom and the theoretical impact function of Equation (2.15), for different values of β. For $\beta > \beta_c = 0.42$ the function becomes negative at long times, as indeed observed empirically for $\ell > 5000$. Furthermore, we show the empirical scaled diffusion constant $\mathcal{V}(\ell)/\ell$ and the predicted function by the model.

2.4.4 Permanent but variable impact: the asymmetric liquidity mechanism

A different interpretation of the previous formalism was proposed by Lillo and Farmer (2004), Farmer et al. (2006), and Gerig (2007), for which the price impact is permanent, but it is variable in time and depends on the past history of the order flow.

They generalize the MMR model proposed by Madhavan et al. (1997), where it is postulated that price moves only because of the unpredictable external shocks (such as "news") and because of the surprise component in the order flow. By construction of the model, predictability of returns is removed and informational efficiency is ensured. Within this framework and neglecting volume fluctuations, the immediate impact of a transaction at a given time t_n is

$$r_n = p_{n+1} - p_n = Af(v)(\epsilon_n - \hat{\epsilon}_n) + \eta_n, \qquad \hat{\epsilon}_n = \mathbb{E}_{n-1}[\epsilon_n | \Omega_{n-1}], \tag{2.20}$$

where $\hat{\epsilon}_n$ is the sign predictor computed at time t_{n-1}, using the information set Ω_{n-1} available at time t_{n-1}. In the Chapter 3, we will discuss in details which are the mechanisms that ensure informationally efficient prices, by removing any predictability of returns.

We can state that only two possibilities exist: either the sign of the n-th transaction matches the sign predictor, or it is opposite. Let us call $s_n = \text{sign}(\hat{\epsilon}_n)$ the sign of the predictor, we define in the two cases the expected response function at lag one after the n-th transaction as

$$\begin{aligned} \mathcal{R}^r(1) &= \mathbb{E}[(p_{n+1} - p_n) \cdot \epsilon_n | \epsilon_n = s_n], \\ \mathcal{R}^w(1) &= \mathbb{E}[(p_{n+1} - p_n) \cdot \epsilon_n | \epsilon_n = -s_n], \end{aligned} \tag{2.21}$$

where in the first case we have an expected return $s_n \mathcal{R}^r(1)$, and in the second case $-s_n \mathcal{R}^r(1)$.

We can also define the probability that the sign predictor $\hat{\epsilon}_n$ matches (or dis-

agrees) with the realized order sign ϵ_n,

$$\begin{aligned}\mathbb{P}(\epsilon_n = s_n|\Omega_{n-1}) &= \mathbb{P}(\epsilon_n = +1|\Omega_{n-1})\frac{1+s_n}{2} + \mathbb{P}(\epsilon_n = -1|\Omega_{n-1})\frac{1-s_n}{2} \\ &= \frac{1+|\hat{\epsilon}_n|}{2} \geqslant \frac{1}{2}, \\ \mathbb{P}(\epsilon_n = -s_n|\Omega_{n-1}) &= 1 - \mathbb{P}(\epsilon_n = s_n|\Omega_{n-1}) = \frac{1-|\hat{\epsilon}_n|}{2} \leqslant \frac{1}{2}. \end{aligned} \tag{2.22}$$

To ensure the unpredictability of the next return, we must have that

$$\mathbb{E}_{n-1}[r_n|\Omega_{n-1}] = s_n\left[\frac{1+|\hat{\epsilon}_n|}{2}\mathcal{R}^r(1) - \frac{1-|\hat{\epsilon}_n|}{2}\mathcal{R}^w(1)\right] \equiv 0. \tag{2.23}$$

Therefore, the informational efficiency of prices imposes that

$$\frac{\mathcal{R}^r(1)}{\mathcal{R}^w(1)} = \frac{1-|\hat{\epsilon}_n|}{1+|\hat{\epsilon}_n|} \leqslant 1. \tag{2.24}$$

This inequality implies that the most likely outcome, when $\epsilon_n = s_n$, has the smallest impact. This is the mechanism called *asymmetric liquidity*, where the impact of trades is permanent, but it is variable and dependent on the past history of the order flow.

We want to emphasize the crucial role of the predictor used by a liquidity provider to forecast order flow and the information set Ω to ensure informational efficiency.

The simplest model and probably the most widely used to forecast the order flow is a linear model of past order signs, called p-th order autoregressive model *AR(p)*. In the next Chapter 3 we will deeply discuss this process and we will propose a more sophisticated discrete valued process. For now, let us assume that the predictor is computed from a simple *AR(p)* model,

$$\hat{\epsilon}_n = \sum_{k=1}^{p} a_k \epsilon_{n-k}, \tag{2.25}$$

where a_i are real numbers provided by calibration of historical data (via resolution of Yule-Walker equations, or ordinary least square).

Using as sign predictor an *AR(p)* model when $p \to \infty$ and neglecting volume fluctuations, the generalized MMR model of Equation (2.20) is *equivalent* to the previous model of transient impact of Equation (2.14)

$$r_n = p_{n+1} - p_n = G_0(1)f(v)\epsilon_n + \sum_{k>0}\left[G_0(k+1) - G_0(k)\right]f(v)\epsilon_{n-k} + \eta_n, \tag{2.26}$$

if one imposes this equivalence

$$-Aa_k = G(k+1) - G(k) \qquad \text{or} \qquad G(\ell) = A\left[1 - \sum_{j=1}^{\ell-1} a_j\right]. \tag{2.27}$$

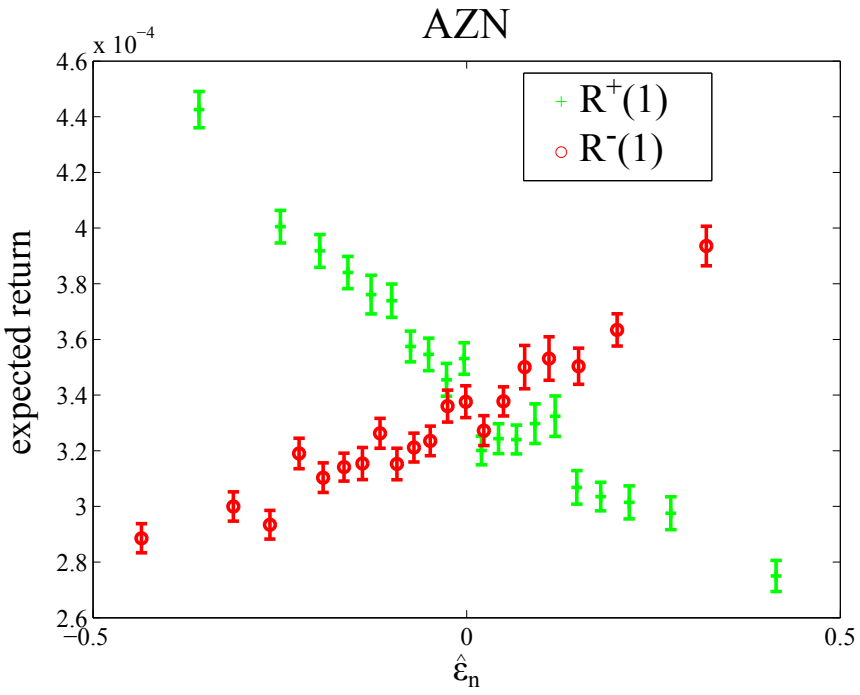

Figure 2.4: The expected return of buyer $\mathcal{R}^+(1)$ and seller $\mathcal{R}^-(1)$ initiated transactions as a function of the sign predictor $\hat{\epsilon}_n$ for the stock AZN traded on London Stock Exchange. This predictor is computed using the membership code of each trade. The data is binned by $\hat{\epsilon}_n$ such that each bin contains an equal number of points (this bin sampling will be explained and intensively used in the Chapter 3), and the mean values of $\mathcal{R}^+(1)$ and $\mathcal{R}^-(1)$ are plotted for each bin with the error bars showing the standard error of this average. From (Gerig, 2007).

Empirical evidences of the mechanism of asymmetric liquidity have been found in several papers (Lillo and Farmer, 2004; Farmer et al., 2006; Gerig, 2007). In particular, Figure 2.4 shows the conditional returns $\mathcal{R}^+(1)$ and $\mathcal{R}^-(1)$, defined as

$$\begin{aligned} \mathcal{R}^+(1) &= \mathbb{E}[(p_{n+1} - p_n) \cdot \epsilon_n | \epsilon_n = +1, \hat{\epsilon}_n], \\ \mathcal{R}^-(1) &= \mathbb{E}[(p_{n+1} - p_n) \cdot \epsilon_n | \epsilon_n = -1, \hat{\epsilon}_n], \end{aligned} \tag{2.28}$$

as a function of the sign predictor $\hat{\epsilon}_n$. A large absolute value of the predictor implies high predictability of the next order sign. The two curves indicate that if a buy market order is more likely ($\hat{\epsilon}_n > 0$) the market impact of a sell market order is larger than that of a buy market order, whereas when a sell order is more likely ($\hat{\epsilon}_n < 0$) the impact of a buy order is larger than that a sell order. This is a strong empirical proof of asymmetric liquidity. Furthermore, the two curves of $\mathcal{R}^+(1)$ and $\mathcal{R}^-(1)$ can be nearly approximated by linear functions of $\hat{\epsilon}_n$ with different slopes. This leads to the conclusion that we have to relax the implicit assumption of symmetry between buys and sells of Equation (2.20). Gerig (2007) has used a *private* information set (the membership code of each trade) to forecast order flow. We consider instead in the Chapter 3 an *anonymous* information set, where we will use only the past order flow to predict order signs.

An important question, which arises from the previous formalism, is what are the microstructural mechanisms responsible for asymmetric liquidity. In the next chapter we try to answer to this question, but in the literature there have been some attempts in this direction. For example, Lillo and Farmer (2004) showed that when the order flow becomes more predictable, the probability that a market order triggers a price change is larger for market orders with the unexpected sign than for those with the expected one. Moreover, the same authors showed that the ratio between the volume of the market order and the volume at the opposite best is lower (higher) for market orders with an expected (unexpected) sign. In the Section 3.3 we will deeply analyse this effect.

Another related basic mechanism is "stimulated refill": The imbalance in the order flow of market orders in one direction, trigger an opposing flow of limit orders of the opposite sign (Bouchaud et al., 2006). The result is the piling up of a liquidity wall, which decreases the probability of further upward moves of the price. This dynamical feedback between market orders and limit orders is therefore fundamental for the stability of markets and for enforcing efficiency.

We have understood that the interplay between liquidity takers and liquidity providers is crucial to reconcile correlation in order flow with the diffusive nature of price changes. If we start from the market ecology proposed before, we can go deeply in the analysis of this interplay. In fact, the previous market participants can be characterized by their trading frequencies. Medium to long term investors contribute to the presence of latent demand in the markets, which create long-term correlations in the sign of the trades. On the other hand, higher-frequency traders try to profit from short term predictability providing liquidity to lower-frequency traders and covering temporal mismatching between them. Therefore, there exist traders that await favourable conditions, in terms of both price and quantity, to be executed on the market by monitoring what happens in the market. A process whereby market orders trigger limit orders and limit orders attract market orders. In other words, optimized execution strategies that look for micro-opportunities impose strong correlations between market order flow of one sign and limit order flow of the opposite sign. In the next chapter we will discuss and propose a model which tries to capture this feedback mechanism ensuring market efficiency.

Chapter 3

The adaptive liquidity model

3.1 Introduction

As stated before, a well established property of financial markets is that the order flow, defined as the process assuming value one for buyer initiated trades and minus one for seller initiated trades, displays a very slowly decaying autocorrelation function (Lillo and Farmer, 2004; Bouchaud et al., 2004). Since a buyer initiated trade moves on average the price up and a seller initiated trade moves it down, one would naively expect that a correlated flow induces a correlated return time series. However this latter correlation is not observed in real data because it would allow to easily predict price movements, and therefore would provide arbitrage opportunity. Reconciling correlated order flow with uncorrelated price returns is therefore a subtle issue, which is subject of current research (see also Bouchaud et al., 2009, for a recent review). The autocorrelation of order flow is strictly connected with the fact that large trading volumes are typically fragmented in small trades and executed incrementally (see Tóth et al., 2015). In this way, investors are able to execute much of the large order, which is called *metaorder*, minimizing the price impact and the reveal of information on their trading activity.

A possible mechanism for efficiency is the *asymmetric liquidity* (Lillo and Farmer, 2004). Defining price impact of a trade as the difference between the log-price before the next trade and the price before the current trade, asymmetric liquidity states that the price impact of a type of order (buy or sell) is inversely related to the probability of its occurrence. This means that if at a certain point in time it is more likely that the next trade is a buy rather than a sell, a buyer initiated trade will have a smaller impact than a seller initiated trade. There is therefore a compensation between probability of an event and its effect on the price.

An empirical demonstration of the asymmetric liquidity is given in Figure 3.1. We indicate with ϵ_n the sign of the n-th trade, where $\epsilon_n = +1$ (-1) for a buyer (seller) initiated trade. Moreover r_n is the observed price impact due to the $n-$th trade (according to the definition above). We construct an autoregressive predictor

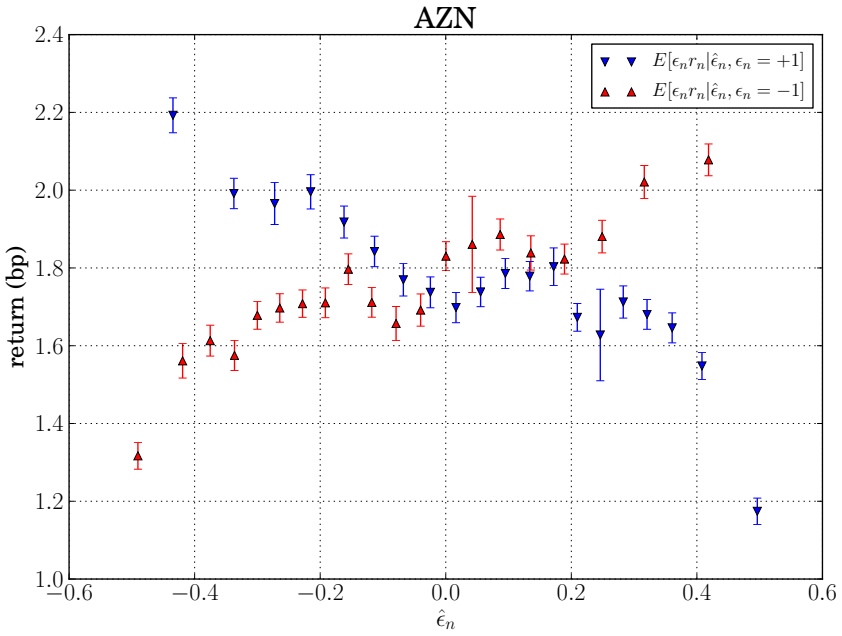

Figure 3.1: Expected value of the product of the tick by tick return times the sign of the triggering order as a function of the order sign predictor $\hat{\epsilon}_n$ for the asset Astrazeneca in 2004.

$\hat{\epsilon}_n = \mathbb{E}[\epsilon_n|\Omega_{n-1}, \mathcal{M}]$ of the order flow, where Ω_{n-1} and $\mathcal{M}$ are, respectively, the information set used and the particular model used to describe the order flow (see below for more details on the predictor). We compute the average signed stock return $\epsilon_n r_n$ conditional on the sign predictor $\hat{\epsilon}_n$ and on being triggered by a buy ($\epsilon_n = +1$) or sell ($\epsilon_n = -1$) initiated trades. The investigated stock is Astrazeneca traded on the London Stock Exchange in the whole year 2004. From the buyer initiated trades curve (blue triangles), we observe that when the next order is more likely to be a buy (essentially due to an excess of buys in the recent past), *i.e.* $\hat{\epsilon}_n > 0$, a buy trade moves on average the price less than a sell trade. The opposite occurs when the next order is more likely to be a sell ($\hat{\epsilon}_n < 0$). This is exactly what asymmetric liquidity prescribes and in this case the mechanism is at work even at the level of individual transactions.

The asymmetric liquidity mechanism is conceptually clear, but it does not give any indication about the microstructural mechanisms which are responsible of it. In other words, why does a highly predictable trade impact very marginally prices? There are several possible explanations, which can be for the sake of convenience classified into two categories: The first one includes those mechanisms due to the action of the initiators of the trade and the second one where the liquidity providers are responsible. In fact, in an electronic double auction market, the initiator of the trade (the *liquidity taker*) can control the volume of the market order initiating the trade. In this way she can decrease the probability that her order triggers a price change by using small volumes when her order sign is more predictable. On the other hand, other agents submitting and cancelling limit orders (the *liquidity providers*)

can control the price adjustment between two trades[1]. This can be done, for example by reverting, at least partly, the price when a predictable order arrives and moving the price in the same direction of the trade when its sign is unpredictable. In reality both types of agents are partly responsible of for asymmetric liquidity. A first contribution of this chapter is the empirical investigation of which microstructural mechanisms enforce efficiency of prices. We will present an extensive empirical analysis aimed at identifying the main contributions to asymmetric liquidity and therefore to price efficiency.

The persistence of the order flow leads to significant challenges also in the modeling of the order book dynamics. Order book modeling is a complex task, especially if one wants to take into account the strategic behaviour of economic agents. For this reason, in recent years there has been a growing interest toward the statistical modeling of order book. This type of modeling, pioneered by Daniels et al. (2003); Smith et al. (2003), drops agent rationality almost completely and describes the different types of orders as random variables. Although no one would dispute that agents in financial markets behave strategically, and that for some purposes taking this into account is essential, there are some problems where other factors might be more important. For example, this approach has the merit that can be calibrated and tested against real data (Farmer et al., 2005; Cont and De Larrard, 2013), because it presents simple quantitative laws that relate one set of market properties to another, placing restrictions on the allowed values of variables. The simplest class of these models are the so called "zero-intelligence" models, where one assumes that limit and market orders arrive randomly according to Poisson processes. Moreover queued limit orders are cancelled according to a Poisson process. To keep the model as simple as possible, there are equal rates for buying and selling, and all these processes are independent. The model just described is a prototypical queueing model of limit order book dynamics which consists in specifying the arrival rates of different types of order book events and the rules of execution of these orders. To the same class of models belong the Markovian queueing models discussed in (Cont et al., 2010; Cont and De Larrard, 2013).

However, all these modeling approaches neglect the persistence of the order flow, which destroys the Markovian feature of the modeling and leads to unrealistic behaviour and wrong predictions on the dynamics of prices. To be specific, we have calibrated a zero-intelligence model (Daniels et al., 2003) on the stock Astrazeneca in the whole year 2004. We have then replaced in the model the Poisson market order flow with the one extracted from the real data. We have then studied the diffusivity properties of the generated prices. To this end we computed the "signature plot" of the model, *i.e.* the plot of the quantity

$$\sigma(\ell) = \sqrt{\frac{\mathbb{E}[(p_{n+\ell} - p_n)^2]}{\ell}} \tag{3.1}$$

where the average $\mathbb{E}[(p_{n+\ell} - p_n)^2]$ is done over different instants of time t_n, which is

[1] In electronic markets the distinction between liquidity takers and providers is a bit artificial since most of the agents use a combination of limit and market orders. However, for exposition convenience we will stick to this terminology to indicate the two types of agents.

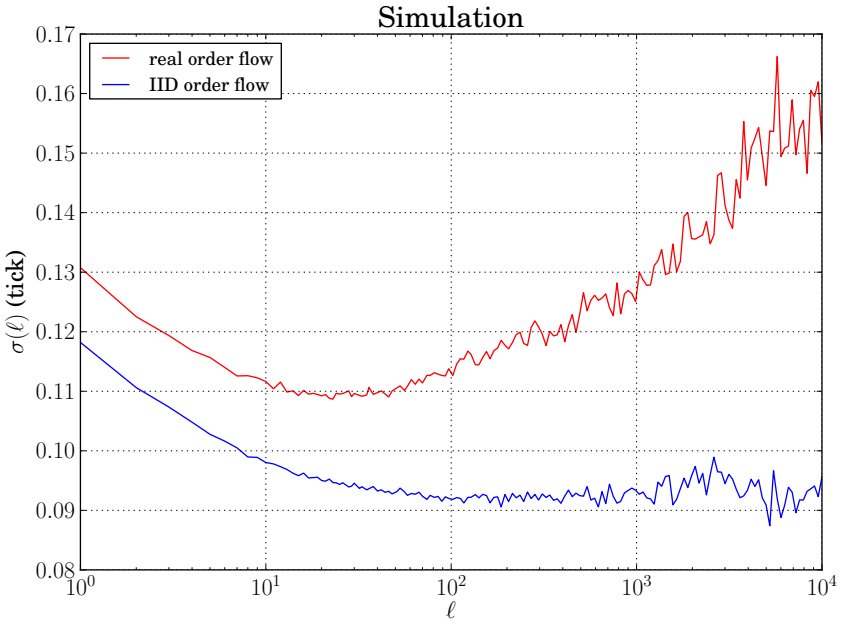

Figure 3.2: Signature plot (see Equation (3.1)) of a simulation of the model discussed in (Smith et al., 2003), for which the parameters are calibrated on Astrazeneca in the whole year 2004. We used the real market order flow as input of the model (red line) and an IID order flow (blue line)

the time that immediately precedes the n-th transaction. This quantity is a measure of the volatility of the price process on time scale ℓ. For a purely diffusive process, $\sigma(\ell) = D$ is constant and independent of ℓ. If $\sigma(\ell)$ decays when ℓ increases, the price motion is "sub-diffusive" and it has a mean-reverting behaviour (there exist negative correlations between lagged returns). On the other hand, if $\sigma(\ell)$ increases when ℓ increases, the price process is super-diffusive and it shows a trending behaviour (positive correlations between lagged returns). Thus, a necessary condition for price efficiency is that $\sigma(\ell) = D$ is constant. It is well known that real price time series show a sub-diffusive behaviour for very short lags, and then the price is diffusive at the other lags.

In Figure 3.2 we present the result of the above described Monte Carlo simulation. If the real order flow is used, we observe that price is initially sub-diffusive and for lags larger than ~ 30 trades it becomes super-diffusive. It is evident that embedding a persistent order flow in the framework developed in (Smith et al., 2003) induces a super-diffusive behaviour of prices. An analogous pattern should be expected from any Markovian model. Recent attempts of modeling the limit order book with strongly persistent order flow include (Mike and Farmer, 2008; Tóth et al., 2011; Mastromatteo et al., 2014). In these papers, however, either the diffusivity is not guaranteed (Mike and Farmer, 2008), or it is attained by fine tuning the value of a parameter describing the counterbalancing reaction to the order flow persistence Tóth et al. (2011). More importantly, in this last case (detailed below) diffusivity is recovered up to the time scale of the lifetime of limit orders, while for longer time scales the price becomes super-diffusive.

In the second part of this chapter we propose a new statistical model of the limit order book which is able to give diffusive prices at all time scales and to reproduce the empirical statistical properties observed to explain the underlying mechanisms of the asymmetric liquidity. The key ingredient of the modeling is a liquidity dynamics that adapts itself to the degree of predictability of the order flow. In other words, instead of having a fine tuning of a parameter that guarantees (approximate) diffusivity, we model liquidity as an adaptive process that responds to the local predictability of the order flow and gives exact diffusivity.

3.2 Dataset description

The data used in our empirical analysis belong to two distinct datasets spanning different time periods and recorded on different markets. The first dataset corresponds to the trading activity of two stocks traded on the London Stock Exchange (LSE) during the whole year 2004. The second one is more recent and records the activity of two stocks traded on the NASDAQ stock exchange in New York. This dataset covers only a short period of time, namely July and August 2009, but the higher trading frequency partially compensates for the shorter horizon.

The LSE dataset includes the limit order book information about the Astrazeneca (AZN) and Vodafone (VOD) stocks. The data come from the Stock Exchange electronic Trading Service (SETS), the LSE's flagship electronic order book, and contain the detailed description of all order book events (submissions of limit and market orders and cancellations of outstanding orders) which occurred in the whole year of 2004 (254 trading days). In particular, the information concerning the market order events report the execution time of the event, the sign of the order (*i.e.* if it is buyer or seller initiated), the traded volume and price. We select AZN and VOD because of the sensible difference in the discretization of the prices. AZN has a tick size-price ratio of few basis points, whereas VOD is characterized by a very large tick size-price ratio (see Table 3.1). For this reason, we refer to the former as a small-tick stock, while to the latter as a large-tick stock.

The second dataset includes all the executed trades and order book updates of stocks traded at the NASDAQ market in New York. In particular, we analyse two liquid stocks, namely a small-tick stock, Apple (AAPL), and, a large-tick stock (relatively to AAPL), Microsoft (MSFT). The data cover 42 days of trading activity during July and August of 2009. For the two datasets, we have taken care of the possibility that the execution of a single market order hitting several existing limit orders produces many records with the same timestamp. We have aggregated them in a single market order, whose volume is the cumulative volume of the components. A summary of the properties of the four stocks is collected in Table 3.1.

The empirical analysis has been performed using a code written in the *Python* programming language. Specifically, we have used the scientific *SciPy* libraries, the statistical library *StatsModels*, while all the graphs have been generated by the

Symbol	Year	Number of trades	Average intertrade time	Average stock price	Average tick size-price ratio
AAPL	2009	857,925	1.1 s	157.17 USD	0.6 bp
MSFT		575,040	1.7 s	23.74 USD	4.2 bp
AZN	2004	405,481	23.1 s	24.38 GBP	4.1 bp
VOD		411,736	22.9 s	1.34 GBP	18.7 bp

Table 3.1: Summary of the investigated stocks. The average stock price is expressed in U.S. Dollars for AAPL and MFST, whereas it is expressed in Great Britain Pounds for AZN and VOD. The average intertrade time and tick size-price ratio are given in seconds and in basis points, respectively.

plotting library *Matplotlib*.

3.3 Empirical evidences of the origin of asymmetric liquidity

In this section we investigate empirically the mechanisms responsible for restoring efficiency and, as a consequence, diffusivity of prices. More specifically, we perform an empirical analysis in order to investigate the origin of the asymmetric liquidity mechanism. We consider the variables of the order book defined in Section 1.3 at the instant of time t_n which immediately precedes the n-th transaction. In particular, the best ask price A_n and the corresponding log-price $a_n = \log A_n$, the best bid price B_n and the corresponding log-price $b_n = \log B_n$, the midpoint price $P_n = (A_n + B_n)/2$ and the log-midprice $p_n = \log P_n$; the ask gap $g_n^A = a_n^{2nd} - a_n$, the bid gap $g_n^B = b_n - b_n^{2nd}$, the shares available at the best ask v_n^A, and the shares available at the best bid v_n^B.

3.3.1 Predictability of market order flow

Given the crucial role played by the order flow and following (Lillo and Farmer, 2004) we introduce the *sign predictor* and study the variables characterizing the state of the order book conditioned on its value. Assuming a model for the order flow process, we can compute at each time t_{n-1} the expected value of the future market order sign, $\hat{\epsilon}_n = \mathbb{E}[\epsilon_n|\Omega_{n-1}, \mathcal{M}]$, conditional to the information set Ω_{n-1} (typically the past order flow) and the particular model $\mathcal{M}$ used to describe the order flow. In early works, the order flow was modelled in terms of a real valued autoregressive process, but clearly the order flow takes only discrete values. Thus, we model it by means of a p-th order Discrete Autoregressive process (*DAR(p)*), which is an integer valued process easy to calibrate on real data. The *DAR(p)* was introduced in a series of papers (Jacobs and Lewis, 1978, 1983) and describes a sequence of stationary discrete random variables with the properties of a Markov process of order p. In the following section, we define in more details the process.

3.3.2 A predictor for the order flow sign: the *DAR(p)* model

The *DAR(p)* model defines a general class of simple models for discrete variate time series $\{X_n\}$. It generates a sequence of stationary discrete random variables with the properties of a p-th order Markov process. These properties are reflected by the fact that the distribution of X_n only depends on $\Omega_{n-1} = \{X_{n-1}, \ldots, X_{n-p}\}$. The process is specified by the stationary marginal distribution of X_n and by the correlation structure of the sequence.

The p-th order discrete autoregressive model *DAR(p)* is given by

$$X_n = V_n X_{n-A_n} + (1 - V_n) Z_n \,. \tag{3.2}$$

The sequence $\{Z_n\}_{\mathbb{N}_m}$ is composed by IID random values drawn by a marginal distribution Ξ, whose sample space is a subset of the integers $\mathbb{N}_m$, where m is the cardinality of the sample space. Furthermore, $\{V_n\}$ is a sequence of IID random values following a Bernoulli distribution $\mathcal{B}(1, \varphi)$. Therefore we have

$$\mathbb{P}(V_n = 1) = 1 - \mathbb{P}(V_n = 0) = \varphi, \qquad \text{with } 0 \leq \varphi < 1 \,.$$

Finally, $\{A_n\}$ is a sequence of IID random values drawn from a multinomial distribution $\mathcal{M}(1, \lambda)$, with states $\{1, 2, \ldots p\}$ and probabilities

$$\mathbb{P}(A_n = k) = \lambda_k \geq 0, \qquad k \in \{1, 2, \ldots p\} \,,$$

where the parameter vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ is normalised to unity, $\sum_{k=1}^{p} \lambda_k = 1$.

Let us explain in a less formal way the *DAR(p)* process of Equation (3.2): The value X_n is either taken from the history of $\{X_n\}$ (with probability φ) or drawn randomly from the Ξ distribution (with probability $1 - \varphi$). The random values V_n have the function of a switch between the two cases. In the case of a positive Bernoulli trial ($V_n = 1$), X_n is determined by moving A_n steps back in the past observations of $\{X_n\}$, with A_n assuming values in $\{1, 2, \ldots p\}$ with probability given by the parameter vector $\lambda = (\lambda_1, \ldots, \lambda_p)$. Therefore, with probability $\varphi\lambda_i$, $X_n = X_{n-i}$, for $i = 1, 2, \ldots, p$. In the second case, when $V_n = 0$, $X_n = Z_n$ is drawn randomly from the specific marginal distribution Ξ, which has a discrete state space.

It is possible to select an initial distribution of X_0 which yields a stationary sequence $\{X_n\}_{\mathbb{N}_m}$ with marginal distribution Ξ, and it is possible to prove that this initial distribution coincides with the marginal distribution Ξ.

This procedure differs essentially from the representation of a Markov process as a matrix of transition probabilities, where the number of parameters to estimate is $m^{p+1} - m$. Here a smaller number $p + 1$ of effective parameters allows a better control of the statistical properties of the sequence, while in the case of a transition probability matrix one has many independent parameters, each of which regulates only a minor aspect of the process.

In Appendix 3.A we review the auto-covariance structure of *DAR(p)* process and how to forecast it.

Symbol	MSE DAR(p)			$1 + \mathbb{E}[\hat{\epsilon}_n^2]$ DAR(p)		
	$p = 100$	$p = 500$	$p = 700$	$p = 100$	$p = 500$	$p = 700$
AAPL	0.7692 ± 0.0009	0.7686 ± 0.0009	0.7684 ± 0.0009	1.2308	1.2313	1.2315
MSFT	0.5660 ± 0.0012	0.5651 ± 0.0012	0.5649 ± 0.0012	1.4336	1.4344	1.4346
AZN	0.9332 ± 0.0008	0.9321 ± 0.0008	0.9317 ± 0.0008	1.0667	1.0678	1.0683
VOD	0.8722 ± 0.0010	0.8709 ± 0.0010	0.8705 ± 0.0010	1.1198	1.1211	1.1215

Table 3.2: MSE values and standard errors for AAPL, MSFT, AZN, VOD, and for three different values $p = 100, 500, 700$. Last three columns: upper bound for MSE in case of absence of predictability.

3.3.3 Analysis of the order flow with the *DAR(p)* model

The first step in our analysis is the estimation of the *DAR(p)* model by using the observed order flow ϵ_n. When performing estimation we discard the first p trades of each trading day to avoid spurious overnight effect. The idea is that a sign predictor obtained with the order flow partially observed the previous trading day is less significant than a sign predictor computed with the order flow which belongs entirely to the same trading day. Therefore, this procedure involves the estimation of the model for each trading day.

In order to evaluate the predictability of order flow, we rely on the sign predictor defined by Equation (3.16) in Appendix 3.A, $\hat{\epsilon}_n = \mathbb{E}[\epsilon_n | \Omega_{n-1}, \mathrm{DAR(p)}]$, and specify a loss function indicating how much our prediction is correct. We employ the Mean Square Error (MSE) defined as

$$\mathrm{MSE}(\hat{\epsilon}_n) = \mathbb{E}\left[(\epsilon_n - \hat{\epsilon}_n)^2\right] . \tag{3.3}$$

This function has an upper bound equal to $\mathrm{MSE}(\hat{\epsilon}_n) = 4$, when the prediction is always wrong, and a lower bound $\mathrm{MSE}(\hat{\epsilon}_n) = 0$ when the prediction is systematically correct. When $\mathbb{E}[\epsilon_n \cdot \hat{\epsilon}_n] = 0$, the MSE is equal to $1 + \mathbb{E}[\hat{\epsilon}_n^2]$ and this value is the upper bound in case of no predictability of the model.

In Table 3.2 we list the MSE values computed for AAPL, MSFT, AZN, VOD, and three different values $p = 100, 500, 700$. As anticipated, we see that the MSE values obtained for each stock are almost independent from the order p of the auto-regression[2]. Thus, for the rest of the analysis we fix it equal to 500.

We notice that the MSE values of small-tick stocks are always higher than the values of large-tick stocks within the same dataset. More interestingly, there are substantial differences between the MSE of the stocks belonging to the LSE and NASDAQ datasets. The MSE for AAPL and MSFT are smaller than those of AZN and VOD, which are closer to the value $1 + \mathbb{E}[\hat{\epsilon}_n^2]$. We also compute the lagged sign predictor $\hat{\epsilon}_{n+s} = \mathbb{E}[\epsilon_{n+s} | \Omega_{n-1}, \mathrm{DAR}(500)]$ for $s \geq 0$ and in Figure 3.3 we show the distributions of the sign predictor values corresponding to $s = 0, 3, 10$ trades. Two distinct regions characterize the predictor distributions: The region where the values of the sign predictor are close to the extrema of the support, and the region where the

[2]This might depend on the choice of MSE as a loss function.

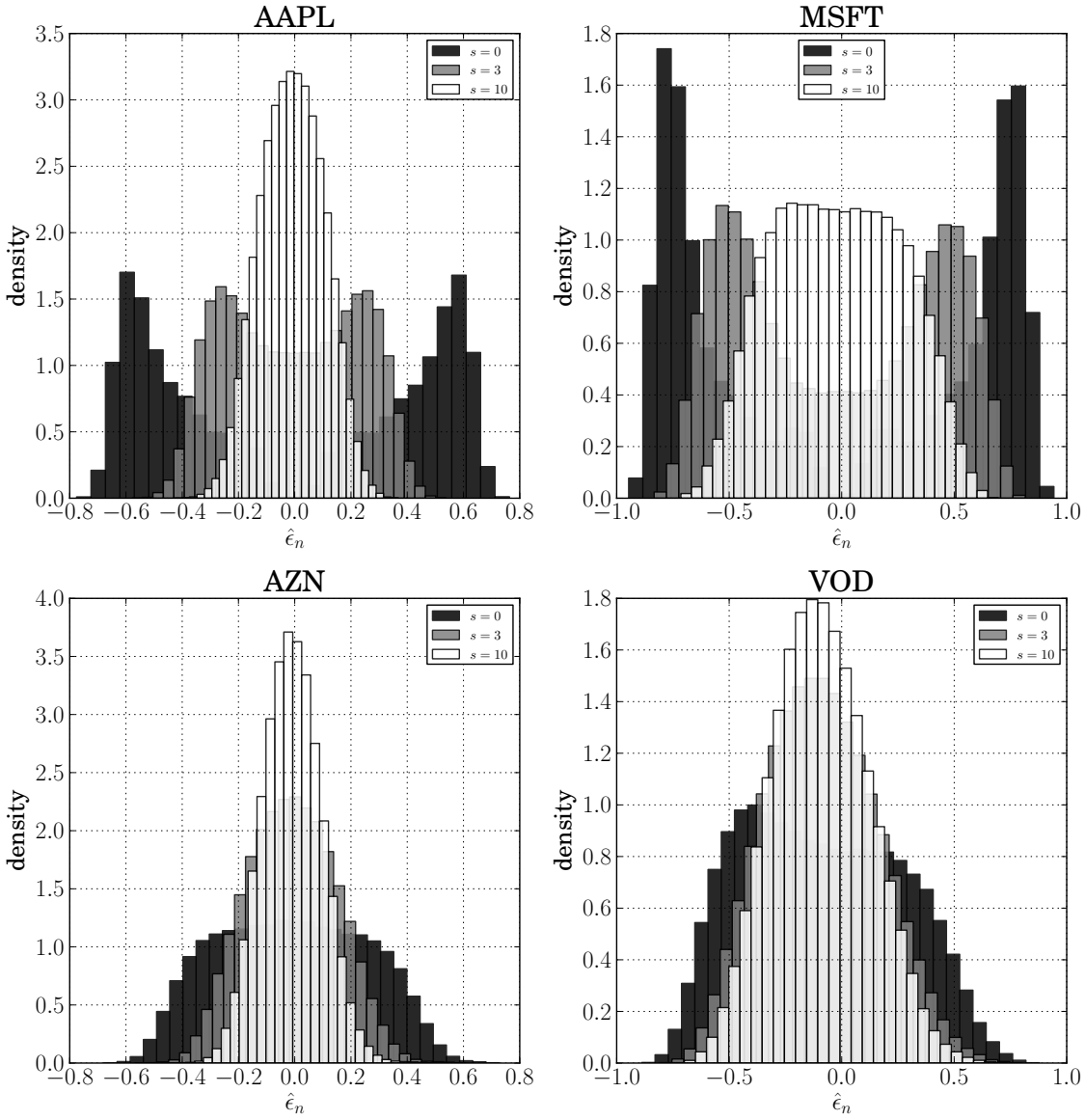

Figure 3.3: Distributions of the sign predictor for the stocks AAPL, MSFT, AZN, VOD and $s = 0, 3, 10$ trades.

predictor is close to zero. The former is the high predictability region and indicates that the corresponding market operates in a high predictable regime. This is the case for the NASDAQ dataset for $s = 0$ and the effect is more intense for the large-tick stock MSFT than for AAPL. The latter is the low predictability region where typically $\hat{\epsilon}_{n+s} \approx 0$. The LSE market for both AZN and VOD operates under this regime confirming the previous findings about the MSE values. It is worth noticing that when the value of s increases from 0 to 10, the predictor distribution converges to the low predictability region also for the assets belonging to the NASDAQ dataset. This convergence is faster for the small-tick asset AAPL than for MSFT, thus we expect that possible divergences from an efficient behaviour should be more evident for large-tick stocks.

3.3.4 Best bid and ask volume conditional expectation

Equipped with the order sign predictor, we measure the order book state variable at the instant of time which immediately precedes the n-th transaction (t_n^-) conditional on the predictor value $\hat{\epsilon}_n$. We use the *DAR(500)* model to construct a sign predictor and we split the range of $\hat{\epsilon}_n$ into a finite number of bins. We do not evenly sample the bins, but we fix the bins according to the empirical quantiles requiring that the number of empirical sign predictors falling within each bin is the same.

The first quantity that we consider is the volume outstanding at the best quotes. The best bid and ask volumes are natural indicators of the liquidity available on each side of the order book. We condition the volume at the best ask v_n^A and the volume at the best bid v_n^B on the level of the sign predictor, and we take their conditional expectation, $\mathbb{E}[v_n^A|\hat{\epsilon}_n]$ and $\mathbb{E}[v_n^B|\hat{\epsilon}_n]$. In Figure 3.4 we show the conditional average of the best volumes as a function of $\hat{\epsilon}_n$ for the assets AAPL, MSFT, AZN, and VOD.

