RETHINKING MARKET IMPACT: QUANTITATIVE ANALYSIS OF TRADE EXECUTION COSTS

A thesis presented for the degree of
Doctor of Philosophy of Imperial College London
and the
Diploma of Imperial College
by

Francesco Capponi

Department of Mathematics
Imperial College London
London, SW7 2AZ

December, 2020

Declaration of Originality

I hereby certify that the present thesis and the research to which it refers are the product of my own work. All research, ideas and work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Copyright

Thesis advisor: Professor Rama Cont Francesco Capponi

Rethinking Market Impact:
Quantitative Analysis of Trade Execution Costs

Abstract

This thesis is a quantitative study of the notion of 'market impact' and develops a new quantitative framework to understand the main determinants of execution costs.

In the first part, we develop a market-microstructure model in which the market impact of an order is a latent variable that depends on the aggregate activity of the order book, as summarised by the *order flow imbalance*. The model stems from the observation that at each point in time, the liquidity in the limit order book available for immediate execution is very small, and most of the liquidity remains hidden. We empirically test this model using the Kalman filter and show that it delivers good out-of-sample forecasts.

The second part is devoted to the identification of the main determinants of execution costs. Using a large dataset of institutional investors' transactions in the US equity market, we find that, for *orderly* trade executions, the main determinant of execution costs is not the 'market impact' of trades but price volatility. Our *causal* analysis shows that price changes during execution are mainly driven by the *aggregate order flow imbalance*, rather then by the characteristics of individual trades, and that *causal* claims that directly relate individual trade characteristics with execution costs may be spurious.

In the third part, we investigate the notion of *cross-impact*. We show that the observed correlation between the returns of an asset and the order flow of other assets is due to the presence of common components in order flow across stocks, which may naturally arise from multi-asset trading strategies. We provide empirical evidence from order flow and returns of NASDAQ-100 stocks to supports this explanation. This leads to a parsimonious approach for *causal* modelling of multi-asset impact, which does not require introducing any concept of cross impact.

To my family, friends and mentors.

Sapere Aude.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Rama Cont. He has been a constant source of inspiration during my PhD and a real mentor. His guidance, his contagious enthusiasm, and his unwavering support and encouragement have helped me become a better version of myself, both personally and professionally. Without him, this thesis would not have been possible. I will always fondly remember our meetings and discussions where he would inspire me to challenge things and to aim high. It has been a privilege to write my PhD thesis under his supervision and to know him personally. Thank you Rama!

I am very grateful to Professor Walter Distaso and Doctor Sasha Stoikov for kindly accepting to be on my examination committee.

I would like to thank Doctor Eyal Neuman for his support and insights, and for helping me navigate my PhD in difficult times.

Many thanks go to my fellow PhD students, and to the academic and administrative staff at the Mathematical Finance Section of Imperial College London for the knowledge I gained from them and for their assistance.

Special thanks to all my friends, and in particular to Paolo Calello, Alicia Higueruelo, Kenneth Querns Langley, Luca Pozzi, and Tom Smet for supporting me during these years and for believing in me.

My final deep gratitude goes to my family: my parents Irma and Giuseppe, my sister Valeria, and my late grandmother Marianna, for always encouraging me, and helping me along this journey.

Contents

0 Introduction

0.1 A changing landscape

Understanding price formation in financial markets is one of the key goals of academic research in finance. In the classic *Walrasian auction* (Walras, 1900), buyers and sellers communicate their trading intentions to an auctioneer, who collects this information into a book. The traders' offers are reflected in the *supply* curve $\mathcal{S}(p)$, which, for each given price p, represents the total volume of sell orders with a price smaller than or equal to p. Similarly, the traders' bids are reflected in the *demand* curve $\mathcal{D}(p)$, which, for each given price p, represents the total volume of buy orders with a price greater than or equal to p. The traders' bids and offers are firm commitments to trade, and are the main way liquidity is provided to the market. In this setting, the auctioneer solves the auction by setting a market price p^* so as to maximise the total volume exchanged for the asset. According to *Walras' law*, this clearing price is set at the intersection of the supply and demand curves. This mechanism is illustrated in Figure 1, where the intersection of the upward-sloping supply curve and the downward-sloping demand curve gives rise to an equilibrium characterised by the price p^* and quantity $Q^* = Q(p^*)$.

In Walrasian auctions, traders cannot update their decisions after observing other market participants' actions, since the auctioneer's book remains invisible to them. This lack of coordination among traders may lead to undesirable outcomes, such as the sub-optimal allocation of the asset, or a situation where a clearing price cannot be found. One solution adopted by most exchanges to obviate this problem was to designate *market-makers* or *specialists* who, in exchange for some privileges, would execute the Walrasian auctioneer's role of maintaining fair and orderly markets. Market-makers had the legal obligation to quote, at each point in time, bid and ask prices and quantities, thus facilitating trading and becoming the markets' *liquidity providers*.

Modern financial markets have largely moved away from the Walrasian auctioneer and from the designation of market-makers and specialists. Floor trading has been replaced by *electronic trading*, and designated market-makers or specialists by a *continuous double auction*

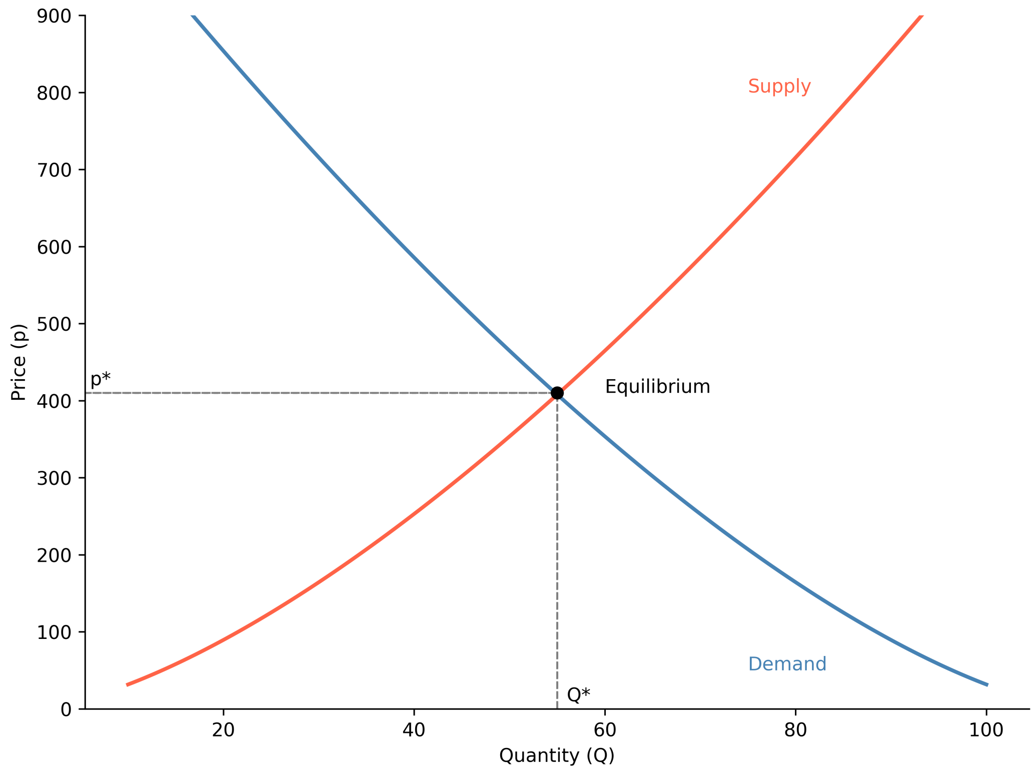

Figure 1: Illustration of the demand $\mathcal{D}(p)$ and supply $\mathcal{S}(p)$ curves. According to Walras' law, the equilibrium price p^* and the total quantity exchanged Q^* are given by the intersection of the two curves.

mechanism. Incoming buy and sell orders, often generated electronically, are now recorded in the *limit order book*, and are matched automatically through *matching engines* whenever a buyer and a seller agree on a price. The distinction between 'liquidity providers' and 'liquidity takers' has also been blurred, since all traders may now take and provide liquidity at the same time. This market structure evolution is most advanced in equity markets, but other markets, such as foreign exchange, futures, and fixed income, seem to be headed in the same direction.

Although continuous double auctions are close to the Walrasian auctioneer, they differ from it in one fundamental way: agents do not necessarily reveal their trading intentions in the limit order book for fear of revealing private information. Thus, at any point in time, the visible liquidity in the limit order book reveals very little about the true underlying supply and demand curves, which remain mostly *latent*.

The widespread adoption of electronic trading has led to an extraordinary increase in the speed at which transactions are executed, with decisions taken at sub-millisecond frequency. It has also consequently led to a 'data revolution', with large amount of high-resolution, high-quality data generated each day by exchanges. This has opened up new opportunities to better understand the intricate workings of financial markets and, more specifically, it has made it possible to analyse at a microscopic level the complex interactions between buyers and sellers in the *price formation process*.

The research area that deals with the analysis of these interactions between buyers and sellers at transaction level, is called *market microstructure*. Market microstructure is, by nature, a multi-disciplinary research subject, where finance, mathematics, and statistics intersect. Despite this 'hybrid' nature, its ultimate goals are quite specific: understanding the liquidity-provision mechanism and the price-formation process. Practitioners use this research in order to formulate quantitative models that help them predict price changes, formulate optimal trading strategies and monitor and reduce *transaction costs*. Regulators try to use the insights from market microstructure research to ensure the fair and orderly functioning of financial markets.

0.2 Limit order books

Today, in the most liquid markets, the standardisation of trading protocols and the evolution towards electronic trading have given rise to the adoption of the limit order book (LOB) as a key tool to facilitate the coordination between buyers and sellers.

In LOBs, the price formation process is not in the hands of a market-maker or specialist, as in *quote-driven* markets, but is decentralised. Market participants can manifest their intentions to trade in the LOB by submitting *limit orders* or *market orders*. Limit orders specify a *limit price*, which is the maximum (minimum) price at which the trader is willing to buy (sell) a certain quantity of the asset. Once issued, if the new limit order can be matched immediately by the matching engine with a pre-existing *passive* order from the other side of the book, a trade occurs. In this case, the order is called *marketable* or *aggressive* order, since it initiates the trade. If the limit order cannot be matched immediately, it enters a *queue* of existing orders for the same price, and remains in that queue unless otherwise *cancelled* or *matched*. In contrast to limit orders, market orders do not specify a limit price. They are orders to buy or sell a certain quantity of an asset at the best price available when the order is issued.

Limit orders can also be seen as option contracts through which a trader offers the market the option to buy or sell a certain quantity of the asset for a specified price. By offering this free option to the market, the trader hopes to execute the trade at a more favourable price than by submitting a market order. However, when submitting a limit order, the trader faces execution uncertainty, since the limit order may never be *filled*. Market orders remove the execution uncertainty of limit orders at the cost of being matched at a worse price. While limit orders provide liquidity to the market, market orders take liquidity from the market.

The distinction between liquidity providers and liquidity takers, which was applicable in markets where market-makers where the only market participants in charge of quoting prices and volumes, has been blurred in markets organised around limit order books. Traders, and their electronic execution algorithms, can now be both liquidity providers and liquidity takers.

In limit order books, the price dynamics are governed by the rules that dictate how orders are matched. The most common rule currently used is the *price-time priority*, where priority is given to buy (sell) limit orders with the highest (lowest) prices that have been submitted earlier. Another mechanism, used for example in Eurodollar futures, is the *pro-rata priority*, in which, in case of a tie, each limit order receives a share equal to the fraction of the submitted volume to the total available volume at the best (Harris, 2003). While price-time priority encourages the early submission of limit orders, pro-rata priority encourages traders to place large limit orders, and therefore encourages the provision of liquidity.

Figure 2 shows an LOB snapshot for Apple Inc. (AAPL) on the NASDAQ stock exchange on September 1st, 2020, with data taken from the LOBSTER database (Huang and Polak, 2011).[1] In an LOB, the *best bid* refers to the queue of buy limit orders with the highest price, p^b, called the *bid price*. Similarly, the *best ask* refers to the queue of sell limit orders with the lowest price, p^a, called the *ask price*. The average of the best bid and ask prices at time t is called the *mid price*:

$$p_t \equiv \frac{1}{2}(p_t^b + p_t^a), \tag{1}$$

while the difference between the ask and bid prices called the *bid-ask spread*:

$$s_t \equiv p_t^a - p_t^b. \tag{2}$$

[1] https://lobsterdata.com.

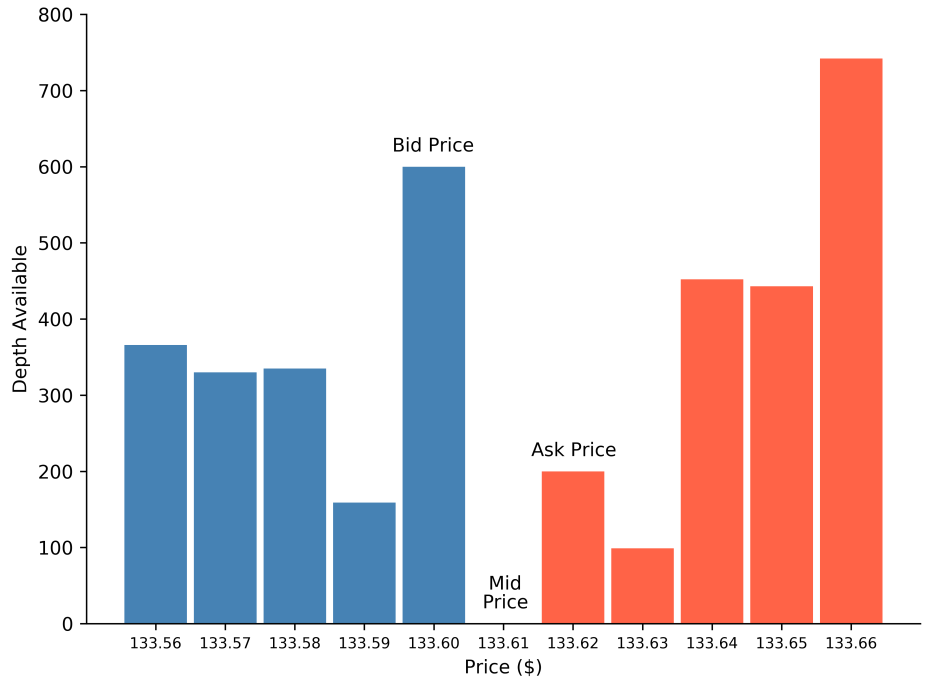

Figure 2: Snapshot of the limit order book for Apple Inc. (AAPL) on the NASDAQ stock exchange on September 1st 2020 at 13:30:33.06. Author computations based on data from LOBSTER (Huang and Polak, 2011).

The bid-ask spread is sometimes considered a measure of market liquidity. Another measure of LOB liquidity is the market *depth*, $q_t^i, i = a, b$, which, for each price level $p_t^i, i = a, b$, denotes the total volume available for execution at that price level.