We start commenting on the stocks which belong to the LSE dataset. We recall from Figure 3.3 that for AZN and VOD the sign predictor is mainly distributed in the low predictability region. We focus on the behaviour of the volumes at the best ask, for those at the bid side similar comments apply. When buy orders are more likely than sell orders ($\hat{\epsilon}_n > 0$) the average volume outstanding at the ask side is smaller than the volume outstanding at the bid side. Moreover, when the sign predictor increases, the best ask volumes decrease and the best bid volumes increase. This behaviour is compatible with a model where liquidity takers mechanically erode the liquidity available at the opposite side of the book. Indeed, a positive sign predictor means that the recent order flow has been dominated by a sequence of buy orders and the volume outstanding at ask side of the book has been eroded by market orders. However, when the predictor is approaching the upper bound ($\hat{\epsilon}_n = 1$), the volume at the ask side starts to increase. Indeed, high predictability of the order flow means that significant information about the intentions of the liquidity taker has been released to the market and a large fraction of her metaorder has been executed. Thus, the probability that the metaorder is close to expiration is high and it becomes pressing for liquidity providers to refill the ask side of the order book at the best price. For the LSE dataset both the large-tick and the small-tick assets manifest

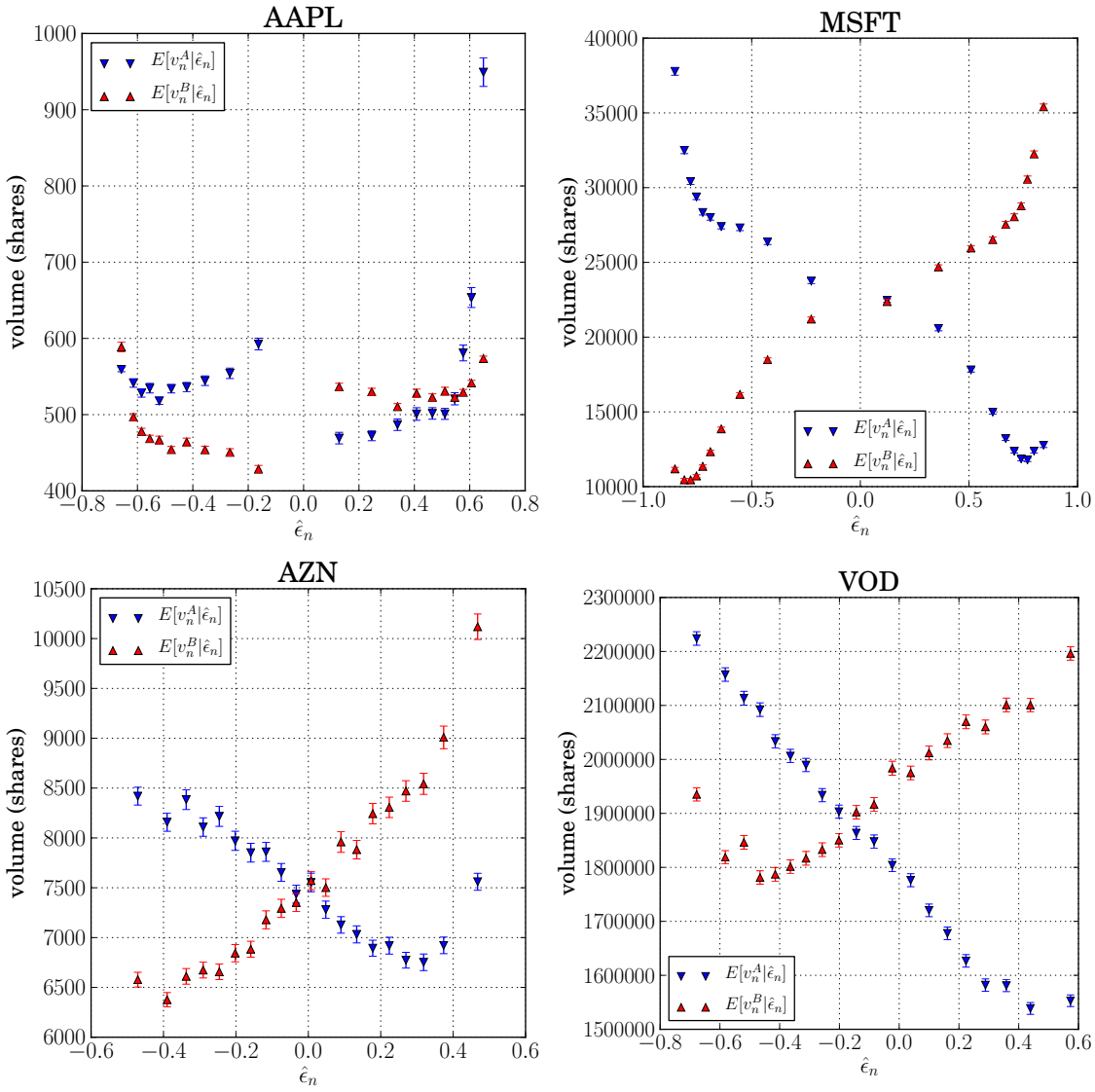

Figure 3.4: Conditional best ask volumes $\mathbb{E}[v_n^A|\hat{\epsilon}_n]$ and conditional best bid volumes $\mathbb{E}[v_n^B|\hat{\epsilon}_n]$ on different sign predictor values, for four stocks (AAPL, MSFT, AZN, VOD). The error bars are standard errors.

the same behaviour. When we switch to the NASDAQ dataset the picture is less clear. While the large-tick asset MSFT follows the same pattern of the LSE assets, for AAPL the situation is different. When buy orders are very likely ($\hat{\epsilon}_n \approx 1$) the volume refill of the liquidity providers dominate and the average outstanding volume increases with $\hat{\epsilon}_n$. However, since we know that the sign predictor is concentrated in the high predictable region conclusions about the behaviour of the average volume in the central region are less definitive.

Finally, we notice that the volumes at the best quotes are higher in large-tick stocks than in small-tick stocks, not only in the number of shares but also in dollar (pound) value. In fact, if we multiply the average volume by the average price, we find that there is a difference of one (two) order of magnitude between large-tick and small-tick stocks of the NASDAQ (LSE) dataset. This difference is likely caused by the discretization effect of limit prices since liquidity providers suffer less availability of quotes in large-tick than in small tick stocks and they pile up volumes at the same price.

3.3.5 Bid and ask gap conditional expectation

Following the same procedure described in the previous section, we consider the conditional distribution of the bid (ask) logarithmic gap between the best bid price and the second best quote (the best ask price and the second best ask quote) immediately before the transaction time t_n. As before we compute the expectation of these quantities conditioning on the level of the sign predictor, *i.e.* we compute $\mathbb{E}[g_n^A|\hat{\epsilon}_n]$ and $\mathbb{E}[g_n^B|\hat{\epsilon}_n]$.

In Figure 3.5 we plot the conditional mean of the bid and ask gap as a function of the sign predictor for the stocks AAPL, MSFT, AZN, and VOD. We observe that for large-tick stocks, independently of the dataset, the bid and ask gap are approximately constant and equal to one tick for all sign predictor values. This is largely expected and is due to the high level of discretization of limit prices. For small-tick stocks the bid gap is larger than the ask gap when sell orders are more likely ($\hat{\epsilon}_n < 0$), whereas the ask gap is larger than the bid gap when buy orders are more likely ($\hat{\epsilon}_n > 0$). The slope of the curves strongly change if we move from the LSE asset to the NASDAQ asset. For AZN the bid gap monotonically decreases and the ask gap monotonically increases when the sign predictor value increases. For AAPL when the sign predictor increases and a buy order is more probable the ask gap decreases, whereas when the sign predictor goes from zero to the minimum value the ask gap increases. The opposite behaviour holds for the bid gaps.

In conclusion, for large-tick assets the conditional distribution of the gaps is not informative. Conversely, for small-tick assets figures show an interesting behaviour. If buy orders are very likely at a given time it means that many buy orders have taken place in the recent past and they have eroded liquidity and increased the sparsity of the ask side of the book. Therefore, the slope of the gap distribution for AZN could be consistent with a purely mechanical effect due to the erosion of the

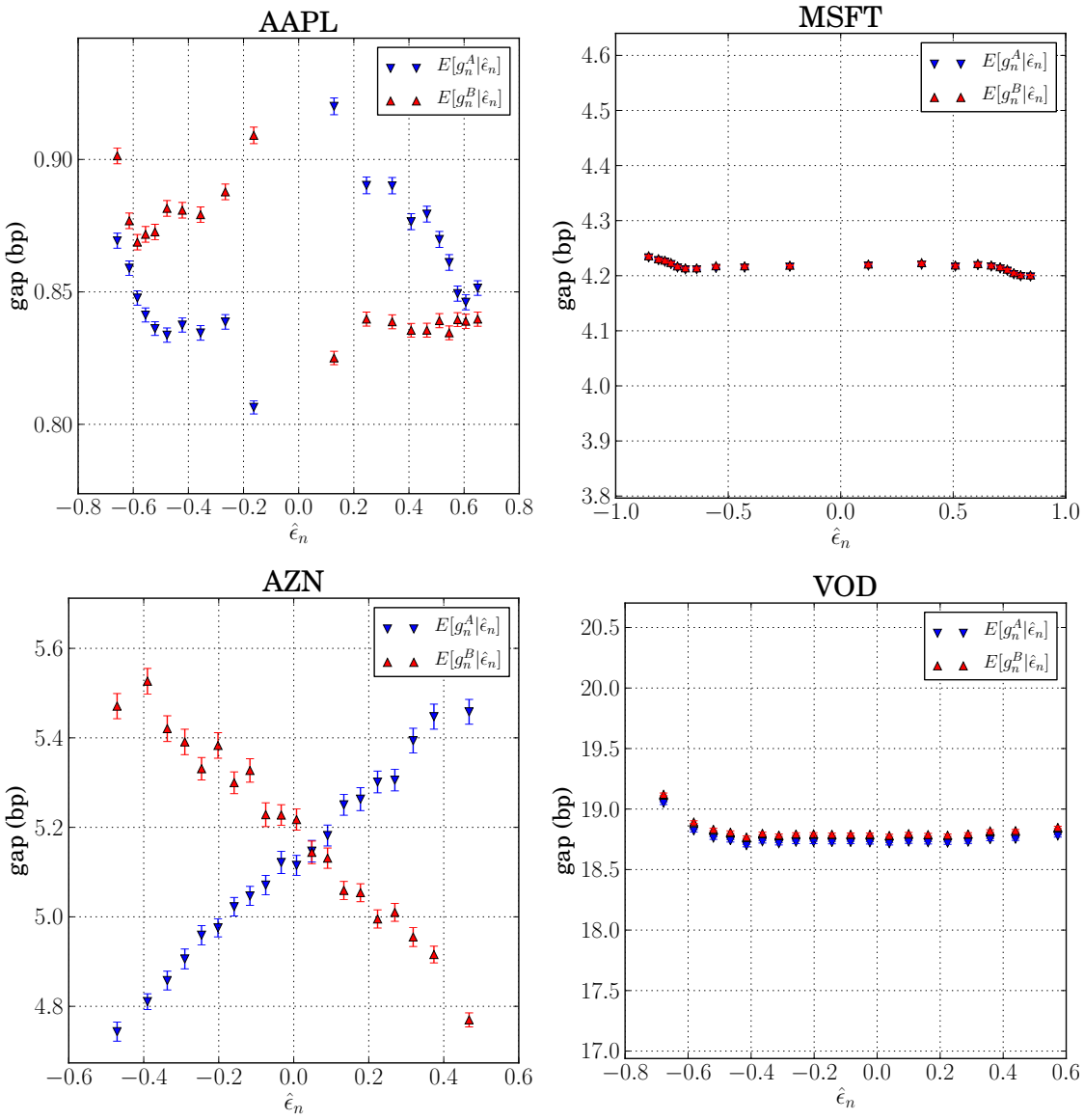

Figure 3.5: Conditional ask gap $\mathbb{E}[g_n^A|\hat{\epsilon}_n]$ and conditional bid gap $\mathbb{E}[g_n^B|\hat{\epsilon}_n]$ as a function of the sign predictor for the four stocks (AAPL, MSFT, AZN, VOD). The error bars are standard errors.

market orders. However, this explanation neglects the possible presence of liquidity providers refilling the order book. Moreover, as it will be clarified in the next section, liquidity takers adjust their trades in order to minimize the price impact, they do not penetrate the opposite side of the order book, and thus the impact of the erosion can not be the only mechanism which determines sparsity of the order book. Finally, the slope of the curves for AAPL cannot be explained without considering the interplay with market makers. Indeed, the negative slope of the curve for the ask gap when predictability increases suggests that the extreme probability of a buy order stimulates the liquidity providers to refill the ask side of the book. For AAPL the figure is consistent with a refill taking place not only at the opposite best, as already confirmed by the volume curve, but also at quotes inside the order book and close to the best price.

3.3.6 Mechanical and quote revision impact

We now ask how market orders, limits and cancellations determine the price impact. We define the returns as the difference of the logarithmic mid-prices measured immediately before the n-th and the $n+1$-th trades, $r_n = p_{n+1} - p_n$, and we decompose them in two components. The first component is due to the mechanical impact of market orders, r_n^M, and is given by the difference between the log-price observed immediately after and the one observed before the trade. The second component is the aggregate effect of the quote revision r_n^Q and cumulates the effect of all the limit orders and cancellations placed in the order book immediately after the n-th trade and before the next trade. Thus, we have

$$r_n = r_n^M + r_n^Q.$$

Then, we introduce the quantity $\epsilon_n \cdot \hat{\epsilon}_n$ which measures the correctness of a prediction at a given trade time t_n and quantifies the surprise of the transaction sign given the level of the predictor. The former information is delivered by the sign of $\epsilon_n \cdot \hat{\epsilon}_n$ since when $\epsilon_n \cdot \hat{\epsilon}_n$ is positive we can conclude that the prediction was correct, whereas when $\epsilon_n \cdot \hat{\epsilon}_n$ is negative the prediction was wrong. The amount of surprise associated to the realized order sign is instead related to the absolute value of $\epsilon_n \cdot \hat{\epsilon}_n$. For instance, a large negative value of $\epsilon_n \cdot \hat{\epsilon}_n$ is more informative than a negative value close to zero since it implies that the order sign was largely unexpected by the market. We are interested in the conditional expectation of the return components

$$\mathbb{E}[\epsilon_n r_n^M | \epsilon_n \cdot \hat{\epsilon}_n], \qquad \mathbb{E}[\epsilon_n r_n^Q | \epsilon_n \cdot \hat{\epsilon}_n],$$

which show how the correctness of the sign prediction determines the mechanical and quote revision impact, respectively.

The first term is the conditional expectation of the mechanical impact and depends on the probability of an order to penetrate the opposite best price and on the distribution of the gaps on the opposite side of the book, g_n^{OB}. This expectation

satisfies the approximate relation

$$\begin{aligned}
\mathbb{E}[\epsilon_n r_n^M|\epsilon_n\cdot\hat{\epsilon}_n] &\simeq \sum_{r_n^M\neq 0}\epsilon_n r_n^M\mathbb{P}(v_n\geqslant v_n^{OB},\epsilon_n g_n^{OB}\simeq 2r_n^M|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)\\
&= \sum_{r_n^M\neq 0}\epsilon_n r_n^M\mathbb{P}(v_n\geqslant v_n^{OB}|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)\\
&\qquad\qquad\cdot\mathbb{P}(\epsilon_n g_n^{OB}\simeq 2r_n^M|v_n\geqslant v_n^{OB},\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)\\
&\simeq \mathbb{P}(v_n\geqslant v_n^A|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v,\epsilon_n=1)\mathbb{E}\left[\frac{g_n^A}{4}\middle| v_n\geqslant v_n^A,\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v,\epsilon_n=1\right]\\
&+\mathbb{P}(v_n\geqslant v_n^B|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v,\epsilon_n=-1)\mathbb{E}\left[\frac{g_n^B}{4}\middle| v_n\geqslant v_n^B,\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v,\epsilon_n=-1\right]\\
&\equiv r_n^M(\hat{\epsilon}_n)\,, \qquad (3.4)
\end{aligned}$$

where v_n^{OB} and v_n^{OB-2nd} are the opposite best and second opposite best volumes, respectively, v_n is the volume of the market order, and $\mathcal{C}_n^v$ corresponds to the condition $v_n < v_n^{OB}+v_n^{OB-2nd}$.

The approximate equality in the first line is due to two distinct effects. First, we neglect the possibility that the market order volume is greater than the sum of the best and second opposite best volumes. In real datasets this condition is verified for the large majority of the transactions and only in a very small fraction of trades ($< 1\%$) a market order penetrates the opposite side of the book deeper than the first price level (see also Farmer et al., 2004). This is in particular true for large-tick assets. Then, we assume that $\epsilon_n g_n^{OB}\simeq 2r_n^M$ when $v_n\geqslant v_n^{OB}$, which is exactly true for linear gaps and linear returns and holds only approximately for logarithmic quantities. For instance, for buy orders the relation between log-gaps and mechanical log-returns is given by

$$\begin{aligned}
r_n^M &= \log\left(A_n+B_n+A_n^{2nd}-A_n\right)-\log\left(A_n+B_n\right)=\log\left(1+\frac{A_n^{2nd}-A_n}{A_n+B_n}\right)\\
&\approx\frac{A_n^{2nd}-A_n}{A_n+B_n}\approx\frac{g_n^A}{1+B_n/A_n}\approx\frac{g_n^A}{2}\,,
\end{aligned}$$

where $B_n/A_n\approx 1$. Finally, the approximation in the third line of Equation (3.4) follows from the realistic assumption that $\mathbb{P}(\epsilon_n=1|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)=\mathbb{P}(\epsilon_n=-1|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)\simeq 1/2$. The quantity $\mathbb{P}(v_n\geqslant v_n^{OB}|\epsilon_n\cdot\hat{\epsilon}_n,\mathcal{C}_n^v)$ corresponds to the conditional probability that the volume of a market order is larger than the liquidity available at the opposite best. Thus, it represents the probability that a market order immediately triggers a mid-price change. We estimate the penetration probability on the real datasets and condition it on the correctness of the order sign predictor $\epsilon_n\cdot\hat{\epsilon}_n$, but we remove the mild conditioning $\mathcal{C}_n^v$. Starting from the quantities

$$\mathbb{P}(v_n\geqslant v_n^A|\epsilon_n\cdot\hat{\epsilon}_n,\epsilon_n=+1),\qquad \mathbb{P}(v_n\geqslant v_n^B|\epsilon_n\cdot\hat{\epsilon}_n,\epsilon_n=-1)\,,$$

we express the total penetration probability of market orders as

$$\mathbb{P}(v_n\geqslant v_n^{OB}|\epsilon_n\cdot\hat{\epsilon}_n)\simeq\frac{1}{2}\left[\mathbb{P}(v_n\geqslant v_n^A|\epsilon_n\cdot\hat{\epsilon}_n,\epsilon_n=1)+\mathbb{P}(v_n\geqslant v_n^B|\epsilon_n\cdot\hat{\epsilon}_n,\epsilon_n=-1)\right]\,,$$

where we have assumed that $\mathbb{P}(\epsilon_n = +1|\epsilon_n \cdot \hat{\epsilon}_n) = \mathbb{P}(\epsilon_n = -1|\epsilon_n \cdot \hat{\epsilon}_n) \simeq 1/2$. We also compute the conditional average fraction of liquidity eroded by a market order $\mathbb{E}[f|\epsilon_n \cdot \hat{\epsilon}_n]$ with $f = v_n/v_n^{OB}$.

Figure 3.6 shows the conditional penetration and fraction for AAPL, MSFT, AZN and VOD. We see that for all the stocks the eroded fraction tracks quite well the behaviour of the penetration probability, and we observe the largest discrepancies for the NASDAQ assets in a region $\epsilon_n \cdot \hat{\epsilon}_n < 0$ which is scarcely populated. For the LSE dataset the penetration probability is consistent with previous findings in the literature (Lillo and Farmer, 2004), *i.e.* when the order sign predictability increases and the prediction is correct, the probability of penetration drops. The stocks of the NASDAQ dataset (AAPL, MSFT) show deviations from a monotonic behaviour. MSFT shows an increasing penetration probability when the order sign predictability increases and the prediction is correct up to $\epsilon_n \cdot \hat{\epsilon}_n \simeq 0.7$ then drops, whereas for AAPL deviations from a decreasing behaviour are relevant in the region where $\epsilon_n \cdot \hat{\epsilon}_n \lesssim 0.3$. This effect leads to inefficiencies of the market that we will comment about more extensively in the next section. Finally, as expected the penetration probability of large-tick stocks (MSFT,VOD) is lower than the probability of small-tick stocks (AAPL, MSFT). Indeed, for large-tick stocks market orders eroding the opposite liquidity are less frequent since they trigger a large impact on the mid-price.

In Figure 3.7 we examine in more details the empirical behaviour of the opposite best volume and market order volume whose ratio leads to the fraction of eroded liquidity. For MSFT, AZN, and VOD two aspects are common: The conditional average market order volume decreases with $\epsilon_n \cdot \hat{\epsilon}_n$. The second striking feature is the behaviour of the average amount of volume available at the opposite side. It decreases when the correctness increases, then, for $\epsilon_n \cdot \hat{\epsilon}_n \approx 1$ it increases quickly. Thus, up to moderate values of $\epsilon_n \cdot \hat{\epsilon}_n$ liquidity is removed from the opposite best either because of a mechanical erosion or because liquidity providers revise their limit orders. Then, finally, the high predictability stimulates the liquidity refill, a liquidity barrier piles up at the opposite best and the penetration probability drops. For AAPL the conditional average market order volume is independent from the predictability of the order flow. However, there are clear signs of the liquidity refill effect. We can therefore interpret these effects concluding that liquidity takers adapt their orders to the outstanding liquidity only when correctness is not too high, because in the extreme region of predictability liquidity takers shrink the volume of their markets orders, though the available volume at the opposite side is high.

In light of above considerations about outstanding and market volumes, and gap distributions, we can now discuss the observed behaviour of the price impact. In Figure 3.8 we show the conditional mechanical impact r_n^M, the approximate expression $r_n^M(\hat{\epsilon}_n)$ derived in Equation (3.4), and the whole impact r_n conditioned on $\epsilon_n \cdot \hat{\epsilon}_n$ for AAPL, MSFT, AZN, and VOD.

From this figure we notice that the approximate quantity $r_n^M(\hat{\epsilon}_n)$ reproduces extremely well the mechanical impact. This evidence supports the idea that the mechanical component of the impact is mainly determined by the gap distribution and by the penetration probability of a market order. The second relevant message

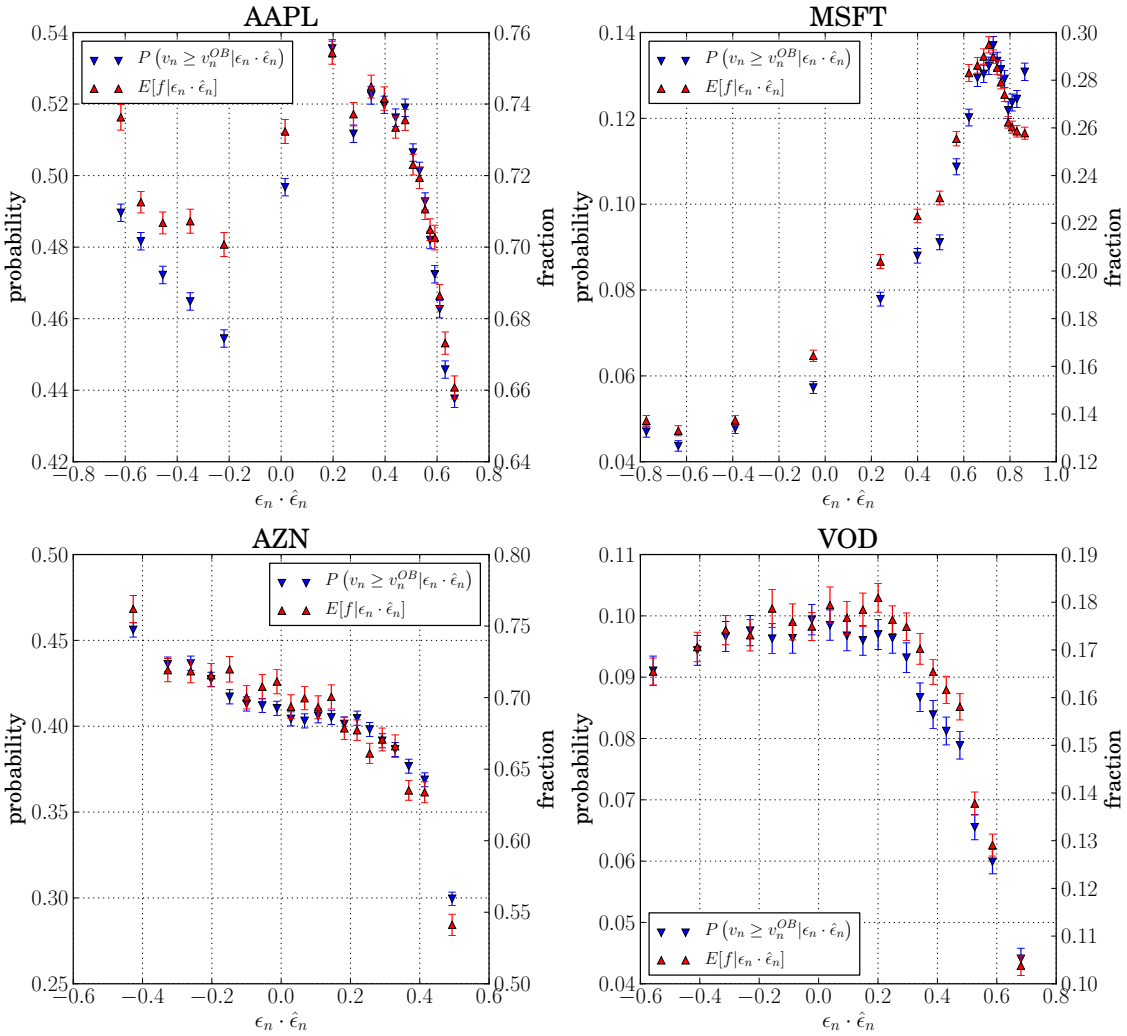

Figure 3.6: Conditional penetration probabilities of the market orders and conditional average ratio between market order volumes and best opposite volumes for AAPL, MSFT, AZN, VOD as a function of $\epsilon_n \cdot \hat{\epsilon}_n$. The error bars are standard errors.

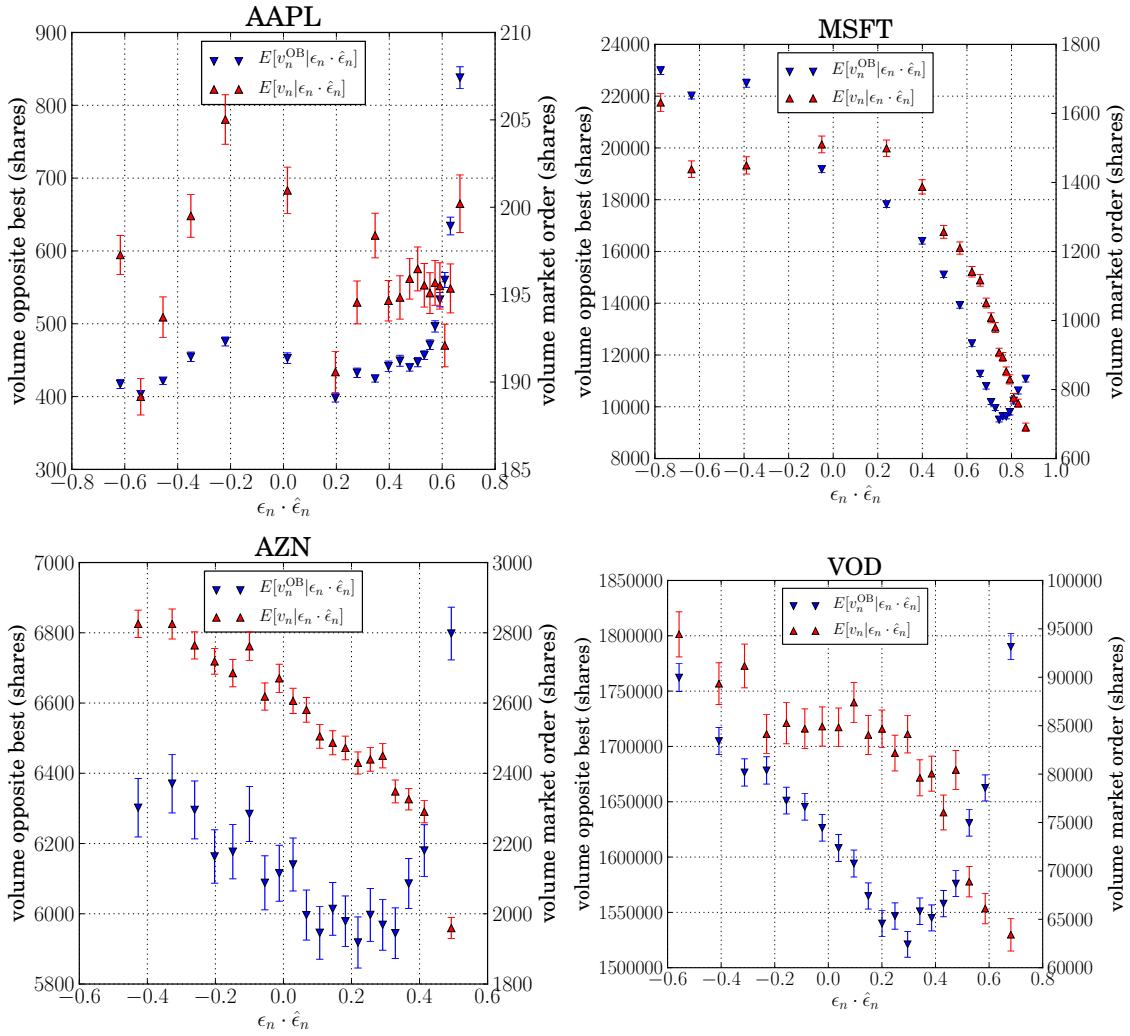

Figure 3.7: Conditional best opposite volumes $\mathbb{E}[v_n^{OB}|\epsilon_n \cdot \hat{\epsilon}_n]$ and conditional market order volumes $\mathbb{E}[v_n|\epsilon_n \cdot \hat{\epsilon}_n]$ for AAPL, MSFT, AZN, and VOD. The error bars are standard errors.

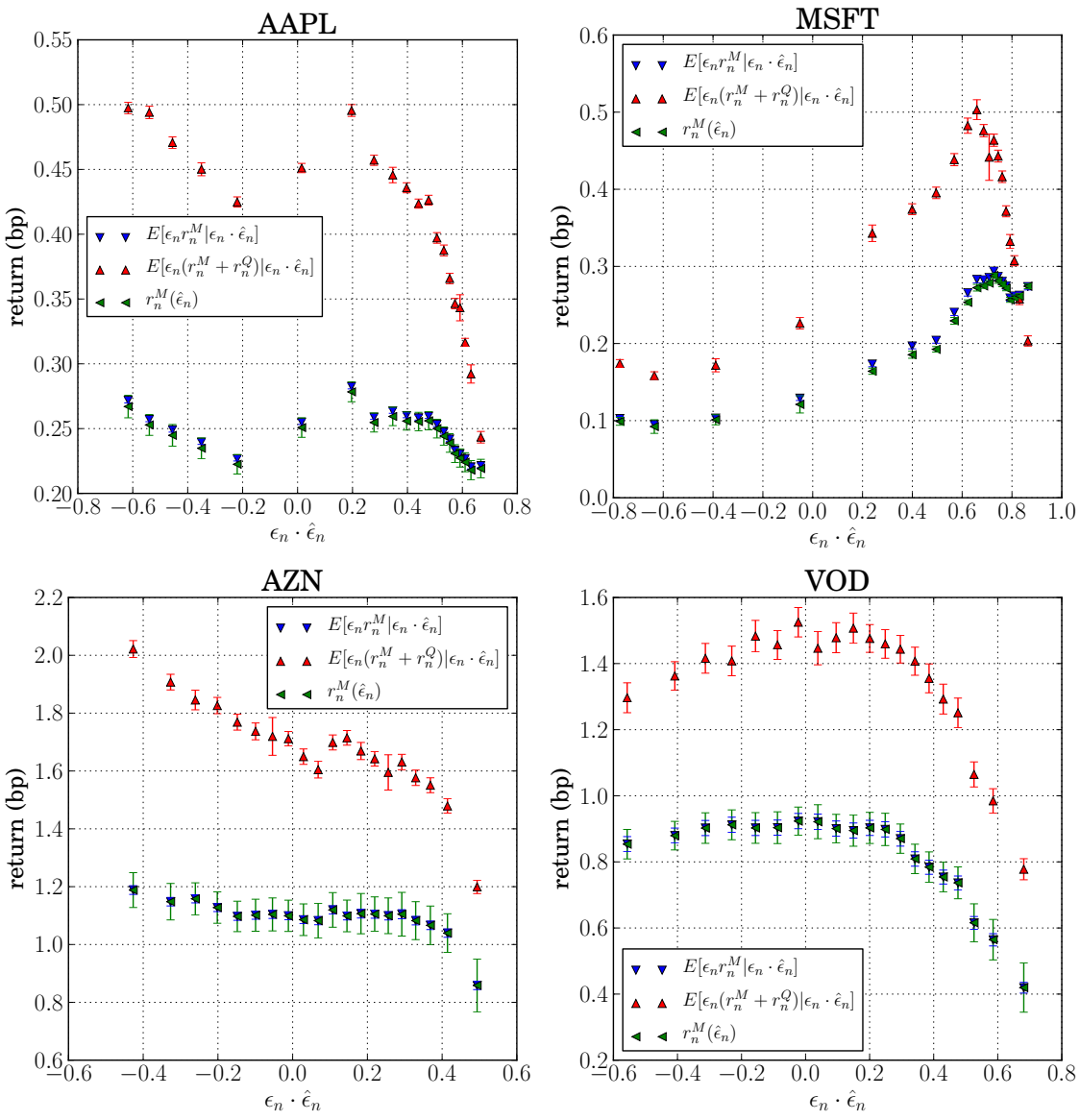

Figure 3.8: Conditional mechanical impact $\mathbb{E}[\epsilon_n r_n^M | \epsilon_n \cdot \hat{\epsilon}_n]$, the approximate expression $r_n^M(\hat{\epsilon}_n)$, and conditional returns $\mathbb{E}[\epsilon_n (r_n^M + r_n^Q) | \epsilon_n \cdot \hat{\epsilon}_n]$ for the stocks AAPL, MSFT, AZN, VOD. The error bars are standard errors.

is that whenever the correctness of the order sign increases the mechanical impact of the order decreases. This result confirms the results in (Lillo and Farmer, 2004), but now we can better understand what is happening in the order book. Indeed, for large-tick assets the gap is basically constant, so the main determinant of the impact is the penetration probability. While for VOD the decrease of the mechanical impact is evident for all values of $\epsilon_n \cdot \hat{\epsilon}_n$, for MSFT the drop of the impact becomes clearer when we reach large values of $\epsilon_n \cdot \hat{\epsilon}_n$. The penetration probability is determined by the interplay between the volume outstanding at the opposite side and the volume of the market order. From Figure 3.7 we know that the volume at the opposite best drops when $\epsilon_n \cdot \hat{\epsilon}_n$ increases and increases again only for very large values of the correctness. Thus, the reduction of the impact has to be given by the decrease of the penetration probability which is determined for high values of $\epsilon_n \cdot \hat{\epsilon}_n$ by liquidity takers placing market orders of decreasing volumes and by liquidity providers placing limit orders at best opposite quotes. From the difference between the return impact and the mechanical return we can also infer the impact of quote revisions and draw conclusions about the adaptive behaviour of liquidity providers.

Figure 3.8 shows that the quote revision always acts in the same direction of the mechanical impact (the only exception is represented by the two extreme bins in the MSFT plot, but such evidence should be confirmed with a more systematic analysis of large-tick stocks from the NASDAQ dataset). This suggests that when the correctness increases, liquidity providers tend to cancel their old limit orders and place new orders at quotes beyond the best price. However, this effect becomes less and less severe when the correctness of the sign predictor is very high, since the impact of the quote revision shrinks to zero, and liquidity providers increase the volume of limit orders outstanding at the opposite best. For small-tick assets the empirical analysis gives similar results. The mechanical impact still decreases when $\epsilon_n \cdot \hat{\epsilon}_n$ increases for both AZN and AAPL. From the analysis of the best volume and market order volume profiles we conclude that this effect is due to liquidity takers which adjust their order volume to the outstanding liquidity and thus drop the penetration probability. The quote revision acts as for the large-tick assets in a similar way: For moderate levels of $\epsilon_n \cdot \hat{\epsilon}_n$ the liquidity providers revise their position, whereas for extreme values the revision stops and liquidity piles up at the opposite best. The major difference between AZN and AAPL emerges looking at the gap distribution. Indeed, for AZN it increases monotonically with $\epsilon_n \cdot \hat{\epsilon}_n$, whilst for AAPL it diminishes (see Figure 3.5). Since liquidity takers act in a similar fashion for both assets, the cause of the different behaviour has to be attributed to the different way liquidity providers revise their position. However, a precise answer to this question is beyond the scope of the current analysis and should deserve further investigation.

3.3.7 The route to market efficiency

The analysis of empirical data discussed in the previous sections largely confirms that asymmetric liquidity is present in financial markets at the transaction by transaction

level. However, our analysis clarifies that the drop of the price impact when the order sign predictability increases is the result of the adaptive behaviour of both liquidity takers and liquidity providers acting on the market. In fact, the former adjust their market order volume at the outstanding liquidity, while the latter revise their limit orders and refill liquidity at the best quotes or within the order book as an adaptive answer to the order flow predictability. How are these results related to market efficiency? By observing Figure 3.8, we notice that for AZN and VOD the return is a non increasing function of $\epsilon_n \cdot \hat{\epsilon}_n$, *i.e.* more predictable trades have a smaller impact[3]. AAPL shows more significant deviations around $\epsilon_n \cdot \hat{\epsilon}_n = 0$, while MSFT shows a pattern, which is clearly inconsistent with market efficiency. We therefore argue that there is room for some inefficiency in the market. More quantitatively, from the figures for the NASDAQ assets we observe that for some $\epsilon_n \cdot \hat{\epsilon}_n > 0$

$$\mathbb{E}[\epsilon_n(r_n^M + r_n^Q)| - \epsilon_n \cdot \hat{\epsilon}_n] < \mathbb{E}[\epsilon_n(r_n^M + r_n^Q)|\epsilon_n \cdot \hat{\epsilon}_n] \,. \tag{3.5}$$

This inequality means that if we use the information set Ω_{n-1} at time t_{n-1}, a non vanishing predictability of return r_n, $\mathbb{E}[r_n|\Omega_{n-1}, \mathcal{M}] \neq 0$, still persists. Obviously this condition is necessary but not sufficient for the inefficiency of markets. In the following we use this condition to test for inefficiency of returns.

Given that at the individual transaction level clear signs of inefficiency exist, one can ask whether these inefficiencies are removed when one considers the expected signed returns s steps (transactions) ahead, conditional to the present information set. To this end we use the information set Ω_{n-1} at time t_{n-1} to build the predictor of market order sign at time t_{n+s}. We then compute expectations at time $n+s$ conditional to the variable $\epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}$. In particular we measure

$$\mathbb{E}[\epsilon_{n+s} r_{n+s}|\epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}], \qquad \mathbb{P}(v_{n+s} \geqslant v_{n+s}^{OB}|\epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}).$$

The results of these analysis are shown in Figures 3.9 and 3.10. We consistently observe qualitatively the same behaviour as s increases. The conditional return as a function of $\epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}$ shows evidences of inefficiency for small values of s. For larger values of s the curves become closer to a linear relation. A similar transition is observed when one considers the lagged probability of penetration. The linear behaviour is consistent with a linear model of market impact, *i.e.* a model where

$$r_n = A(\epsilon_n - \hat{\epsilon}_n) + \eta_n \,,$$

where for simplicity we have neglected any dependence on the volume and η_n is an idiosyncratic component. In this model we have that

$$\mathbb{E}[\epsilon_n r_n|\epsilon_n \cdot \hat{\epsilon}_n] = A(1 - \epsilon_n \cdot \hat{\epsilon}_n) \,,$$

i.e. a linear behaviour. The data shows that this linear behaviour is not observed for $s = 0$, but rather for intermediate values of s. We postulate therefore that a linear

[3]For VOD there is an anomalous behaviour for large negative values of $\epsilon_n \cdot \hat{\epsilon}_n$, but this effect is relatively small.

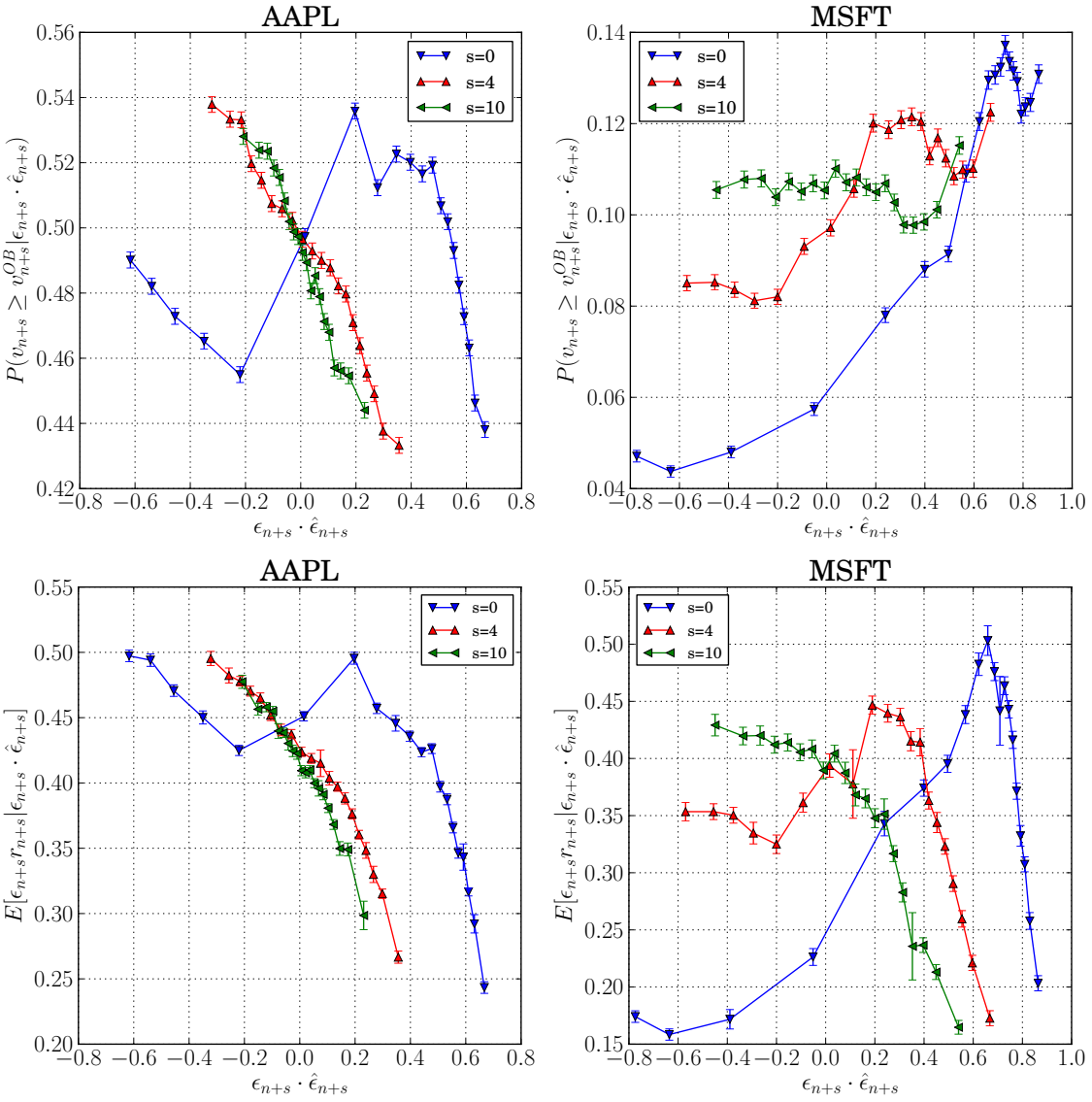

Figure 3.9: Top row: Lagged probability of penetration $\mathbb{P}(v_{n+s} \geqslant v^{OB}_{n+s} | \epsilon_{n+s} \cdot \hat{\epsilon}_{n+s})$. Bottom row: Returns $\mathbb{E}[\epsilon_{n+s} r_{n+s} | \epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}]$ for AAPL and MSFT and different step values s. The error bars are standard errors.

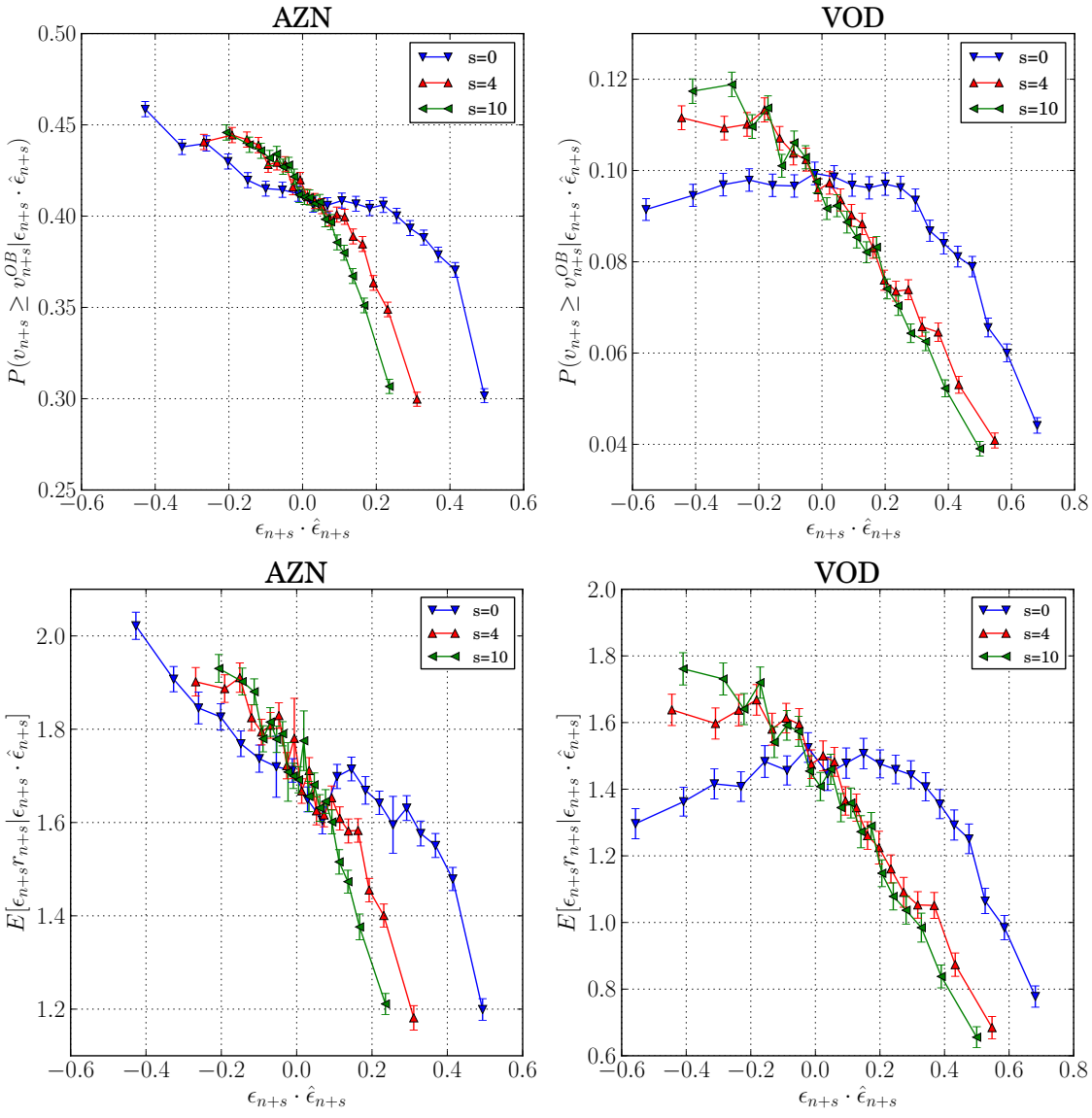

Figure 3.10: Top row: Lagged probability of penetration $\mathbb{P}(v_{n+s} \geqslant v^{OB}_{n+s} | \epsilon_{n+s} \cdot \hat{\epsilon}_{n+s})$. Bottom row: Returns $\mathbb{E}[\epsilon_{n+s} r_{n+s} | \epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}]$ for AZN and VOD and different step values s. The error bars are standard errors.

model of market impact could be developed to describe returns on an aggregated time scale.

When s is very large, the conditional return curves become flatter and flatter. The flat behaviour can be understood by considering that when the lag s is very large, the value of the predictor is typically very close to zero and its predictive power is very low. In fact in the limit of no predictability, it is $\mathbb{E}[\epsilon_{n+s} r_{n+s} | \epsilon_{n+s} \cdot \hat{\epsilon}_{n+s}] = \mathbb{E}[\epsilon_{n+s} r_{n+s}]$.

It is important to stress that the number of transactions needed to observe a transition from the behaviour of returns which shows inefficiencies to the linear behaviour is different in the two datasets. Specifically, by observing the figures, we note that the stocks of NASDAQ market reach an approximately linear behaviour of the return for a value of s which is larger than the corresponding s for the LSE stocks. We interpret this fact as a sign that NASDAQ market needs more trade time to process past information than LSE market. This may seem surprising at first view, because modern financial markets are supposed to be more efficient and faster in processing information when compared to several years ago. This is surely true in physical time, but it might be false in trade time. High frequency trading decreases the physical time needed to restore efficiency in the market, but it might increase the trade time.

In order to test this hypothesis more quantitatively, for each stock we estimated the minimal value of s such that the inefficiency condition of Equation (3.5) is not observed for any value of $\epsilon_n \cdot \hat{\epsilon}_n > 0$. We then multiply this value of s by the average time in seconds between trades, already shown in Table 3.1, in order to get an average minimal time needed to not observe inefficiency as the one of Equation (3.5). By spanning different s values for each stock, we find that this physical time is 5.5 s (5 lags) for AAPL, for MSFT it is 18.7 s (11 lags), for AZN it is 23.1 s (1 lags) and for VOD it is 114.5 s (5 lags). By considering separately large and small tick size stocks, we conclude that recent NASDAQ stocks become efficient in a smaller physical time than older LSE stocks.

3.4 Statistical models of order book and order flow

Modeling the dynamics of the order book is in general quite complicated and challenging. This is due to its multidimensional nature and to the non trivial coupling between the different components of the process. In recent years there has been a growing interest toward the statistical modeling of the order book (Smith et al., 2003), *i.e.* a modeling approach where the different components of the order flow (limit orders, market orders, and cancellations) are treated as independent Poisson processes and the state of the order book emerges as the result of the interplay between these different components. This type of modeling is sometimes termed "zero-intelligence", because the flow of each type of orders follows the simplest un-

conditional process. Despite being conceptually simple, this modeling approach has proven to be surprisingly useful in giving testable predictions of some very short time (Cont and De Larrard, 2013) or long time (Farmer et al., 2005) properties of the order book.

From the previous empirical section we have seen that a key element describing the microstructure of financial markets is the fact that the order flow is long range correlated. As we have seen in the introduction of this chapter, when a strongly autocorrelated order flow (such as the one of real markets) is used as input of an empirically calibrated statistical model of the order book, an unrealistic price time series emerges. In particular, strong predictability of returns and super-diffusivity of prices are observed. This can be easily understood by considering that we use a strongly correlated input in a Markovian model of the order book.

In recent years there have been few notable attempts of modeling the limit order book dynamics in presence of a strongly correlated order flow. Mike and Farmer (2008) introduced a model with correlated order flow. However their main task was to reproduce fat tails of returns and not to reproduce diffusive prices and uncorrelated returns, and in fact in their model these last two properties are not verified.

The attempt of understanding how diffusivity can be recovered in a limit order book model with long memory market order flow has been discussed in (Tóth et al., 2011). As we will detail more below, the proposed model is a variation of the basic zero-intelligence model where the market order signs are long memory. In order to limit the effect of the persistence of market orders on prices, authors proposed that market order volume is a random fraction f of the volume at the opposite best price. They claimed that for a fixed level of market order persistence, there is a critical value of the mean value of f such that the price is diffusive (see also Mastromatteo et al., 2014).

Our model is a variation of the zero-intelligence model. More specifically, the order book is modeled as a discrete price grid of constant minimum price increment w (the tick size, that we set to $w = 1\,\text{tick}$). In the simulations this grid must be sufficiently large in order to consider it as an infinite support. Each price level is populated by buy limit orders, if the price level is below the current midpoint, or sell limit orders, if the price level is above the current midpoint. This is the instantaneous snapshot of the order book state, whereas its time evolution is dominated by three different types of stochastic processes: Limit order placement, market orders arrival, and cancellations of existing limit orders. Limit order placement follows a Poisson process of rate λ per tick and unit event time, which for simplicity is uniform across the discrete price grid. For each event time and each price level we draw the number of limit orders of size s (in our simulations $s = 100\,\text{shares}$) from a Poisson distribution. Market orders arrival triggers an immediate transaction with limit orders at the opposite side of the book. Market orders arrive at a rate μ per unit event time, following a Poisson process, which is independent from limit orders and cancellations. Finally, each existing limit order has the same probability ν per unit event time to be cancelled by liquidity providers.

These features are present also in the zero-intelligence model proposed by Smith et al. (2003); Cont and De Larrard (2013). The modification to the zero-intelligence model affects mostly the market order stochastic process. We assume that the market order flow sign process has long-range correlations, reflecting the order splitting strategy of large orders used by liquidity takers. In the next section we detail how we model this correlated process. Moreover, as in (Tóth et al., 2011) we set the market order volume executed at time t_n as a fraction f of the best opposite volume, $v_n = f \cdot v_n^{OB}$. The scalar f is a random variable drawn from a specific distribution taking values in $f \in [0, 1]$, whose shape plays a crucial role in the model. Tóth et al. (2011) proposed that the random scalar $f \in [0, 1]$ is drawn from a beta distribution $P_\zeta(f) = \zeta(1 - f)^{\zeta - 1}$. The parameter $\zeta > 0$ determines the typical relative volume of market orders and the aggressiveness of liquidity takers. In fact, for $\zeta \to 0$, the distribution peaks around $f = 1$; $\zeta = 1$ corresponds to a uniform distribution; finally, the limit $\zeta \to \infty$ corresponds to unit volumes, because we fix a lower bound for market order volumes to $\min(f \cdot v_n^{OB}) = 1$.

In order to test price diffusivity, we investigate the "signature plot" of the model using Equation (3.1). Tóth et al. (2011) have found numerically that there exists a critical value ζ_c for which the resulting price process of the model is "diffusive" in the intermediate time scale region $\mu^{-1} \ll t \ll \nu^{-1}$, where t is the event time of the model. Thus the lifetime of limit orders ν^{-1} is a critical ingredient for the diffusivity of the price process. For times longer than this value, the long-range correlation of the order signs dominates and the lagged returns are positively correlated.

An illustration of this fact is shown in the top panels of Figure 3.11. The left top panel shows the signature plot of the Tóth et al. (2011) model for different values of the parameter ζ. From the figure, where we set $\nu^{-1} = 100\,\mathrm{s}$, it is clear that prices are asymptotically super-diffusive. Moreover when $\zeta \to \infty$ volatility goes to zero. This is due to the fact that in this limit, volume at the best is never eroded by market orders and price remains constant. The top right panel of Figure 3.11 shows the signature plot of the model in Tóth et al. (2011) for different values of cancellation rates, but keeping fixed the asymptotic order book depth $\rho_\infty = \lambda w/\nu = 50\,\mathrm{shares}$. The critical value of the cancellation rate is $\nu = 10^{-4}$ and we observe an approximately diffusive behaviour for lags ℓ between 1 and 100 (we convert event time to trade times, like lags ℓ, by using the relation $\ell \approx t \cdot \mu\,\mathrm{trades}$). As expected, after this value the price becomes highly super-diffusive. By increasing the cancellation rate, the intermediate region for which prices are diffusive shrinks to zero.

In conclusion, the Tóth et al. (2011) model reproduces diffusive prices in a range of lags which strongly depends on the cancellation rate. In order to extend the range of diffusivity one needs to decrease the cancellation rate to very low values. These values are unrealistically small if one wants to consider the model as describing the *real* order book. The authors in (Tóth et al., 2011) used this model to describe the latent order book instead. This means that this is a hidden liquidity model, where the values of the rates of cancellation are explicitly chosen to be much smaller than the ones observed in the visible limit order book.

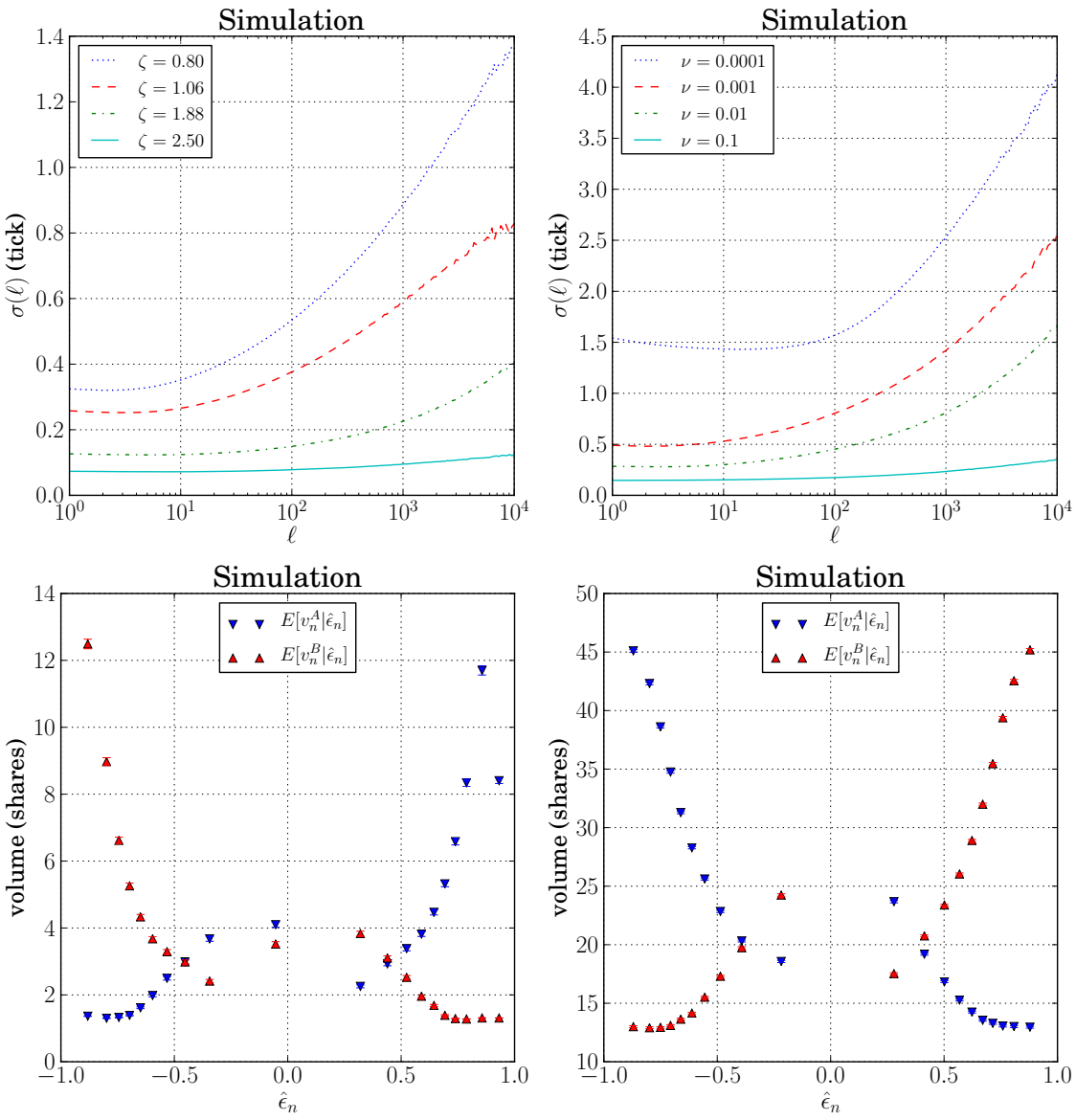

Figure 3.11: (Top left) Signature plot $\sigma(\ell)$ for the parameter choice $\mu = 0.1\,\mathrm{s}^{-1}$, $\lambda = 0.5\,\mathrm{s}^{-1}\mathrm{w}^{-1}$, $\nu = 0.01\,\mathrm{s}^{-1}$, $\gamma = 0.5$ and different values of ζ. The resulting curves show a strong super-diffusive behaviour for large values of ℓ. When $\zeta \to \infty$, the volatility converges to zero. (Top right) Signature plot $\sigma(\ell)$ for the parameter choice $\mu = 0.1\,\mathrm{s}^{-1}$, $\zeta = 0.95$, $\gamma = 0.5$, different values of ν and fixed asymptotic order book depth $\rho_\infty = \lambda w/\nu = 50\,\mathrm{shares}$. The resulting curves show a strong super-diffusive behaviour for large values of ℓ, whereas for low values of ν the price process has sub-diffusive behaviour for an intermediate time scale region ℓ. (Bottom) Conditional volumes at the best quotes on different values of the sign predictor, for parameters $\nu = 10^{-4}\,\mathrm{s}^{-1}$, $\zeta_c = 0.95$ (left) and $\nu = 10^{-2}\,\mathrm{s}^{-1}$, $\zeta_c = 2.5$ (right). The result is compatible with real markets when the cancellation rate is high, whereas for low values of ν volumes at the best quotes behave in the opposite way.

We notice that low cancellation rates lead to wrong predictions of stylized facts of the order book. In the bottom panels of Figure 3.11 we show the volume at the best bid and ask conditional to the value of the predictor $\hat{\epsilon}_n$. In all cases we have selected the value of ζ that gives "diffusive" prices, according to the method in (Tóth et al., 2011). The left panel refers to the low cancellation rate regime, while the right panel refers to the high cancellation rate regime (compare with the empirical results of Figure 3.4). We observe that in the low cancellation rate regime, the conditional volume at the best is opposite to the one observed in real data, *i.e.* there is more volume at the ask (bid) when it is more likely that the next order is a buy (sell). On the other hand, when the cancellation rate is high (as in real markets) the conditional volume at the best is in agreement with real data.

Thus if we want diffusive prices on a large range of lags we need low cancellation rates, but in this case the conditional properties of the order book have the wrong sign. If we want to reproduce the latter, we need high and realistic cancellation rates, but in this case the range of diffusivity will be very small.

We therefore conclude that current statistical models of the order book are unable to reproduce the observed stylized facts when one considers a strongly persistent order flow and is interested in how order book quantities change as a function of order flow predictability as well as efficiency and diffusivity. Even when one uses mechanisms for counterbalancing the persistence of order flow, such as by fine tuning the value of a parameter (e.g. the penetration probability in the Tóth et al. (2011) model), diffusivity is reproduced up to the maximal time scale of the cancellation rate. By decreasing the cancellation rate one obtains a very low volatility and it is not able to reproduce other stylized facts, such as the volume imbalance at bid and ask as a function of order flow predictability.

In the following section we present a statistical order book model with long memory order flow where we are able to simultaneously obtain exact diffusivity of prices and the correct conditional properties of the order book as a function of the order flow predictability. The key intuition behind our modeling scheme is that order book and flow dynamics depend on the predictability of the order flow itself. In other words, instead of fine tuning the value of a parameter (such as the ζ in (Tóth et al., 2011) model), we assume that this "parameter" adapts itself dynamically, depending on the predictability of order flow. This kind of adaptation guarantees diffusivity and the correct dependence of order book quantities on order flow predictability.

The version of the model we present here aims at modeling how liquidity takers adapt their order flow, keeping the price diffusive and efficient. In this sense our model is close to the one in (Tóth et al., 2011), since the limit order and cancellation processes are totally random. The adaptation occurs inside the market order flow. However we believe that the idea of adaptation could be exported for modeling also the liquidity providers.

3.5 Adaptive liquidity model, market efficiency, and price diffusivity

The main intuition behind our modeling approach is that the mechanism restoring efficiency (and therefore diffusivity) must depend on the local level of predictability of the order flow. In the previous sections on the empirical analysis we have shown that many quantities of the order flow and of the limit order book depend in fact from the degree of predictability of the order flow, as well as from the fact that the next market order is in agreement or not with the predictor. In particular we have seen that the penetration probability, *i.e.* the probability that the market order volume is larger or equal to the volume at the opposite best, strongly depends on $\epsilon_n \cdot \hat{\epsilon}_n$ (see Figures 3.6 and 3.7). When this quantity is large, *i.e.* the predictability is high and order executed agrees with the predictor, the volume at the opposite best is small, but the penetration probability also declines, suggesting that liquidity takers adjust the volume of their market order to reduce market impact. This behaviour clearly counterbalances the persistence of order flow, making prices more diffusive and efficient. We now show that it is possible to *exactly* counterbalance the super-diffusivity of order flow and to give the correct conditional behaviour of limit order book quantity. Before describing the mechanism a caveat is in order. We do not believe that this is the only mechanism responsible for efficiency and diffusivity. We believe that liquidity providers, through the so-called "stimulated refill" mechanism (see Eisler et al., 2012b; Mastromatteo et al., 2014), are also responsible in part of the restoration of efficiency. However, we think that this mechanism should also be adaptive, depending on the local level of predictability of order flow.

Our model takes as a starting point the model of Tóth et al. (2011). We assume that there exist two type of traders which execute market orders in the limit order book. The first one is an informed trader, I, for which the distribution of f_I, the ratio between the volume of the market order and the volume at the opposite best, depends parametrically on $\epsilon_n \cdot \hat{\epsilon}_n$. The second one is an uniformed or noise trader, U, which decide the volume to trade by drawing the ratio f_U from an unconditional beta distribution of parameter β. The participation rate of the informed trader is π, which means that $1 - \pi$ is the participation rate of the noise trader.

In particular, we assume that

$$\begin{aligned} P_{f_I}(f|\epsilon_n \cdot \hat{\epsilon}_n) &= g(\epsilon_n \cdot \hat{\epsilon}_n)(1-f)^{g(\epsilon_n \cdot \hat{\epsilon}_n)-1} \\ P_{f_U}(f) &= \beta(1-f)^{\beta-1}, \end{aligned} \tag{3.6}$$

where $g(\epsilon_n \cdot \hat{\epsilon}_n)$ is the exponent of the beta distribution, which in (Tóth et al., 2011) model is the constant ζ, and fine tuned to recover "diffusivity". In our model this exponent is a function which depends on the predictability of market order flow and the degree of surprise of the market order (*i.e.* if it agrees with the predictor).

In our model, when $f_{I,U} \in [1-\delta, 1]$, where $\delta \in (0,1)$ is a small parameter, the market order volume is equal to the volume at the opposite best and penetrates

the book. Thus the conditional probability of penetration for the informed and the noise trader are

$$\begin{aligned} P\left(v_n = v_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, I\right) &= \int_{1-\delta}^{1} P_{f_I}(f|\epsilon_n \cdot \hat{\epsilon}_n) df \\ &= \int_{1-\delta}^{1} g(\epsilon_n \cdot \hat{\epsilon}_n)(1-f)^{g(\epsilon_n \cdot \hat{\epsilon}_n)-1} df = \delta^{g(\epsilon_n \cdot \hat{\epsilon}_n)}\,, \\ P\left(v_n = v_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, U\right) &= \int_{1-\delta}^{1} P_{f_U}(f) df = \delta^{\beta}\,. \end{aligned}$$

Since $\delta < 1$, $P\left(v_n = v_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, I\right)$ is a decreasing function of $\epsilon_n \cdot \hat{\epsilon}_n$ if g is an increasing function of its argument.

This framework reproduces the strategic behaviour of liquidity takers against liquidity providers which operates in a completely random setting. Those who place market orders and are informed adjust locally the requirement of liquidity on the level of predictability of the order signs. This mechanism is captured by the model through the adaptive dependence of $P_{f_I}(f|\epsilon_n \cdot \hat{\epsilon}_n)$ on the sign predictor value. An informed liquidity taker knows exactly the past history of the market order sign process and the sign of the next order (buy or sell) executed in the market, therefore the choice of the local dependence of P_{f_I} appears to us reasonable. We can explain this strategic behaviour of the traders in this way: high predictability of the order flow means that liquidity takers reveal to the market information about their intentions, and in order to control the market impact of their trades, they reduce the volumes of the market orders progressively during the execution of the whole metaorder. Finally, the uniformed liquidity takers place orders of random signs, with volume drawn by an unconditional distribution, and their orders do not contribute to the persistence of the order flow.

Under some conditions, the penetration probability can be connected to the market impact. In fact, we can decompose the market impact function in the contribute of the order flow of informed traders and the one of the noise traders,

$$\mathbb{E}[\epsilon_n r_n | \epsilon_n \cdot \hat{\epsilon}_n] = \mathbb{P}(I|\epsilon_n \cdot \hat{\epsilon}_n)\mathbb{E}[\epsilon_n r_n | \epsilon_n \cdot \hat{\epsilon}_n, I] + \mathbb{P}(U|\epsilon_n \cdot \hat{\epsilon}_n)\mathbb{E}[\epsilon_n r_n | \epsilon_n \cdot \hat{\epsilon}_n, U]\,, \quad (3.7)$$

where the two conditional impact functions are

$$\begin{aligned} \mathbb{E}[\epsilon_n r_n | \epsilon_n \cdot \hat{\epsilon}_n, I] &\simeq \sum_{r_n \neq 0} \epsilon_n r_n P\left(v_n = v_n^{OB}, \epsilon_n g_n^{OB} \simeq 2r_n \middle| \epsilon_n \cdot \hat{\epsilon}_n, I\right) \\ &\simeq \sum_{r_n \neq 0} \epsilon_n r_n P\left(v_n = v_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, I\right) \\ &\qquad \cdot P\left(\epsilon_n g_n^{OB} \simeq 2r_n \middle| \epsilon_n \cdot \hat{\epsilon}_n, v_n = v_n^{OB}, I\right) \\ &\simeq \frac{1}{2} P\left(v_n = v_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, I\right) \mathbb{E}\left[g_n^{OB} \middle| \epsilon_n \cdot \hat{\epsilon}_n, v_n = v_n^{OB}, I\right] \end{aligned}$$

$$\begin{aligned}\mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n,U] &\simeq \sum_{r_n\neq 0}\epsilon_n r_n P\left(v_n=v_n^{OB},\epsilon_n g_n^{OB}\simeq 2r_n\middle|\epsilon_n\cdot\hat{\epsilon}_n,U\right)\\ &\simeq \sum_{r_n\neq 0}\epsilon_n r_n P\left(v_n=v_n^{OB}\middle|U\right)P\left(\epsilon_n g_n^{OB}\simeq 2r_n\middle|\epsilon_n\cdot\hat{\epsilon}_n,v_n=v_n^{OB},U\right)\\ &\simeq \frac{1}{2}P\left(v_n=v_n^{OB}\middle|U\right)\mathbb{E}\left[g_n^{OB}\middle|\epsilon_n\cdot\hat{\epsilon}_n,v_n=v_n^{OB},U\right]\end{aligned}$$

We consider a configuration of our model where the price gaps are constant and equal to $w=1$ tick (this is the case of large-tick stocks, where each price level behind the best quotes is populated by limit orders), so that the previous equations reduces to

$$\begin{aligned}\mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n,I] &\simeq \frac{w}{2}P\left(v_n=v_n^{OB}\middle|\epsilon_n\cdot\hat{\epsilon}_n,I\right)=\frac{w}{2}\delta^{g(\epsilon_n\cdot\hat{\epsilon}_n)},\\ \mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n,U] &\simeq \frac{w}{2}P\left(v_n=v_n^{OB}\middle|U\right)=\frac{w}{2}\delta^{\beta}.\end{aligned}\tag{3.8}$$

The expression in the left hand side of the previous equations are exactly the probability of penetration of market orders, which is the probability that the volume of the market order is equal to the volume at the opposite best.

It is possible to choose the function g is such a way that the total market impact, $\mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n]$, is a linear function of the random variable $\epsilon_n\cdot\hat{\epsilon}_n$. For this purpose, we derive in the Appendix 3.B the expression of the probability that an order sign comes from an informed trader or from a noise trader conditioned on the level of predictability,

$$\mathbb{P}(I|\epsilon_n\cdot\hat{\epsilon}_n)=\frac{\pi+\epsilon_n\cdot\hat{\epsilon}_n}{1+\epsilon_n\cdot\hat{\epsilon}_n},\qquad \mathbb{P}(U|\epsilon_n\cdot\hat{\epsilon}_n)=\frac{1-\pi}{1+\epsilon_n\cdot\hat{\epsilon}_n}$$

Then, the total market impact of Equation (3.7) is

$$\mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n]=\frac{w}{2}\frac{\pi+\epsilon_n\cdot\hat{\epsilon}_n}{1+\epsilon_n\cdot\hat{\epsilon}_n}\delta^{g(\epsilon_n\cdot\hat{\epsilon}_n)}+\frac{w}{2}\frac{1-\pi}{1+\epsilon_n\cdot\hat{\epsilon}_n}\delta^{\beta}$$

and the model reproduces efficient prices,

$$\mathbb{E}[\epsilon_n r_n|\epsilon_n\cdot\hat{\epsilon}_n]\equiv A(1-\epsilon_n\cdot\hat{\epsilon}_n),\tag{3.9}$$

if and only if the function g is

$$g(\epsilon_n\cdot\hat{\epsilon}_n)=\frac{\log\left(\frac{1-(\epsilon_n\cdot\hat{\epsilon}_n)^2}{\pi+\epsilon_n\cdot\hat{\epsilon}_n}A-\frac{1-\pi}{\pi+\epsilon_n\cdot\hat{\epsilon}_n}\delta^{\beta}\right)}{\log\delta}$$

where $A\leq\frac{1-\pi}{1-\pi^2}\delta^{\beta}$, because $\delta^{g(\epsilon_n\cdot\hat{\epsilon}_n)}\leq 1$ if $\epsilon_n\cdot\hat{\epsilon}_n\to-\pi$. This model can be rewritten as

$$r_n=p_{n+1}-p_n=A(\epsilon_n-\hat{\epsilon}_n)+\eta_n,\qquad \hat{\epsilon}_n=\mathbb{E}_{n-1}[\epsilon_n|\Omega_{n-1},\mathcal{M}],\tag{3.10}$$

where η_n is an idiosyncratic IID component of variance Σ^2. The model is completely defined when we assign the model for the time series of order signs as well as the predictor and the information set used.

We have thus found a reduced statistical model for price returns starting from a structural model of the order book. As shown in (Bouchaud et al., 2009), using as sign predictor an *AR(p)* model when $p \to \infty$, the statistical model of Equation (3.10) proposed by Farmer et al. (2006) is equivalent to the propagator model (see Bouchaud et al., 2004; Bacry and Muzy, 2014). We expand this result to the case of the *DAR(p)* model, for which we remind the analytical expression of the sign predictor (see Equation (3.16))

$$\hat{\epsilon}_n = \varphi \sum_{k=1}^{p} \lambda_k \epsilon_{n-k},$$

where for simplicity we restrict the model to the case of $\mu_Z = \mathbb{E}[\epsilon_n] = 0$. In the propagator model Bouchaud et al. (2004), prices are written as a superposition of past order signs, weighted by a propagator $G(\ell)$, and external shocks

$$p_n = \sum_{k<n} \left[G(n-k)\epsilon_k + \eta_k \right],$$

which leads to the expression of the tick by tick returns of the model,

$$r_n = p_{n+1} - p_n = G(1)\epsilon_n + \sum_{k>0} \left[G(k+1) - G(k) \right] \epsilon_{n-k} + \eta_n \,.$$

If we impose the equivalence

$$A\varphi\lambda_k = G(k) - G(k+1) \qquad \text{or} \qquad G(\ell) = A\varphi \left[1 - \sum_{i=1}^{\ell-1} \lambda_i \right],$$

we find the relation between the coefficient of the statistical model of Equation (3.10) and the functional form of the propagator $G(\ell)$ of the model of Bouchaud et al. (2004).

Let us emphasize the statistical properties of returns of the model of Equation (3.10). We clearly see that prices are efficient, since

$$\mathbb{E}_{n-1}[r_n | \Omega_{n-1}, \mathcal{M}] = 0 \,.$$

This means that returns are uncorrelated, $\mathbb{E}[r_n r_{n+\ell}] = 0$, for all $\ell > 1$. Since $\mathbb{E}[r_n r_{n+\ell}] = \mathbb{E}[\mathbb{E}[r_n r_{n+\ell} | \Omega_{n-1}, \mathcal{M}]]$ it is enough to prove that $\mathbb{E}[r_n r_{n+\ell} | \Omega_{n-1}, \mathcal{M}] = 0$, which follows making use of the law of iterated expectations.

Furthermore, prices of the model are diffusive for all lags. Let us consider the quantity $p_{n+\ell} - p_n$ and compute its variance, using the last result of uncorrelated returns of the model

$$\mathbb{E}[(p_{n+\ell} - p_n)^2] = \mathbb{E}\left[\left(\sum_{i=0}^{\ell-1} r_{n+i} \right)^2 \right] = \sum_{i=0}^{\ell-1} \mathbb{E}\left[r_{n+i}^2 \right] = (\Sigma^2 + A^2)\ell - A^2 \sum_{i=0}^{\ell-1} \mathbb{E}\left[\hat{\epsilon}_{n+i}^2 \right] \,.$$

The quantity $\mathbb{E}\left[\hat{\epsilon}^2_{n+i}\right]$ depends only on the particular choice of the driving model of the order flow. In the case of the *DAR(p)* model it is constant and independent from ℓ,

$$\sum_{i=0}^{\ell-1}\mathbb{E}\left[\hat{\epsilon}^2_{n+i}\right] = \left(\varphi^2\sum_{k=1}^{p}\lambda_k^2 + 2\varphi^2\sum_{1\le k<h}^{p}\lambda_r\lambda_s C(h-k)\right)\ell\,,$$

where $C(\ell) = \mathbb{E}[\epsilon_n\epsilon_{n+\ell}]$ is the empirical autocorrelation of order signs. The unconditional variance finally reads

$$\mathbb{E}[(p_{n+\ell}-p_n)^2] = \left[\Sigma^2 + A^2\left(1-\varphi^2\sum_{k=1}^{p}\lambda_k^2 - 2\varphi^2\sum_{1\le k<h}^{p}\lambda_r\lambda_s C(h-k)\right)\right]\ell\,,$$

which scales perfectly as a diffusive process with the lag ℓ.

3.6 Results

Since in our model the market order flow plays a crucial role, in this section we present the specific model for the time series describing it. Moreover we shall discuss the different predictors that can be built for this time series. In the final subsection we shall present in detail numerical simulations of the model.

3.6.1 Models of the market order flow

Order flow is strongly autocorrelated in time. As shown in (Tóth et al., 2015), correlation of order flow is mostly due to order splitting, rather than herding. This was originally suggested by Lillo et al. (2005) on the basis of indirect empirical evidences. In this paper, authors proposed a simple model where the correlation of order flow is a consequence of order splitting and the very heterogeneous distribution of metaorder sizes. Here we use a variation of this model to generate the market order flow which enters the limit order book.

According to the model of Lillo et al. (2005), there are M funds that want to trade one metaorder each of a size L_i $(i = 1, .., M)$ taken from a distribution p_L, where for simplicity $L_i \in \mathbb{N}^+$. The sign of the each metaorder is taken randomly and at each trade time step, one fund is picked randomly with uniform probability. The selected fund initiates a trade of the sign of its metaorder, and the size of the metaorder is reduced by one unit. When a metaorder is completely traded, a new one is drawn from p_L and assigned a random sign.

In (Lillo et al., 2005) it is shown how to connect the distribution p_L of metaorder size with the autocorrelation function of trade signs. In particular, if the distribution is Pareto

$$p_L = \frac{1}{\zeta(\alpha)}\frac{1}{L^{1+\alpha}} \tag{3.11}$$

where $\zeta(\alpha)$ is the Riemann zeta function, then the autocorrelation function of trade signs decays asymptotically as

$$\rho_s(\ell) = \mathbb{E}[\epsilon_n \epsilon_{n+\ell}] \sim \frac{M^{\alpha-2}}{\ell^{\alpha-1}} \sim \frac{1}{\ell^{\gamma}} \tag{3.12}$$

This model connects the exponent of the autocorrelation function of order signs with the tail exponent of metaorder distribution, since $\gamma = \alpha - 1$. The market order sign is a long memory process if $\alpha < 2$, *i.e.* if the variance of the metaorder size diverges.

There is a growing empirical evidence that the distribution of metaorder size is asymptotically Pareto distributed with a tail exponent close to $\alpha = 1.5$. Gabaix et al. (2006) and Lillo et al. (2005) argue that block trades (*i.e.* traded off book) could be used as a proxy of metaorders and find that an exponent very close to 1.5 describes the tail of the trade size distribution. Vaglica et al. (2008) use trade data of the Spanish Stock Exchange with an identifier of the broker to statistically reconstruct the metaorders. They find that the size of the metaorder is asymptotically Pareto distributed with an exponent $\alpha \approx 1.7$. Finally, Bershova and Rakhlin (2013) use proprietary data of a set of large institutional metaorders executed at AllianceBernstein's buy-side trading desk in the US equity market and find that the tail of metaorder size is Pareto with exponent $\alpha = 1.56$.

Here in order to have an analytically tractable expression of the predictor of the order flow, we shall consider a slight modification of the above model. First of all, we will consider that only one metaorder is present at each time. This is similar to what is done in (Tóth et al., 2011; Mastromatteo et al., 2014). The second modification is that we will assume that other traders are present and that they contribute with a random background of signs. This can be considered as a large set of metarorders of size 1.

More specifically, we introduce the participation ratio π of the metaorder, which is the probability that a trade is initiated by the metaorder (of size larger than one), while $1 - \pi$ is the probability that the trade is initiated by the noise traders.

The introduction of the noise traders does not change the long memory properties of the autocorrelation of the order flow. Their only effect is to reduce the global level of the autocorrelation. More specifically, if $\rho_s(\ell)$ is the autocorrelation function of the order flow when one considers only the trades of the metaorder (*i.e.* Equation (3.12) with $M = 1$), one has in presence of noise

$$\rho(\ell) = \mathbb{E}[\epsilon_n \epsilon_{n+\ell}] \simeq \pi^2 \rho_s(\pi\ell)\,,$$

because the probability that the two trades at time t_n and $t_{n+\ell}$ both come from a metaorder (not necessarily the same) is π^2 and a time lag of ℓ trades corresponds on average to a time lag of $\pi\ell$ trades from the metaorder.

If the metaorder size distribution is Pareto (see Equation (3.11) and 3.12), we have

$$\rho(\ell) \sim \frac{\pi^2}{(\pi\ell)^{\alpha-1}} = \frac{\pi^{3-\alpha}}{\ell^{\alpha-1}}\,,$$

i.e. the autocorrelation function is dampened by a factor $\pi^{3-\alpha}$, but it is still long memory with the same Hurst exponent.

3.6.2 Predictors of the order flow

In our model market order volume depends on the predictability of market order flow. Given the time series model described above, we will consider two predictors of the order flow.

The first predictor is the one associated with the *DAR(p)* model discussed in the empirical section and reviewed in the appendix. The p signs of past market orders are used to build the expected value of the next sign. Clearly this predictor does not have any direct information on how many metaorders were present in the estimation window, thus we call it the "public" sign predictor. Given the fact that our order flow model is composed by one metaorder at a time (plus the noise background), it is likely that the estimation window of p past signs includes orders that are coming from past (*i.e.* not anymore active) metaorders. This adds of course noise and decreases the forecasting ability of the predictor. On the opposite side, if the metaorder is longer than πp, from a certain point on, the predictor is using information of the most recent part of the metaorder and it is discarding information of the first part of the metaorder. As we will see below, this will have an effect on the diffusivity of price at time scales longer than p trades.

The second predictor is the one which makes use of the information allowing to discriminate the orders due to the active metaorder to those due to the noisy background. This information is not typically of public domain and therefore cannot be used by the liquidity providers, therefore we call it the "private" sign predictor. In our model the liquidity taker adjusts the volume of their market orders to the degree of predictability of the order flow. Given their active role, they are able to use a predictor that takes into account the history of the recent order flow and the information on the current length of the metaorder.