The *tick size*, τ, is the smallest possible price interval between orders and determines the price a trader has to pay to achieve priority. Since assets trade at very different prices, a more useful quantity is represented by the *relative tick size*, which is the tick size divided by the mid-price of the asset:

$$\tilde{\tau}_t \equiv \frac{\tau}{p_t}. \tag{3}$$

Small-tick stocks are defined as stocks with a very small relative tick size, while *large-tick* stocks are defined as stocks with a large relative tick size.

A useful quantity that summarises the interplay between limit orders, market orders, and cancellations in an LOB is the *order flow imbalance* (OFI), introduced by Cont et al. (2014), which represents the net order flow at the best bid and ask levels. Consider a time interval $[t-\Delta t, t]$ and denote by L_t^b, C_t^b the volume of new limit buy orders submitted to and cancelled from the current best bid during that time interval, respectively. Also denote by M_t^b the volume of marketable sell orders that arrived at the current best bid. Similarly, for the ask side, we have the analogous quantities L_t^a, C_t^a, M_t^a. Then, we can define the OFI as (Cont et al., 2014):

$$OFI_t \equiv (L_t^b - C_t^b - M_t^b) - (L_t^a - C_t^a - M_t^a), \tag{4}$$

$$\equiv V_t^b - V_t^a, \tag{5}$$

where V_t^b and V_t^a denote, the volume of orders posted at the best bid and ask queues during the time interval $[t - \Delta t, t]$, respectively. OFI is a measure of excess demand.

0.3 Execution costs

Understanding price dynamics in the limit order book is key for market participants. In particular, the analysis of *execution costs* has generated renewed interest in recent years, mainly as a result of the shift from floor to electronic trading, and the subsequent increase in the speed and number of transactions executed. When implementing a trading strategy, traders and portfolio managers are now increasingly aware of the potential drag on the performance of their trading strategy due to execution costs, and have moved towards incorporating this analysis into their decision-making toolbox.

In this sense, the analysis of execution costs, sometimes called *transaction cost analysis (TCA)*, can be seen as the tool that traders and portfolio managers use to achieve *best execution*. Indeed, best execution can be described as: "the process of maximising the investment idea" (Wagner and Edwards, 1993), and managing transaction costs during all phases of the investment decision process is a key component of this. It consists of *pre-trade*, *intraday*, and *post-trade* analysis to ensure that the execution strategies are in line with the trading decisions (Kissell, 2014).

According to Wagner and Edwards (1993) and Kissell (2006), execution costs can be classified into several separate components and categorised according to the phase of the investment cycle in which they occur. *Investment-related costs* arise during the investment decision phase, which is the period of time from the investment decision to the time the order is released to the market (Kissell, 2006). *Trading-related costs* arise during the implementation of the investment decision, which is the time period from the start of trading to the end of trading (Kissell, 2006). And the *opportunity cost* represents the foregone profit (or avoided loss) which arise from not being able to execute the entire order within the trading period (Kissell, 2006).

Investment-related costs comprise *taxes* on realised earnings, and the *delay cost*. The last one is the loss in value of the investment between the time the portfolio manager makes the investment decision and the time the order is released. Trading-related costs include *commissions*, *fees*, *rebates*, *spreads*, *price appreciation*, *market impact*, and *volatility or timing risk* (Kissell, 2006). Commissions are the payments made to brokers for executing trades, and vary depending on broker and trading difficulty. Fees, charged by brokers, and often included into the total commission cost, include exchange fees, clearing and settlement costs etc. Rebates are provided by some exchanges to market participants who provide liquidity. The bid-ask spread (2) compensates the brokers for matching buyers with sellers, and for the risk they face by accumulating or depleting their inventory while waiting to unwind the position (Kissell, 2006). The bid-ask spread also compensate the brokers for the potential of being *adversely selected* by informed investors. All these costs, with the exception of the delay cost, are *visible costs*. Some of them are *fixed* (commissions, fees, and rebates), while others are *variable* (taxes and spreads).

Among the trading-related costs, special attention has been given to price appreciation, market impact and volatility risk. These are all *hidden* and variable costs. Price appreciation, also called price trend or drift, represents the 'natural' movement of the price of the asset in absence of uncertainty (Kissell, 2006). It represents a cost in as much as the investor buys when the price of the asset is rising, or sells when the price of the asset is falling. It is a saving if the investor trades in the opposite direction of the trend of the asset price. Market impact refers to the fact that a trader's order might move the price in an unfavourable direction thereby causing higher execution costs. Market impact can be thought of as the deviation of the transaction price from the "unperturbed price" movement of the price, had the trade not occurred (Keim and Madhavan, 1998). Although theoretically simple, market impact is in fact difficult to measure because the unperturbed price is not observable (Keim and Madhavan, 1998). Finally, *volatility risk* is the cost paid, or savings made, due to the fact that, regardless of the investor's order, prices move during trade execution given the activity of other market participants.

Market impact costs have further been divided into a *temporary* and a *permanent* component. Temporary impact arises from the liquidity and urgency needs of the investor. It is the premium buyers have to pay, and the discount sellers have to provide, to attract additional investors to take the other side of the trade, since the investor's liquid needs cannot be fulfilled at the price specified in the order (Kissell, 2014). Permanent impact arises from the information content of the trade. Market participants adjust their prices after what they perceive to be an informed trade. Therefore, they might adjust their prices downward after a perceived informed sell trade, thinking the asset is overvalued, or revise them upward after an informed buy trade, thinking the asset is undervalued.

0.4 Literature review

Interest in understanding price formation in markets is not new. Already in the seventeenth century, Vega (1688) provided an accurate account of the functioning of the Amsterdam Stock Exchange. It is much more recently, however, that theoretically sound, and empirically validated research has been carried out in this realm. O'Hara (1995) and Madhavan (2000) provide comprehensive surveys of the market microstructure literature, while Harris (2003) gives a good overview of the trading mechanisms and the organisation of exchanges.

In contrast with the standard frictionless model of efficient markets, market microstructure research is interested in how frictions and asymmetric information affect trading and prices. Much of the early literature is therefore concerned with understanding these frictions and with developing models of price formation. In early studies, a central role is given to market-makers, who provide liquidity to the market by quoting the bid and ask prices and quantities in exchange for a compensation given by the bid-ask spread.

Demsetz (1968) presented an early empirical work on the bid-ask spread for NYSE stocks. The author affirms that the ask-bid spread is the markup that is paid to market-makers because, by standing ready and waiting to trade with the incoming orders, they provide immediacy to those who demand the immediate execution of their order (Demsetz, 1968). Following Demsetz (1968), most empirical studies focused on the determinants of the bid-ask spread, followed by theoretical studies that tried to explain bid-ask spreads in relation to prices. Smidt (1971) suggested that market-makers are not only compensated for providing immediacy, as proposed by Demsetz (1968), but they also actively set the bid-ask spread by taking into account their inventory positions. Following this intuition, several models of inventory control were proposed by Garman (1976), Stoll (1978), and Amihud and Mendelson (1980) among others.

The quoted bid-ask spread, however, may provide an incorrect estimate of the cost of trading with a market-maker, since trades are often executed inside the quoted spread; because bid (ask) prices tend to rise (fall) following a buy (sell) order, so that the true round-trip costs are smaller than the quoted spread, and because large trades do not necessarily occur at the quoted bid or ask prices (Keim and Madhavan, 1998). Several authors have therefore proposed alternative cost measures based on actual transaction prices called *effective bid-ask spread* (Roll, 1984; George et al., 1991; Madhavan et al., 1997).