The key point is that the correlation of the order flow comes from the presence of the metaorder. If m trades of the current metaorder has been already traded, the probability that the metaorder continues is (Farmer et al., 2013)

$$\mathcal{P}_m = \frac{\sum_{i=m+1}^{\infty} p_i}{\sum_{i=m}^{\infty} p_i} .$$

For example, if the metaorder size distribution is Pareto (see Equation (3.11)) this continuation probability is

$$\mathcal{P}_m = \frac{\zeta(1+\alpha, 1+m)}{\zeta(1+\alpha, m)} \simeq \left(\frac{m}{m+1}\right)^{\alpha} \sim 1 - \frac{\alpha}{m} ,$$

where $\zeta(s, a)$ is the generalized Riemann zeta function (also called the Hurwitz zeta function). The approximations are valid in the large m limit. This means that the longer the metaorder has been active, the more likely is that it continues.

Let us suppose that the active metaorder is a buy and the participation rate is π. The probability that the next order is a buy is

$$p_m^+ = \frac{1-\pi}{2} + \pi\left(\mathcal{P}_m + \frac{1-\mathcal{P}_m}{2}\right) = \frac{1+\pi\mathcal{P}_m}{2}\,.$$

The first term describes the event in which the next order is from a noise trader, which with probability 1/2 will place a buy. The second term describes the event in which the next order comes from a metaorder. Moreover, the first term in parenthesis gives the probability that the active metaorder is not finished (and one trade from it will be surely a buy), while the second term in brackets describes the possibility that the active metaorder is finished and that the order comes from a new metaorder, which with 1/2 probability is a buy. Similarly if the active metaorder is a sell, rather than a buy, then $p_m^+ = (1-\pi\mathcal{P}_m)/2$. If we indicate with s_n the sign of the active metaorder at time t_n, we can rewrite

$$p_n^+ = \frac{1+s_n\pi\mathcal{P}_m}{2}\,.$$

In general, since the sign predictor $\hat{\epsilon}_n \equiv p_n^+ - p_n^-$ and obviously $p_n^+ + p_n^- = 1$, we have

$$p_n^+ = \frac{1+\hat{\epsilon}_n}{2}\,, \qquad p_n^- = \frac{1-\hat{\epsilon}_n}{2}\,,$$

or, in other words, it is

$$\hat{\epsilon}_n^{\mathrm{LMF}} = \mathbb{E}[\epsilon_n|\Omega_{n-1}, \mathrm{LMF}] = 2p_n^+ - 1\,.$$

where LMF refers to the model developed by Lillo et al. (2005) and described above. This means that the predictor which allows to discriminate the trades from the metaorder is

$$\hat{\epsilon}_n^{\mathrm{LMF}} = s_n\pi\mathcal{P}_m\,. \tag{3.13}$$

3.6.3 Numerical results

Numerical simulations of the model confirm the theoretical prediction explained above. We have measured the signature plot (see Equation (3.1)) of the price process as result of the interaction of market orders, limit orders, and cancellation of the model. We have used the "private" sign predictor of Equation (3.13). Figure 3.12 shows the signature plot of the model. As one can observe, the volatility as a function of the lag ℓ is almost constant, $\sigma(\ell) = D$, and it is compatible with a diffusive process. The vertical line is the lifetime of limit orders $\nu^{-1}\mu = 10$ trades, which is in the model of (Tóth et al., 2011) the maximum time scale for which prices are still diffusive. Furthermore, by construction the resulting prices of the model are informationally efficient. This characteristic does not depend on the particular choice of the participation π and of the parameter δ. In fact we observe that, as expected from the theoretical analysis, volatility does not depend on δ for a fixed

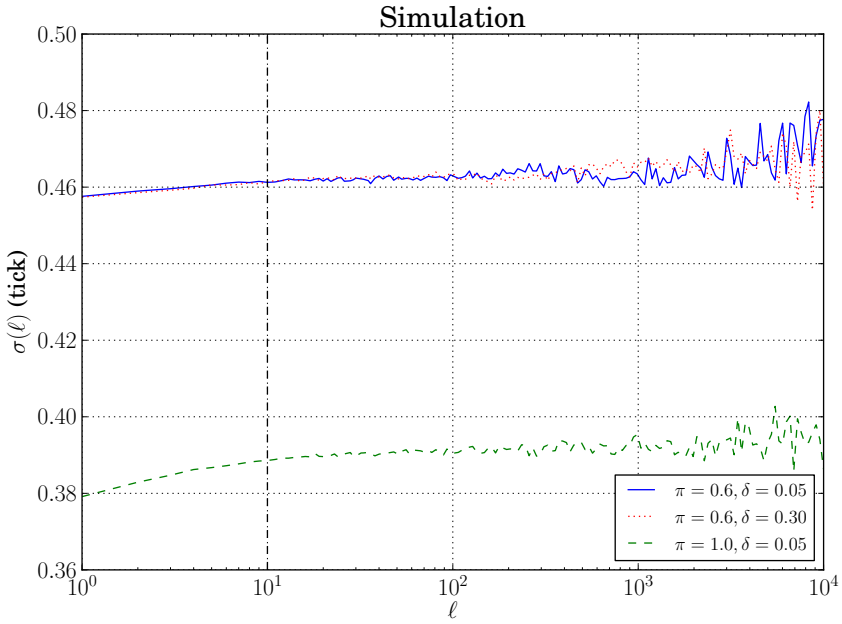

Figure 3.12: Signature plot of the model as a function of the (tick) time lag for different values of participation rate π and δ, and for the parameter choice $\mu = 0.1\,\mathrm{s}^{-1}$, $\lambda = 0.5\,\mathrm{s}^{-1}\mathrm{w}^{-1}$, $\nu = 0.01\,\mathrm{s}^{-1}$, $\gamma = 0.5$, $\alpha = 0.5$. The vertical line is the lifetime of limit orders $\nu^{-1}\mu$ trades.

value of π. On the other hand, volatility is lower for higher values of π, which is not surprising because high levels of participation rate lead to lower uncertainty in the order flow and lower volatility. We have repeated the same simulations by using the sign predictor of the *DAR(p)* process, which uses only the signs of past order flow, and no information on metaorders.

Figure 3.13 shows the signature plot of the resulting price process in such a setting. The time scale for which the price process is diffusive depends on the chosen order p of the *DAR(p)* process, and $\sigma(\ell)$ is constant for $\ell < p$. This is not surprising because if one considers a time window of length $\ell > p$, there might exist non vanishing positive correlations of order signs due to metaorders longer than πp, but the predictor considers only the past p trades. By taking longer windows for the *DAR(p)* predictor (or by considering models with shorter metorders) one recovers diffusivity at all scales. An interesting result of the simulations is that the volatility using the two different sign predictor has approximately the same value, *i.e.* it is almost independent from the choice of the particular set of information used for the sign predictor.

We therefore conclude that our model is able to give exactly diffusive prices. Our model is also able to reproduce the empirically observed dependencies of order book quantities from the predictability of the order flow. Specifically, we have measured the same order book quantities of section 3.3 in our synthetic market, using in P_g the "private" sign predictor. However, the conditional expected values of volume at best quotes, price gaps, probability of penetration, fraction, and returns are computed

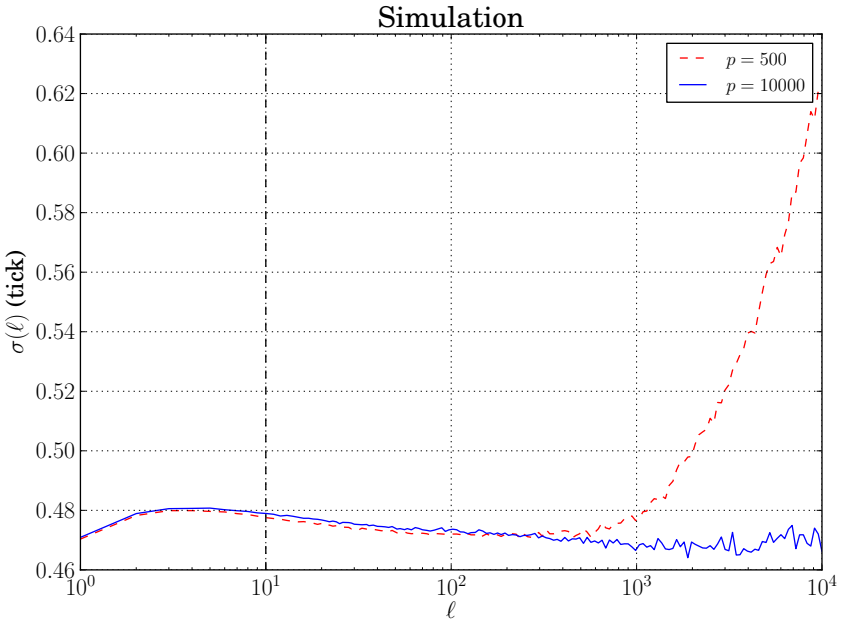

Figure 3.13: Signature plot of the model as a function of the (tick) time lag for different values of order p of the *DAR(p)* process used for the computation of the sign predictor, using $\mu = 0.1\,\mathrm{s}^{-1}$, $\lambda = 0.5\,\mathrm{s}^{-1}\mathrm{w}^{-1}$, $\nu = 0.01\,\mathrm{s}^{-1}$, $\gamma = 0.5$, $\pi = 0.6$, $\delta = 0.05$, $A = 0.5$ and $A = \frac{1-\pi}{1-\pi^2}\delta^\beta$. The vertical line is the lifetime of limit orders $\nu^{-1}\mu$ trades.

conditioning them on the sign predictor of the *DAR(p)* process. In this way, we can compare the results from our model with the evidences from real markets.

The results are shown in Figure 3.14. The model reproduces quite well the behaviour of real order books. In particular, volume at the bid is higher (smaller) than the volume at the ask when the most likely next market order is a buy (sell), as observed in Figure 3.4. Gaps are constant because we are working in the large tick approximation, where all the level of the order book are occupied (see the right panels of Figure 3.5). The penetration probability and the average fraction f both decline with $\epsilon_n \cdot \hat{\epsilon}$, as seen in Figure 3.6[4]. Finally, the impact $\epsilon_n r_n$ declines with $\epsilon_n \cdot \hat{\epsilon}$, as postulated by the asymmetric liquidity mechanism, and as observed in real data in the region where the mass of the distribution of sign predictors is concentrated (see Figure 3.8).

Let us comment the property of efficiency of our model. The synthetic market simulated by the model is more efficient than the real ones. In fact, the signature plot is almost constant for every time scale ℓ and conditional returns are linear in $\epsilon_n \cdot \hat{\epsilon}_n$ like the efficient model in Equation (3.10). In modern markets linearity is recovered after few trades (see Figures 3.9 and 3.10). We are confident that introducing some inefficiency in the model, it can reproduce some effects measured during our empirical analysis of real stocks.

[4]In the case of NASDAQ stocks, one can observe this behaviour in the region of high predictability where the majority of the mass of the distribution of predictors is concentrated.

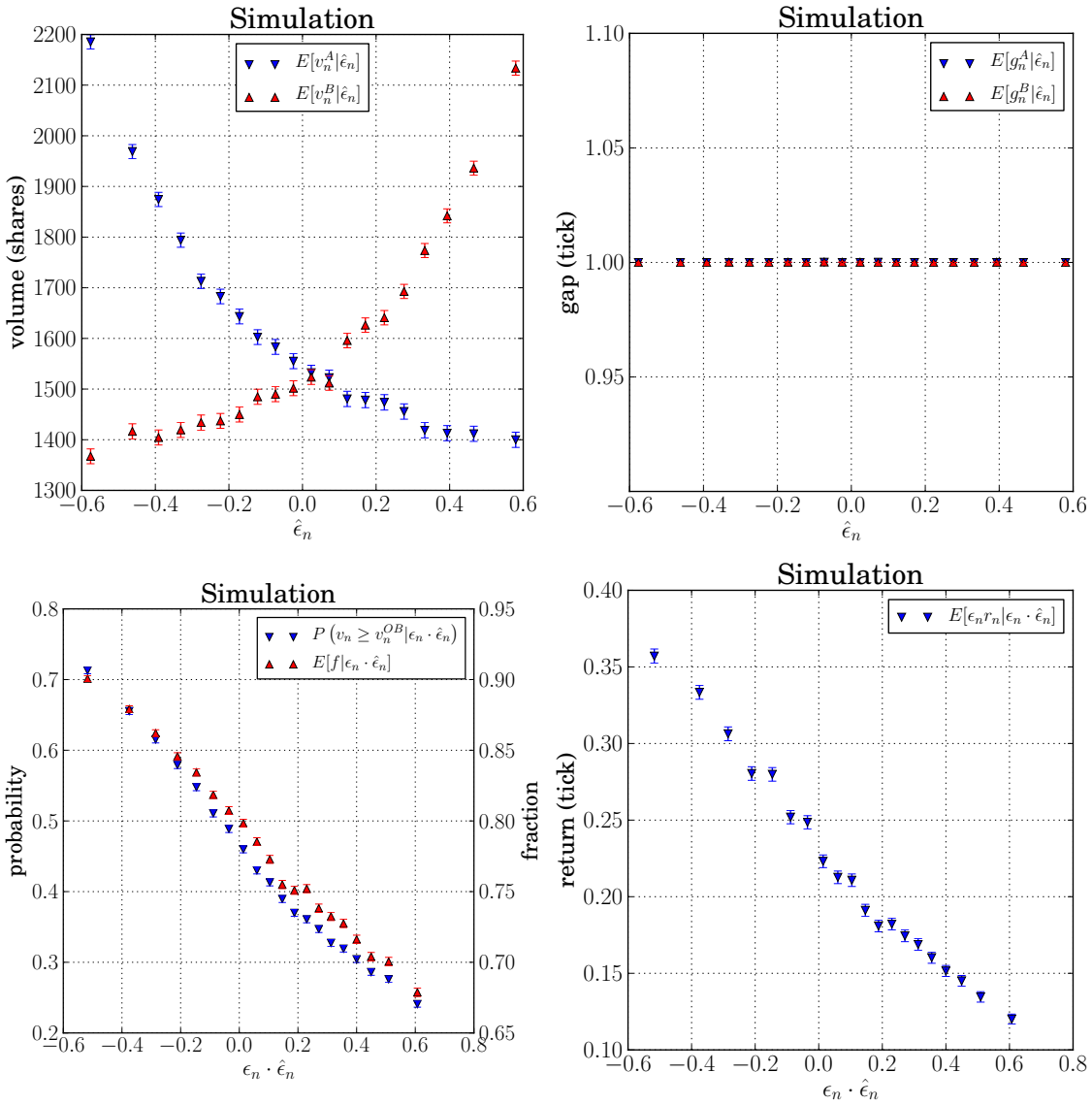

Figure 3.14: (Top left) Conditional volumes at the best, (top right) conditional price gaps on different sign predictor values, (bottom left) probability of penetration and conditional fraction, and (bottom right) conditional return on $\epsilon_n \cdot \hat{\epsilon}_n$ of the model. The parameters of the model were $\mu = 0.1\,\mathrm{s}^{-1}$, $\lambda = 0.5\,\mathrm{s}^{-1}\mathrm{w}^{-1}$, $\nu = 0.01\,\mathrm{s}^{-1}$, $\gamma = 0.5$, $\pi = 0.6$, $\delta = 0.05$, $A = 0.5$ and $A = \frac{1-\pi}{1-\pi^2}\delta^\beta$.

3.7 Discussion and partial conclusions

In this chapter we have considered the subtle issue of reconciling the persistence of order flow with price efficiency and return diffusivity. Since on average a buyer initiated trade pushes the price up while a seller initiated trade pushes it down, in a naive view the strong positive correlation between trades measured empirically would lead to strongly correlated returns. However, the empirical evidence of price efficiency clashes with this view.

We have investigated the microstructural mechanisms able to reconcile both evidences. In the first part of our analysis we have performed an empirical study of the behaviour of four stocks, Astrazeneca, Vodafone, Apple, and Microsoft, which have been selected in light of their different features. While the order book data for the former two stocks were recorded at the London Stock Exchange in 2004, for the latter two stocks the data sample is relatively more recent and was recorded at NASDAQ in July/August 2009. Moreover, while Vodafone and Microsoft are stocks whose tick-to-price ratio is on average large, for Astrazeneca and Apple the ratio is very small. Our choice should guarantee independence of our results on the specificity of the stocks and of the market place. Nonetheless, we have planned a future extension of our analysis to a wider data sample.

A possible mechanism able to reconcile the persistence of trade signs and price efficiency is the asymmetric liquidity mechanism proposed by Lillo and Farmer (2004): The price impact of an order is inversely related to the probability of its occurrence. This means that if at a certain point in time it is more likely that the next trade is a buy rather than a sell, a buyer initiated trade will have a smaller impact than a seller initiated trade. There is therefore a compensation between the probability of an event and its effect on the price. In spite of its conceptual simplicity there are many possible microstructural mechanisms responsible for it. Among the several explanations of the drop of the impact one could consider the case where efficiency is guaranteed by the agents initiating the trade and adjusting the volume of their trades to the volume outstanding on the opposite side of the order book. A second explanation would focus on the leading role of liquidity providers revising their quotes after a trade in order to compensate for the impact due to liquidity takers. Our empirical analysis evidences that when the order flow predictability increases in one direction (buy or sell) the volume outstanding at the opposite best decreases, the opposite side of the book becomes more and more sparse, but the probability that a trade moves the price decreases significantly. While the last mechanism is able to counterbalance the persistence of order flow and restore efficiency and diffusivity, the first two act in the opposite direction. Moreover, disentangling each return in a component due to a mechanical impact and in a second aggregated component due to the revision of liquidity providers, we have measured a positive correlation between impact and quote revision. However, this effect tends to disappear when the order sign predictability increases.

The above empirical evidences lead to significant challenges in the modeling of the order book dynamics. A growing strand of literature is dealing with this issue,

and in the second part of the chapter we have introduced a statistical model designed for large tick stocks which is able to successfully recover the empirical findings in the presence of a strongly persistent order flow. The main intuition behind our approach is that the mechanism restoring efficiency must depend on the local level of predictability of the order flow. More precisely, the agent placing a market order knows exactly the past history of the market order sign process and the sign of the next order (buy or sell) she is going to execute and adapts her order volume to the level of predictability of the order sign. We explain this strategic behaviour in this way: High predictability of the order flow means that liquidity takers reveal to the market information about their intentions, and in order to control the market impact of their trades, they reduce the volumes of the market orders progressively during the execution of the whole metaorder. We have supported our conclusions with extensive Monte Carlo simulations.

The adaptive liquidity taking mechanism described above is, however, only part of the story. In spite of its effectiveness, it is indeed evident that a determinant role has to be also played by liquidity providers. While here we have focused on modeling the strategic behaviour of liquidity takers, it is worth working on the extension of the statistical model in order to include the strategic behaviour of market makers.

Appendix

3.A Autocovariance structure and forecasting of *DAR(p)* model

Autocovariance structure. Let $\{X_n\}_{\mathbb{N}_m}$ be a stationary *DAR(p)* process with marginal distribution Ξ, parameter φ and parameter vector $\lambda = (\lambda_1, \ldots, \lambda_p)$. From Equation (3.2), we immediately find that $\mu_X = \mathbb{E}[X_n] = \mathbb{E}[Z_n] = \mu_Z$.

We center the X_n's with the unconditional mean, $\widetilde{X}_n = X_n - \mu_X$, multiply Equation (3.2) by $\widetilde{X}_{n-\ell}$, $\ell > 0$ and we take the expectation of both sides

$$\gamma(\ell) = \mathbb{E}[\widetilde{X}_n \widetilde{X}_{n-\ell}] = \varphi \sum_{k=1}^{p} \lambda_k \mathbb{E}[\widetilde{X}_{n-k} \widetilde{X}_{n-\ell}] + (1-\varphi)\mathbb{E}[(Z_n - \mu_X)\widetilde{X}_{n-\ell}]\,.$$

Dividing both sides by the variance of $\widetilde{X}_n$, we obtain the corresponding relation for the autocorrelations

$$\rho(\ell) = \varphi \sum_{k=1}^{p} \lambda_k \rho(\ell - k), \qquad \ell \geq 1\,, \tag{3.14}$$

which are the usual Yule-Walker equations (Hamilton, 1994). This linear system can be solved recursively after the computation of the sample autocorrelations from the time series. Given $\rho_1, \rho_2, \ldots, \rho_p$, the first p equations can be solved for the p parameters $\lambda_1, \ldots, \lambda_{p-1}$ and φ. The parameter λ_p is given by $(1 - \lambda_1 - \ldots - \lambda_{p-1})$. The estimations of the components of the parameter vector $\vec{\lambda}$ can lead to negative values, but probabilities must be always greater than or equal to zero, $\lambda_i \geq 0$. This problem is important when we perform simulations of the process, in this case we smooth the empirical coefficients performing a moving average which spans ten points and finally we set the negative elements to zero.

The advantages of the *DAR(p)* model is that it is intrinsically autoregressive, and its parameters can be easily computed by means of the sample autocorrelation.

Forecasting. We can now construct the best predictor of variable X_n within this model. We recall the fact that all sequences $\{V_n\}$, $\{A_n\}$ and $\{Z_n\}_{\mathbb{N}_\lambda}$ are independent one from each other. We take the expected values both unconditional and conditional on Ω_{n-1} and we calculate the second conditional moment,

$$\begin{aligned}
\mathbb{E}[X_n] &= \mu_Z\,, \\
\mathbb{E}[X_n|\Omega_{n-1}] &= \varphi \sum_{k=1}^{p} \lambda_k X_{n-k} + \mu_Z(1-\varphi) \equiv \hat{X}_n^{DAR}, \\
\mathbb{E}[X_n^2|\Omega_{n-1}] &= \varphi \mathbb{E}[X_{n-A_n}^2|\Omega_{n-1}] + \mu_Z(1-\varphi) \\
&= \varphi \sum_{k=1}^{p} \lambda_k \mathbb{E}[X_{n-k}^2|\Omega_{n-1}] + \mu_Z(1-\varphi) \\
&= \varphi \sum_{k=1}^{p} \lambda_k X_{n-k}^2 + \mu_Z(1-\varphi)\,. \qquad (3.15)
\end{aligned}$$

The expression of the predictor can be extended by computing the conditional expected value of X_{n+s}, for $s = 1, 2, \ldots$. After simple calculations, we find that

$$\hat{X}_{n+s}^{DAR} = \varphi \sum_{k=1}^{p} \lambda_k Y_{n+s-k} + \mu_Z(1-\varphi)\,, \qquad (3.16)$$

where

$$Y_{n+s-k} = \begin{cases} \hat{X}_{n+s-k}^{DAR} & \text{if } k \le s \\ X_{n+s-k} & \text{if } k > s \end{cases}.$$

3.B Probability of informed and noise order sign

We need the analytical expression of $\mathbb{P}(I|\epsilon_n \cdot \hat{\epsilon}_n)$ and $\mathbb{P}(U|\epsilon_n \cdot \hat{\epsilon}_n)$, which are the probability that an order sign comes from the order flow of the informed traders or from the one of the noise traders.

First, we have that:

$$\mathbb{P}(I|\epsilon_n \cdot \hat{\epsilon}_n) = \frac{\mathbb{P}(\epsilon_n \cdot \hat{\epsilon}_n|I)\mathbb{P}(I)}{\mathbb{P}(\epsilon_n \cdot \hat{\epsilon}_n)}.$$

We expand the probabilities in the denominator:

$$\mathbb{P}(\epsilon_n \cdot \hat{\epsilon}_n|I) = \frac{1}{2}\left[\mathbb{P}(\hat{\epsilon}_n|I, \epsilon_n = 1) + \mathbb{P}(-\hat{\epsilon}_n|I, \epsilon_n = -1)\right]$$

For the order signs which come from the informed traders we have:

$$\begin{aligned}
\mathbb{P}(\hat{\epsilon}_n|I,\epsilon_n=1) &= \frac{\mathbb{P}(\hat{\epsilon}_n,I,\epsilon_n=1)}{\mathbb{P}(I,\epsilon_n=1)} = \frac{\mathbb{P}(\epsilon_n=1|\hat{\epsilon}_n,I)\mathbb{P}(I)\mathbb{P}(\hat{\epsilon}_n)}{\pi/2} \\
&= 2\mathbb{P}(\epsilon_n=1|\hat{\epsilon}_n,I)\mathbb{P}(\hat{\epsilon}_n) = 2\left(\frac{\hat{\epsilon}_n}{\pi}+\frac{1-\frac{\hat{\epsilon}_n}{\pi}}{2}\right)\mathbb{P}(\hat{\epsilon}_n) \\
&= \left(1+\frac{\hat{\epsilon}_n}{\pi}\right)\mathbb{P}(\hat{\epsilon}_n) \\
s\mathbb{P}(\hat{\epsilon}_n|I,\epsilon_n=-1) &= 2\left(-\frac{\hat{\epsilon}_n}{\pi}+\frac{1+\frac{\hat{\epsilon}_n}{\pi}}{2}\right)\mathbb{P}(\hat{\epsilon}_n) = \left(1-\frac{\hat{\epsilon}_n}{\pi}\right)\mathbb{P}(\hat{\epsilon}_n)
\end{aligned}$$

and we can write:

$$\mathbb{P}(\epsilon_n\cdot\hat{\epsilon}_n) = \pi\left(1+\frac{\epsilon_n\cdot\hat{\epsilon}_n}{\pi}\right)\mathbb{P}(\hat{\epsilon}_n)+(1-\pi)\mathbb{P}(\hat{\epsilon}_n) = \mathbb{P}(\hat{\epsilon}_n)(1+\epsilon_n\cdot\hat{\epsilon}_n)$$

Finally, the conditional probabilities on the predictability of the order flow that the order signs come from the informed or the noise trades are

$$\begin{aligned}
\mathbb{P}(I|\epsilon_n\cdot\hat{\epsilon}_n) &= \frac{\mathbb{P}(\epsilon_n\cdot\hat{\epsilon}_n|I)\mathbb{P}(I)}{\mathbb{P}(\epsilon_n\cdot\hat{\epsilon}_n)} = \frac{\pi+\epsilon_n\cdot\hat{\epsilon}_n}{1+\epsilon_n\cdot\hat{\epsilon}_n} \\
\mathbb{P}(U|\epsilon_n\cdot\hat{\epsilon}_n) &= 1-\mathbb{P}(I|\epsilon_n\cdot\hat{\epsilon}_n) = \frac{1-\pi}{1+\epsilon_n\cdot\hat{\epsilon}_n}
\end{aligned} \tag{3.17}$$

Chapter 4

Linear models for market impact

4.1 Introduction

Understanding how the order flow affects the dynamics of prices in financial markets is of utmost importance, both from a theoretical point of view (why and how prices move?) and for practical/regulatory applications (i.e trading costs, market stability, high frequency trading, "Tobin" taxes, etc.). The availability of massive data sets has triggered a spree of activity around these questions (Hasbrouck, 1988, 1991; Jones et al., 1994; Biais et al., 1995; Dufour and Engle, 2000; Cont et al., 2014; Bacry and Muzy, 2014) and for a review see Bouchaud et al. 2009. One salient (and initially unexpected) stylized fact is the long-memory of the order flow, i.e. the fact that buy/sell orders are extremely persistent, leading to a slowly decaying correlation of the sign of the order imbalance (Bouchaud et al., 2004; Lillo and Farmer, 2004). This immediately leads to two interesting questions: first, why is this so? Is it the result of large "metaorders" being split in small pieces and executed incrementally, or is it due to herding or copy-cat trades, i.e. trades induced by the same external signal or by some traders following suit, hoping that the initial trade was informed about future price movements? Second, how is it possible that a highly predictive order flow impacts the price in such a way that very little predictability is left in the time series of price changes?

Several empirical investigations, as well as order of magnitude comparisons between the typical total size of metaorders and the immediately available liquidity present in the order book, strongly support the "splitting" hypothesis (Lillo et al., 2005; Tóth et al., 2015). Since the metaorder has to be executed over some predefined time scale (typically several days for stocks), the structure of the order flow is expected to be, in a first approximation, independent of the short term dynamics of the price and can be treated as exogenous – see below. The idea then naturally leads to a class of so-called "propagator" models, where the mid-point price p_n (just before the n-th trade at time t_n) can be written as a linear superposition of the impact of all past trades, considered as given, plus noise Bouchaud et al. (2004,

2006):

$$p_n = \sum_{k<n} [G(n-k)\epsilon_k + \eta_k] + p_{-\infty} \tag{4.1}$$

where ϵ_k is the sign of trade at time t_k (± 1 for buy/sell market orders), η_k is a noise term which models any price changes not induced by the trades (e.g. limit orders/cancellations inside the spread, jumps due to news, etc.). The function $G(\ell)$ is called the "propagator" and describes the decay of impact with time. The crucial insight of this formulation is precisely that this impact decay may counteract the positive auto-correlation of the trade signs and eventually lead to a diffusive price dynamics (see Bouchaud et al., 2004, and below). Although highly simplified, the above framework leads to an interesting approximate description of the price dynamics. Still, many features are clearly missing (see Eisler et al., 2012b):

- First, the above formalism posits that all market orders have the same impact, in other words G only depends on $n-k$ and not on n and k separately, which is certainly very crude. For example, some market orders are large enough to induce an immediate price change, and are expected to impact the price more than smaller market orders. One furthermore expects that depending on the specific instant of time and the previous history, the impact of market orders is different.

- Second, limit orders and cancellations should also impact prices, but their effect is only taken into account through the time evolution of $G(\ell)$ itself that phenomenologically describes how the flow of limit orders opposes that of market orders and reverts the impact of past trades.

- Third, the model assumes a *linear* addition of the impact of past trades and neglect any non-linear effects which are known to exist. For example, the total impact of a metaorder of size Q is now well known to grow as $\approx \sqrt{Q}$, a surprising effect that can be traced to non-linearities induced by the deformation of the underlying supply and demand curve (see Tóth et al., 2011; Mastromatteo et al., 2014; Donier et al., 2015).

However, before abandoning the realm of linear models, it is interesting to see how far one can go within the (possibly extended) framework of propagator models, in order to address point 1 and 2 above.

The aim of this chapter is to explore generalised linear propagator models, in the spirit of (Eisler et al., 2012b), with a fully consistent description of the impact of different market events and of the statistics of the order flow. In particular, we investigate in detail two possible generalisations of Equation (4.1) above, where price-changing and non price-changing market orders are treated differently. We show that separating these two types of events already leads to a significant improvement of the predictions of the model, in particular for large tick stocks. We revisit the difference between the "transient impact model" (TIM) and the "history dependent impact model" (HDIM) introduced in (Eisler et al., 2012a), correct some

misprints in that paper, and show that HDIM is always (slightly) better than TIM for small tick stocks, as expected intuitively.

4.2 The one-event propagator model

The propagator model defined by Equation (4.1) above can alternatively be written in its differential form, where instead of the price process we consider the return process, $r_n = p_{n+1} - p_n$:

$$r_n = G(1)\epsilon_n + \sum_{k<n} \mathcal{G}(n-k)\epsilon_k + \eta_n, \qquad \mathcal{G}(\ell) \equiv G(\ell+1) - G(\ell), \tag{4.2}$$

where $G(\ell \leq 0) \equiv 0$. In the following we will call this model Transient Impact Model (as in Eisler et al., 2012a) and we label the predicted values according to the above model with TIM1 where the "1" refers to the fact that one propagator function, $G(\ell)$, characterizes the model.

Empirical results show (Bouchaud et al., 2004; Eisler et al., 2012b) that for small ticks $G(\ell)$ is a decreasing function with time, therefore the kernel $\mathcal{G}(\ell > 0)$ is expected to be a negative function. This means that the impact of a market order is smaller if it follows a sequence of trades of the same sign than if it follows trades of the opposite sign. As stated before, Lillo and Farmer (2004) call this behaviour the "asymmetric liquidity" mechanism: The price impact of a type of order (buy or sell) is inversely related to the probability of its occurrence. We recall that the reason for this mechanism is that liquidity providers tend to pile up their limit orders in opposition of a specific trend of market orders (Bouchaud et al., 2006; Mastromatteo et al., 2014), whereas liquidity takers tend to reduce the impact of their trades by adapting their request of liquidity to the available volume during the execution of their metaorders, as seen in Chapter 3.

4.2.1 Calibration of the model

In order to calibrate the above model, we can measure the empirical response function $\mathcal{R}(\ell) = \mathbb{E}[(p_{n+\ell} - p_n) \cdot \epsilon_n]$ and the empirical correlation function of the order signs $C(\ell) = \mathbb{E}[\epsilon_n \epsilon_{n+\ell}]$. These two functions form a linear system of equations

$$\mathcal{R}(\ell) = \sum_{0<k\leq\ell} G(k)C(\ell-k) + \sum_{k>0} \left[G(k+\ell) - G(k)\right] C(k), \tag{4.3}$$

whose solution is the propagator function $G(\ell)$, for $\ell > 0$.

An alternative method of estimation, which is less sensitive to boundary effects, uses the return process of Equation (4.2), such that the associated response function $\mathcal{S}(\ell) = \mathbb{E}[r_{n+\ell} \cdot \epsilon_n]$ and $C(\ell)$ are related through:

$$\mathcal{S}(\ell) = \sum_{k\geq 0} \mathcal{G}(k)C(k-\ell), \tag{4.4}$$

whose solution represents the values of the kernel $\mathcal{G}(\ell)$. The relation between $\mathcal{R}(\ell)$ and $\mathcal{S}(\ell)$ is:

$$\mathcal{R}(\ell) = \sum_{0 \le i < \ell} \mathcal{S}(i) \tag{4.5}$$

allowing to recover the response function from its differential form.

Once the propagator $G(\ell)$ is calibrated on the data, the model is fully specified by the statistics of the noise η_n. For simplicity, we will assume that η_n has a low-frequency, white noise part of variance D_{LF}, describing any "news" component not captured by the order flow itself, and a fast mean-reverting component of variance D_{HF} describing e.g. high frequency activity inside the spread (affecting the position of the mid-point p_n) or possible errors in the data itself.

4.2.2 Direct tests of the model

Once the model is fully calibrated on data, we examine its performance by considering the prediction of two quantities, namely the negative lag response function and the signature plot. The former is the extension of the price response function, $\mathcal{R}(\ell)$, to $\ell < 0$ values, measuring the correlation between the present sign of the market order and the past price changes:

$$\mathcal{R}(-\ell) = -\sum_{0 < i \le \ell} \mathcal{S}(-i) = -\mathbb{E}[(p_n - p_{n-\ell}) \cdot \epsilon_n]. \tag{4.6}$$

$\mathcal{R}(-\ell)$, with $\ell > 0$, is fully specified by the model, independently of D_{LF} and D_{HF}. Naturally the one propagator model assumes a "rigid" order flow that does not adapt to price changes and leads to:

$$\mathcal{R}^{\text{TIM1}}(-\ell) = -\sum_{0 < i \le \ell} \sum_{k \ge 0} \mathcal{G}(k) C(k+i) < 0. \tag{4.7}$$

where TIM1 reminds us that this is the prediction according to the one propagator model. Empirically, however one expects that the order flow should be adapting to past price changes, and an upward movement of the price should attract more sellers (and vice-versa). In Section 4.4.2 we will compare the prediction of Equation (4.7) to empirical results.

The second prediction of the propagator model concerns the scale-dependent volatility of price changes, or "signature plot", defined as:

$$D(\ell) = \frac{1}{\ell} \mathbb{E}[(p_{n+\ell} - p_n)^2]. \tag{4.8}$$

Using the propagator model, one finds the following exact expression:

$$D^{\text{TIM1}}(\ell) = \frac{1}{\ell} \sum_{0 \le k < \ell} G^2(\ell - k) + \frac{1}{\ell} \sum_{k > 0} [G(\ell + k) - G(k)]^2 + 2\Delta(\ell) + \frac{D_{\text{HF}}}{\ell} + D_{\text{LF}}, \tag{4.9}$$

where $\Delta(\ell)$ is the correlation-induced contribution to the price diffusion:

$$\begin{aligned}\ell\Delta(\ell) = &\sum_{0\le k<h<\ell} G(\ell-k)G(\ell-h)C(h-k)\\ &+\sum_{0\le k<h} \left[G(\ell+k)-G(k)\right]\left[G(\ell+h)-G(h)\right]C(h-k)\\ &+\sum_{0\le k<\ell}\sum_{h>0} G(\ell-k)\left[G(\ell+h)-G(h)\right]C(h+k). \end{aligned} \tag{4.10}$$

Hence, once $G(\ell)$ is known, the signature plot of the price process can be computed and compared with empirical data.

4.2.3 Transient impact vs. history dependent impact

The above model describes trades that impact prices, but with a time dependent, decaying impact function $G(\ell)$. One can in fact interpret the same model slightly differently, by writing as an identity:

$$r_n = G(1)(\epsilon_n - \hat{\epsilon}_n) + \eta_n, \qquad \hat{\epsilon}_n = -\sum_{k>0}\frac{\mathcal{G}(k)}{G(1)}\epsilon_{n-k}. \tag{4.11}$$

This can be read as a model where the deviation of the realized sign ϵ_n from an expected level $\hat{\epsilon}_n$ impacts the price linearly and permanently. If $\hat{\epsilon}_n$ is the best possible predictor of ϵ_n, then the above equation leads by construction to an exact martingale for the price process (i.e. the conditional average of r_n on all past information is zero as in Madhavan et al. 1997). Since the impact depends on the past history of order flow, following (Eisler et al., 2012a), we refer to the model on the left of Equation (4.11) as the History Dependent Impact Model and since only one type of past events is considered in the predictor we label it with HDIM1. When the best predictor is furthermore *linear* in the past order signs (as in the right equation of Equations 4.11), then the TIM1 defined by Equation (4.2) is *equivalent* to the HDIM1, Equation (4.11). We will see below that as soon as one attempts to generalize the propagator model to multiple event types, TIM and HDIM become no longer equivalent.

4.2.4 The DAR process for trade signs

When is the best predictor of the future price a linear combination of past signs, such that TIM and HDIM are equivalent when restricted to one type of market orders only? The answer is that this is true whenever the string of signs is generated by a so-called Discrete Autoregressive (DAR) process (see Jacobs and Lewis, 1978). DAR processes are constructed as follows (our description here lays the ground for the more general MTD models described in the following chapter). The sign at time

t_n is thought of as the "child" of a previous sign $n-k$, where the distance k is a random variable distributed according to a certain discrete distribution λ_k, with:

$$\sum_{k=1}^{\infty} \lambda_k = 1. \tag{4.12}$$

If $\lambda_{k>p} \equiv 0$, the model is called as DAR(p), and involves only p lags. Once the "father" sign is chosen, one postulates that:

$$\begin{aligned} \epsilon_n &= \epsilon_{n-k} \qquad \text{with probability } \rho \\ \epsilon_n &= -\epsilon_{n-k} \qquad \text{with probability } 1-\rho. \end{aligned} \tag{4.13}$$

One can then show that in the stationary state, the signs $\pm$ are equiprobable, and the sign auto-correlation function $C(\ell)$ obeys the following Yule-Walker equation:

$$C(\ell) = (2\rho - 1) \sum_{k=1}^{\infty} \lambda_k C(\ell - k). \tag{4.14}$$

There is therefore a one-to-one relation between λ_k and $C(\ell)$. Note that in the empirical case where $C(\ell)$ decays as a power-law $\ell^{-\gamma}$ with exponent $\gamma < 1$, one can show that $\lambda_k \sim k^{(\gamma-3)/2}$ and $\rho \to 1^-$.

Now, from the very construction of the process, the conditional average of ϵ_t is given by:

$$\hat{\epsilon}_n = (2\rho - 1) \sum_{k=1}^{\infty} \lambda_k \epsilon_{n-k}, \tag{4.15}$$

such that one can indeed identify the HDIM1 with a TIM1, with:

$$\mathcal{G}(\ell) = -(2\rho - 1) G(1) \lambda_\ell. \tag{4.16}$$

When $C(\ell) \overset{\ell \gg 1}{\sim} \ell^{-\gamma}$, one finds as expected $G(\ell) = G(1) + \sum_{k=1}^{\ell} \mathcal{G}(k) \overset{\ell \gg 1}{\sim} \ell^{-\beta}$ with $\beta = (1-\gamma)/2$ (see Bouchaud et al., 2004).

4.3 An extended propagator model with two types of market orders

In order to develop the idea that large market orders (compared to the volume at the opposite best) may have a different impact than small ones, we need to extend the above propagator model to different events π_n at a given time t_n, where we choose here two types of events π_n defined as:

$$\pi_n = \begin{cases} \text{NC} & \text{if } r_n = p_{n+1} - p_n = 0 \\ \text{C} & \text{if } r_n = p_{n+1} - p_n \neq 0. \end{cases} \tag{4.17}$$

We follow the general framework of (Eisler et al., 2012b), but here the definition of price changing events is different. They refer to the total returns until the next transaction and they include the behaviour of liquidity takers and liquidity providers. These different events are discriminated by using indicator variables denoted as $I(\pi_n = \pi)$. The indicator, $I(\pi_n = \pi)$, is 1 if the event at n is of type π and zero otherwise. The time average of the indicator function is the unconditional probability of event π, $\mathbb{P}(\pi) = \mathbb{E}[I(\pi_n = \pi)]$. The usage of the indicator function simplifies the calculation of the conditional expectations, which will be intensively used in the following. For example, if a quantity $X_{\pi_n,n}$ depends on the event type π and the time t_n, then its conditional expectation is

$$\mathbb{E}[X_{\pi_n,n}|\pi_n = \pi] = \frac{\mathbb{E}[X_{\pi_n,n} I(\pi_n = \pi)]}{\mathbb{P}(\pi)}. \tag{4.18}$$

By definition of the indicator function we have that

$$\sum_{\pi} I(\pi_n = \pi) = 1; \qquad \text{and} \qquad \sum_{\pi} X_{\pi,n} I(\pi_n = \pi) = X_{\pi_n,n}. \tag{4.19}$$

4.3.1 Generalisation of the TIM

At this stage, the natural generalisation of the TIM is to write the return process as

$$r_n = \sum_{\pi} G_{\pi}(1) I(\pi_n = \pi)\epsilon_n + \sum_{k<n} \sum_{\pi'} \mathcal{G}_{\pi'}(n-k) I(\pi_k = \pi')\epsilon_k + \eta_n \,,$$
$$\mathcal{G}_{\pi'}(\ell) \equiv G_{\pi'}(\ell+1) - G_{\pi'}(\ell) \,, \tag{4.20}$$

where $\pi = \{\mathrm{NC}, \mathrm{C}\}$. Therefore we call this model TIM2. The resulting price process is a linear superposition of the decaying impact of different (signed) events:

$$p_n = \sum_{k<n} \left[\sum_{\pi} G_{\pi}(n-k) I(\pi_k = \pi)\epsilon_k + \eta_k \right] + p_{-\infty}. \tag{4.21}$$

which can be used to compute the signature plot $D(\ell)$ of model (see Appendix 4.A).

The TIM2 can be calibrated very similarly as the TIM1 above, by noting that the differential response function

$$\mathcal{S}_{\pi}(\ell) = \mathbb{E}[r_{n+\ell} \cdot \epsilon_n | \pi_n = \pi] = \frac{\mathbb{E}[r_{n+\ell} \cdot \epsilon_n I(\pi_n = \pi)]}{\mathbb{P}(\pi)}, \tag{4.22}$$

and the conditional correlation[1] of order signs of a pair of events π_1 and π_2

$$C_{\pi_1,\pi_2}(\ell) = \frac{\mathbb{E}[\epsilon_n I(\pi_n = \pi_1) \cdot \epsilon_{n+\ell} I(\pi_{n+\ell} = \pi_2)]}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)} \tag{4.23}$$

[1]It should be noted that $C_{\pi_1,\pi_2}(\ell)$ is not bounded in $[-1, 1]$ because we normalize the expectation in the numerator by the product $\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)$ rather than by the joint probability $\mathbb{P}(\pi_t = \pi_1, \pi_{t+\ell} = \pi_2)$. This choice is done for speeding the computations and we have verified that the difference is very small.

are related through:

$$S_{\pi_1}(\ell) = \sum_{\pi_2} \mathbb{P}(\pi_2) \sum_{k \geq 0} \mathcal{G}_{\pi_2}(k) C_{\pi_1,\pi_2}(\ell - k). \tag{4.24}$$

We use these quantities to evaluate the conditional response function $\mathcal{R}_\pi(\ell) = \sum_{0 \leq i < \ell} S_\pi(\ell)$, the total impact function $\mathcal{S}(\ell) = \sum_\pi \mathbb{P}(\pi) \mathcal{S}_\pi(\ell)$ and the corresponding response function $\mathcal{R}(\ell)$. As for the TIM1, once we have calibrated $\mathcal{G}_\pi(\ell)$, we compute the predicted values of these response functions for negative lags, $\mathcal{R}_\pi^{\text{TIM2}}(\ell)$ and $\mathcal{R}^{\text{TIM2}}(\ell)$, and the predicted signature plot $D^{\text{TIM2}}(\ell)$.

4.3.2 Generalisation of the HDIM

However, this is not the only generalisation of the propagator model. In fact, the HDIM formulation, Equation (4.11), lends itself to the following, different extension:

$$r_n = \sum_\pi G_\pi(1) I(\pi_n = \pi) \epsilon_n + \sum_{k<n} \sum_{\pi',\pi} \kappa_{\pi',\pi}(n-k) I(\pi_n = \pi) I(\pi_k = \pi') \epsilon_k + \eta_n, \tag{4.25}$$

meaning that the expected sign for an event of type π is a linear regression of past signed events, with an "influence kernel" κ that depends on both the past event type π' and the current event π. This model is the HDIM2. It is clear that TIMs are actually special cases of HDIMs, with the identification:

$$\kappa_{\pi',\pi}(\ell) = \mathcal{G}_{\pi'}(\ell), \qquad \forall \pi, \tag{4.26}$$

i.e. the influence kernel κ does not depend on the present event type π: Only the type of the past event π' matters. The calibration of this model turns out to be more subtle and is discussed in Appendix 4.B (where some errors and misprints appearing in the text of (Eisler et al., 2012a) are corrected).

As above, we may ask when it is justified to consider that the expected sign for an event of type π is a linear regression of past signed events. This requires to generalize the DAR model described in Section 4.2.4 above to a multi-event framework. This will precisely be the aim of the following chapters of this thesis, where we introduce MTDs as a natural generalisation of DAR for order book events.

4.3.3 Tests of the two families of models

Much as for the simple propagator model, one can test the predictive power of the TIM and HDIM framework by comparing the conditional response functions for negative lags $\mathcal{R}_\pi(-\ell) = -\mathbb{E}[(p_n - p_{n-\ell}) \cdot \epsilon_n | \pi_n = \pi]$, $\pi = \{\text{NC}, \text{C}\}$ with empirical data, as well as the signature plot $D(\ell)$ of the price process. In the following section we will investigate the results of the estimation of the above models, and compare these predicted quantities with their empirical determination. Our conclusion, in a nutshell, is that introducing two types of events substantially increases the performance of the propagator models and that – perhaps expectedly – the HDIM fares better than TIM, but only very slightly.

4.4 Empirical calibration

4.4.1 Dataset description

We have analysed the trading activity of the 50 most traded stocks at NYSE and NASDAQ stock exchanges, during the period February 2013 - April 2013 with a total of 63 trading days. We have chosen a wide panel of stocks of different types in order to perform a deep analysis of the two markets. We have considered only the trading activity in the period 9:30–15:30 in all the days under analysis, in order to reduce intraday patterns of activity, such as volume traded, average spread, etc. In particular we try to avoid the trading activity just after the pre-auction and the closing period of the end of the trading day. After trimming the beginning and the end of each trading day, for each stock we concatenate the data on different trading days and carry out our analysis on these time series. The tick size of all the stocks is 0.01 USD.

In Table 4.1 we list the details of the stocks analysed. In particular, we have listed the volatility in basis points, the average daily traded amount in USD, the average bid-ask spread in ticks, and the average tick size-price ratio and we ranked the stocks by these values. We can divide the sample in two different groups, which are the large and small tick stocks. The bid-ask spread of a large tick stock is most of the times equal to one tick, whereas small tick stocks have spreads that are typically a few ticks. We will emphasise in the following sections the very different behaviour of these two groups of stocks. There exist also a number of stocks in the intermediate region between large and small tick stocks, which have the characteristics of both types.

For the period studied, the stock of Apple Inc. (AAPL) had on average a bid-ask spread of 9.14 ticks, clearly making it a small tick stock. On the other hand, Microsoft Inc. (MSFT), with average bid-ask spread being 1.00 ticks is a good candidate for a large tick stock. To illustrate our empirical analysis, we chose to show results for these two stocks in the following.

4.4.2 The one-event propagator model: calibration and tests

The top panels of Figure 4.1 show the estimation of the propagators $G(\ell)$ for MSFT and AAPL. For both large and small tick stocks the decay of the propagator is slow, well above the noise level after 1000 transactions. We can see that for MSFT (as well as for other large tick stocks) the propagator function first increases for a few time lags, and starts decreasing only after that. Thus, the derivative $\mathcal{G}(\ell)$ is positive for small lags, and since $G(1) > 0$ too, the market impact should be reinforced by a sequence of orders on the same side of the order book. This should lead to violations of the market efficiency on short time scales. This is a direct symptom of the inadequacy of the one-event propagator formalism for large ticks: in fact, we will see that the order flow cannot be considered to be independent of the price

Average traded volume (M$)		Volatility (bp)	Average spread (tick)		Average tick size price ratio	
AAPL	1695.13	1.05	PCLN	38.40	MU	11.24
FB	935.17	1.86	GOOG	19.13	BAC	8.38
GOOG	764.73	1.58	AAPL	9.14	INTC	4.74
MSFT	451.80	1.21	NFLX	9.07	CSCO	4.71
AMZN	420.61	1.82	AMZN	8.39	YHOO	4.54
TSLA	373.28	7.20	IDPH	5.12	GE	4.31
XOM	337.04	0.95	V	2.78	EMC	4.12
BAC	324.83	2.20	TSLA	2.73	FB	3.59
BEL	304.55	1.28	GS	2.68	GMZ	3.58
GILD	294.27	1.83	IBM	2.55	PFE	3.58
NFLX	280.16	3.12	BIDU	2.48	MSFT	3.56
C	255.12	1.57	CELG	1.95	ORCL	2.93
CSCO	248.78	1.86	BRK	1.49	WFC	2.76
PCLN	247.75	3.18	MMM	1.42	SBC	2.76
CMCSA	241.75	1.78	CHV	1.36	TSLA	2.60
GE	239.93	1.65	PM	1.30	KO	2.57
QCOM	238.70	1.39	BA	1.27	CMCSA	2.50
JNJ	236.30	0.84	SLB	1.27	GILD	2.35
EBAY	227.06	1.69	AMGN	1.23	MRK	2.32
CMB	221.06	1.40	XOM	1.07	C	2.27
INTC	220.21	1.41	WMT	1.06	BEL	2.13
CHV	218.52	1.16	HD	1.05	CMB	2.04
PFE	217.23	1.37	SBUX	1.04	EBAY	1.85
GMZ	204.79	2.13	PG	1.02	DIS	1.80
SBC	204.01	1.20	EBAY	1.02	SBUX	1.77
IBM	203.51	1.14	PEP	1.02	QCOM	1.51
PG	202.61	1.03	GILD	1.02	HD	1.47
WFC	196.04	1.32	QCOM	1.02	WMT	1.38
V	195.18	1.40	DIS	1.01	PEP	1.31
MU	193.90	3.96	JNJ	1.00	JNJ	1.30
YHOO	187.33	2.15	GMZ	1.00	SLB	1.30
BIDU	185.04	3.03	C	1.00	PG	1.30
KO	172.95	1.38	MRK	1.00	BA	1.27
DIS	163.36	1.41	CMB	1.00	AMGN	1.13
MRK	161.24	1.57	BEL	1.00	XOM	1.12
CELG	157.84	2.04	CMCSA	1.00	PM	1.09
IDPH	151.86	3.33	KO	1.00	BIDU	1.09
BRK	151.23	1.75	WFC	1.00	BRK	0.99
SBUX	150.43	1.66	ORCL	1.00	MMM	0.96
EMC	146.03	1.84	FB	1.00	CELG	0.96
AMGN	141.99	1.72	SBC	1.00	CHV	0.85
PEP	140.86	1.05	EMC	1.00	GS	0.66
WMT	140.10	1.08	PFE	1.00	V	0.63
BA	137.41	1.65	CSCO	1.00	IDPH	0.60
PM	135.10	1.30	YHOO	1.00	NFLX	0.55
SLB	132.91	1.76	INTC	1.00	IBM	0.49
GS	130.26	1.81	MSFT	1.00	AMZN	0.38
ORCL	129.98	1.91	GE	1.00	AAPL	0.22
HD	128.21	1.62	MU	1.00	PCLN	0.14
MMM	125.19	1.47	BAC	1.00	GOOG	0.12

Table 4.1: Details of analysed stocks: rank by average traded daily amount (M$), volatility, rank by average spread over tick size, and by average tick size (bp).

changes in this case. After an uptick move, there is a high probability that the next order will be in the opposite direction, reinstalling price efficiency. This will be well captured by the two-event propagator below.

For AAPL and other small tick stocks we only see a monotone the decay of the propagator. The assumption of a rigid order flow, insensitive to price moves, will be approximately correct in that case (see Tóth et al., 2012), the relaxation of the propagator alleviating the correlation of the signs. We can already anticipate that the two-event propagator framework will be much more beneficial for large tick stocks than for small tick stocks.

The bottom panels of Figure 4.1 show the price response for both positive and negative lags. The dashed lines in the plots show the theoretical prediction of the one-event propagator model by using the estimated kernels. In the case of MSFT the measured response function for negative lags ℓ is well above the prediction of the propagator model (solid line), that, as we discussed assumes a rigid order flow not depending on price changes. As anticipated above, this means that in the data there exists an additional anti-correlation between past returns and the subsequent order flow, which is not captured by the model. A similar, though much weaker deviation can be seen in the case of AAPL. In general, this effect is very pronounced in the case of large tick stocks, whereas in the case of small tick stocks it exists but is much weaker. In fact, in Figure 4.2 we plot the ratio $[\mathcal{R}(-\ell) - \mathcal{R}^{\mathrm{TIM1}}(-\ell)]/\sigma$ for $\ell = 1, 10, 100$, σ being the volatility per trade, by ranking the stocks in the x-axis by the average spread. We observe that for small tick stocks (left part of the plot) the difference is relatively small, while for large tick stocks (right part of the plot) the prediction error on the negative lag response of the TIM1 becomes quite large, especially for large lags ℓ.

Turning now to the signature plot $D(\ell)$, we see in Figure 4.3 that small tick and large tick stocks behave very differently. For small tick stocks, we see that $D(\ell)$ increases with ℓ as soon as $\ell \geq 3$, corresponding to a "trend-like" behaviour. The decreasing behaviour of $D(\ell)$ for smaller lags corresponds to high frequency activity with the spread, leading to a minimum in $D(\ell)$. For large tick stocks this is absent and one finds "mean-reverting" behaviour, with a steadily decreasing signature plot. The prediction of the one-event propagator model fares quite well at accounting for the trending behaviour of small tick stocks, provided the two extra fitting parameters D_{LF} and D_{HF} are optimized with OLS in order to minimize the distance between the empirical and the theoretical curves of the model. We note for example that choosing $D_{\mathrm{LF}} = 0$ would underestimate (in the case of AAPL) the long-term volatility by a factor of two. For large tick stocks, however, the mean-reverting behaviour is completely missed. We now turn to propagator models that distinguish between price-changing and non price-changing market orders, and see how the situation for large tick stocks indeed greatly improves.

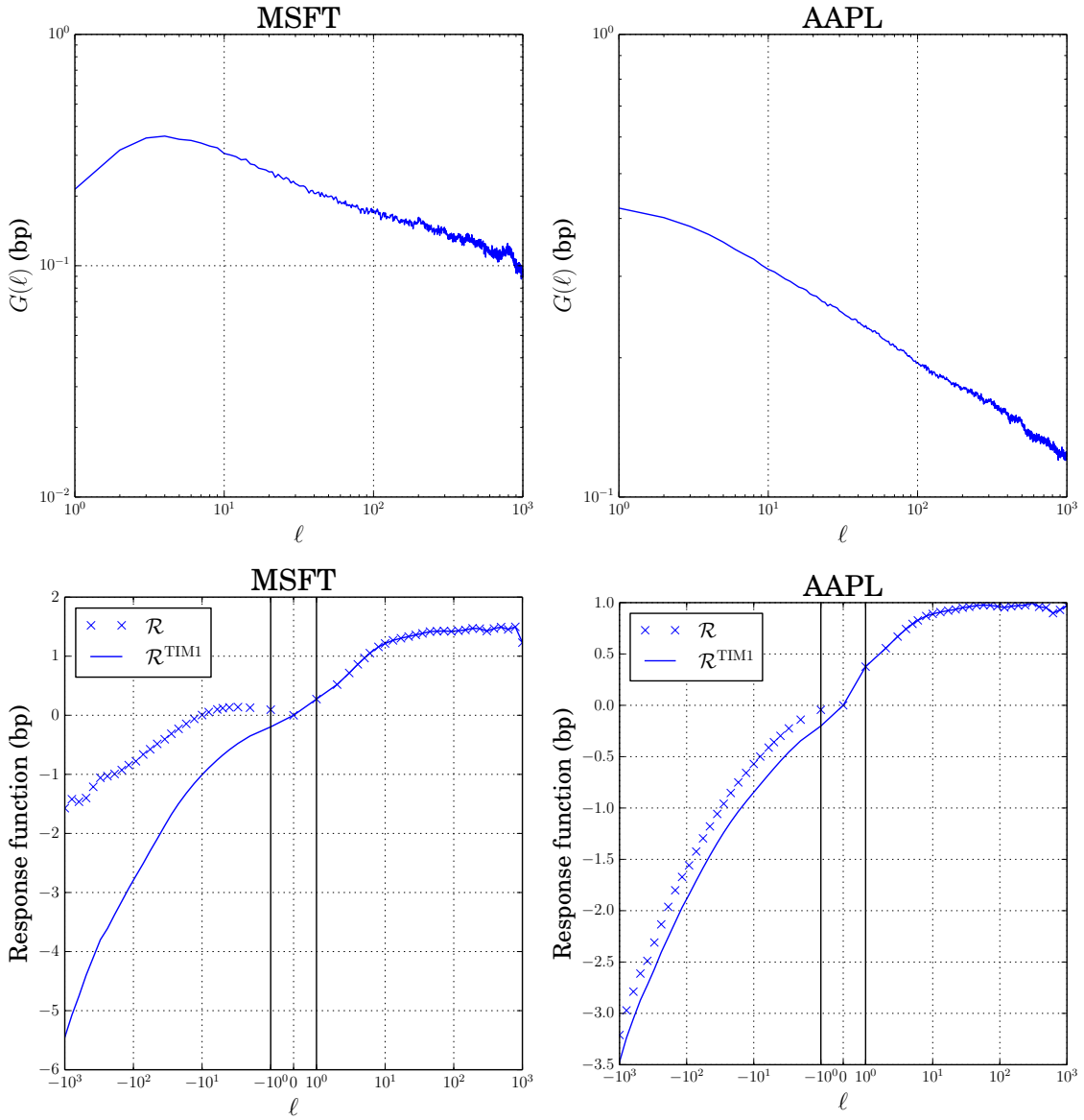

Figure 4.1: (Top panels) The estimated propagator functions for MSFT and AAPL. (Bottom panels) Response functions for positive and negative lags (blue markers) and the theoretical prediction of the estimated TIM1 (solid lines) for MSFT and AAPL. The scale for ℓ close to zero and bounded by the two vertical lines is linear, whereas outside this region the scale is logarithmic.

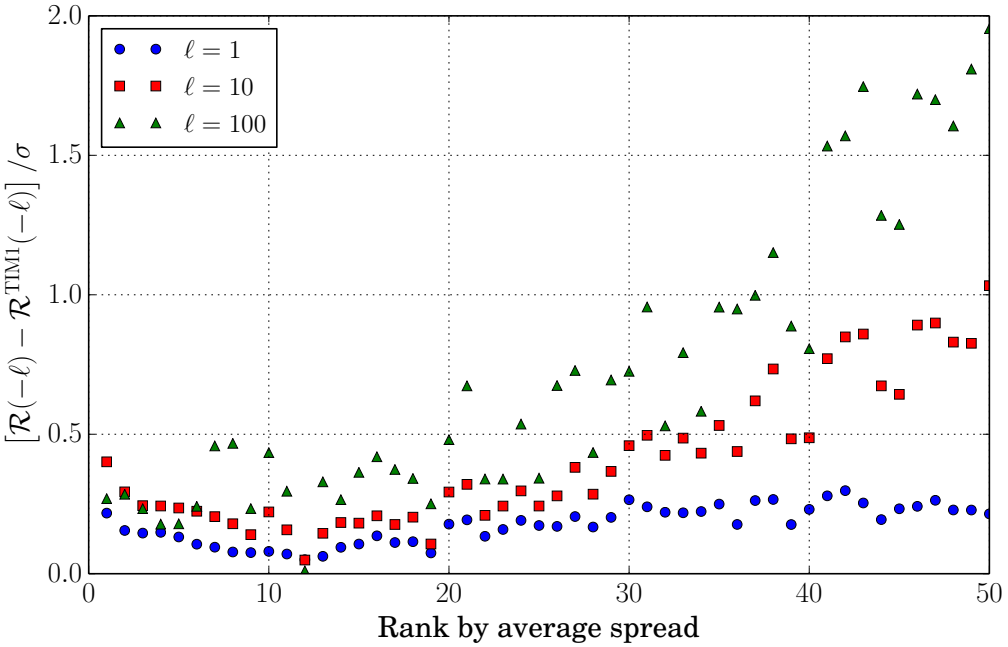

Figure 4.2: Deviation from the TIM1 theoretical prediction of the response function at negative lags for 50 stocks under analysis ranked by the average spread. Small tick stocks are in the left side of the figure, whereas large tick stocks are in the right side

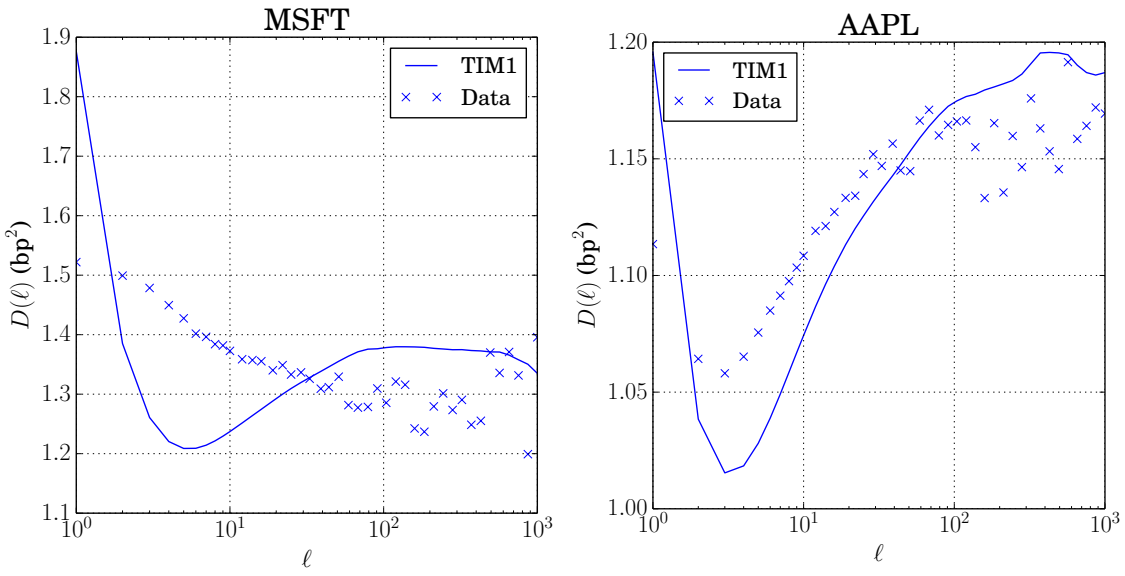

Figure 4.3: The empirical signature plot $D(\ell)$ and the theoretical curves of the estimated TIM1 for MSFT ($D_{\mathrm{LF}} = 0.65$ and $D_{\mathrm{HF}} = 1.13$) and AAPL ($D_{\mathrm{LF}} = 0.58$ and $D_{\mathrm{HF}} = 0.46$).

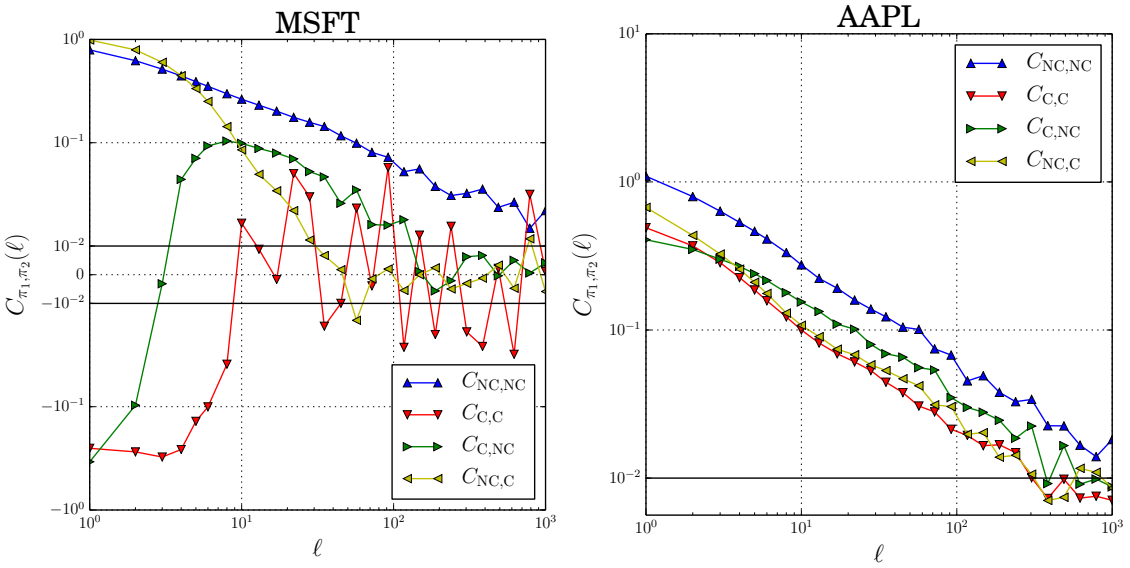

Figure 4.4: Conditional correlations function of signed events $C_{\pi_1,\pi_2}(\ell)$ measured on MSFT and APPL data. Note that the first subscript corresponds to the event that happened first chronologically. The scale for values of the correlations close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

4.4.3 Two-event propagator model

The aim of this section is to show that an extended propagator model allows us to reproduce satisfactorily the additional anti-correlations between past returns and subsequent order signs (revealed by the discrepancy between $\mathcal{R}(\ell < 0)$ and $\mathcal{R}^{\mathrm{TIM1}}(\ell < 0)$) by including an implicit coupling between past returns and order flow. We will also require that the signature plot $D(\ell)$ is correctly accounted for, in particular for large tick stocks.

The extended version of the propagator model with two events $\pi = \{\mathrm{NC}, \mathrm{C}\}$ can follow two routes, as discussed above. One is the TIM2, which can be estimated much as the one-event model, by solving the linear system of Equation (4.24). The second is the HDIM2, whose estimation involves determining the influence kernels $\kappa_{\pi_1,\pi_2}(\ell)$ for $\pi_2 = \mathrm{C}$, because $\kappa_{\pi_1,\mathrm{NC}}(\ell) = 0$ by construction. The calibration requires estimating three-point correlation functions or approximating them in terms of two-point correlations – as detailed in Section 4.4.3 we will follow the latter approximation. Thus, the correlation $C_{\pi_1,\pi_2}(\ell)$ of the different signed events, defined in Equation (4.23) is an important input of the calibration for both generalised linear models. Note that the first subscript corresponds to the event that happened first chronologically. We start by showing its empirical estimation for the two typical stocks (see Figure 4.4).

For AAPL, all auto-correlation and cross-correlation functions have almost the same power-law decay and they are all positive. This is expected since C and NC events are not radically different for small tick stocks. Note that the unconditional

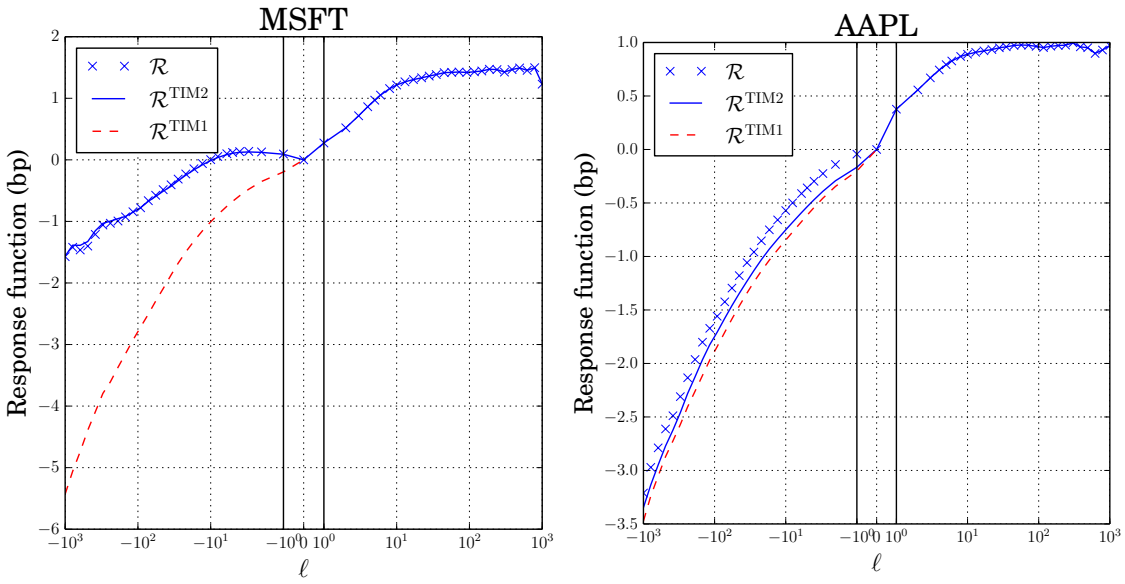

Figure 4.5: Conditional response function for positive and negative lags (blue markers) and the theoretical prediction of the TIM2 calibrated on MSFT and APPL (solid lines). Theoretical prediction of response function for negative lags for TIM1 (red dashed lines). The scale for ℓ close to zero and bounded by vertical solid lines is linear, whereas outside this region the scale is logarithmic.

probability of price changing market orders is $\mathbb{P}(\pi = \mathrm{C}) = 0.69$. Correlation functions look similar for other small tick stocks too.

For MSFT the curves reveal a different behaviour. For example the $C_{\mathrm{NC,NC}}$ auto-correlation has the familiar power-law shape possibly due to order splitting. The $C_{\mathrm{NC,C}}$ correlation is also positive but decays faster. Note that it starts at $C_{\mathrm{NC,C}}(1) \approx 0.95$, which means that a C order immediately following a NC order is in the same direction with very high probability. This describes NC orders that leave a relatively small quantity at the best offer, which is then immediately "eaten" by the next market orders. Its relatively fast decay suggests that agents splitting their metaorders avoid being aggressive and nearly only send NC orders. The other two correlations $C_{\mathrm{C,C}}$ and $C_{\mathrm{C,NC}}$ both start negative and capture the effect we are interested in: After a price changing event, it is highly likely that the subsequent order flow (either C or NC) will be in the other direction. Note however that $\mathbb{P}(\pi = \mathrm{C}) = 0.08$ and that it is exceedingly rare to observe a succession of two C events separated by a small lag. This type of behaviour is the one that can be seen in general for large tick stocks.

Tests on the TIM2

The estimation procedure involves the empirical determination of the response function for positive lags, and allows us to calculate the theoretical prediction of the response function for negative lags, as well as the signature plot.

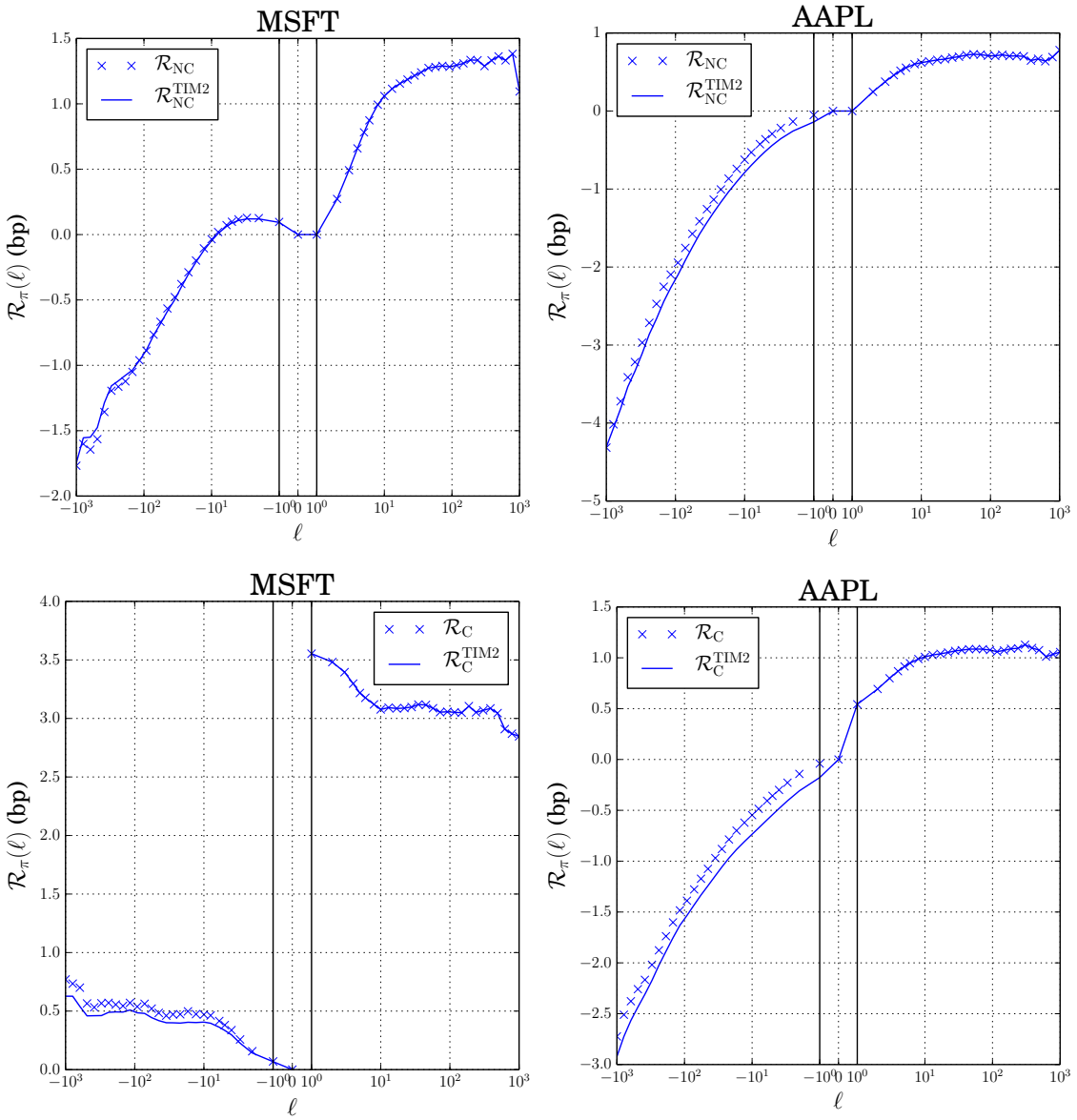

Figure 4.6: (Top panels) Conditional response function for NC events (blue markers) and the theoretical predictions of the TIM2 (solid lines). (Bottom panels) Conditional response function for C events (blue markers) and the theoretical predictions of the TIM2 (solid lines). Left: MSFT (large tick), Right: APPL (small tick). The scale for ℓ close to zero and bounded by vertical solid lines is linear, whereas outside this region the scale is logarithmic.

Figure 4.5 shows the empirical response function for positive lags $\mathcal{R}(\ell > 0)$ and negative lags $\mathcal{R}(\ell < 0)$, together with the predicted response function $\mathcal{R}^{\text{TIM2}}(\ell)$, according to the calibrated TIM2. In the case of large tick stocks the empirical curves are perfectly reproduced, whereas for small tick stocks some little deviation still persists. The improvement with respect to the TIM1 is quite remarkable. This can be seen the from comparison of the prediction of the response function for negative lags of the TIM1, $\mathcal{R}^{\text{TIM1}}(\ell < 0)$, also plotted in Figure 4.5.

Let us now discuss the observed response functions for positive lags, and the resulting calibrated propagators for small tick stocks, as for AAPL, shown in Figures 4.6 and 4.7 (right panels). The conditional response function $\mathcal{R}_{\text{C}}(\ell)$ after an event of type $\pi = \text{C}$ is a rigid shift of the $\mathcal{R}_{\text{NC}}(\ell)$ curve. The reaction of market agents to the two types of events is therefore very similar. The shift indeed is due to the very definition of event types, that leads to a non-zero value of $\mathcal{R}_{\text{C}}(\ell = 1)$, comparable to the average spread. Turning now to the conditional response function for negative lags, we observe a small deviation between the model and the empirical data: There exists an additional anti-correlation between past returns and future order signs which is not captured by the model. The curves $\mathcal{R}_{\text{NC}}(-\ell)$ and $\mathcal{R}_{\text{C}}(-\ell)$ behave in similar way, but in the latter case the anti-correlation is stronger than in the former case.

The propagator functions $G_{\text{C}}(\ell)$ can be fit by a power-law, but the G_{NC} curves are non monotonic (see Figure 4.7). Note that, as a result of the non-trivial structure of the correlation, the calibration of the TIM2 leads to $G_{\text{NC}}(\ell = 1) > 0$. This is inconsistent with the interpretation of the model – which would require $G_{\text{NC}}(\ell = 1) = 0$ – and shows the theoretical limitations of the TIM framework. In the case of the HDIM framework, by construction, we have that $\kappa_{\pi\text{NC}}(\ell = 1) = 0$.

The results of the estimation of the model for large tick stocks are completely different. Figures 4.6 and 4.7 (left panels) show the results for MSFT. The $\mathcal{R}_{\text{NC}}(\ell)$ curve is a positive and increasing function which starts, as expected, from zero and reaches a plateau for large lags. The $\mathcal{R}_{\text{C}}(\ell)$ curve starts from the value of the spread in basis point and slightly decreases, which means that the reaction of the market after price change events consists in a mean reversion of the price. For negative lags, the curve $\mathcal{R}_{\text{NC}}(-\ell)$ shows that if an event occurs that does not change the price, then for small lags the past returns are on average anti-correlated with the present order sign. The case of the $\mathcal{R}_{\text{C}}(-\ell)$ is quite interesting, because it shows that if a price changing event occurs, then the past returns are on average anti-correlated with the present order sign.

The propagator functions G_π are almost constant with different values: G_{C} is equal to the spread, whereas G_{NC} is equal to zero. The fact that the two propagators are constant means that the price process in Equation (4.21) is simply a sum of non-zero price changes, all equal to the spread, and for which the impact is permanent. Therefore, as noted in (Eisler et al., 2012b) the dynamics of the price is completely determined by the sequence of random variables $\{(\epsilon_t, \pi_t)\}_{t\in\mathbb{N}}$, and the temporal structure of their correlations. More precisely, if spread fluctuations can

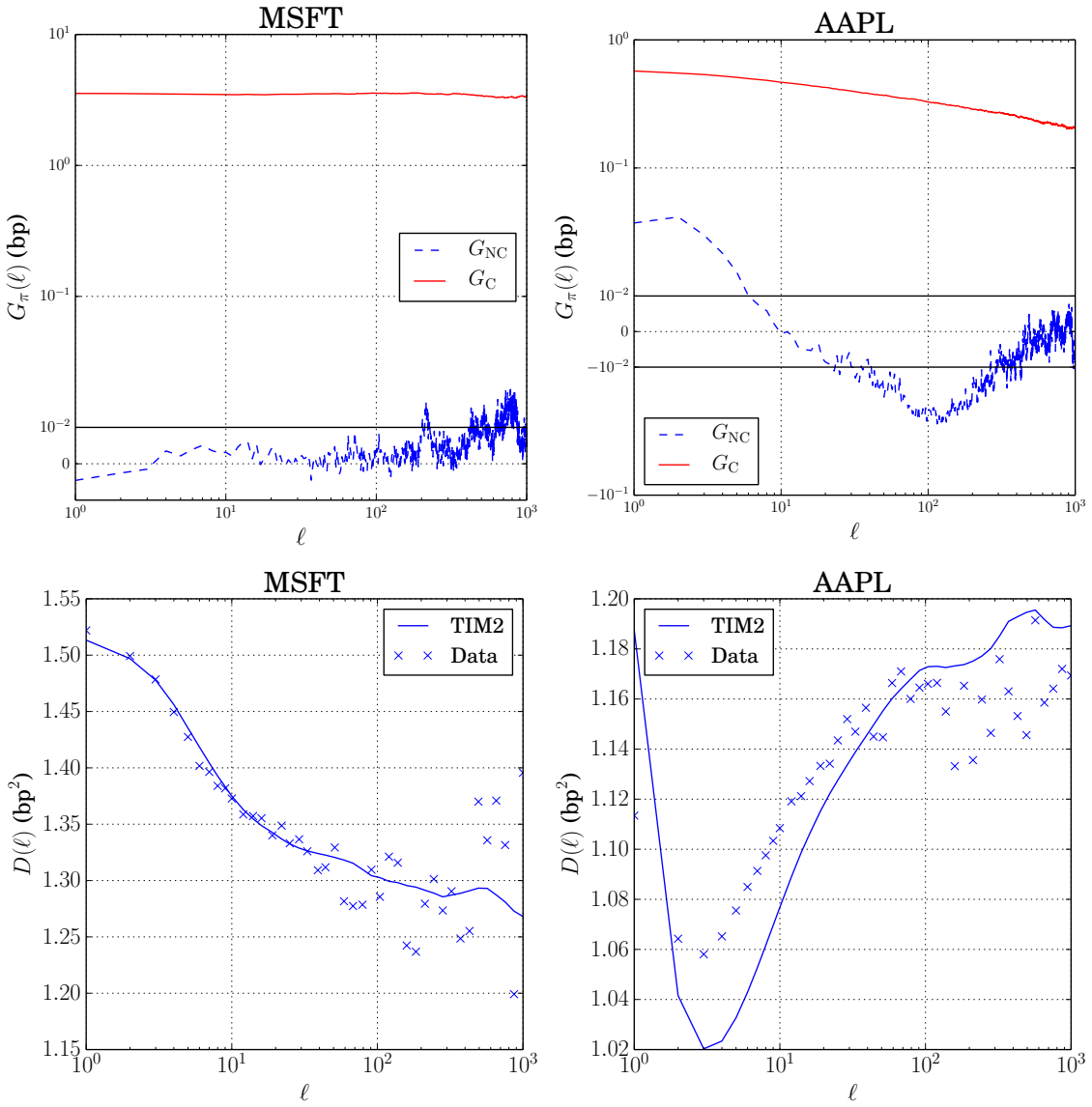

Figure 4.7: (Top panels) The estimated propagator functions $G_\pi(\ell)$ of TIM2. The scale for ℓ close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic. (Bottom panels) Signature plots, empirical and predicted by the calibrated TIM2. Left: MSFT with $D_{\mathrm{LF}} = 0.54$ and $D_{\mathrm{HF}} = 0$, Right: APPL with $D_{\mathrm{LF}} = 0.56$ and $D_{\mathrm{HF}} = 0.41$.

be neglected, TIM2 lead to the following simple predictions:

$$\begin{aligned}\mathcal{R}_\pi^{\mathrm{TIM2}}(\ell>0) &\approx \sum_{0\le k<\ell}\sum_{\pi_1}\mathbb{P}(\pi_1)G_{\pi_1}(1)C_{\pi,\pi_1}(k)\\ &= G_{\mathrm{C}}(1)\left[1+\sum_{0<k<\ell}\mathbb{P}(\mathrm{C})C_{\pi,\mathrm{C}}(k)\right],\\ \mathcal{R}_\pi^{\mathrm{TIM2}}(\ell<0) &\approx -\sum_{0<k\le\ell}\sum_{\pi_1}\mathbb{P}(\pi_1)G_{\pi_1}(1)C_{\pi_1,\pi}(k)\\ &= -G_{\mathrm{C}}(1)\sum_{0<k\le\ell}\mathbb{P}(\mathrm{C})C_{\mathrm{C},\pi}(k).\end{aligned} \tag{4.27}$$

and:

$$\begin{aligned}D^{\mathrm{TIM2}}(\ell) &\approx D_{\mathrm{LF}}+\sum_\pi G_\pi(1)^2\mathbb{P}(\pi)\\ &\quad+\frac{2}{\ell}\sum_{0\le k<h<\ell}\sum_{\pi_1,\pi_2}\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)G_{\pi_1}(1)G_{\pi_2}(1)C_{\pi_1,\pi_2}(h-k)\\ &= D_{\mathrm{LF}}+G_{\mathrm{C}}(1)^2\mathbb{P}(\mathrm{C})+2\frac{G_{\mathrm{C}}(1)^2}{\ell}\sum_{0\le k<h<\ell}\mathbb{P}(\mathrm{C})^2C_{\mathrm{C},\mathrm{C}}(h-k).\end{aligned} \tag{4.28}$$

Note that the both the empirical response for negative lags and the signature plot are now perfectly reproduced. The improvement from the TIM1 is quite remarkable.