But even effective bid-ask spread estimates, although informative, fail to give a complete picture of total execution costs. Large aggressive trades, whose size exceeds the quantity available at the best, may move prices in the direction of the trade and therefore incur in greater execution costs. In a very influential paper, Perold (1988) introduced the notion of *implementation shortfall* (*IS*) as a measure to represent the total cost of implementing an investment idea. The implementation shortfall is defined as the difference between the paper return of the portfolio, where all shares are assumed to have been executed at the portfolio manager's decision price, and the actual return of the portfolio, computed using actual transaction prices and size of the order executed.

To understand how the *IS* measure is constructed and used, suppose that a portfolio manager (PM) wants to buy Q shares of a stock at time t_{start}, which is the *authorisation time*. In a perfectly liquid market, without transaction costs, the PM's order would be filled at p_{start}, the price prevailing at the time the PM authorised the order, and the PM's (paper) return R_p over the investment horizon ending at time t_{end}, would be given by:

$$R_p \equiv Q(p_{end} - p_{start}) \tag{6}$$

Foucault et al. (2013), where p_{end} is the price at the end of the portfolio manager's investment horizon. However, given that the market is not perfectly liquid, the portfolio manager will normally buy the shares at a higher price than p_{start}. Suppose, only a fraction κ of the order is filled, at an average price p_{avg} (Foucault et al., 2013). Then, the actual return R_a, is given by:

$$R_a \equiv \kappa Q(p_{end} - p_{avg}). \tag{7}$$

The *IS* for this order is given by:

$$\begin{aligned} IS &\equiv R_p - R_a && (8)\\ &= \underbrace{\kappa Q(p_{avg} - p_{start})}_{\text{Execution Cost}} + \underbrace{(1-\kappa)Q(p_{end} - p_{start})}_{\text{Opportunity Cost}}. && (9) \end{aligned}$$

The first term in (9) is the execution cost, and captures the cost of transacting at an average price p_{avg} instead of the authorisation price p_{start}. The second term in (9) is the opportunity cost, and captures the foregone returns on the unexecuted part of the order.

A variant to Perold's implementation shortfall was proposed by Wagner and Edwards (1993). This 'expanded' version, in the words of Kissell (2006), breaks down execution costs based on where these costs occur in the investment cycle, and is constituted of a delay-related cost, a trading-related cost, and an opportunity cost. Implementation shortfall is still widely used by portfolio managers and traders to measure trading costs, and the performance of brokers and their algorithms.

The availability of transaction-level data, especially from institutional investors' trades, has greatly increase our understanding of transaction costs since this data became first available in the early 1990s. Hasbrouck (1991), Chan and Lakonishok (1993), Chan and Lakonishok (1995), Wagner and Edwards (1993), Hasbrouck and Sofianos (1993), Keim and Madhavan (1996), Keim and Madhavan (1997), Madhavan and Cheng (1997) are examples of early empirical studies that provide estimates of trading costs and their determinants. Empirical studies of market impact are difficult for a number of reasons. The first is the inherent *endogeneity* of the datasets. The reported trades might have been executed because certain market conditions were available at the time, or some of their characteristics, such as size or speed, might have been conditioned on the price trajectory or the available liquidity. Moreover, the unexecuted portion of the trade is very rarely available, so that the intended trade size is not known and the opportunity cost of the trade cannot be computed. Publicly available data also usually lack the necessary details to be able to re-construct the trade, sometimes called *parent order* or *meta-order* in the literature, executed by an investor. This is because trades are usually executed through several transactions of smaller sizes, sometimes called *child orders*, in order to minimise market impact. Without a trade identifier it is therefore impossible to link all child orders that are part of the same parent order. Some of these shortcomings are missing in proprietary datasets, which contain unexecuted trades, the intended trade size, and therefore the portion of the trade that has not been executed, and, of course, information that allows to reconstruct the trade or parent order. The drawbacks with these datasets is that they usually contain trades of only one institution, and that they are not publicly available for independent verification of the results.

Over time, empirical studies have uncovered some regularities in market impact. There seems to be a consensus that market impact is dependent on the *size* of the trade (Kraus and Stall, 1972; Loeb, 1983; Holthausen et al., 1987; Chan and Lakonishok, 1993, 1995; Torre, 1997, 1999; Lillo et al., 2003; Lillo and Farmer, 2004; Bouchaud et al., 2009; Toth et al., 2011; Bershova and Rakhlin, 2013; Cont et al., 2014; Bacry et al., 2015; Gomes and Waelbroeck, 2015; Toth et al., 2016); *volatility* and *market capitalisation* (Stoll, 1978; Amihud and Mendelson, 1980; Madhavan and Sofianos, 1998; Chan and Lakonishok, 1995; Keim and Madhavan, 1997); *market conditions* (Wagner and Edwards, 1993; Perold and Sirri, 1993; Huang and Stoll, 1994) ; and *the trade execution strategy* (Bertsimas and Lo, 1998; Almgren and Chriss, 2001).

In particular, several empirical studies that examine trades executed through several transactions (i.e. parent or meta-orders) point to a concave dependence of the amplitude of price variations on trade size, sometimes represented as a square root function. These findings are sometimes referred to as the 'square-root law of market impact' (Lillo et al., 2003; Almgren et al., 2005; Moro et al., 2009; Toth et al., 2011; Bacry et al., 2015; Bershova and Rakhlin, 2013; Donier and Bonart, 2015; Gomes and Waelbroeck, 2015; Zarinelli et al., 2015; Toth et al., 2016) and may be summarised by the following relation between the size Q of the trade, as a percentage of the daily market volume of the traded asset from open to close V_D, and its expected 'market impact' $\mathcal{I}(\frac{Q}{V_D})$ (Toth et al., 2011):

$$\mathcal{I}\left(\frac{Q}{V_D}\right) = Y\sigma_D\sqrt{\frac{Q}{V_D}}, \tag{10}$$

where $\mathcal{I}(\frac{Q}{V_D})$ is defined as the average percentage execution cost C, conditioned on trade size as a percentage of daily volume:

$$\mathcal{I}\left(\frac{Q}{V_D}\right) \equiv E\left(C\middle|\frac{Q}{V_D}\right), \tag{11}$$

where:

$$C \equiv \varepsilon\frac{(p_{vwap} - p_{\text{start}})}{p_{start}}, \tag{12}$$

with p_{vwap} being the volume-weighted average execution price ('trade VWAP'), and ε being the trade direction ($\varepsilon = 1$ for buy trades and $\varepsilon = -1$ for sell trades). In (10), Y is a constant, and σ_D is the daily volatility of the traded asset. Pohl et al. (2017) give arguments based on dimensional analysis as to why one should expect a square root dependence.

In fact, there is considerable variation in these empirical findings. Many studies have pointed to deviations from this square root dependence (Lillo et al., 2003) and propose modifications of this 'square-root law' where the square root is replaced by a power law with some exponent $\alpha < 1$. The literature on *optimal execution* has instead emphasised the *speed* at which the trade is executed, often parameterised through the *participation rate*, as the key variable determining execution costs Almgren and Chriss (2001).

However, a closer inspection leads to many questions regarding the methodology and the conclusions of these studies. Most of these studies are based on the idea that one can separate one's own demand or supply contribution from the aggregate contribution of the rest of the market, and, consequently, decompose execution costs into two separate and independent components: *market impact* and *volatility risk*. In particular, starting with the seminal paper of Almgren and Chriss (2001), the problem of minimising execution costs has been studied through the angle of a trade-off between these two components, and, with a few notable exceptions (Almgren et al., 2005), the prevailing view in the literature is that the market impact of a trade is the main determinant of execution costs, while the contribution of price volatility averages to zero if one examines a large number of trades.