Tests on the HDIM2

The calibration of the HDIM2 model requires the determination of the influence matrix κ_{π_1,π_2}, which can be done from the empirical knowledge of the response matrices $\mathcal{S}_{\pi_1,\pi_2}(\ell)$ since

$$\mathcal{S}_{\pi_1,\pi_2}(\ell)=G_{\pi_2}(1)C_{\pi_1,\pi_2}(\ell)+\sum_{k>0}\sum_\pi\mathbb{P}(\pi)\kappa_{\pi,\pi_2}(k)C_{\pi,\pi_1,\pi_2}(k,\ell), \tag{4.29}$$

where

$$\begin{aligned}\mathcal{S}_{\pi_1,\pi_2}(\ell) &= \frac{\mathbb{E}[I(\pi_{n-\ell}=\pi_1)\epsilon_{n-\ell}\cdot I(\pi_n=\pi_2)r_n]}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)},\\ C_{\pi,\pi_1,\pi_2}(h,\ell) &= \frac{\mathbb{E}[I(\pi_{n-h}=\pi)\epsilon_{n-h}\cdot I(\pi_{n-\ell}=\pi_1)\epsilon_{n-\ell}\cdot I(\pi_n=\pi_2)]}{\mathbb{P}(\pi)\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)}.\end{aligned} \tag{4.30}$$

Actually the previous equation is not convenient to be used for the estimation of the model, because it includes the empirical determination of the three-point correlation functions $C_{\pi,\pi_1,\pi_2}(h,\ell)$. Therefore, Eisler et al. (2012a) employed a Gaussian assumption which leads to the factorization of the three-point correlation functions in terms of two-point correlation functions:

$$\mathcal{S}_{\pi_1,\pi_2}(\ell)\approx G_{\pi_2}(1)C_{\pi_1,\pi_2}(\ell)+\sum_{\substack{k>0\\k\neq\ell}}\sum_\pi\mathbb{P}(\pi)\kappa_{\pi,\pi_2}(k)C_{\pi,\pi_1}(k-\ell)+\kappa_{\pi_1,\pi_2}(\ell)[\Pi_{\pi_1,\pi_2}(\ell)+1],$$

where

$$\Pi_{\pi_1,\pi_2}(\ell) = \frac{\mathbb{E}[I(\pi_{n-\ell} = \pi_1) \cdot I(\pi_n = \pi_2)]}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)} - 1. \tag{4.31}$$

The resulting formula for the signature plot $D^{\text{HDIM2}}(\ell)$ is considerably more complicated. We report it for completeness in Appendix 4.B.

On purely theoretical grounds, HDIMs are better founded than TIMs and we have extended the above analysis to HDIMs as well. In the case of large tick stocks, there is no gain over the TIM framework since the influence kernels are found to be extremely small. Any gain is therefore only possible for small tick stocks. We show the empirical determination of the two influence kernels $\kappa_{\pi_1,\text{C}}(\ell)$ as well as the resulting predicted response $\mathcal{R}_\pi^{\text{HDIM2}}(\ell)$ for AAPL in Figure 4.8. As can be noted, the estimated kernels differ whether the sequence of events which precede the price-changing trade is composed of price-changing or non price-changing orders. We can argue that Equation (4.26) – which neglects the role of the realised event – is too restrictive. It is worth to comment that, when statistically different from zero, the influence kernel $\kappa_{C,C}$ is negative. Then, a sequence of price-changing orders on the same side of the final C trade is going to impact the market less than a C order preceded by a sequence of price-changing events of the opposite sign. Thus we see the same asymmetric liquidity mechanism described in (Lillo and Farmer, 2004). As a sole difference with the picture described in Section 4.2, the influence kernel $\kappa_{NC,C}$ is positive for the very last NC event occurring before a price-changing event. This implies that the impact of the C market order is larger if it follows a sequence of NC trades whose last event occurs on the same side of the C event.

We see some further improvement over the TIM2 for the conditional response functions at negative lags. It seems that HDIM2 performs slightly better than TIM2 in capturing the excess anti-correlation measured from the data between past returns and future order signs. We also observe an improvement – albeit in a marginal way – for the signature plot in Figure 4.8. We recall here that in the 6-event extension of the propagator model considered in (Eisler et al., 2012a), HDIMs appeared to fare slightly worse than TIMs for small tick stocks, for a reason that is still not well understood, and that would deserve further scrutiny.

4.5 Discussion and partial conclusions

The above study attempts to build the most accurate linear model of price dynamics based on the only observation of market orders.

We have seen that treating all market orders on the same footing, as in the first version of the propagator model, leads to systematic discrepancies that increase with the tick size. For large tick sizes, the predictions of this simple framework are qualitatively erroneous, both for the price response at negative lags and for the diffusion properties of the price. This can be traced to the inability of the model to describe the feedback of price changes on the order flow, which is strong

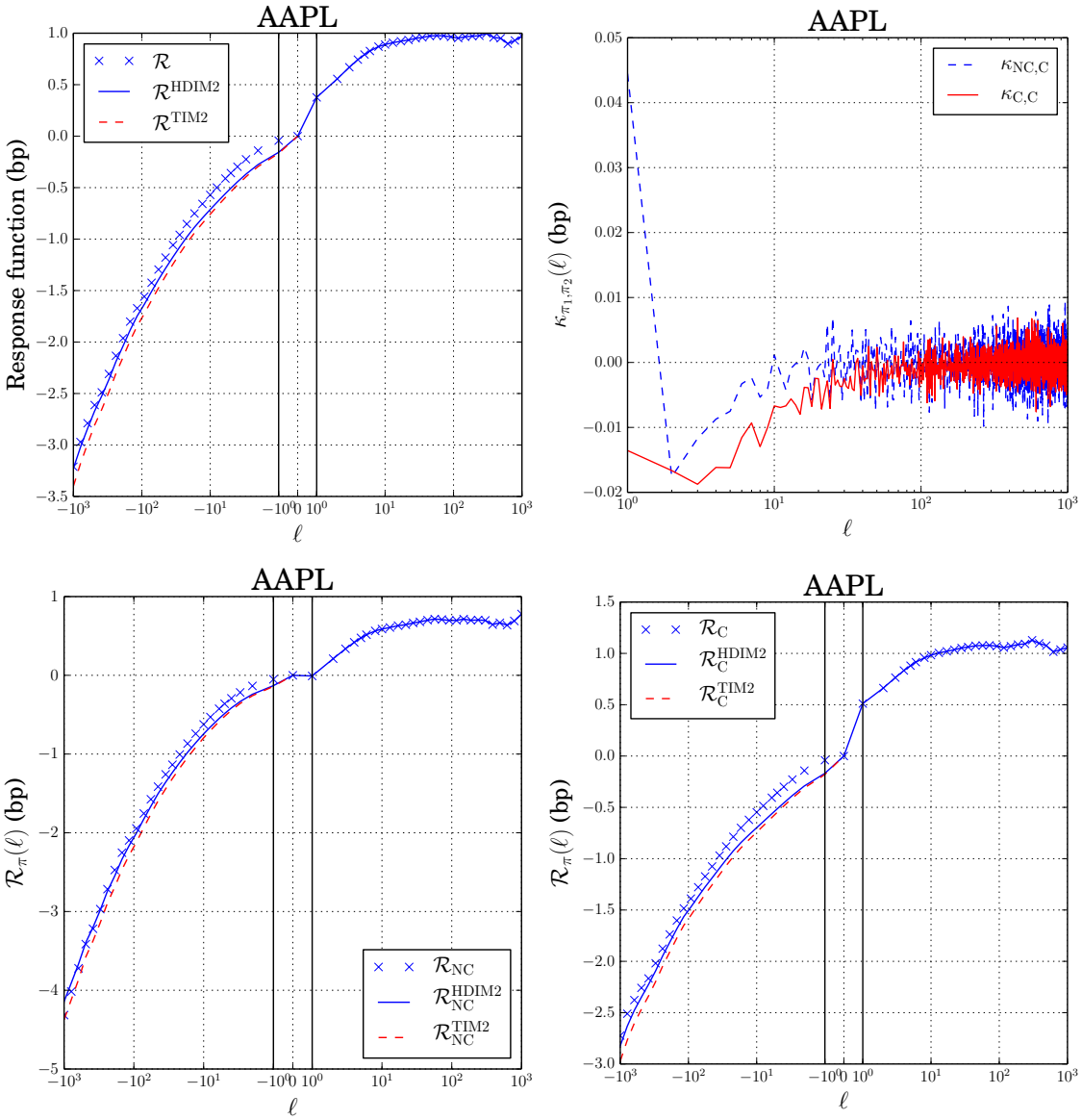

Figure 4.8: (Top left) Response function (blue markers), theoretical prediction of the HDIM2 (blue solid line), and of the TIM2 (red dashed lines) for AAPL. Top right. Influence kernels κ_{π_1,π_2} of the HDIM2 calibrated on AAPL. (Bottom panels) Conditional response functions (blue markers), theoretical predictions of the HDIM2 (blue solid line), and of the TIM2 (red dashed lines) calibrated on AAPL data. The scale for ℓ close to zero and bounded by vertical solid lines is linear, whereas outside this region the scale is logarithmic.

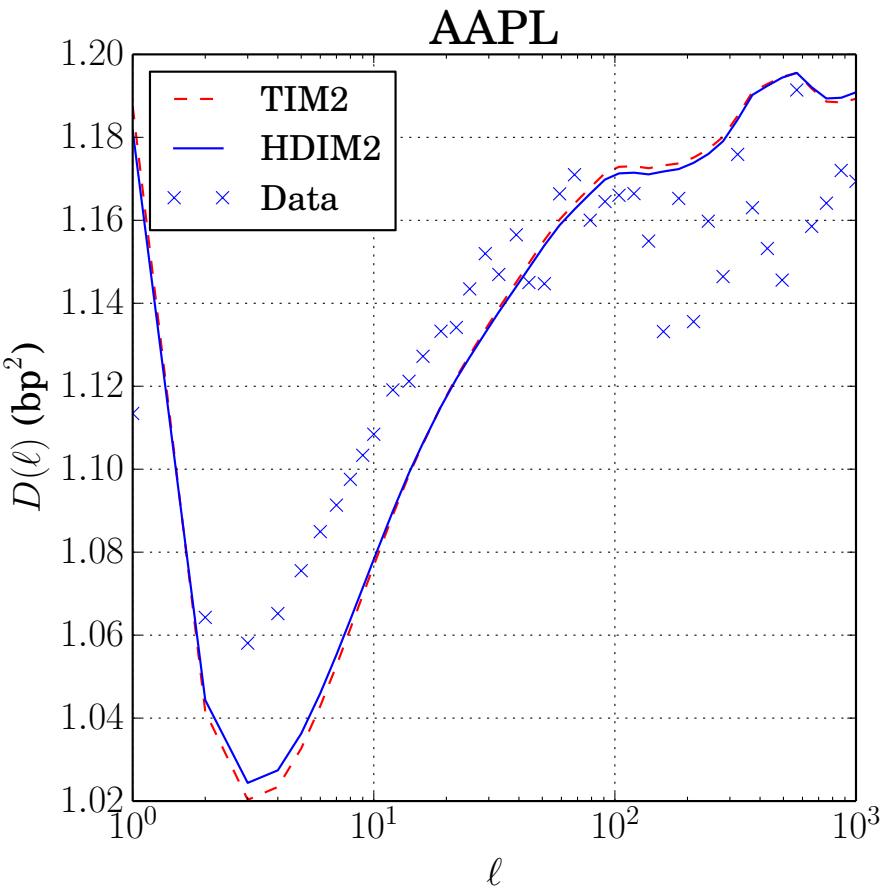

Figure 4.9: Signature plots for AAPL, namely empirical, predicted by the HDIM2 (with $D_{\mathrm{LF}} = 0.6$ and $D_{\mathrm{HF}} = 0.39$), and by the TIM2.

for large tick stocks. Generalizing the model to two types of market orders, those which leave the price unchanged and those which lead to an immediate price change, considerably improves the predictive power of the model, in particular for large ticks for which the above inadequacy almost entirely disappears, leading to a remarkable agreement between the model's predictions and empirical data. We have also seen that, although better justified theoretically, the "history dependent'' impact models (HDIM) fare only slightly better than the "transient'' impact models (TIM) when only two event types are considered.

Still, we are left with two important questions about the order flow itself, which we considered "rigid" in the above formalism, in the sense that it is entirely described by its correlation structure and does not explicitly react to past events (at variance with the price itself). It would be desirable to develop a more dynamic description of the order flow, for at least two reasons. One is that linear models are best justified in a context where the best predictor of the order flow is itself linear, as is the case of DAR processes for the sign of market orders. We therefore need to generalize DAR processes to a multi-event context, and see how well the corresponding so-called MTD models account for the statistics of the order flow, i.e. the string of $\{(-1, \mathrm{C})\}$, $\{(-1, \mathrm{NC})\}$, $\{(+1, \mathrm{NC})\}$, $\{(+1, \mathrm{C})\}$ events. The second reason is that the "true" impact of an additional market order, not present in the past time series, should include the mechanical contributions captured by the TIMs or HDIMs, but also the possible change of the order flow itself due to an extra order in the market, an effect clearly not captured by our assumption of a rigid order flow. We thus need to define and calibrate the equivalent of the influence kernels defined above, but for the order flow itself. This is what we do in the following chapters.

Appendix

4.A Diffusion properties of TIMs

The exact expression of the diffusive curve $D(\ell)$, given in (Eisler et al., 2012a), is:

$$\begin{aligned} D^{\mathrm{TIM}}(\ell)\ell = {} & D_{\mathrm{LF}}\ell + D_{\mathrm{HF}} + \sum_{0\le k<\ell}\sum_{\pi}\mathbb{P}(\pi)G_{\pi}(\ell-k)^2 \\ & + \sum_{k>0}\sum_{\pi}\mathbb{P}(\pi)\left[G_{\pi}(\ell+k)-G_{\pi}(k)\right]^2 \\ & + 2\sum_{0\le k<h<\ell}\sum_{\pi_1,\pi_2}\mathbb{P}(\pi_{1,2})G_{\pi_1}(\ell-k)G_{\pi_2}(\ell-h)C_{\pi_1,\pi_2}(h-k) \\ & + 2\sum_{0\le k<h}\sum_{\pi_1,\pi_2}\mathbb{P}(\pi_{1,2})\left[G_{\pi_1}(\ell+k)-G_{\pi_1}(k)\right] \\ & \qquad\qquad \cdot\left[G_{\pi_2}(\ell+h)-G_{\pi_2}(h)\right]C_{\pi_2,\pi_1}(h-k) \\ & + 2\sum_{0\le k<\ell}\sum_{h>0}\sum_{\pi_1,\pi_2}\mathbb{P}(\pi_{1,2})G_{\pi_1}(\ell-k)\left[G_{\pi_2}(\ell+h)-G_{\pi_2}(h)\right]C_{\pi_2,\pi_1}(h+k), \end{aligned} \tag{4.32}$$

where $\mathbb{P}(\pi_{i,\dots,j}) = \mathbb{P}(\pi_i)\cdots\mathbb{P}(\pi_j)$.

4.B Diffusion properties of HDIMs

Knowing the κ_{π_1,π_2}'s and using the factorization of three-point and four-point correlations in terms of two-point correlations, one can finally estimate the diffusion curve, which is given by the following approximate equation:

$$
\begin{aligned}
D^{\mathrm{HDIM}}(\ell)\ell \approx \Bigg[& D_{\mathrm{LF}} + \frac{D_{\mathrm{HF}}}{\ell} + \sum_{\pi_1} G_{\pi_1}(1)^2 \mathbb{P}(\pi_1) \\
& + \sum_{k>0} \sum_{\pi_1,\pi_2} \mathbb{P}(\pi_{1,2}) \kappa_{\pi_1,\pi_2}(k)^2 [\Pi_{\pi_1,\pi_2}(k) + 1] \\
& + 2 \sum_{0<k<h} \sum_{\pi_1,\pi_2,\pi_3} \mathbb{P}(\pi_{1,2,3}) \kappa_{\pi_1,\pi_3}(h) \kappa_{\pi_2,\pi_3}(k) C_{\pi_1,\pi_2}(h-k) \\
& + 2 \sum_{k>0} \sum_{\pi_1,\pi_2} G_{\pi_2}(1) \kappa_{\pi_1,\pi_2}(k) C_{\pi_1,\pi_2}(k) \Bigg] \ell \\
& + 2 \sum_{0<k<\ell} \sum_{\pi_1,\pi_2} (\ell - k) \mathbb{P}(\pi_{1,2}) G_{\pi_1}(1) G_{\pi_2}(1) C_{\pi_1,\pi_2}(k) \\
& + 2 \sum_{0<k<\ell} \sum_{i>0} \sum_{\pi_1,\pi_2,\pi_3} (\ell - k) \mathbb{P}(\pi_{1,2,3}) G_{\pi_1}(1) \kappa_{\pi_2,\pi_3}(i) C_{\pi_1,\pi_2}(k+i) \\
& + 2 \sum_{0<k<\ell} \sum_{i>0} \sum_{\pi_1,\pi_2,\pi_3} (\ell - k) \mathbb{P}(\pi_{1,2,3}) G_{\pi_1}(1) \kappa_{\pi_2,\pi_3}(i) C_{\pi_1,\pi_2}(k-i) \\
& \qquad i \neq k \\
& + 2 \sum_{0<k<\ell} \sum_{\pi_1,\pi_2} (\ell - k) \mathbb{P}(\pi_{1,2}) G_{\pi_1}(1) \kappa_{\pi_1,\pi_2}(k) [\Pi_{\pi_1,\pi_2}(k) + 1] \\
& + 2 \sum_{0<k<\ell} \sum_{i,j>0} \sum_{\pi_3,\pi_4} (\ell - k) \mathbb{P}(\pi_{1,2,3,4}) \kappa_{\pi_1,\pi_2}(i) \kappa_{\pi_3,\pi_4}(j) \\
& \qquad j \neq k \\
& \qquad \cdot C_{\pi_1,\pi_3}(k+i-j) [\Pi_{\pi_2,\pi_4}(k) + 1] \\
& + 2 \sum_{0<k<\ell} \sum_{i>0} \sum_{\pi_1,\pi_2,\pi_3} (\ell - k) \mathbb{P}(\pi_{1,2,3}) \kappa_{\pi_1,\pi_2}(i) \kappa_{\pi_2,\pi_3}(k) C_{\pi_1,\pi_2}(i), \qquad (4.33)
\end{aligned}
$$

where $\mathbb{P}(\pi_{i,\dots,j}) = \mathbb{P}(\pi_i) \cdots \mathbb{P}(\pi_j)$ and

$$
\Pi_{\pi_1,\pi_2}(\ell) = \frac{\mathbb{E}[I(\pi_{n-\ell} = \pi_1) \cdot I(\pi_n = \pi_2)]}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)} - 1. \qquad (4.34)
$$

Chapter 5

The Mixture Transition Distribution model

5.1 Introduction

In the previous chapter, we discussed the differences and similarities between two linear models describing the impact of order flow on prices, namely the Transient Impact Model (TIM) and the History Dependent Impact Model (HDIM). In these models, the sign of the order flow is considered to be an exogenous, time correlated process that affects price dynamics either through a "propagator", i.e. a linear combination of past values (TIM) or via a "surprise" mechanism, i.e. the deviation between the realised order flow and its expected level (HDIM). In reality, however, order flow is not exogenous and is itself affected by the past history of price. As shown before, we partly overcame this issue by enhancing the description of the order flow to account for price changing events and non price changing events, in the spirit of (Eisler et al., 2012b,a). This allows one to encode the propensity of the order flow to invert its sign after a price change, an effect that is particularly important for large tick stocks. This extended model improves significantly the description of the price process, both in terms of the lag-dependent volatility (i.e. the signature plot) and in terms of the response function computed for negative lags. However this approach is incomplete as it does not specify the data generating process for the order flow itself, which is only described through two-point correlation functions. This does not allow one to *forecast* the future order flow itself, for example whether a trade is likely to change the price or not.

Here we attempt to model the joint dynamics of order flow and prices. This family of models has a long tradition in market microstructure, starting from the seminal work of Hasbrouck (1988, 1991), who proposed a Vector Autoregressive (VAR) model for the joint dynamics of order flow and prices[1]. There are two main related limitations of this approach. The first is that VAR models are adequate

[1]More recent modeling in continuous time makes use of Hawkes processes (Bacry and Muzy, 2014), which bear some degree of similarity with the models considered in the present chapter.

for variables with continuous support (e.g. Gaussian), while the order flow (signs and events) and tick by tick price changes are more naturally described by discrete variables. Second, the standard VAR approach prescribes a linear relation between the variables, while a broader definition includes the possibility of a linear relation between past variables and the *probability* of observing in the future the value of a given variable.

A natural way to describe the joint dynamics of discrete valued variables (such as the order flow sign and price changes) in a linear setting is with a Markov process of large order. In fact, we have shown in Chapter 4 that for large tick stocks the model with two propagators (TIM2) corresponding to price changing and not-changing trades gives constant (in time) propagators when calibrated on real data (see top left panel of Figure 4.7). This means that the knowledge of the order flow and the information on whether a trade changes the price completely characterises the price dynamics. Thus, in the framework of linear models, it is natural to describe the system with a Markov process with $m = 4$ states, $(\epsilon_t, \pi_t) \in \{(-1, \mathrm{C}), (-1, \mathrm{NC}), (+1, \mathrm{NC}), (+1, \mathrm{C})\}$, corresponding to buys ($\epsilon_t = +1$) and sells ($\epsilon_t = -1$) and price changing ($\pi_t = \mathrm{C}$) and not changing ($\pi_t = \mathrm{NC}$) trades.

However, the main limitation of Markov models comes from the long memory of the order flow (Bouchaud et al., 2004; Lillo and Farmer, 2004). Since the order flow sign is very persistent, a low order Markov process cannot be suitable to describe real markets. On the other hand, Markov processes of high order p depend in general on a very large number of parameters ($O(m^p)$) and might result in inefficient estimation when a limited amount of data is available. For this reason in this Thesis we propose to use a parsimonious, yet versatile class of high order Markov processes termed the Mixed Transition Distribution (MTD) model (Raftery, 1985) and its generalization (MTDg) (Berchtold, 1995). Thanks to a simple structure, where each lag contributes to the prediction of the current state in a separate and additive way, the dimension of model parameter space grows only linearly with the order of the MTDg model, i.e. as $O(m^2 p)$. The model can be calibrated via Maximum Likelihood or via the Generalized Method of Moments. Moreover in the case of $m = 2$ states (such as the signs of the order flow), the version of the MTDg model proposed here reduces to the Discrete Autoregressive (DAR) model (Jacobs and Lewis, 1978), which has been used to model the order flow in Chapter 3. Hence MTD and MTDg aim at providing a natural generalisation of the DAR(p) model to account for an arbitrary number of $m \geq 2$ states, while avoiding the exponential increase (m^p) of the number of parameters of the full Markov model. Perhaps surprisingly, this class of models has not been applied to financial data and our work attempts to fill this gap. In fact, the main methodological innovation of our work is a weakly restricted MTDg model which can be estimated even when the number of parameters is very large, as required to account for the correlation structure of financial data. The restriction consists in constraining the MTDg model within the class of ergodic Markov model. Ergodicity allows to write all the transition matrices in terms of a first component, which depends linearly on the stationary distribution, and a second term, whose kernel contains the stationary distribution. Exploiting the buy-sell symmetry present in the data, the latter term significantly simplifies

and, as a result, this translates in a feasible estimation procedure.

5.2 Markov chains

The Markov chain is a probabilistic model used to represent dependences between successive observations of a random variable. This model was introduced by Andrej Andreevič Markov at the beginning of the 20th century and it is used in many disciplines, including meteorology, geography, biology, chemistry, physics, behavior, social sciences and music. For comprehensive treatments of Markov chains and their applications, see, for example Kemeny and Snell (1976); Karlin and Taylor (1981); Brémaud (1999).

We consider a discrete-time random variable X_t taking values in the finite set $\mathcal{X} = \{1, \ldots, m\}$. Our goal is to predict or explain the value taken by X_t as a function of the values taken by previous observations of this same variable. The first-order Markov hypothesis says that the present observation at time t is conditionally independent of those up to and including time $(t-2)$ given the immediate past [time $(t-1)$]. Thus we can write

$$\begin{aligned}\mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_0 = i_t) &= \mathbb{P}(X_t = i | X_{t-1} = i_1) \\ &= q_{i_1,i}(t)\,,\end{aligned}$$

where $i_t, \ldots, i_1, i \in \{1, \ldots, m\}$. If we suppose that the probability $q_{i_1,i}(t)$ is time-invariant, it is replaced by $q_{i_1,i}$ and we have a homogeneous Markov chain. Considering all combinations of i_1 and i, we construct a transition matrix $\boldsymbol{Q}$, each of whose rows sums to 1:

$$\boldsymbol{Q} = \begin{array}{c} \\ X_{t-1} \\ 1 \\ \vdots \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} & & X_t & \\ 1 & \ldots & \ldots & m \end{array} \\ \begin{pmatrix} q_{1,1} & \ldots & \ldots & q_{1,m} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ q_{m,1} & \ldots & \ldots & q_{m,m} \end{pmatrix} \end{array} .$$

Let

$$\chi_t = \big(x_t(1), ..., x_t(m)\big) \tag{5.1}$$

be a row vector such that $x_t(i) = 1$ if $X_t = i$ and zero otherwise and let $\hat{\chi}_t$ be the probability vector

$$\hat{\chi}_t = \big(\mathbb{P}(X_t = 1), ..., \mathbb{P}(X_t = m)\big)\,. \tag{5.2}$$

Then the following relationships hold:

$$\begin{aligned}\hat{\chi}_t &= \hat{\chi}_{t-1}\boldsymbol{Q}\,, \\ \hat{\chi}_t &= \hat{\chi}_0\boldsymbol{Q}^t\,.\end{aligned} \tag{5.3}$$

The process is fully defined once we know the initial vector $\hat{\chi}_0$ and the transition matrix $\boldsymbol{Q}$.

In some situations, the present depends not only on the first lag, but on the last p observations. We have then an pth-order Markov chain whose transition probabilities are

$$\begin{aligned}\mathbb{P}(X_t = i|X_{t-1} = i_1, \ldots, X_0 = i_t) &= \mathbb{P}(X_t = i|X_{t-1} = i_1, \ldots, X_{t-p} = i_p) \\ &= q_{i_p,\ldots,i_1,i}\end{aligned}$$

For instance, if we set $p = 2$ and $m = 3$, the corresponding transition matrix is

$$\boldsymbol{Q} = \begin{array}{cc} & X_t \\ X_{t-2} & X_{t-1} \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 1 & 2 \\ 2 & 2 \\ 3 & 2 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 \end{array} \\ \begin{pmatrix} q_{1,1,1} & 0 & 0 & q_{1,1,2} & 0 & 0 & q_{1,1,3} & 0 & 0 \\ q_{2,1,1} & 0 & 0 & q_{2,1,2} & 0 & 0 & q_{2,1,3} & 0 & 0 \\ q_{3,1,1} & 0 & 0 & q_{3,1,2} & 0 & 0 & q_{3,1,3} & 0 & 0 \\ 0 & q_{1,2,1} & 0 & 0 & q_{1,2,2} & 0 & 0 & q_{1,2,3} & 0 \\ 0 & q_{2,2,1} & 0 & 0 & q_{2,2,2} & 0 & 0 & q_{2,2,3} & 0 \\ 0 & q_{3,2,1} & 0 & 0 & q_{3,2,2} & 0 & 0 & q_{3,2,3} & 0 \\ 0 & 0 & q_{1,3,1} & 0 & 0 & q_{1,3,2} & 0 & 0 & q_{1,3,3} \\ 0 & 0 & q_{2,3,1} & 0 & 0 & q_{2,3,2} & 0 & 0 & q_{2,3,3} \\ 0 & 0 & q_{3,3,1} & 0 & 0 & q_{3,3,2} & 0 & 0 & q_{3,3,3} \end{pmatrix} \end{array}.$$

When the order is greater than 1, notice that the transition matrix $\boldsymbol{Q}$ contains several elements corresponding to transitions that cannot occur. For instance, it is impossible to go from the row defined by $X_{t-2} = 1$ and $X_{t-1} = 2$ to the column defined by $X_{t-1} = 1$ and $X_t = 1$ because of the different value taken by X_{t-1}. The probability of this transition is then 0 and we call this element a structural zero. Since the exact form of the transition matrix is known for any combination of p and m, it is possible to rewrite $\boldsymbol{Q}$ in a more compact form excluding the structural zeros. This way of writing $\boldsymbol{Q}$, as given by Pegram (1980), is called the collapsed or reduced form of $\boldsymbol{Q}$ and is denoted by $\boldsymbol{R}$. The reduced form of the matrix corresponding to $p = 2$ and $m = 3$ is

$$\boldsymbol{R} = \begin{array}{cc} & X_t \\ X_{t-2} & X_{t-1} \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 1 & 2 \\ 2 & 2 \\ 3 & 2 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} q_{1,1,1} & q_{1,1,2} & q_{1,1,3} \\ q_{2,1,1} & q_{2,1,2} & q_{2,1,3} \\ q_{3,1,1} & q_{3,1,2} & q_{3,1,3} \\ q_{1,2,1} & q_{1,2,2} & q_{1,2,3} \\ q_{2,2,1} & q_{2,2,2} & q_{2,2,3} \\ q_{3,2,1} & q_{3,2,2} & q_{3,2,3} \\ q_{1,3,1} & q_{1,3,2} & q_{1,3,3} \\ q_{2,3,1} & q_{2,3,2} & q_{2,3,3} \\ q_{3,3,1} & q_{3,3,2} & q_{3,3,3} \end{pmatrix} \end{array}.$$

Each possible combination of p successive observations of the random variable X is called a *state* of the model. The number of states is equal to m^p ($= 3^2 = 9$

in our example). In the case of a first-order Markov chain, each value taken by the random variable X is also a state of the model.

The relationships of Equations (5.3) defined in the case of a first-order Markov chain still hold. Whatever the order is, there are $(m-1)$ independent probabilities in each row of the matrix $\boldsymbol{Q}$, the last one of which is completely determined by the others since each row is a probability distribution summing to 1. The total number of independent parameters to be estimated is thus equal to $m^p(m-1)$. Given a set of observations, these parameters can be computed as follows. Let $n_{i_p,\ldots,i_1,i}$ denote the number of transitions of the type

$$X_{t-p} = i_p, \qquad \ldots, \qquad X_{t-1} = i_1, \qquad X_t = i$$

in the data. The maximum likelihood estimate of the corresponding transition probability $q_{i_p,\ldots,i_1,i}$ is then

$$\hat{q}_{i_p,\ldots,i_1,i} = \frac{n_{i_p,\ldots,i_1,i}}{n_{i_p,\ldots,i_1,+}}$$

where

$$n_{i_p,\ldots,i_1,+} = \sum_{i=1}^{m} n_{i_p,\ldots,i_1,i}$$

and the log-likelihood function of the parameters set $\hat{\boldsymbol{q}}$ for the entire sequence of observations is written

$$\ell(\hat{\boldsymbol{q}}) = \sum_{i_p,\ldots,i_1,i=1}^{m} n_{i_p,\ldots,i_1,i} \log\left(\hat{q}_{i_p,\ldots,i_1,i}\right) .$$

5.3 The need for parsimonious models of high-order Markov chains

Markov chains are well suited for the representation of high-order dependencies between successive observations of a random variable. Unfortunately, as the order p of the chain and the number m of possible values increase, the number of independent parameters increases exponentially and rapidly becomes too large to be estimated efficiently, or even identifiably, with data sets of the sizes typically encountered in practice.

The mixture transition distribution model was introduced to approximate high-order Markov chains with far fewer parameters than the fully parameterized model. Each element of a transition matrix is the probability of observing an event at time t given the events observed at times $(t-p)$ to $(t-1)$. In the MTD model, the effect of each lag upon the present is considered separately and the conditional probability is modeled by

$$\mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q_{i_g,i} \,, \tag{5.4}$$

where the $q_{i_g,i}$ are the probabilities of an $m \times m$ transition matrix and λ_g is the weight parameter associated with lag g. This model has only $m(m-1)+(p-1)$ independent parameters and each additional lag adds only one additional parameter. In fact, when the order is greater than 1, the MTD model is far more parsimonious than the corresponding fully parameterized Markov chain. Therefore, it can be used to estimate high-order transition matrices, even when the amount of data is relatively small.

Parsimonious modeling can also make interpretation easier. A high-order Markov chain can have hundreds or thousands of parameters and it can be difficult to interpret the estimates. On the other hand, a MTD model is generally composed of only one small transition matrix and a vector of lag parameters which are easier to interpret.

5.4 The Mixture Transition Distribution model

5.4.1 Definition

We start from a simple, but restrictive, definition of MTD models and below we will extend it in a more general class of MTD models. Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of random variables taking values in the finite set $\mathcal{X} = \{1,\ldots,m\}$. This random sequence is said to be a p-th order MTDg sequence if for all $t > p$ and for all $(i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1}$,

$$\mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q^g_{i_g,i}\,, \tag{5.5}$$

where the vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ is subject to the constraints:

$$\lambda_g \geq 0, \qquad \forall g \in \{1,\ldots,p\}\,, \tag{5.6}$$

$$\sum_{g=1}^{p} \lambda_g = 1\,. \tag{5.7}$$

The matrices $\left\{\boldsymbol{Q}^g = \left[q^g_{i,j}\right];\, i,j \in \mathcal{X}; 1 \leq g \leq p\right\}$ are positive $m \times m$ stochastic matrices, i.e. they satisfy

$$q^g_{i,j} \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} q^g_{i,j} = 1 \qquad \forall g \in \{1,\ldots,p\}, \forall i,j \in \mathcal{X}\,. \tag{5.8}$$

Raftery (1985) has originally defined the model with the same transition matrix $\boldsymbol{Q}^g \equiv \boldsymbol{Q}$ for each lag $g = 1,\ldots,p$ and this model is called the MTD. Later, Berchtold (1995) has introduced the more general definition of MTD models as a mixture of transitions from subsets of lagged variables $\{X_{t-1},\ldots,X_{t-p}\}$ to the present one X_t. In other words, the order of the transition matrices $\boldsymbol{Q}^g$ can be larger than one.

Berchtold and Raftery (2002) have published a complete review of the MTD model. They recall theoretical results on the limiting behavior of the model and on its auto-correlation structure. In particular, they proved that if conditions of Equations (5.6), (5.7), and (5.8) are satisfied, then the model of Equation (5.5) is a well defined high-order Markov chain and its stationary distribution $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_m)$ exists and it is unique. The above Mixture Transition Distribution models are Markov models where each lag $X_{t-1}, X_{t-2}, \ldots$ contributes additively to the distribution of the random variable X_t. Hence the model is linear in the sense described in the introduction.

In words, this class of models means the following: in order to determine the type of event X_t, occurring at time t, start choosing a reference time $t - g$ in the past, where g is drawn at random with probability λ_g. If the event X_{t-g} that occurred at time $t - g$ is of type j, then choose the event at time t to be of type i with probability $q_{i,j}$. This model can thus be interpreted as a probabilistic mixture of Markov processes. For this interpretation the fact that $(\lambda_g)_{g=1,\ldots,p}$ is a probability vector and $\boldsymbol{Q}^g$ are stochastic matrices is critical. However, as already noted in the original papers (Raftery, 1985; Berchtold, 1995), the MTDg model can be also defined when these parameters are negative or larger than one, provided that the conditions

$$0 \leq \sum_{g=1}^{p} \lambda_g q^g_{i_g,i} \leq 1, \qquad \forall (i, i_1, \ldots i_p) \in \mathcal{X}^{p+1}\,, \tag{5.9}$$

are satisfied, in such a way that all transition probabilities are well defined. As we shall see below, calibrated parameters do not necessarily abide to the probabilistic interpretation.

5.4.2 Limiting behavior of the MTDg model

In this Thesis we will consider a specific class of MTDg models where the matrices $\boldsymbol{Q}^g$ share the same stationary state, i.e. the same left eigenvector $\hat{\eta}$ corresponding to the eigenvalue 1. Under this assumption, generalizing a result of (Berchtold, 1995), we can prove the following theorem of the existence and uniqueness of the stationary distribution.

Theorem 1. *Suppose that a sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ taking values in the finite set $\mathcal{X} = \{1, \ldots, m\}$ is defined by*

$$\mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q^g_{i_g,i}\,, \tag{5.10}$$

where $\boldsymbol{Q}^g = \left[q^g_{i,j}\right]_{i,j\in\mathcal{X}}$ are matrices with normalized row, $\sum_j q^g_{i,j} = 1$, $\sum_{g=1}^{p} \lambda_g = 1$, and assume that $\hat{\eta}\boldsymbol{Q}^g = \hat{\eta}, \forall g$. If the vector $\hat{\eta}$ is such that $\hat{\eta}_i > 0, i \in \mathcal{X}$ and $\sum_i \hat{\eta}_i = 1$, and

$$0 < \sum_{g=1}^{p} \lambda_g q^g_{i_g,i} < 1, \qquad \forall (i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1}\,, \tag{5.11}$$

then

$$\lim_{\ell\to\infty} \mathbb{P}(X_{t+\ell} = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \hat{\eta}_i \,. \tag{5.12}$$

Proof. Let $\boldsymbol{T}$ be the $m^p \times m^p$ transition matrix for the Markov chain with the m^p possible values of $(X_{t-1}, \ldots, X_{t-p})$ as states. The elements of $\boldsymbol{T}$ are

$$\begin{aligned} &\mathbb{P}(X_t = i, X_{t-1} = i_1, \ldots, X_{t-p+1} = i_{p-1} | X_{t-1} = j_1, \ldots, X_{t-p} = j_p) \,, \\ &\qquad = \begin{cases} \sum_{g=1}^{p} \lambda_g q^g_{j_g,i} & \text{if } i_g = j_g, \text{ for } g = 1, 2, \ldots, p-1 \,, \\ 0 & \text{otherwise} \,. \end{cases} \end{aligned} \tag{5.13}$$

Each column of $\boldsymbol{T}$ represents the p-vector $(i, \ldots, i_{p-1})$ of arrival states, which are ordered in such a way that i varies most slowly, i_1 second most slowly, and so on. Similarly, the rows of $\boldsymbol{T}$ represents the values of $(j_1, \ldots, j_p)$ with j_1 varies most slowly, and so on.

The assumption of Equation (5.11) guarantees that all states of $\boldsymbol{T}$ intercommunicate, so $\boldsymbol{T}$ is irreducible. Amongst the diagonal elements of $\boldsymbol{T}$ m are aperiodic, then, since $\boldsymbol{T}$ is irreducible, all states are aperiodic. Hence, $\boldsymbol{T}$, being finite, specifies an ergodic Markov chain and has a unique equilibrium distribution ξ satisfying $\xi\boldsymbol{T} = \xi$ with elements

$$\xi_{j_1,\ldots,j_p} = \lim_{t\to\infty} \mathbb{P}(X_{t-1} = j_1, \ldots, X_{t-p} = j_p) \tag{5.14}$$

where the p-vector $(j_1, \ldots, j_p)$ is ordered in the same way of the matrix $\boldsymbol{T}$. We call $\omega = (\omega_1, \ldots, \omega_m)$ the corresponding one-dimensional marginal equilibrium distribution. Also let $\boldsymbol{R}$ be the 'collapsed form' of $\boldsymbol{T}$ as defined in (Pegram, 1980), which is the $m^p \times m$ matrix of the non-zero elements of $\boldsymbol{T}$. Clearly, in general

$$\xi\boldsymbol{R} = \omega \,. \tag{5.15}$$

We write the same matrix for the model of Equation (5.32)

$$\boldsymbol{R} = \sum_{g=1}^{p} \lambda_g \boldsymbol{U}_g \,, \tag{5.16}$$

where $\boldsymbol{U}_g = \boldsymbol{A}_{g,1} \otimes \cdots \otimes \boldsymbol{A}_{g,p}$ where

$$\boldsymbol{A}_{g,k} = \begin{cases} \boldsymbol{Q}^g & \text{if } g = k \\ 1^T & \text{if } g \neq k \end{cases} \tag{5.17}$$

and $\otimes$ is the Kronecker product and 1^T is a $m \times 1$ vector of ones.