As a result, most of the theoretical models on optimal execution have focused on optimising the market impact component of execution costs, and the empirical studies on market impact have tried to 'explain' execution costs using trade features (such as size or speed) as (sole) explanatory variables. In contrast with the more traditional view regarding the source of price fluctuations, the *aggregate* supply and demand, represented by the net order flow (see e.g. Blume et al. (1989); Hasbrouck (1991); Brown et al. (1997); Hasbrouck and Seppi (2001); Chordia et al. (2002); Cont et al. (2014)) has been taken out of the picture, and correlation analyses and regression studies merely based on trade characteristics have formed the basis of *causal* claims regarding the determinants of execution costs.

More recently, interest in the market microstructure literature has shifted towards trying to understand and model how prices in an asset may be influenced by trades in other assets. This has led to the notion of *cross-impact* and has prompted the development of multivariate models of market impact.

In their simplest form, such models may be seen as multi-asset extensions of the Kyle (1985) model, parameterised by a matrix of impact coefficients whose off-diagonal elements are meant to capture how trades in one asset influence the price of other assets (Caballé and Krishnan, 1994). Such multi-asset impact models have been the subject of several recent theoretical and empirical studies (Andrade et al., 2008; Schied et al., 2010; Boulatov et al., 2013; Pasquariello and Vega, 2015; Alfonsi et al., 2016; Schöneborn, 2016; Benzaquen et al., 2017; Mastromatteo et al., 2017; Tsoukalas et al., 2017; Min et al., 2018; Schneider and Lillo, 2019; Tomas et al., 2020).

A common argument used to justify such models is based on the covariance of returns and order flow across different assets. The existence of positive covariance between the returns of one asset and the order flow imbalance of other assets has been hailed as evidence in favour of cross-impact, leading to multi-asset market-impact models with off-diagonal *cross-impact coefficients* (Alfonsi et al., 2016; Benzaquen et al., 2017; Mastromatteo et al., 2017; Min et al., 2018). While some authors *assume* cross-impact without detailing the underlying causal mechanism that might give rise to this phenomenon (e.g. Schied et al., 2010; Schöneborn, 2016), other authors present models of sequential trades in which market makers' quote revisions give rise to cross-impact (Pasquariello and Vega, 2015).

Despite the wide use of such models, the theoretical and empirical evidence underlying their justification is not entirely convincing for a number of reasons. The mere presence of positive covariation between the returns of an asset and the order flow of another asset might not be sufficient to provide a causal justification for the existence of cross-impact as a separate phenomenon from, for example, correlation between the order flow of the two assets.

Common components in order flow, which are well documented empirically (Hasbrouck and Seppi, 2001), naturally arise from cross-asset trading (Bernhardt and Taub, 2008), index trading (Harford and Kaul, 2005) and portfolio rebalancing (Cont and Wagalath, 2016) may lead to similar effects without the need to invoke the notion of cross-impact.

0.5 Main contributions

In Chapter 1, we study the issue of liquidity in limit order books. Most studies on this topic have emphasised measures of 'visible' liquidity, such as order book depth or trading volume. However, as we have mentioned in Section 0.2, the liquidity available for immediate execution in LOBs at every point in time is very small and most of the liquidity remains *latent* since agents try to conceal their intentions. In Section 1.3, we propose a market microstructure model where the market impact of an individual order is an internal *latent* state of the order book and depends on the aggregate activity of the order book, as summarised by a single variable: the *order flow imbalance* (4).

We start from the model proposed in Cont et al. (2014), which shows that there is a linear relation between the order flow imbalance (4) and price changes that holds locally for short intervals of time $[t_{k-1,i}, t_{k,i}] \subset [T_{i-1}, T_i]$, where $[T_{i-1}, T_i]$ are longer intervals:

$$\Delta p_{k,i} = \beta_i OFI_{k,i} + \varepsilon_{k,i}, \tag{13}$$

where $\Delta p_{k,i} \equiv p_{k,i} - p_{k-1,i}$ are the mid-price changes over the intervals $[t_{k-1,i}, t_{k,i}]$ for an i-th time interval $[T_{i-1}, T_i]$, $OFI_{k,i}$ are the OFIs over the same interval i, β_i is the *market-impact coefficient*, and $\varepsilon_{k,i}$ is a noise term summarising influences of other factors, such as, deeper levels of the order book (Cont et al., 2014, eqn. 7). We then define an econometric model in which the relation between price changes $\Delta p_{k,i}$ and the order flow imbalance $OFI_{k,i}$ is governed by a *latent* market-impact process $\beta_{k,i}$, which is modelled as the sum of a deterministic ($\mu_{k,i}$) and a stochastic component ($\eta_{k,i}$):

$$\begin{aligned} \Delta p_{k,i} &= \beta_{k,i} OFI_{k,i} + \varepsilon_{k,i} \qquad \varepsilon_{k,i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon,i}), & (14) \\ \beta_{k,i} &= \mu_{k,i} + \eta_{k,i} \quad , & (15) \\ \eta_{k,i} &= \rho_i \eta_{k-1,i} + \xi_{k,i} \qquad \xi_{k,i} \sim \mathcal{N}(0, \sigma^2_{\xi,i}), & (16) \end{aligned}$$

where the deterministic component $\mu_{k,i}$ is meant to capture the diurnal effects of the market-impact coefficient, as found in Cont et al. (2014). We assume $cor(\varepsilon_{s,i}, \xi_{t,i}) = 0$.

Equation (16) describes an autoregressive process of order one, so that, if $|\rho_i| < 1$, the process is stationary, and the *average* deviation of the market-impact coefficient from its deterministic component is zero. After a liquidity shock, market impact reverts to its equilibrium level $E[\beta_{k,i}] = \mu_{k,i}$ at a speed regulated by the parameter ρ_i. Kyle (1985) defines resiliency as "the speed with which prices recover from a random, uninformative shock", and so we can interpret ρ_i as the order book *resiliency coefficient*. The instantaneous market-impact process is related to the order book depth by the following equation:

$$D_i = \frac{1}{2\beta_i}, \tag{17}$$

which captures the intuitive concept that, other things equal, market impact is higher (lower) the 'shallower' ('deeper') the order book (the factor 2 being linked to the fact that we examine impact on mid price).

We empirically test this model on large and medium tick stocks and ETFs from the NASDAQ stock exchange, using the Kalman filter. We use two uniform time grids $\{T_0, \ldots, T_I\}$ and $\{t_{0,0}, \ldots, t_{I,K}\}$ with time steps $T_i - T_{i-1} = 30$ minutes and $t_{k,i} - t_{k-1,i} = 10$ seconds, and use the Kalman Filter for estimation. In Section 1.4, we show that this simple model is able to capture the market impact dynamics rather well, and that real-time filtered estimates of the latent market impact outperform historical estimates out of sample. We find that once the seasonality of the market impact coefficient is accounted for, the deviations around this seasonal component are not very persistent, which contradicts previous findings in the literature. We also find that the smoothed market impact coefficients are closely related to the average order book depth for large tick stocks and ETFs. The evidence of this relationship for medium-tick stocks is less strong. These results suggest that the model proposed in this paper may be useful for the design of optimal execution strategies since it provides an efficient tool for real-time estimation of market impact in the presence of latent liquidity. In particular, optimal execution strategies could use real-time estimation of market impact to dynamically adapt their execution schedule so that trades are executed at times when market impact is lower.

In Chapter 2 we analyse the main determinants of execution costs using a data set of millions of institutional trade executions, and propose a radical rethink of what it means to optimally execute a trade. In particular, we propose a *causal* approach (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009) to understand the key drivers of execution costs, which differs from the standard paradigm used in many 'market impact' models.

A common approach to the quantification of execution costs has been to separate one's own demand or supply contribution from the aggregate contribution of the rest of the mar-

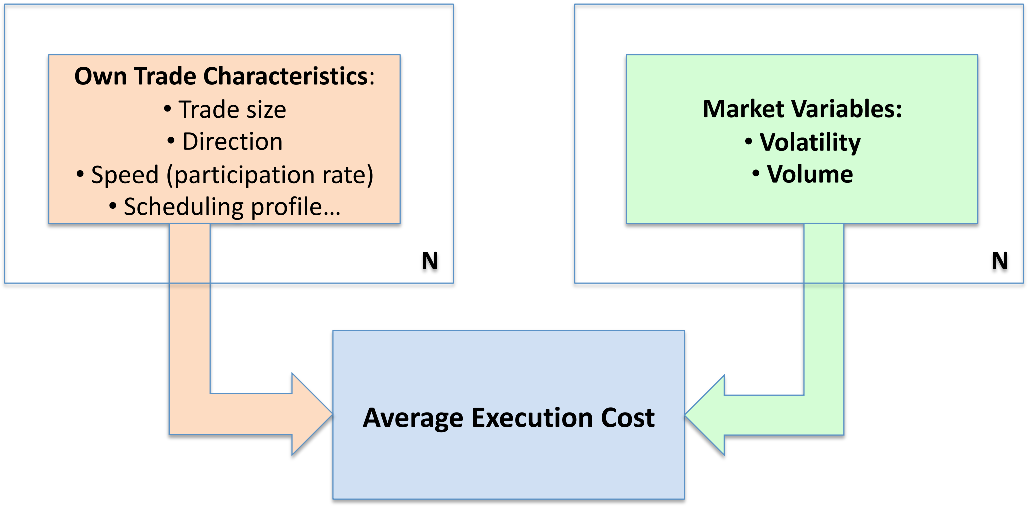

Figure 3: Causal representation of standard market impact models. The *average* execution cost (over N trades or 'meta-orders') is solely explained in terms of trade characteristics, volatility and trading volume, without any reference to *aggregate* demand and supply. Volatility is treated as *independent* from the market participant's trade and its characteristics.

ket, and, consequently, to decompose execution costs into two components: *market impact* and *volatility risk* (Collins and Fabozzi, 1991). With a few notable exceptions (Almgren et al., 2005), the prevailing view in the literature is that the market impact of a trade is the main determinant of execution costs, while the contribution of price volatility averages to zero if one examines a large number of trades. This viewpoint is summarised in the causal graph in Figure 3 where the *average* execution cost (over N trades or 'meta-orders') is solely explained in terms of trade characteristics, volatility and trading volume, without any reference to *aggregate* demand and supply.

Unfortunately, unconditional correlation analysis has little to say with respect to causal relations; in fact, it may yield spurious 'relations' between quantities even when there is none. A key aspect of our analysis is therefore to focus on *conditional* relations, in particular conditioning on trade duration, and aggregate order flow imbalance, to elicit the role of different variables. This approach relies on the work of Haavelmo (1943, 1944); Heckman and Pinto (2015); Pearl (2009). In particular, we adopt the idea of Pearl (2009) of expressing causality through the language of graphs. Using causal graphs, we can express the fact that X causes Y as follows:

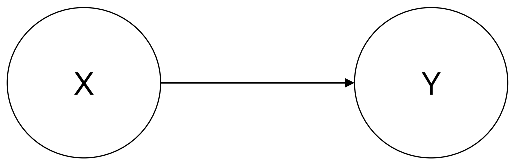

Figure 4: Causal graph: X causes Y.

More formally, the graph in Figure 4 implies that the joint distribution of X and Y, $P(X, Y)$, can be simplified as $P(X, Y) = P(Y|X)P(X)$. Similarly, in the three-variable case, we can express the fact that X only causes Z through its effect on Y as follows:

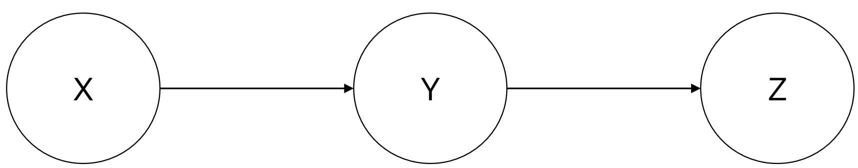

Figure 5: Causal graph: X only causes Z through its effect on Y.

This means that Y contains all of the information about X that is relevant for determining the value of Z, and therefore that knowing the value of Y makes X irrelevant. The graph in Figure 5 can be expressed mathematically by saying that the *conditional distribution* of Z

given Y and X is equal to the conditional distribution of Z given Y: $P(Z|Y, X) = P(Z|Y)$. Pearl (2009) expresses this relationship by saying that Y *blocks*, or *d-separates* X from Z.

It is also easy to show that when $P(Z|Y, X) = P(Z|Y)$, then $P(Z, X|Y) = P(Z|Y)P(X|Y)$. That is, X and Z are *conditionally independent* given Y. Or, in other words, the dependence of X on Z has been 'explained away' by Y.

Blocking the path of a variable through conditioning is therefore seen by Pearl (2009) as a way to test for the presence of a causality. In a regression framework, this means that the regression coefficient from a regression of Z on X should become insignificant once conditioned on Y, and that the regression coefficient from a regression of Z on Y alone, should be equal to the regression coefficient from a regression of Z on Y and X.

Figure 6 shows the causal chain of relations proposed in Chapter 2: a market participant's order merge into the aggregate order flow (the conditioning variable in our model), which, together with the market depth provided by standing (passive) orders in the limit order book, determines the market price and the execution cost. This approach emphasises the relation between each participants order flow and the market order flow imbalance in the determination of the resulting price move, which is consistent with traders' experience of the importance of market timing in trade execution. Moreover, in contrast with standard 'market impact' models, in Chapter 2 we examine the influence of these variables on price variations on a trade-by-trade basis, rather than on average.

Our *causal* analysis in Chapter 2 shows that the dominant component of execution costs for orderly trade executions is not the 'market impact' of trades but price volatility. Using a dataset of millions of institutional trade executions in the US equity market provided by ANcerno, we show in Section 2.3.2, that execution costs are mainly driven by the *aggregate*