We now calculate $\xi\boldsymbol{R}$ in another way. The k-th column of $\xi\boldsymbol{U}_g$ is

$$\begin{aligned} \sum_{i_1,\ldots,i_p}^{m} q^g_{i_g,k} \xi_{i_1,\ldots,i_p} &= \sum_{i_g}^{m} q^g_{i_g,k} \sum_{i_h, h\neq g}^{m} \xi_{i_1,\ldots,i_p} \\ &= \sum_{i_g}^{m} q^g_{i_g,k} \omega_{i_g} \end{aligned} \tag{5.18}$$

which is also the k-th column of $\omega \boldsymbol{Q}^g$. Thus

$$\xi \boldsymbol{R} = \sum_{g=1}^{p} \lambda_g \omega \boldsymbol{Q}^g \,. \tag{5.19}$$

Equating Equations (5.15) and 5.19, we have that $\omega = \hat{\eta}$, by uniqueness of ω and $\hat{\eta}\boldsymbol{Q}^g = \hat{\eta}, \forall g$. □

Notice that in this theorem we do not need to assume that the parameters $(\lambda_g, \boldsymbol{Q}^g)_{1 \leq g \leq p}$ are between zero and one, but the probabilistic interpretation is guaranteed by the condition of Equation (5.11). Finally notice that the condition on $\hat{\eta}$ implies that $\forall g$ we can write $\boldsymbol{Q}^g = \boldsymbol{Q} + \tilde{\boldsymbol{Q}}^g$, where $\hat{\eta}\boldsymbol{Q} = \hat{\eta}$ and $\hat{\eta}\tilde{\boldsymbol{Q}}^g = 0$.

5.4.3 Estimation

Despite being parsimonious with respect to full Markov models, the MTDg parameters $\boldsymbol{\theta} = (\lambda_g, \boldsymbol{Q}^g)_{1 \leq g \leq p}$ are difficult to estimate because they have to comply with the normalization constraints of transition matrices. In the literature many different estimation methods have been proposed (Berchtold and Raftery, 2002), but here we will focus on two specific methodologies: the maximum likelihood estimation (MLE) and the generalized method of moments (GMM). Let us introduce the details of these methods.

Maximum likelihood estimation

For a given data sequence with length n, $\{X_t = x_t\}_{t=1,\ldots,n}$, we define $(X_{t_1}^{t_2} = x_{t_1}^{t_2})$ as the sequence of events $(X_{t_1} = x_{t_1}, X_{t_1+1} = x_{t_1+1}, ..., X_{t_2} = x_{t_2})$ and $\mathbb{P}(X_1^p = x_1^p)$ is the joint distribution of $\{X_t = x_t\}_{t=1,\ldots,p}$. From the definition of MTDg models of order p, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}(X_1^n = x_1^n) = \mathbb{P}(X_1^p = x_1^p)\mathbb{P}_{\boldsymbol{\theta}}(X_{p+1}^n = x_{p+1}^n | X_1^p = x_1^p) \\ &= \mathbb{P}(X_1^p = x_1^p) \prod_{t=p+1}^{n} \left\{ \sum_{g=1}^{p} \lambda_g q^g_{x_{t-g}, x_t} \right\} , \end{aligned} \tag{5.20}$$

To estimate the parameters of MTDg model, we have excluded $\mathbb{P}(X_1^p = x_1^p)$ from the likelihood function. Therefore, the log-likelihood function that we consider is

$$\ell(\boldsymbol{\theta}) = \log \mathbb{P}_{\boldsymbol{\theta}}(X_{p+1}^n = x_{p+1}^n | X_1^p = x_1^p) = \sum_{t=p+1}^{n} \log \left\{ \sum_{g=1}^{p} \lambda_g q^g_{x_{t-g}, x_t} \right\} , \tag{5.21}$$

where $\boldsymbol{\theta} = (\lambda_g, \boldsymbol{Q}^g)_{1 \leq g \leq p}$ satisfies all the constraints of Equations (5.6), (5.7), (5.8) or (5.11). Hence, the maximum likelihood estimation of the parameters

$$\hat{\boldsymbol{\theta}} = \left(\hat{\lambda}_g, \hat{\boldsymbol{Q}}^g \right)_{1 \leq g \leq p}$$

is the solution of the following constrained non-linear optimization problem:

$$\left(\hat{\lambda}_g, \hat{\boldsymbol{Q}}^g\right)_{1\leq g\leq p} = \underset{(\lambda_g,\boldsymbol{Q}^g)_{1\leq g\leq p}}{\operatorname{argmax}} \sum_{t=p+1}^{n} \log\left\{\sum_{g=1}^{p} \lambda_g q^g_{x_{t-g},x_t}\right\},$$

$$\text{s.t} \quad \sum_{g=1}^{p} \lambda_g = 1\,,$$

$$\lambda_g \geq 0, \qquad \forall g \in \{1,\ldots,p\}$$

$$q^g_{i,j} \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} q^g_{i,j} = 1 \qquad \forall g \in \{1,\ldots,p\}, \forall i,j \in \mathcal{X}\,. \tag{5.22}$$

Clearly the solution of the previous optimization problem is very hard due to the high number of constraints. Berchtold (2001) proposes an efficient iterative process with the boundary adjustment in the MLE process which leads to a modification of the Newton's method. Under the constraints of Equations (5.6) and (5.7), Lèbre and Bourguignon (2008) introduce a hidden process for the coefficients of the MTDg and propose an Expectation-Maximization approach for the parameters estimation. Chen and Lio (2009) note that all the previous constraints can be rewritten in a box-constrained form, which is easier to handle.

Generalized Method of Moments

Raftery (1985) shows that the bivariate distributions of the MTD model satisfy a linear system of equations similar to the Yule-Walker equations. Here we extend this result to the MTDg case, i.e. when transition matrices $\boldsymbol{Q}^g$ differ at each lag g. We prove the following proposition:

Proposition 1. *Suppose that a sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ taking values in the finite set $\mathcal{X} = \{1,\ldots,m\}$ is defined by Equation (5.5) and is stationary. Let $\boldsymbol{B}(k)$ be a $m \times m$ matrix with elements*

$$b^k_{i,j} = \mathbb{P}(X_t = i, X_{t+k} = j), \qquad i,j \in \mathcal{X}; k \in \mathbb{Z} \tag{5.23}$$

and $\boldsymbol{B}(0) = diag(\hat{\eta}_1,\ldots,\hat{\eta}_m)$. Then

$$\boldsymbol{B}(k) = \sum_{g=1}^{p} \lambda_g \boldsymbol{B}(k-g)\boldsymbol{Q}^g\,. \tag{5.24}$$

Proof. First consider the case where $k = 1,\ldots,p$. Let

$$Y^k_t = \{X_{t+k-g} : g = 1,\ldots,p; g \neq k\}, \tag{5.25}$$

then

$$\begin{aligned}
b_{i,j}^k &= \mathbb{P}(X_t = i, X_{t+k} = j) \\
&= \sum_{Y_t^k} \mathbb{P}(X_t = i, X_{t+k} = j | Y_t^k) \mathbb{P}(Y_t^k) \\
&= \sum_{Y_t^k} \mathbb{P}(X_{t+k} = j | X_t = i, Y_t^k) \mathbb{P}(X_t = i | Y_t^k) \mathbb{P}(Y_t^k) \\
&= \sum_{Y_t^k} \sum_{g=1, g\neq k}^{p} \lambda_g q^g_{X_{t+k-g},j} \mathbb{P}(X_t = i | Y_t^k) \mathbb{P}(Y_t^k) + \sum_{Y_t^k} \lambda_k q_{i,j}^k \mathbb{P}(X_t = i | Y_t^k) \mathbb{P}(Y_t^k) \\
&= \sum_{g=1, g\neq k}^{p} \lambda_g \sum_{X_{t+k-g}} q^g_{X_{t+k-g},j} \mathbb{P}(X_t = i | X_{t+k-g}) \mathbb{P}(X_{t+k-g}) + \lambda_k \hat{\eta}_i q_{i,j}^k \\
&= \sum_{g=1, g\neq k}^{p} \lambda_g \sum_{h=1}^{m} b_{i,h}^{k-g} q_{h,j}^g + \lambda_k \hat{\eta}_i q_{i,j}^k
\end{aligned} \tag{5.26}$$

which is the (i, j)-th element of

$$\sum_{g=1}^{p} \lambda_g \boldsymbol{B}(k-g) \boldsymbol{Q}^g \tag{5.27}$$

as required. □

The system of Equation (5.24) consists in $m^2 p$ different equations which can be employed as orthogonality conditions of the GMM applied to the MTDg model. However, these equations are not all independent, because the matrices of the bivariate distributions $\boldsymbol{B}(k)$ satisfy the usual normalization conditions. In fact, the rows and the columns of each matrix sum up to the corresponding unconditional probability, $\sum_j b_{i,j}^k = \hat{\eta}_i$ and $\sum_i b_{i,j}^k = \hat{\eta}_j$. By using these relations, the number of independent equations is reduced to $p(m^2 - 2m + 1)$. The uniqueness of the solution of the system of Equation (5.24) requires that the number of independent parameters of the model has to be equal to the number of independent equations. We want to emphasize the fact that in literature the previous system of matrix equations has not be used to estimate the model, because the simple inversion of this linear system requires the definition of a class of MTDg models which makes feasible the previous method of estimation. For this purpose, in the following section we will propose a version of the MTDg models which is very close to the general one.

5.5 A general class of MTDg models

Here we introduce the main methodological innovation of this chapter, namely a parametrization of the MTDg model which can be estimated with GMM even when the number of parameters is very large. To motivate it, let us consider the DAR(p)

process with m states (employed for example in Chapter 4) as a model for the order flow)[2]. The model can be seen as a particular case of the MTD(p) model, where the transition matrices are the same for all g, $\boldsymbol{Q}^g \equiv \boldsymbol{Q}$ and such that

$$\boldsymbol{Q} = 1^T\hat{\eta} + \varphi(\mathbb{I} - 1^T\hat{\eta})\,, \tag{5.29}$$

where 1 is a row of m ones and the parameter φ ranges between zero and one. The left eigenvector of $\boldsymbol{Q}$ corresponding to the eigenvalue 1 is $\hat{\eta}$, since it belongs to the kernel of $\mathbb{I} - 1^T\hat{\eta}$.

Following the same idea, we introduce MTD(p) models where

$$\boldsymbol{Q}^g = 1^T\hat{\eta} + \tilde{\boldsymbol{Q}}^g \tag{5.30}$$

and $\hat{\eta}\tilde{\boldsymbol{Q}}^g = 0$. Moreover normalization of $\boldsymbol{Q}^g$ imposes that each row of $\tilde{\boldsymbol{Q}}^g$ sums to zero, hence these matrices will have negative elements. Thus, we have that the parameters of the model of order p consist in the vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ and the matrices $\boldsymbol{Q}^g$, such that

$$\begin{aligned}
&\boldsymbol{Q}^g = 1^T\hat{\eta} + \tilde{\boldsymbol{Q}}^g\,,\\
&\tilde{\boldsymbol{Q}}^g = \begin{pmatrix}
\tilde{q}^g_{1,1} & \cdots & \tilde{q}^g_{1,m-1} & -\sum_{i=1}^{m-1}\tilde{q}^g_{1,i}\\
\vdots & \ddots & \vdots & \vdots\\
\tilde{q}^g_{m-1,1} & \cdots & \tilde{q}^g_{m-1,m-1} & -\sum_{i=1}^{m-1}\tilde{q}^g_{m-1,i}\\
-\sum_{i=1}^{m-1} c_i\tilde{q}^g_{i,1} & \cdots & -\sum_{i=1}^{m-1} c_i\tilde{q}^g_{i,m-1} & \sum_{i,j=1}^{m-1} c_i\tilde{q}^g_{i,j}
\end{pmatrix},
\end{aligned} \tag{5.31}$$

where $\hat{\eta}\tilde{\boldsymbol{Q}}^g = 0, \forall g$ and $c_i = \hat{\eta}_i/\hat{\eta}_m$.

As in Theorem 1, all the $\boldsymbol{Q}^g$ share the same left eigenvector $\hat{\eta}$ with eigenvalue 1. It is easy to show that the conditional probabilities of this model can be written as

$$\mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \hat{\eta}_i + \sum_{g=1}^{p} a^g_{i_g,i}\,, \tag{5.32}$$

where $a^g_{i_g,i} \equiv \lambda_g(\tilde{\boldsymbol{Q}}^g)_{i_g,i}$. Thus the matrices $\boldsymbol{A}^g \equiv \lambda_g\tilde{\boldsymbol{Q}}^g$ describe the deviations of the $p-$order transition probability from the stationary value given by $\hat{\eta}$. It is important to emphasize the fact that the total number of independent parameters $a^g_{i_g,i}$ is $m^2 - 2m + 1$ for each g.

As stated before, we write the independent elements of the bivariate distributions of the random variable X_t, $\boldsymbol{B}(k)$. Let $\boldsymbol{B}(k)$ be an $m \times m$ matrix whose elements are

$$b^k_{i,j} = \mathbb{P}(X_t = i, X_{t+k} = j), \qquad i, j = 1, \ldots, m, k \in \mathbb{Z}, \tag{5.33}$$

[2]The case $m = 2$ considered in Section 4.2.4 corresponds to a MTD(p) model with transition matrices that are the same for all g, $\boldsymbol{Q}^g \equiv \boldsymbol{Q}$ and

$$\boldsymbol{Q} = \begin{pmatrix} \rho & 1-\rho \\ 1-\rho & \rho \end{pmatrix}. \tag{5.28}$$

In the stationary condition the two states are equiprobable, as can be verified solving the left eigenvalue problem $\hat{\eta}\boldsymbol{Q} = \hat{\eta}$.

where $\boldsymbol{B}(0) = diag(\hat{\eta}_1, \ldots, \hat{\eta}_m)$. Then, we have that

$$\boldsymbol{B}(k) = \begin{pmatrix} b^k_{1,1} & \cdots & b^k_{1,m-1} & \hat{\eta}_1 - \sum_{i=1}^{m-1} b^k_{1,i} \\ \vdots & \ddots & \vdots & \vdots \\ b^k_{m-1,1} & \cdots & b^k_{m-1,m-1} & \hat{\eta}_{m-1} - \sum_{i=1}^{m-1} b^k_{m-1,i} \\ \hat{\eta}_1 - \sum_{i=1}^{m-1} b^k_{i,1} & \cdots & \hat{\eta}_{m-1} - \sum_{i=1}^{m-1} b^k_{i,m-1} & 2\hat{\eta}_m - 1 + \sum_{i,j=1}^{m-1} b^k_{i,j} \end{pmatrix}, \tag{5.34}$$

where the total number of independent elements is $m^2 - 2m + 1$ for each k.

Within this framework, the bivariate distributions and the matrices $\tilde{\boldsymbol{Q}}^g$ satisfy the following system of matrix equations

$$\boldsymbol{B}(k) - \hat{\eta}^T \hat{\eta} = \sum_{g=1}^{p} \boldsymbol{B}(k-g)\boldsymbol{A}^g \,, \tag{5.35}$$

where $\boldsymbol{A}^g \equiv \lambda_g \tilde{\boldsymbol{Q}}^g$.

This linear system can be used to estimate the model, i.e. the matrices $\boldsymbol{A}^g$, from the knowledge of the stationary probabilities $\hat{\eta}$ and the bivariate distributions $\boldsymbol{B}(k)$. In fact, the previous system of matrix equations can be inverted, because the number of independent equations (the equations related to each independent elements of the bivariate distributions) matches exactly the number of independent parameters of the model.

There are however two technical problems:

- The estimated model might not have a probabilistic interpretation, i.e. the estimated model might generate transition probabilities larger than one or smaller than zero;
- The solution of Equation (5.35) gives the matrix $\boldsymbol{A}^g$, while one might need λ_g and $\tilde{\boldsymbol{Q}}^g$ separately, thus the identifiability problem must be solved by fixing arbitrarily one parameter. Note however that the dynamics of the model is independent from this choice.

In the following we will tackle these points.

To fix the first problem, and to be able to use Theorem 1, it must also hold that

$$0 < \hat{\eta}_i + \sum_{g=1}^{p} a^g_{i_g,i} < 1, \qquad \forall (i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1} \,, \tag{5.36}$$

which correspond to $2m^{p+1}$ constraints. Following Proposition 1 in (Raftery and Tavaré, 1994), they can all be satisfied simultaneously by imposing the necessary

and sufficient conditions

$$\hat{\eta}_i + \sum_{g=1}^{p} \max_{i_g} \left(a^g_{i_g,i} \right) < 1, \qquad \forall i \in \mathcal{X} \tag{5.37}$$

$$\hat{\eta}_i + \sum_{g=1}^{p} \min_{i_g} \left(a^g_{i_g,i} \right) > 0, \qquad \forall i \in \mathcal{X} \tag{5.38}$$

which reduce to $2m$ inequality constraints. Under these conditions the process is well defined and possesses a unique stationary solution. The estimation of the model can be performed solving the optimization program

$$\begin{aligned} \hat{q} = & \underset{\mathbf{q} \in \mathbb{R}^{p(m^2-2m+1)}}{\operatorname{argmin}} \left\| \mathbf{d} - \boldsymbol{K} \cdot \mathbf{q} \right\|^2 \\ \text{s.t.} \quad & \hat{\eta}_i + \sum_{g=1}^{p} \max_{i_g} \left(a^g_{i_g,i} \right) < 1, \qquad \forall i \in \mathcal{X} \\ & \hat{\eta}_i + \sum_{g=1}^{p} \min_{i_g} \left(a^g_{i_g,i} \right) > 0, \qquad \forall i \in \mathcal{X} \end{aligned} \tag{5.39}$$

where the elements of the $p(m^2 - 2m + 1)$-dimensional vector $\mathbf{d}$ correspond to left hand side of Equation (5.35), namely

$$\begin{aligned} \mathbf{d} = (\overline{b}^1_{1,1}, \ldots, \overline{b}^1_{1,m-1}, \ldots, \overline{b}^1_{m-1,1}, \ldots, \overline{b}^1_{m-1,m-1}, \ldots \\ \ldots, \overline{b}^p_{1,1}, \ldots, \overline{b}^p_{1,m-1}, \ldots, \overline{b}^p_{m-1,1}, \ldots, \overline{b}^p_{m-1,m-1}) \end{aligned} \tag{5.40}$$

with

$$\overline{b}^k_{i,j} = b^k_{i,j} - \hat{\eta}_i \hat{\eta}_j, \tag{5.41}$$

the vector $\mathbf{q}$ corresponds to the parameters of the model $\lambda_g \tilde{q}^g_{i,j}$

$$\begin{aligned} \mathbf{q} = (a^1_{1,1}, \ldots, a^1_{1,m-1}, \ldots, a^1_{m-1,1}, \ldots, a^1_{m-1,m-1}, \ldots \\ \ldots, a^p_{1,1}, \ldots, a^p_{1,m-1}, \ldots, a^p_{m-1,1}, \ldots, a^p_{m-1,m-1}) \end{aligned} \tag{5.42}$$

and the elements of the matrix $\boldsymbol{K}$ are linear combinations of $b^k_{i,j}$, according to Equation (5.35) (we do not report the matrix since its form is not transparent).

The reason for the choice of the constraints in Equation (5.39) is that we prove in Appendix 5.A the following proposition:

Proposition 2. *If $\boldsymbol{K}$ is not singular, the optimization program of Equation (5.39) is strictly convex in $\mathbb{R}^{p(m^2-2m+1)}$.*

Therefore if a local minimum exists, then it is a global minimum. The convexity property solves the issue of the high dimensionality of the problem and the model can be estimated also for large order p.

Appendix

5.A Convexity of the optimization problem

Proposition 3. *If $\boldsymbol{K}$ is not singular, the following constrained optimization problem*

$$\begin{aligned} \hat{q} = & \underset{\mathbf{q}\in\mathbb{R}^{p(m^2-2m+1)}}{\operatorname{argmin}} \|\mathbf{d}-\boldsymbol{K}\cdot\mathbf{q}\|^2 \\ s.t. \qquad & \hat{\eta}_i + \sum_{g=1}^{p} \max_{i_g}\left(a^g_{i_g,i}\right) < 1, \qquad \forall i \in \mathcal{X} \\ & \hat{\eta}_i + \sum_{g=1}^{p} \min_{i_g}\left(a^g_{i_g,i}\right) > 0, \qquad \forall i \in \mathcal{X} \end{aligned} \tag{5.43}$$

is convex in $\mathbb{R}^{p(m^2-2m+1)}$.

Proof. This is true if the objective function and all the constraints are convex functions. First of all, it is straightforward to show that the Hessian of the objective function $2\boldsymbol{K}\boldsymbol{K}^T$ is a positive semi-definite matrix. The constraints are convex in q, if they are convex in the parameters $a^g_{i,j}$ because they are affine functions of the components of q. Let a be the vector of parameters $\left(a^g_{i,j}\right)_{i,j\in\mathcal{X};1\leq g\leq p}$, we need to prove that the function

$$f(a) = \sum_{g=1}^{p} \max_{i_g}\left(a^g_{i_g,i}\right), \qquad \forall i \in \mathcal{X} \tag{5.44}$$

is convex in $\mathbb{R}^{p(m^2-2m+1)}$.

If we prove it for a fixed i, then it is true for all $i \in \mathcal{X}$ and also for the constraints with the minimum function. The function $f(a)$ satisfies, for $0 \leq \theta \leq 1$, different vectors of parameters $a, b \in \mathbb{R}^{m^2p}$, and fixed i

$$\begin{aligned} f(\theta a + (1-\theta)b) &= \sum_{g=1}^{p} \max_{i_g}\left(\theta a^g_{i_g,i} + (1-\theta)b^g_{i_g,i}\right) \\ &\leq \theta \sum_{g=1}^{p} \max_{i_g}\left(a^g_{i_g,i}\right) + (1-\theta)\sum_{g=1}^{p} \max_{i_g}\left(b^g_{i_g,i}\right) \\ &= \theta f(a) + (1-\theta) f(b)\,. \end{aligned} \tag{5.45}$$

Therefore, we conclude that the function $f(a)$ is convex in $\mathbb{R}^{p(m^2-2m+1)}$. □

Chapter 6

MTD for order flow and price impact

6.1 Introduction

We consider in this chapter the MTD and MTDg models proposed before as promising models for the joint dynamics of order flow and price changes for large tick stocks. Compared to the models investigated in the Chapter 4, we provide here an explicit model for the order flow, and in particular its response to past price dynamics. Thus we aim at reproducing with the MTD model the complex conditional correlation functions of signed events for large tick stocks (see left panel of Figure 4.4 whose curves are reproduced also in Figure 6.1). Moreover our modeling approach allows to perform out of sample analyses of the MTD's forecasting ability of the order flow and future price changes. Still, this framework has limitations when calibrated on anonymized order flow because one cannot easily disentangle order flow correlations coming from "herding" and coming from "order splitting". In other words, although MTDs give explicit predictions for the response of the order flow to a single event (impulse response), one has to be careful in interpreting the result, as it might not describe the true reaction of the market to an isolated, exogenous order (see Tóth et al., 2012, 2015, 2017)

Specifically, we consider the joint dynamics of order flow and price changes in transaction time $t \in \mathbb{N}$. Each event is a transaction which has a positive sign ($\epsilon_t = +1$) if it is buyer initiated or negative ($\epsilon_t = -1$) if is seller initiated. For the price we distinguish two possibilities, namely that the trade changes the price ($\pi_t = \mathrm{C}$) or not ($\pi_t = \mathrm{NC}$). Notice that we are not considering the amplitude if the immediate price changes. For large tick stocks this is a minor problem, since price changes almost always of ± 1 tick, while for small tick stocks this is not true and we lose the information on the size of price change. In this Thesis we use the MTDg model to describe the sequence of signed events

$$\{(\epsilon_t, \pi_t)\}_{t\in\mathbb{N}} \to \{X_t\}_{t\in\mathbb{N}}\,, \tag{6.1}$$

hence the number of states of the model is $m = 4$. The relation between the states of the model and the signed events is obtained with the arbitrary mapping

$$\begin{aligned} \epsilon_t = -1, \pi_t = \mathrm{C} \quad &\rightarrow \quad X_t = 1\,, \\ \epsilon_t = -1, \pi_t = \mathrm{NC} \quad &\rightarrow \quad X_t = 2\,, \\ \epsilon_t = +1, \pi_t = \mathrm{NC} \quad &\rightarrow \quad X_t = 3\,, \\ \epsilon_t = +1, \pi_t = \mathrm{C} \quad &\rightarrow \quad X_t = 4\,. \end{aligned} \tag{6.2}$$

The main quantity of interest is the cross and autocorrelation functions $C_{\pi_1,\pi_2}(\ell)$, already introduced in (Eisler et al., 2012b,a) and in the Chapter 4. Since

$$\hat{\eta} = \mathbb{P}(X_t) \equiv \mathbb{P}(\epsilon_t, \pi_t) \qquad \boldsymbol{B}(\ell) = \mathbb{P}(X_t; X_{t+\ell}) \equiv \mathbb{P}(\epsilon_t, \pi_t; \epsilon_{t+\ell}, \pi_{t+\ell}) \tag{6.3}$$

these correlations

$$\begin{aligned} C_{\pi_1,\pi_2}(\ell) &= \frac{\mathbb{E}[\epsilon_t I(\pi_t = \pi_1) \cdot \epsilon_{t+\ell} I(\pi_{t+\ell} = \pi_2)]}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)} \\ &= \sum_{\epsilon_t \epsilon_{t+\ell}} \sum_{\pi_t \pi_{t+\ell}} \frac{\epsilon_t I(\pi_t = \pi_1) \epsilon_{t+\ell} I(\pi_{t+\ell} = \pi_2) \mathbb{P}(\epsilon_t, \pi_t; \epsilon_{t+\ell}, \pi_{t+\ell})}{\mathbb{P}(\pi_1)\mathbb{P}(\pi_2)}\,, \end{aligned} \tag{6.4}$$

where $I(\pi_t = \pi)$ is the indicator function, can be expressed in terms of $\hat{\eta}$ and $\boldsymbol{B}(\ell)$. For instance, for $\pi_t = \mathrm{NC}$ and $\pi_{t+\ell} = \mathrm{NC}$ the following relations hold

$$\begin{aligned} \mathbb{P}(\mathrm{NC}) &= \hat{\eta}_2 + \hat{\eta}_3\,, \\ C_{\mathrm{NC,NC}}(\ell) &= \frac{b_{2,2}(\ell) - b_{2,3}(\ell) - b_{3,2}(\ell) + b_{3,3}(\ell)}{(\hat{\eta}_2 + \hat{\eta}_3)^2}\,. \end{aligned} \tag{6.5}$$

In the next two sections we will estimate MTDg models on real financial data of the US markets. We will consider two different parametrizations and estimation methods. The first one, used as a benchmark case, is based on MLE and uses a parametrization which preserves the probabilistic interpretation of the mixture, i.e. it assumes that the parameters $(\lambda_g, \boldsymbol{Q}^g)_{1 \le g \le p}$ are between zero and one. Moreover, in order to be able to apply MLE, we will impose a very strong structure of $(\lambda_g, \boldsymbol{Q}^g)_{1 \le g \le p}$, reducing the number of parameters from $p(m^2 - m) + p - 1 \sim 1,300$ for $p = 100$ to 11.

In the second case we apply the framework explained in the Section 5.5, by relaxing the constraint that $(\lambda_g, \boldsymbol{Q}^g)_{1 \le g \le p}$ are between zero and one and we use GMM. We have shown that this suitable parametrization allows to reduce the estimation to the solution of a constrained linear system, which we have proven to be a convex problem. This model is weakly constrained and we are able to estimate reliably 500 parameters, improving significantly the performance of the model with respect to the benchmark case.

6.2 Strongly constrained MTDg model

Estimation methods for the MTDg model proposed so far in literature have dealt with low order models. Unfortunately, our case requires the estimation of an high-

order version of the model to capture the long-ranged dependence measured for the flow of trade signs. The log-likelihood function of Equation (5.21) is highly non-linear and the solution of the optimization problem could be very hard to find for large values of p.

6.2.1 Parametrization

In order to reduce the number of parameters and to avoid non-linear constraints, we impose a functional form for the parameters which automatically satisfies all the constraints. For the λ_g it is very natural to assume a power law scaling, $\lambda_g = N_\beta g^{-\beta}$, where $N_\beta^{-1} = \sum_{i=1}^{p} g^{-\beta}$. The reason behind this choice is that the values of λ_g influence the correlations for large lags ℓ, which empirically decay slowly with the lag. Another significant simplification of the problem can be achieved by assuming a buy/sell symmetry, which leads to the definition of centro-symmetric matrices $\boldsymbol{Q}^g$. This assumption leads to

$$q_{ij}^g = q_{m-i+1,m-j+1}^g, \quad \text{for } i,j = 1,\ldots,m\,, \tag{6.6}$$

and for the first-order stationary distribution of the process

$$\hat{\eta}_i = \hat{\eta}_{m-i+1}, \quad \text{for } i = 1,\ldots,m\,. \tag{6.7}$$

For instance, $q_{12}^g = q_{43}^g$ since the influence of a sell order price changing event at time $t-g$ on the probability of a sell order not price changing event at time t is equal to the influence of a buy order price changing event at time $t-g$ on the probability of a buy order not price changing event at time t.

As mentioned above (see Theorem 1), we consider matrices $\boldsymbol{Q}^g$ sharing the same left eigenvector with eigenvalue one. Writing $\boldsymbol{Q}^g = \boldsymbol{Q} + \tilde{\boldsymbol{Q}}^g$, we make the following strongly parametrized *ansatz*:

$$\begin{aligned}
\boldsymbol{Q} &= \begin{pmatrix} B_1 & A_1 & A_1 & B_1 \\ B_2 & A_2 & A_2 & B_2 \\ B_2 & A_2 & A_2 & B_2 \\ B_1 & A_1 & A_1 & B_1 \end{pmatrix}, \\
\tilde{\boldsymbol{Q}}^g &= \begin{pmatrix} -\mu_1 e^{-\alpha_{11} g} & -\nu_1 e^{-\alpha_{12} g} & \nu_1 e^{-\alpha_{12} g} & \mu_1 e^{-\alpha_{11} g} \\ \mu_2 e^{-\alpha_{21} g} & \nu_2 e^{-\alpha_{22} g} & -\nu_2 e^{-\alpha_{22} g} & -\mu_2 e^{-\alpha_{21} g} \\ -\mu_2 e^{-\alpha_{21} g} & -\nu_2 e^{-\alpha_{22} g} & \nu_2 e^{-\alpha_{22} g} & \mu_2 e^{-\alpha_{21} g} \\ \mu_1 e^{-\alpha_{11} g} & \nu_1 e^{-\alpha_{12} g} & -\nu_1 e^{-\alpha_{12} g} & -\mu_1 e^{-\alpha_{11} g} \end{pmatrix},
\end{aligned} \tag{6.8}$$

where $\alpha_{ij} \geq 0$. Imposing

$$\begin{aligned}
A_1 &= 1/2 - B_1, & A_2 &= 1/2 - B_2\,, \\
0 \leq B_1 &\leq 1/2, & 0 \leq B_2 &\leq 1/2\,, \\
-B_1 \leq \mu_1 &\leq B_1, & -B_2 \leq \mu_2 &\leq B_2\,, \\
B_1 - 1/2 \leq \nu_1 &\leq 1/2 - B_1, & B_2 - 1/2 \leq \nu_2 &\leq 1/2 - B_2\,,
\end{aligned} \tag{6.9}$$

we automatically satisfy all the constraints of the model. Moreover it is immediate to see that $\hat{\eta}\tilde{\boldsymbol{Q}}^g = 0$, as required, where

$$\hat{\eta} = \left(\tfrac{B_2}{1-2B_1+2B_2}, \tfrac{1-2B_1}{2-4B_1+4B_2}, \tfrac{1-2B_1}{2-4B_1+4B_2}, \tfrac{B_2}{1-2B_1+2B_2}\right) , \tag{6.10}$$

Finally, the parametrization in Equation (6.8) with the linear constraints of Equation (6.9) guarantees that the matrices have the right normalization on the rows, $\sum_j q_{ij}^g = 1, \forall g, i$ and $0 < q_{ij}^g < 1, \forall g, i, j$.

The specification of the functional forms of λ_g and of the entries of $\tilde{Q}^g$ is motivated by the following reasonings. The parameters q_{ij}^g determine the correlations between the event i at time t and the event j at time $t-g$. From the left panel in Figure 4.4, reporting the empirical correlations measured for the large tick stock Microsoft, we see a quite different behaviour depending on the conditioning event. For instance, the order flow correlations among non price-changing events is extremely persistent. It is therefore quite natural to model the decay of the pre-factor λ_g in terms of an hyperbolic function. On the contrary, to reproduce the faster decay of the empirical correlations which involve price-changing events, and since λ_g multiplies all entries of the matrices $\boldsymbol{Q}^g$, we include in the definition of the $\tilde{Q}^g$ exponential damping factors with rates α_{ij} $(i, j = 1, 2)$.

The parameters of this model can be obtained via MLE. The optimization problem is non-trivial since the likelihood function is highly non-linear. However the dimensionality is low and, thanks to the parametrization, the constraints of the problem are linear inequalities. The total number of parameters is 11, $\boldsymbol{\theta} = \{\beta, B_i, \mu_i, \nu_i, \alpha_{ij}\}$ with $i, j = 1, 2$.

6.2.2 Results

In Figure 6.1 and 6.2 we plot the correlation functions computed from a Monte Carlo simulation of the MTDg(100) model with parameter values obtained from MLE on Microsoft (MSFT) and Apple (AAPL) data (details about the data set are given in Section 4.4.1). More precisely, we compare the auto and cross-correlations $C_{\pi_1,\pi_2}(\ell)$ for price-changing and non price-changing events with the empirical ones. As can be noted, for the small tick stock the model can reproduce the structure of the correlations for short time scales, but not their persistence. For the large tick stock, the persistence of the empirical correlations is not well reproduced either. The quality of the fit varies across the different correlations. The behaviour of the C,C and NC,C correlations is recovered quite well both at short and long time scales (for the yellow curve, it is important to stress that the scale of the plot is logarithmic on both axes and an apparently large deviation corresponds to a small difference in the linear scale). The MTDg model describes quite well the NC,NC correlation for short lags, but the quality of the fit worsens for larger lag values. The lack in the persistence of the simulated correlations can be explained by the fact that the estimated exponent β is too high. Finally, the behaviour of the C,NC curve is recovered only for the very first lags, then both the amplitude and the sign reproduced by the MTDg do not match the empirical correlations.

The robustness of the numerical results in this and in the next sections has been tested with extensive Monte Carlo simulations. The weakly and strongly constrained models have been estimated on different sub-samples of the data, then simulated and re-estimated on the synthetic time series. All the experiments, whose detailed results are available under request, have shown that both parametrisations – weak and strong – and estimation procedures – MLE and GMM – are robust to the choice of the functional forms of λ_g and $\boldsymbol{Q}^g$. Moreover, only a minor dependence of the parameter values on the estimation periods can be reported.

We conclude that, for both small and large tick stocks, the restrictions imposed on the matrices $\boldsymbol{Q}^g$ are too much binding to reproduce the different decays of the empirical correlations. Nonetheless, it is worth to comment that these modelling assumptions guarantee a fast estimation procedure, even for very high-order Markov models. This fact can motivate the use of the strongly constrained MTDg model in spite of its mild performances.

6.3 Weakly constrained MTDg model

We now consider the application of the above described MTDg model to the $m = 4$ process describing jointly the order flow and the price changes. As done in the previous section, we reduce the dimensionality of the system by exploiting the buy/sell symmetry, which leads to centrosymmetric $\hat{\eta}$ and $\boldsymbol{B}(k)$. In fact, for $m = 4$ we have that

$$b^k_{i,j} = b^k_{m-i+1,m-j+1}, \tag{6.11}$$

and for the stationary distribution

$$\hat{\eta}_i = \hat{\eta}_{m-i+1}, \quad \text{for } i = 1, \dots, m\,. \tag{6.12}$$

The buy/sell symmetry and the normalization of matrices $\boldsymbol{B}(k)$ reduces the number of independent variables in $\boldsymbol{B}(k)$ to $5p$, 5 for each lag k. Thus, we have that

$$\boldsymbol{B}(k) = \begin{pmatrix} b^k_{1,1} & b^k_{1,2} & \hat{\eta}_2 - b^k_{1,2} - b^k_{2,2} - b^k_{3,2} & \hat{\eta}_1 - \hat{\eta}_2 + b^k_{2,2} + b^k_{3,2} - b^k_{1,1} \\ b^k_{2,1} & b^k_{2,2} & b^k_{3,2} & \hat{\eta}_2 - b^k_{2,1} - b^k_{2,2} - b^k_{3,2} \\ \hat{\eta}_2 - b^k_{2,1} - b^k_{2,2} - b^k_{3,2} & b^k_{3,2} & b^k_{2,2} & b^k_{2,1} \\ \hat{\eta}_1 - \hat{\eta}_2 + b^k_{2,2} + b^k_{3,2} - b^k_{1,1} & \hat{\eta}_2 - b^k_{1,2} - b^k_{2,2} - b^k_{3,2} & b^k_{1,2} & b^k_{1,1} \end{pmatrix}. \tag{6.13}$$

In order to find a solution of the problem of Equation (5.39), we assume that the imposed centrosymmetry of $\boldsymbol{B}(k)$ and $\hat{\eta}$ does not change the rank of the matrix $\boldsymbol{K}$. In this case the solution is unique and it is easy to show that also $\tilde{\boldsymbol{Q}}^g$ must be centrosymmetric, as

$$\begin{aligned} &\boldsymbol{Q}^g = 1^T \eta + \tilde{\boldsymbol{Q}}_g\,, \\ &\tilde{\boldsymbol{Q}}^g = \begin{pmatrix} \tilde{q}^g_{1,1} & \tilde{q}^g_{1,2} & -\tilde{q}^g_{1,2} - c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) & -\tilde{q}^g_{1,1} + c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) \\ \tilde{q}^g_{2,1} & \tilde{q}^g_{2,2} & \tilde{q}^g_{2,3} & -\tilde{q}^g_{2,1} - \tilde{q}^g_{2,2} - \tilde{q}^g_{2,3} \\ -\tilde{q}^g_{2,1} - \tilde{q}^g_{2,2} - \tilde{q}^g_{2,3} & \tilde{q}^g_{2,3} & \tilde{q}^g_{2,2} & \tilde{q}^g_{2,1} \\ -\tilde{q}^g_{1,1} + c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) & -\tilde{q}^g_{1,2} - c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) & \tilde{q}^g_{1,2} & \tilde{q}^g_{1,1} \end{pmatrix}, \end{aligned} \tag{6.14}$$

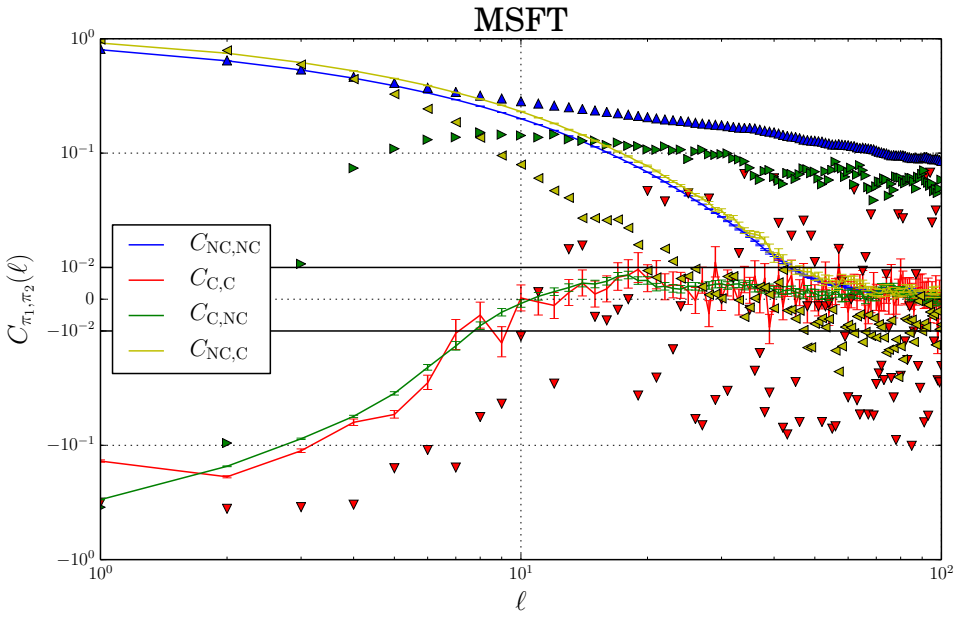

Figure 6.1: *MLE calibration of the strongly constrained MTDg.* Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on MSFT data (solid lines) and the empirical curves (triangles). The error bars correspond to one standard deviation. Estimated parameter values are $\beta = 2.38$, $B_1 = 0.018$, $B_2 = 0.40$, $\mu_1 = 0.018$, $\alpha_{11} = 0.0$, $\nu_1 = 0.48$, $\alpha_{12} = 0.47$, $\mu_2 = 0.04$, $\alpha_{21} = 0.003$, $\nu_2 = 0.42$, and $\alpha_{22} = 0.0$. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

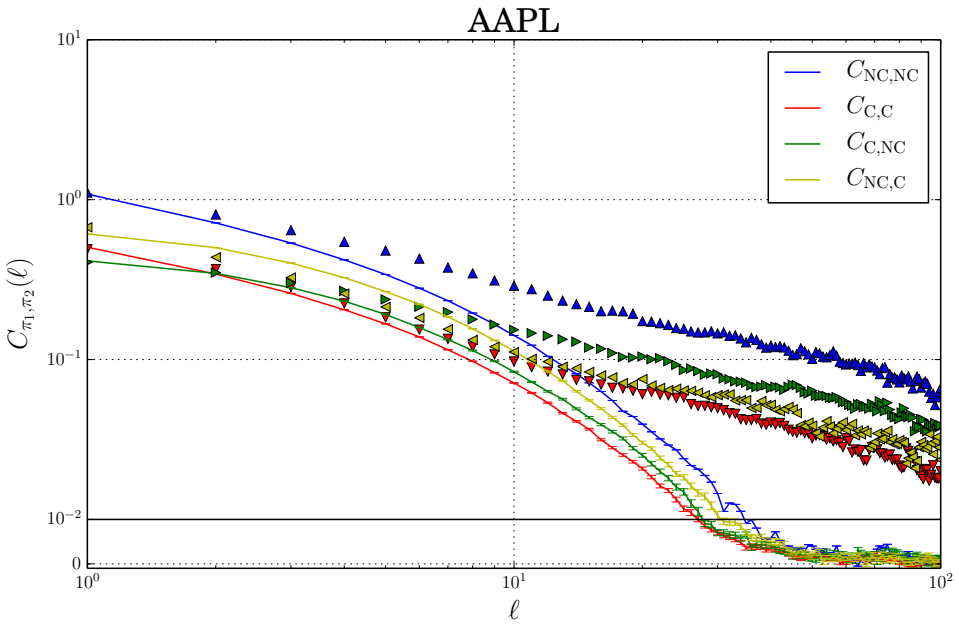

Figure 6.2: *MLE calibration of the strongly constrained MTDg.* Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on AAPL data (solid lines) and the empirical curves (triangles). The error bars correspond to one standard deviation. Estimated parameter values are $\beta = 2.21$, $B_1 = 0.38$, $B_2 = 0.01$, $\mu_1 = -0.22$, $\alpha_{11} = 0.0$, $\nu_1 = -0.07$, $\alpha_{12} = 0.0$, $\mu_2 = 0.27$, $\alpha_{21} = 0.043$, $\nu_2 = 0.21$, and $\alpha_{22} = 0.0$. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

where $c_2 = \hat{\eta}_2/\hat{\eta}_1$. With this definition the number of independent parameters in $\boldsymbol{Q}^g$ is also equal to 5 for each g.