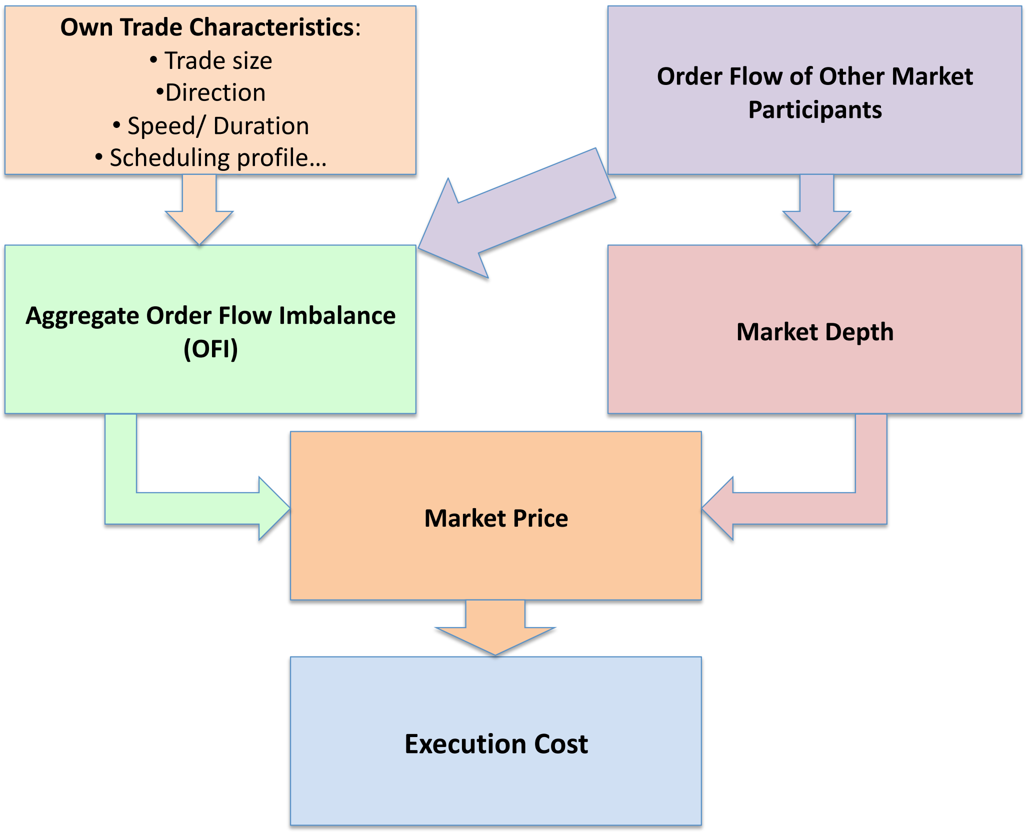

Figure 6: Causal representation of execution costs proposed in Chapter 2. A market participant's order merge into the *aggregate order flow*, which interacts with the limit order book to determine the market depth and, consequently, the market price and the execution cost.

(net) order flow imbalance rather then characteristics of individual trades, even when the trade size is large. In particular, we evaluate the relationship between trade size and price changes *conditional* on a subset of our ANcerno sample containing trades in 77 stocks that overlap the Nasdaq 100 index. We sort the trades into 50 *equally-populated* bins for a standardised measure of *OFI* (4), defined as:

$$OFI \equiv \frac{V_t^b - V_t^a}{V_t^b + V_t^a}, \tag{18}$$

and run, for each bin k, the following conditional regression:

$$\left(\frac{R}{\sigma_D}\right)_k = \alpha_k + \beta_{s,k}\left(\varepsilon\frac{Q}{V_D}\right)_k + \xi_k, \tag{19}$$

where $R \equiv \frac{(p_{vwap} - p_{\text{start}})}{p_{start}}$ is the *unsigned percentage execution cost*, p_{vwap} the volume-weighted average execution price of the trade, σ_D is the daily volatility of the traded stock, Q is the volume of the trade, ε is the direction of the trade, with $\varepsilon = 1$ for buy trades and $\varepsilon = -1$ for sell trades, and V_D is the daily volume of the traded stock from open to close.

We find that with exception of very large imbalances ($|OFI| > 0.30$), the R^2 of these regressions is very low, with an average value of only 3.9% for $OFI \in [-0.30, 0.30]$. These results show that price moves during execution are not determined by the characteristics of individual trades, and that the claim that individual trades *cause* the market price to move, and increase execution costs, might be spurious.

A second finding, presented in Section 2.4, is that the amplitude of price changes during execution is proportional to the square-root of trade *duration* T, rather than trade size, a relation which has a simple explanation in term of the scaling of volatility as a function of the time horizon (Bachelier, 1900). Letting π be the average participation rate,

$$Q = \pi\, V_D T(Q). \tag{20}$$

When returns are uncorrelated, the amplitude of a price move over a horizon τ scales like $\sigma_D\sqrt{\tau}$. So, over the trade duration T, the typical amplitude of a price change will be:

$$\sigma_D\sqrt{T} = \sigma_D\sqrt{\frac{Q}{\pi V_D}} = Y\sigma_D\sqrt{\frac{Q}{V_D}}, \tag{21}$$

where $Y = 1/\sqrt{\pi}$. In other words, the 'square-root law' of market impact (10), may be derived from nothing more than the (well known) square-root scaling of volatility with the (execution) horizon (Bachelier, 1900).

To test this formally, we look at whether or not execution costs depend on trade size when conditioned on duration. We sort the data into 389 bins for the trade duration T, where T is expressed in minutes, and estimate, for each bin k, the conditional logarithmic regression of the absolute execution cost $|C|$ (defined in (12)), standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T:

$$\log\left(\left|\frac{C}{\sigma_D}\right|\right)_k = \alpha_k + \beta_{s,k}\log\left(\frac{Q}{V_D}\right)_k + \xi_k, \tag{22}$$

and the conditional non-linear least square (NLS) regression for the execution cost C, standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T:

$$\left(\frac{C}{\sigma_D}\right)_k = Y_k\left(\frac{Q}{V_D}\right)_k^{\beta_{s,k}} + \xi_k. \tag{23}$$

We find that once we condition on trade duration, the dependence of both absolute and signed price changes on trade size vanishes almost entirely, supporting the explanation provided above for the 'square root law of market impact' as arising from the square-root scaling of volatility with trade duration (Bachelier, 1900). The implication is that the two components of execution costs –'market impact' and 'volatility risk'– are in fact two sides of the same coin: market volatility.

Most optimal execution models consider the participation rate, which is a measure of the speed of trading, as the key driver of execution costs. In Section 2.5, we examine how execution costs depend on the participation rate, using the conditional logarithmic regression of the absolute execution cost $|C|$, standardised by the daily volatility, on the participation rate Q/V_P (where V_P is the volume of the traded stock during execution), given the trade duration T:

$$\log\left(\left|\frac{C}{\sigma_D}\right|\right)_k = \alpha_k + \beta_{p,k}\log\left(\frac{Q}{V_P}\right)_k + \xi_k \tag{24}$$

and the conditional non-linear least square (NLS) regression for the execution cost C, standardised by the daily volatility, on the participation rate Q/V_P, given the trade duration T:

$$\left(\frac{C}{\sigma_D}\right)_k = Y_k\left(\frac{Q}{V_P}\right)_k^{\beta_{p,k}} + \xi_k. \tag{25}$$

Our results show that conditional on trade duration, the *participation rate* has little or no influence on both absolute and signed price changes during execution.

We also run a step-wise multivariate log-regression of the absolute value of execution costs on the participation rate, duration and volatility to further investigate the role of the participation rate on execution costs, once trade duration is controlled for. We estimate the following log-log regressions for the absolute value of execution costs:

$$\log|C| = \alpha + \beta_p \log\frac{Q}{V_P} + \beta_v \log\sigma_D + \xi, \tag{26}$$

$$\log|C| = \alpha + \beta_d \log T + \beta_v \log\sigma_D + \xi, \tag{27}$$

$$\log|C| = \alpha + \beta_p \log\frac{Q}{V_P} + \beta_d \log T + \beta_v \log\sigma_D + \xi., \tag{28}$$

and the following NLS regressions for signed execution costs:

$$C = Y\Big(\frac{Q}{V_P}\Big)^{\beta_p}\sigma_D^{\beta_v} + \xi, \tag{29}$$

$$C = YT^{\beta_d}\sigma_D^{\beta_v} + \xi, \tag{30}$$

$$C = Y\Big(\frac{Q}{V_P}\Big)^{\beta_p}T^{\beta_d}\sigma_D^{\beta_v} + \xi. \tag{31}$$

In contrast to what has normally been assumed in optimal execution models, we find that, for trades of a given size, *slower* execution leads to higher execution costs, which is explainable by higher exposure to volatility risk during execution. In particular, we find that for day-long VWAP trades, a large increase in the traded quantity (and in the participation rate) from $\simeq 18.5\%$ to $\simeq 50\%$ of daily volume leaves absolute price changes almost unchanged, and leads to a *decline* in signed price changes, contradicting the idea that the participation rate is the main determinant of execution costs.