We can now solve the system of Equatiion (5.39) whose unknowns are the components of the matrix $\boldsymbol{A}^g$. This way we obtain the value of the products $\lambda_g \tilde{q}^g_{i,j}$, but not the value of the components λ_g and $\tilde{q}^g_{i,j}$ separately. For this reason we impose that one of the five components among $\tilde{q}^g_{1,1}$, $\tilde{q}^g_{1,2}$, $\tilde{q}^g_{2,1}$, $\tilde{q}^g_{2,2}$, and $\tilde{q}^g_{2,3}$ is independent of the lag g. We arbitrarily fix $\tilde{q}^g_{2,1} \equiv \tilde{q}_{2,1}$. We are left with $4p$ free parameters from $\tilde{\boldsymbol{Q}}^g$ (4 for each g), $p-1$ parameters from λ_g and $\tilde{q}_{21}$. In total we have $5p$ free parameters, which is exactly the same number of independent components $b^k_{i,j}$. The values of the products $\lambda_g \tilde{q}^g_{i,j}$ define the MTDg model. Different choices of $\tilde{q}^g_{i,j} = \tilde{q}_{i,j}$ give different factorizations, but lead to the same high-order Markov chain. The arbitrariness of the choice is an evidence of the well known identifiability problem of all mixture models.

In the literature there exist many algorithms which solve iteratively the constrained optimization problem of Equation (5.39). A widely used class belongs to the Sequential Quadratic Programming (SQP) family (Boggs and Tolle, 1995). However, an issue of our optimization is that constraints are non-smooth functions, which is a necessary condition required by the usual SQP algorithms. In a recent paper, Curtis and Overton (2012) have proposed the Sequential Quadratic Programming Gradient Sampling algorithm (SQP-GS), which can be applied to non-smooth, non-linear objective and constraint functions. We have implemented this algorithm in order to solve our optimization problem.

6.3.1 Results

We estimated the above MTDg(100) model on MSFT and AAPL. Before showing the results, it is worth to stress again an important aspect of our approach. Preliminary, we have estimated the model using Equation (5.35) *without* the additional constraints of Equations (5.37) and (5.38). We have found negative transition probabilities. Thus, the data reject a probabilistic description based on the unconstrained MTD model. To obtain a meaningful, even though approximated, description of the data, the MTD model parameters have to be restricted according to Equations (5.37) and (5.38). Since the constraints are binding, the resulting Markov model of order p is not ergodic, i.e. some of the estimated parameters lie on the boundary of the parameter domain. To apply the results of Theorem 1, which ensures the existence and uniqueness of the stationary distribution, we have restricted the model within the class of ergodic Markov models of order p. Consistently, we have replaced the inequalities of Equations (5.37) and (5.38) with

$$\hat{\eta}_i + \sum_{g=1}^{p} \max_{i_g} \left(a^g_{i_g,i} \right) \leq 1 - \epsilon, \quad \text{and} \quad \hat{\eta}_i + \sum_{g=1}^{p} \min_{i_g} \left(a^g_{i_g,i} \right) \geq \epsilon, \qquad \forall i \in \mathcal{X}$$

where ϵ is a positive constant which guarantees that all Markov states intercommunicate. Clearly the point is to check how the obtained results are sensitive to

the choice of ϵ. By performing extensive numerical simulations we have found that the obtained solution is weakly sensitive to this choice. For three decades of ϵ the squared residuals, which are used to estimate the weakly parametrised version of our model, are essentially constant and the estimation is unaffected by the value of ϵ.

Figures 6.3 and 6.5 show the estimation of $\lambda_q \tilde{q}^g_{i,j}$ for MSFT and AAPL. Despite the large number of estimated parameters, they turn out to be only moderately noisy. Moreover it is interesting to note that negative values of $\lambda_q \tilde{q}^g_{i,j}$ are present, even if, by construction, the transition probabilities of the model are well defined in $[0, 1]$. Clearly the estimation shows that the probabilistic mixture discussed at the beginning is, perhaps meaningfully, not suitable for the present data.

Figures 6.4 and 6.6 show correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events computed from a Monte Carlo simulation of the calibrated model and compared with real data. As can be noted, for small tick stocks we have significantly improved the results of Figure 6.2. Compared with the benchmark, the new estimation method reproduces the high persistence of the correlations of order signs independently from the conditioning events. In the case of the large tick stocks, whose correlations present an highly non-trivial structure, the GMM methodology greatly improves the results with respect to Figure 6.1. In particular, the high persistence of non price-changing events is very well reproduced. Moreover, the $C_{\mathrm{NC,C}}(\ell)$ curve decays faster as compared to the previous estimation method, and is thus closer to data.

6.4 Large tick stock signature plot

Another way to assess the quality of the MTDg model is to analyse how well it describes the volatility of prices. As noted in the top left panel of Figure 4.7, the impact of a price changing event is nearly price independent for large tick stocks (within the TIM2 model). This means that the signature plot is simply given by:

$$D^{\mathrm{TIM2}}(\ell) \approx D_{\mathrm{LF}} + G_{\mathrm{C}}(1)^2 \mathbb{P}(\mathrm{C}) + 2\frac{G_{\mathrm{C}}(1)^2}{\ell} \sum_{0 \le n < m < \ell} \mathbb{P}(\mathrm{C})^2 C_{\mathrm{C,C}}(m-n)\,, \qquad (6.15)$$

which is completely determined by the correlation function $C_{\mathrm{C,C}}(\ell)$ (once the value of $G_{\mathrm{C}}(1)$ has been estimated). This correlation function is, as presented above, only approximately reproduced by the MTDg model, although it is calibrated to minimize the distance to all $C_{\pi_1,\pi_2}(\ell)$. In the context of financial applications, it is therefore interesting to replot the difference between the MTDg $C_{\mathrm{C,C}}(\ell)$ and empirical data in terms of the signature plot $D^{\mathrm{TIM2}}(\ell)$, which involves the integral of the correlation function.

In Figure 6.7 we show the curves corresponding to Equation (6.15) for the strongly and weakly constrained versions of the MTDg model proposed above, where the extra fitting parameter D_{LF} is optimized with OLS in order to minimize the distance between the empirical and the theoretical curves of the model. We see that

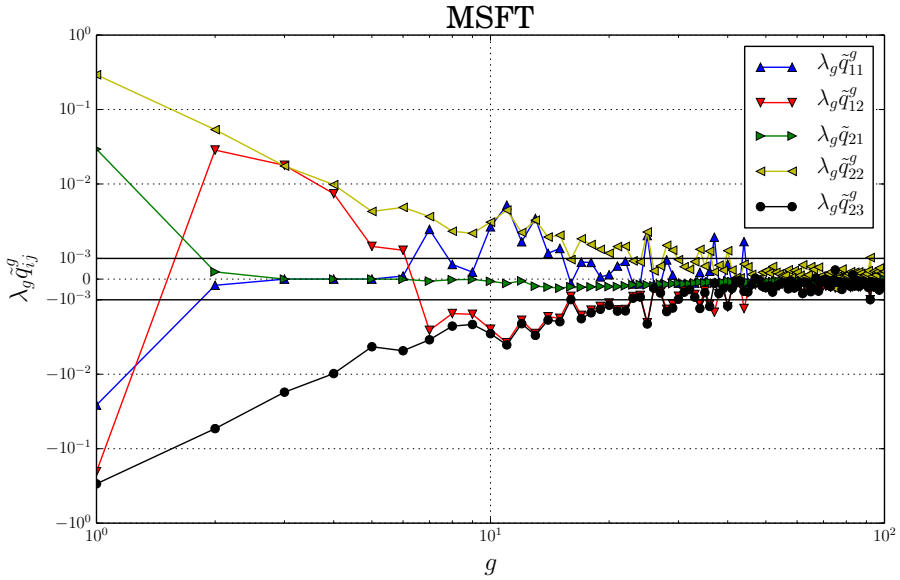

Figure 6.3: *GMM calibration of the weakly constrained MTDg.* Plot of the parameters $a^g_{i,j}$ solution of the optimization problem of Equation (5.39) for an MTDg of order $p = 100$ model estimated from MSFT data. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

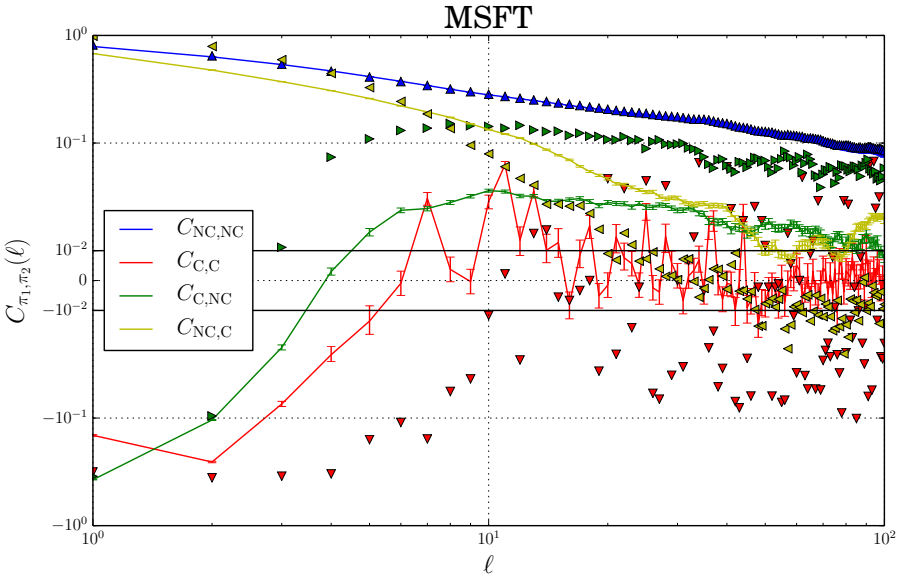

Figure 6.4: *GMM calibration of the weakly constrained MTDg.* Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on MSFT data (triangles) and the empirical curves (solid lines). The error bars correspond to one standard deviation. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

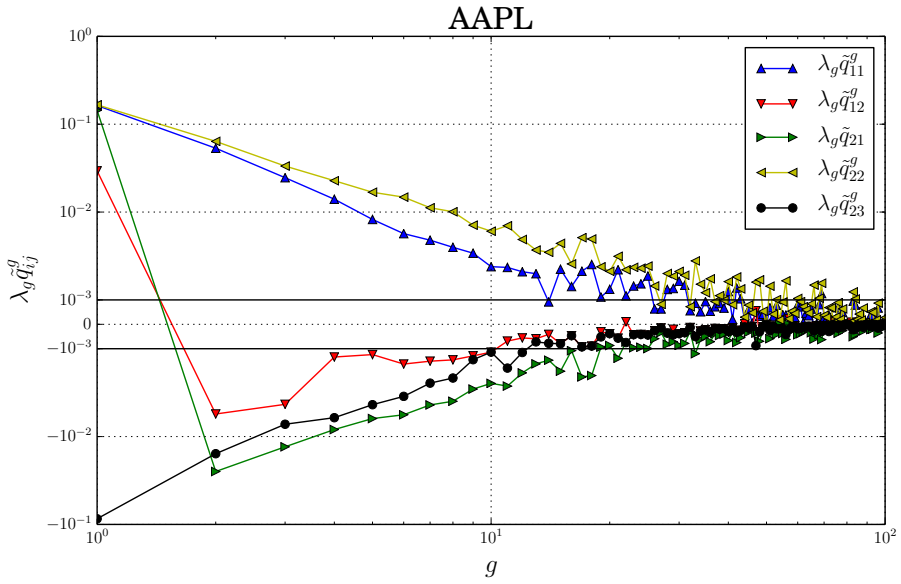

Figure 6.5: *GMM calibration of the weakly constrained MTDg.* Plot of the parameters $a^g_{i,j}$ solution of the optimization problem of Eq. (5.39) for an MTDg of order $p = 100$ model estimated from AAPL data. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

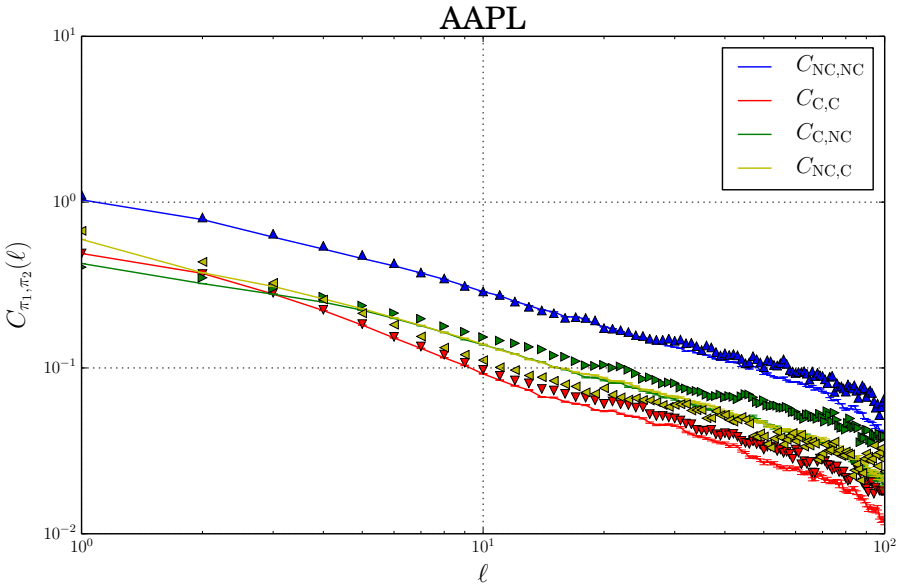

Figure 6.6: *GMM calibration of the weakly constrained MTDg.* Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on AAPL data (triangles) and the empirical curves (solid lines). The error bars correspond to one standard deviation.

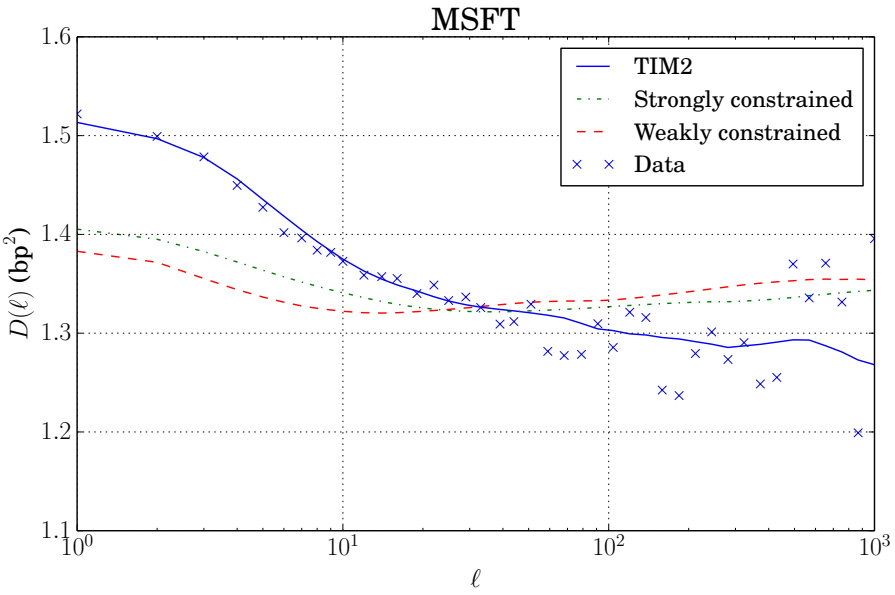

Figure 6.7: Signature plot for MSFT data: Empirical data (crosses), weakly constrained (GMM) MTDg(100) model with $D_{\mathrm{LF}} = 0.41$ (dashed line), strongly constrained (MLE) MTDg(100) model with $D_{\mathrm{LF}} = 0.43$ (dashed-dotted line), and the theoretical prediction of the calibrated TIM2 model, as in the bottom left panel of Figure 4.7.

in terms of the signature plot of the model, the weakly constrained and strongly constrained MTDg fare nearly equally well. We also show the predictions of the TIM2 model that uses the empirical $C_{\mathrm{C,C}}(\ell)$; the nearly perfect fit in this case is a consequence of the fact that $G_{\mathrm{C}}(\ell) \approx G_{\mathrm{C}}(1)$ for large tick stocks.

Note that the TIM2 price process is strictly diffusive only if the quantity $D^{\mathrm{TIM2}}(\ell+1)(\ell+1) - D^{\mathrm{TIM2}}(\ell)\ell$ is a constant independent from ℓ. In fact, we have that

$$D^{\mathrm{TIM2}}(\ell+1)(\ell+1) - D^{\mathrm{TIM2}}(\ell)\ell = D_{\mathrm{LF}} + G_{\mathrm{C}}(1)^2\mathbb{P}(\mathrm{C}) + 2G_{\mathrm{C}}(1)^2\mathbb{P}(\mathrm{C})^2 \sum_{0<n\leq\ell} C_{\mathrm{C,C}}(n)\,, \quad (6.16)$$

which means that the price process becomes diffusive for $\ell > \ell^*$ only if $C_{\mathrm{C,C}}(\ell > \ell^*) = 0$. Figures 6.1 and 6.7 suggest that this is indeed the case for $\ell^* \approx 10$.

6.5 Out-of-sample analysis

In the previous sections we have presented two MTDg models – strongly and weakly constrained – and discussed two alternative estimation methodologies based on MLE and GMM. Since they differ both in the number of parameters and in estimation efficiency, it is important to compare their performances testing the predictive power of the models in an out-of-sample analysis. We consider as a measure of the perfor-

mance the expected prediction error (EPE) defined as

$$\mathrm{EPE}(\boldsymbol{\theta}) = \mathbb{E}[L(X_t, \hat{X}_t^{\boldsymbol{\theta}})]\,, \tag{6.17}$$

where X_t is the observed process, $\hat{X}_t^{\boldsymbol{\theta}}$ is the predictor of X_t based on the model with parameter set $\boldsymbol{\theta}$, and the p past observations of the process X_t. As common in the literature for categorical data, we use as loss function the log-likelihood $L(X_t, \hat{X}_t^{\boldsymbol{\theta}}) = -2\sum_{i=1}^{m} I(X_t = i)\log(\hat{\chi}_t)_i = -2\log(\hat{\chi}_t)_{X_t}$, also called cross-entropy. We remind that $\hat{\chi}_t$ is the m-probability vector describing the prediction of the model and in the previous formula we take the X_t-th component. For the MTDg(p) model this probability vector is

$$\hat{\chi}_t = \sum_{g=1}^{p} \chi_{t-g}\lambda_g \boldsymbol{Q}^g \tag{6.18}$$

where, as before, χ_{t-g} is a m-vector of zeros with the exception of the realized component X_{t-g}. This quantity can be easily computed once the model is calibrated, since it depends on the transition probabilities. EPE values are in the range $[0, +\infty)$, and it is zero if all probabilities $(\hat{\chi}_t)_{X_t}$ of the sample are equal to one (perfect prediction), and it is infinity if all probabilities $(\hat{\chi}_t)_{X_t}$ of the sample are zero (prediction of impossible events).

We evaluate the best performing model as the model with the lowest EPE and benchmark the MTDg with a model with the unconditional probabilities as predictors of future signed events. Table 6.1 reports all the EPE values for different models of the predictor estimated on MSFT, Bank of America-CitiGroup (BAC), General Electric (GE), Cisco (CSCO), AAPL, and Amazon (AMZN) data. The scheme of the out-of-sample analysis is the following: The model is trained on a time period of 10 days, then we compute the loss functions in the following trading day by using the parameter set provided by MLE (strongly constrained) or by GMM (weakly constrained). We repeat the procedure by shifting the estimation by one trading day ahead. Finally, we compute the global loss by averaging all measured loss function. The financial interpretation of the EPE values is clear in the case of the large tick stocks, because a price-changing event moves the price by one tick with probability almost one and thus there exists a direct relation between the states of the MTDg model and the price return. Hence, for large tick stocks the EPE value can be employed as a proxy of the predictability of returns at high frequency time scale.

From Table 6.1 we see that both MTDg models out-perform the benchmark. More importantly, there is a clear evidence that the weakly constrained model with the highest number of parameters (Model C) outperforms the strongly constrained MTDg, for all considered stocks. These results exclude the over-fitting hypothesis, and support the claim that weakly constrained MTDg models are good candidates to capture the high-frequency dynamics of signed events.

		Model A	Model B	Model C
MSFT	EPE	1.928	1.199	1.181
	SE	0.003	0.004	0.004
BAC	EPE	1.744	0.799	0.785
	SE	0.003	0.004	0.004
GE	EPE	1.922	1.169	1.153
	SE	0.004	0.005	0.005
CSCO	EPE	1.919	1.112	1.098
	SE	0.004	0.005	0.005
AAPL	EPE	2.643	2.211	2.192
	SE	0.001	0.002	0.002
AMZN	EPE	2.579	2.196	2.183
	SE	0.002	0.004	0.004

Table 6.1: EPE values and standard errors (SE) for MSFT, BAC, GE, CSCO, AAPL and AMZN data. *Model A*: Unconditional probabilities as predictor. *Model B*: Strongly constrained MTDg(100) estimated via MLE according to Equation (6.8). Total number of parameters: 11. *Model C*: Weakly constrained MTDg(100) model estimated via GMM with matrices as in Equation (6.14). Total number of parameters: 500.

6.6 Discussion and partial conclusions

The last part of this Thesis has established that treating all market orders on the same basis produces erroneous predictions both for the "response functions" (average lagged impact) at negative lags and the signature plot. Single-propagator models and history dependent impact models are not designed to capture the feedback effects between past price returns and future order flow. These serious discrepancies have been significantly reduced by introducing the extended versions of the linear impact models (TIM and HDIM) which consider a richer set of signed events (see Eisler et al., 2012b,a). The argument which has motivated our generalization of the impact models is the observation that price-changing and non price-changing events have to be treated differently. This is particularly evident for large tick stocks, where price moving events are extremely rare but very informative. This apparently minor modification has lead to an extended class of propagator models which describe with remarkable realism the intertwined high-frequency dynamics of prices and order flow. Nonetheless, the linear description of the market dynamics achieved in the Chapter 4 is still too rigid: these models are designed to describe the evolution of the market with an exogenously specified order flow. This fact seriously limits the forecasting capabilities of linear impact models.

The Mixture Transition Distribution model partly solves the above issue by introducing an explicit stochastic model for the order flow, treated as an endogenous component of the dynamics. It is specially designed for variables which are inherently discrete – a feature of great relevance for price returns of large tick stocks. Here we have presented a class of so-called MTDg models as a natural extension of

the Discrete Autoregressive DAR(p) in a multi-event context. Our aim was to test how well a calibrated MTDg model can account for the statistics of the order flow, i.e. the string of 4 events: buy/sell – price changing/non changing events. One of the most interesting aspects of our work is methodological, and concerns the nature of the restrictions to impose on large models. We have restricted the MTD within the class of ergodic Markov models of order p. The existence and uniqueness of the stationary solutions for the ergodic model, and the buy/sell symmetry of the order book data have motivated the introduction in Section 5.5 of the class of weakly constrained MTDg models. They represent a rich family of discrete models, where the number of free parameters equates the number of independent observable correlation functions. This fact allows to introduce a numerical procedure which solves the estimation of the model parameters in a remarkably robust way. This result is rooted on the proof that the optimization problem is convex in the parameter space. From the financial viewpoint we have shown that – perhaps surprisingly – a weakly constrained version of the MTDg models captures the dynamics of signed events with greater realism than alternative and more parsimonious versions. Despite the large number of parameters, the out-of-sample analysis confirms that such good performances are achieved without over-fitting the data.

The improvement brought by the MTDg models and the new estimation methodology is remarkable, but still some discrepancies persist when comparing the model predictions with the empirical correlation functions. Several reasons may be responsible for these deviations. The first one was already pointed out by Raftery (1985) where he has shown that there exist regions of correlations which simply cannot be reproduced by MTDg models. A second reason is that, even though the MTDg model was the correct data generating process, the estimation methodology which involves information only coming from second order conditions, may lack efficiency with respect to the MLE approach. Finally, the MTDg model represents a parsimonious approximation of a full Markov chain of order p. This parsimony may come at expense of the realism of the model.

From a microstructural point of view, we can hypothesize that the string of past signed events $X_{t-1}, \ldots, X_{t-p}$ is not informative enough to predict the value X_t. In particular, for large tick stocks price-changing events $\pi = \mathrm{C}$ are much rarer than non price-changing event $\pi = \mathrm{NC}$. Therefore, a $\pi = \mathrm{C}$ event is by construction difficult to predict with past information based only on realised signs and trades. Hence, the behavior that we observe may be ascribed to a problem of missing explanatory variables. A natural candidate in this respect could be the volume of orders outstanding at the opposite side of the limit order book before the execution of a trade order, i.e. the local order book imbalance. A similar reasoning suggests that an important role – different from that of the trade orders – should be played by limit orders and cancellations. Their impact on quote revision is not taken under consideration explicitly in the present version of the paper. A richer description could be considered in a future implementation of the MTD model, which, however, is beyond the scope of the present investigation. In this respect, it will be also interesting to investigate whether it is convenient to move to a description of the limit order book and price formation process in the physical or event time

– the clock moves every time something changes in the order book – instead of the present description based on the trade time. We live this extension for future research, too.

From a more fundamental point of view, we should also point out that the MTDg calibrated kernel, which gives the probability that an event at time $t = 0$ will trigger similar or opposite events at time $t = g$ later, must be interpreted with care. Indeed, this kernel receives contributions both from order splitting, which increases the probability that an agent places an order of the same sign in the future, and from genuine reactions of the rest of the market to this event (Tóth et al., 2012, 2015). These reactions can be herding (copy cat trades) or, on the contrary, trades in the other direction (coming e.g. from liquidity providers). The response of the order flow to a single, isolated trade is thus expected to be rather different from the impulse function obtained by calibrating an MTDg model to the full order flow since order splitting contributions will be absent in the former, but contribute to the latter. The distinction between the two effects requires trade identification to be resolved as stated in (Tóth et al., 2017).

Conclusions

Market microstructure is a branch of finance concerned with the details of how exchange occurs in markets. The enormous quantity of data available of the trading activity in electronic platforms allows to study in details the microstructural effect of the price formation. The determinants of market liquidity, volatility, and market efficiency can be explored and modeled. Striking proprieties of times series at very low time scales emerge from the analysis of this data. Here, we summarize the main contributions of this Thesis.

In the Chapter 3 we have considered the subtle issue of reconciling the persistence of order flow with price efficiency and return diffusivity. It is empirically verified that on average a buy order pushes the price up while a seller initiated trade pushes it down. The positive correlation of order signs would naively lead to strongly correlated returns. However, the empirical evidence of price efficiency is in conflict with this view. Mechanisms which are able to take into account the persistence of the order flow and price efficiency have been proposed in the past. In this Thesis we focused in the asymmetric liquidity mechanism, which states that the price impact of an order is inversely related to the probability of its occurrence. We found, after a deep empirical analysis of real data, that on the liquidity taking side efficiency is guaranteed by the agents initiating the trade and adjusting the volume of their trades to the volume outstanding on the opposite side of the order book. In the second part of the chapter we have introduced a statistical model for the order book dynamics designed for large tick stocks, which is able to reproduce the empirical findings.

In the Chapter 4 we proposed a generalization of the propagator model in order to better reproduce market impact and the diffusive behaviour of the price process. We have relaxed the assumption that all the orders at any time have the same impact on the market. The simple decoupling of the impact of trades which trigger or not a price change, improves significantly the understating of the relation between the price process and the order flow. In particular, the signature plot of the price process can be reproduced perfectly.

Finally, in the Chapters 5 and 6 we proposed how to deal with the discretization effect of the high frequency price process. We found that the framework of high order Markov chains can be used to model the price process. In particular, the Mixture Transition Distribution model, which is an approximation of the full Markov chains, can be estimate on real data. We proposed, also, a subclass of MTD models for which

the Generalized Method of Moments can be applied to estimate the coefficient of the model. The estimation procedure is proved to be convex, therefore it solves the issue of the high dimensionality of the problem.

Bibliography

Arthur, W. B., S. N. Durlauf, and D. A. Lane (1997). *The economy as an evolving complex system II*, volume 28. Reading, MA: Addison-Wesley. 2, 20

Bachelier, L. (1900). Théorie de la spéculation. In *Annales scientifiques de l'École Normale Supérieure*, volume 17, Pp. 21–86. Elsevier. 12

Bacry, E. and J.-F. Muzy (2014). Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166. 66, 79, 105

Benveniste, L. M., A. J. Marcus, and W. J. Wilhelm (1992). What's special about the specialist? *Journal of Financial Economics*, 32(1):61–86. 10

Berchtold, A. (1995). Autoregressive modelling of Markov chains. In *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling*, Pp. 19–26. Springer-Verlag. 106, 110, 111

Berchtold, A. (2001). Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22(4):379–397. 114

Berchtold, A. and A. E. Raftery (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17(3):328–356. 110, 113

Bershova, N. and D. Rakhlin (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, 13(11):1759–1778. 68

Biais, B., P. Hillion, and C. Spatt (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse. *The Journal of Finance*, 50(5):1655–1689. 79

Boggs, P. T. and J. W. Tolle (1995). Sequential Quadratic Programming. *Acta Numerica*, 4:1–51. 128

Bouchaud, J.-P., J. D. Farmer, and F. Lillo (2009). *How markets slowly digest changes in supply and demand*, volume 4. North-Holland, San Diego, CA: Handbook of Financial Markets. 3, 11, 21, 23, 35, 66, 79

Bouchaud, J.-P., Y. Gefen, M. Potters, and M. Wyart (2004). Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quantitative Finance*, 4(2):176–190. 2, 4, 23, 27, 28, 30, 35, 66, 79, 80, 81, 84, 106

Bouchaud, J.-P., J. Kockelkoren, and M. Potters (2006). Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance*, 6(02):115–123. 23, 27, 28, 34, 80, 81

Bouchaud, J.-P., M. Mézard, and M. Potters (2002). Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256. 15, 16

Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. New York, NY: Springer Science & Business Media. 107

Campbell, J. Y., A. W.-C. Lo, and A. C. MacKinlay (1997). *The econometrics of financial markets*, volume 1. Princeton, NJ: Princeton University Press. 12

Chen, D.-G. and Y.-L. Lio (2009). A novel estimation approach for mixture transition distribution model in high-order Markov chains. *Communications in Statistics - Simulation and Computation*, 38(5):990–1003. 114

Cont, R. and A. De Larrard (2013). Price dynamics in a Markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25. 37, 59, 60

Cont, R., A. Kukanov, and S. Stoikov (2014). The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88. 79

Cont, R., S. Stoikov, and R. Talreja (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563. 37

Curtis, F. E. and M. L. Overton (2012). A Sequential Quadratic Programming algorithm for nonconvex, nonsmooth constrained optimization. *SIAM Journal on Optimization*, 22(2):474–500. 128

Daniels, M. G., J. D. Farmer, L. Gillemot, G. Iori, and E. Smith (2003). Quantitative model of price diffusion and market friction based on trading as a mechanistic random process. *Physical Review Letters*, 90(10):108102. 17, 27, 37

Donier, J., J. Bonart, I. Mastromatteo, and J.-P. Bouchaud (2015). A fully consistent, minimal model for non-linear market impact. *Quantitative Finance*, 15(7):1109–1121. 80

Dufour, A. and R. F. Engle (2000). Time and the price impact of a trade. *The Journal of Finance*, 55(6):2467–2498. 79

Einstein, A. (1905). On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heart. *Annalen der Physik*, 17:549–560. 12

Eisler, Z., J.-P. Bouchaud, and J. Kockelkoren (2012a). Models for the impact of all order book events. In *Market Microstructure: Confronting Many Viewpoints*, F. Abergel, J.-P. Bouchaud, T. Foucault, C.-A. Lehalle, and M. Rosenbaum, eds., Pp. 113–135. Oxford, UK: John Wiley & Sons Ltd. 80, 81, 83, 86, 97, 98, 102, 105, 122, 134

Eisler, Z., J.-P. Bouchaud, and J. Kockelkoren (2012b). The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12(9):1395–1419. 3, 63, 80, 81, 85, 95, 105, 122, 134

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417. 2, 19, 20

Farmer, J. D., A. Gerig, F. Lillo, and S. Mike (2006). Market efficiency and the long-memory of supply and demand: Is price impact variable and permanent or fixed and temporary? *Quantitative Finance*, 6(2):107–112. 31, 33, 66

Farmer, J. D., A. Gerig, F. Lillo, and H. Waelbroeck (2013). How efficiency shapes market impact. *Quantitative Finance*, 13(11):1743–1758. 69

Farmer, J. D., L. Gillemot, F. Lillo, S. Mike, and A. Sen (2004). What really causes large price changes? *Quantitative Finance*, 4(4):383–397. 16, 25, 49

Farmer, J. D., P. Patelli, and I. I. Zovko (2005). The predictive power of zero-intelligence in financial markets. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):2254–2259. 18, 37, 59

Foucault, T., B. Biais, and P. Hillion (1997). *Microstructure des marchés financiers: institutions, modèles et tests empiriques*. Presses Universitaires de France-PUF. 10

Foucault, T., S. Moinas, and E. Theissen (2007). Does anonymity matter in electronic limit order markets? *Review of Financial Studies*, 20(5):1707–1747. 10

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley (2006). Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2):461–504. 68

Garfinkel, J. A. and M. Nimalendran (2003). Market structure and trader anonymity: An analysis of insider trading. *Journal of Financial and Quantitative Analysis*, 38(03):591–610. 10

Gerig, A. (2007). *A Theory for Market Impact: How Order Flow Affects Stock Price*. PhD thesis, University of Illinois. 31, 33

Glosten, L. R. and P. R. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100. 10

Grossman, S. (1976). On the efficiency of competitive stock markets where trades have diverse information. *The Journal of Finance*, 31(2):573–585. 19

Grossman, S. J. and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408. 21

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton, NJ: Princeton University Press. 76

Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press, USA. 10

Harris, L. E. (1990). Liquidity, trading rules and electronic trading systems. Technical report, Southern California-School of Business Administration. 10

Hasbrouck, J. (1988). Trades, quotes, inventories, and information. *Journal of Financial Economics*, 22(2):229–252. 79, 105

Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, 46(1):179–207. 79, 105

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, 50(4):1175–1199. 11, 12

Hasbrouck, J. (1996). Modeling market microstructure time series. *Handbook of Statistics*, 14:647–692. 11

Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press. 1

Jacobs, P. A. and P. A. W. Lewis (1978). Discrete time series generated by mixtures III: Autoregressive processes (DAR(p)). Technical report, Naval Postgraduate School Technical Report, Monterey, CA. 40, 83, 106

Jacobs, P. A. and P. A. W. Lewis (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36. 40

Jones, C. M., G. Kaul, and M. L. Lipson (1994). Transactions, volume, and volatility. *Review of Financial Studies*, 7(4):631–651. 79

Karlin, S. and H. E. Taylor (1981). *A second course in stochastic processes*. Elsevier. 107

Kavajecz, K. A. (1999). A specialist's quoted depth and the limit order book. *The Journal of Finance*, 54(2):747–771. 10

Kemeny, J. G. and J. L. Snell (1976). *Finite Markov chains*, volume 356. New York, NY: Springer-Verlag. 107

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 53(6):1315–1335. 20

Lèbre, S. and P.-Y. Bourguignon (2008). An EM algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation*, 78(8):713–729. 114

Lee, C. M. C., B. Mucklow, and M. J. Ready (1993). Spreads, depths, and the impact of earnings information: An intraday analysis. *Review of Financial Studies*, 6(2):345–374. 10

Lillo, F. and J. D. Farmer (2004). The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3):1. 2, 4, 23, 28, 31, 33, 34, 35, 40, 50, 54, 74, 79, 81, 98, 106

Lillo, F., J. D. Farmer, and R. N. Mantegna (2003). Econophysics: Master curve for price-impact function. *Nature*, 421(6919):129–130. 25

Lillo, F., S. Mike, and J. D. Farmer (2005). Theory for long memory in supply and demand. *Physical Review E*, 71(6):066122. 24, 67, 68, 70, 79

Lintner, J. (1969). The aggregation of investor's diverse judgments and preferences in purely competitive security markets. *Journal of Financial and Quantitative Analysis*, 4(4):347–400. 19

Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258. 1, 10

Madhavan, A., M. Richardson, and M. Roomans (1997). Why do security prices change? A transaction-level analysis of NYSE stocks. *Review of Financial Studies*, 10(4):1035–1064. 31, 83

Mastromatteo, I., B. Tóth, and J.-P. Bouchaud (2014). Agent-based models for latent liquidity and concave price impact. *Physical Review E*, 89(4):042805. 38, 59, 63, 68, 80, 81

Mike, S. and J. D. Farmer (2008). An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control*, 32(1):200–234. 15, 38, 59

O'Hara, M. (1995). *Market microstructure theory*, volume 108. Blackwell Cambridge, MA. 1, 10

O'Hara, M. (2003). Presidential address: Liquidity and price discovery. *The Journal of Finance*, 58(4):1335–1354. 11

Pegram, G. G. S. (1980). An autoregressive model for multilag Markov chains. *Journal of Applied Probability*, 17(02):350–362. 108, 112

Potters, M. and J.-P. Bouchaud (2003). More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 324(1):133–140. 25, 27

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539. 4, 106, 110, 111, 114, 135

Raftery, A. E. and S. Tavaré (1994). Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model. *Applied Statistics*, Pp. 179–199. 117

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6:41–49. 2, 19, 20

Schreiber, P. S. and R. A. Schwartz (1986). Price discovery in securities markets. *The Journal of Portfolio Management*, 12(4):43–48. 11

Smith, E., J. D. Farmer, L. Gillemot, and S. Krishnamurthy (2003). Statistical theory of the continuous double auction. *Quantitative Finance*, 3(6):481–514. 17, 37, 38, 58, 60

Tóth, B., Z. Eisler, and J.-P. Bouchaud (2017). The short-term price impact of trades is universal. *arXiv preprint arXiv:1702.08029*. 121, 136

Tóth, B., Z. Eisler, F. Lillo, J. Kockelkoren, J.-P. Bouchaud, and J. D. Farmer (2012). How does the market react to your order flow? *Quantitative Finance*, 12(7):1015–1024. 89, 121, 136

Tóth, B., Y. Lemperiere, C. Deremble, J. De Lataillade, J. Kockelkoren, and J.-P. Bouchaud (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006. 3, 38, 59, 60, 62, 63, 68, 70, 80

Tóth, B., I. Palit, F. Lillo, and J. D. Farmer (2015). Why is equity order flow so persistent? *Journal of Economic Dynamics and Control*, 51:218–239. 24, 35, 67, 79, 121, 136

Vaglica, G., F. Lillo, E. Moro, and R. N. Mantegna (2008). Scaling laws of strategic behavior and size heterogeneity in agent dynamics. *Physical Review E*, 77(3):036110. 68

Zovko, I. and J. D. Farmer (2002). The power of patience: a behavioural regularity in limit-order placement. *Quantitative Finance*, 2(5):387–392. 15