Our findings in Chapter 2 are consistent with previous studies of Engle et al. (2012); Bershova and Rakhlin (2013) on trade executions in equity markets and Salem et al. (2018) on trade execution in the US Treasury market, which similarly underline volatility risk as the main determinant of execution costs. They question in particular the received notion of 'price impact of a trade' and challenge the view that 'market impact' should be treated as the main determinant of execution costs. Our findings also question the use of participation rate as the key determinant of impact, a feature common to many optimal execution models. Finally, they suggest that instead of focusing on optimising market 'impact' at a trade-by-trade level, investors should focus on modelling the aggregate dynamics of net order flow and on managing the main source of risk: price volatility.

In Chapter 3, we investigate the notion of *cross-impact*, which refers to the fact that the price of a given asset may be influenced not only by its own order flow, but also by the order flow of *other* assets, thus resulting in higher execution costs when more than one security is traded at the same time. Cross-impact has been seen by some as a ”natural extension” of the concept of market impact (Bouchaud et al., 2018), and has been the subject of several theoretical and empirical studies in the last few years (Andrade et al., 2008; Schied et al., 2010; Boulatov et al., 2013; Pasquariello and Vega, 2015; Alfonsi et al., 2016; Schöneborn, 2016; Benzaquen et al., 2017; Mastromatteo et al., 2017; Tsoukalas et al., 2017; Min et al., 2018; Schneider and Lillo, 2019; Tomas et al., 2020).

A common approach to the quantification of cross impact has been to estimate the covariance of returns and order flow across multiple assets:

$$\boldsymbol{\mathcal{M}} \equiv E\Big[(\boldsymbol{R_t} - \mu_{\boldsymbol{R}})^T(\boldsymbol{OFI_t} - \mu_{OFI})\Big]. \tag{32}$$

where R_t and OFI_t are $N \times 1$ vectors of returns and order flow imbalances over the time interval $[t - \Delta t, t]$ for assets $i = 1, ..., N$, $\mu_{\boldsymbol{R}}$ is an $N \times 1$ vector of mean returns and μ_{OFI} is an $N \times 1$ vector of mean OFIs. In (32), the diagonal terms represent the covariance between the returns of asset i and its own order flow, while the off-diagonal terms represent the covariance between the returns of asset i and the order flow of asset $j \neq i$.

The covariance matrix (32) has been the subject of several empirical studies (Wang et al., 2016b,a; Mastromatteo et al., 2017; Tomas et al., 2020). These studies consistently find positive off-diagonal elements in $\boldsymbol{\mathcal{M}}$, and have used this as justification for the extension of univariate impact models to multiple assets with off-diagonal, *cross-impact coefficients*.

In Chapter 3, we propose a simpler and more intuitive explanation for the presence of

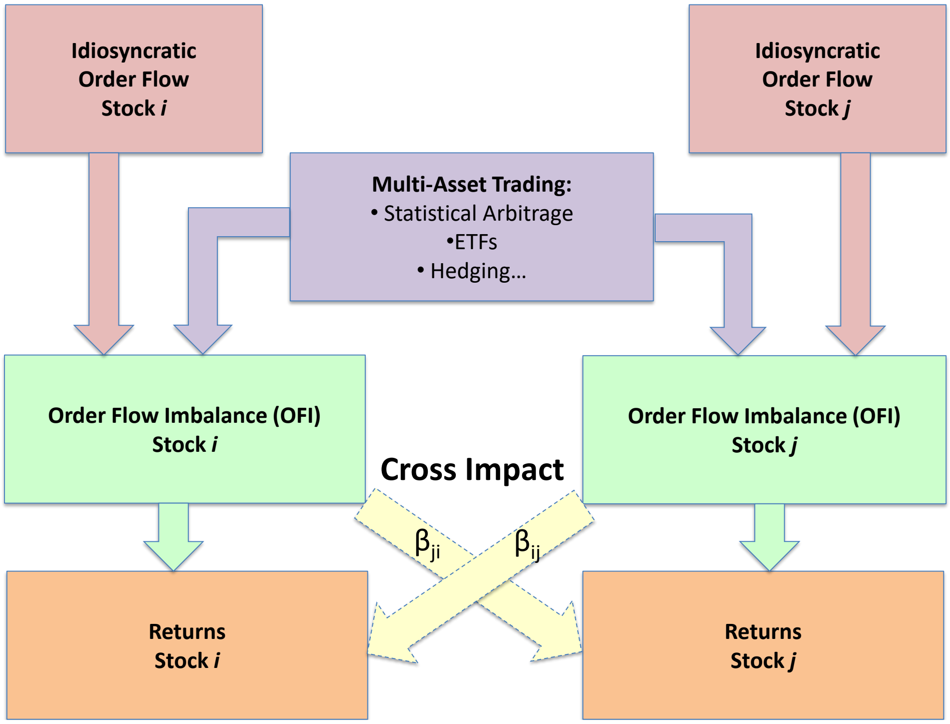

Figure 7: Causal representation of multi-asset impact models. The returns of each asset are influenced by the asset's own order flow imbalance, which is the superposition of an idiosyncratic order flow and of a *common component* across assets. This common component leads to correlation in order flow imbalance across assets. Cross-impact models posit the existence of an additional direct influence of the order flow imbalance of an asset on the returns of other assets, represented by the diagonal arrows in the causal diagram.

positive off-diagonal terms in the matrix $\mathcal{M}$, and suggest a parsimonious approach for modelling multi-asset market impact, which does not require introducing any cross-impact term.

We start from a linear model for market impact that relates stocks' returns over a trade execution horizon to stocks' OFIs:

$$\boldsymbol{R}_t = \mu + \beta \boldsymbol{OFI}_t + \boldsymbol{\varepsilon}_t, \tag{33}$$

and show that cross-impact is not needed to explain the positive covariation between the returns of an asset and the order flow of other assets. The non-zero covariance between the returns of an asset R_i and the order flow of other assets OFI_j ($j \neq i$) may instead be explained by the simpler observation that OFIs across assets are correlated, since these assets are usually jointly traded through statistical arbitrage strategies, ETFs, and other cross-asset strategies, or for portfolio rebalancing and hedging purposes.

We investigate these questions through the angle of causal modelling (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009), which emphasises causal relations among variables, testable through conditioning.

Figure 7 shows the causal chain of relations for two such assets i and j: market participants trade in stocks i and j at the same time, which leads to correlation between the OFIs of the two stocks. The return of each asset is determined by the order flow imbalance, which is the superposition of a *common component* and the idiosyncratic order flow in each security. The presence of this common component leads to correlations in order flow imbalance across assets, and results in correlations between the returns of an asset and order flow of other assets.

Cross-impact models insert an additional mechanism, represented in Figure 7 as a pair of diagonal arrows, through which the order flow of one asset would somehow influence the prices of another asset, bypassing the supply/demand mechanism usually assumed to underpin price formation.

The causal graph in Figure 7 suggests that without conditioning on the common component in the order flow, one cannot distinguish between cross-impact and order flow correlation. In Section 3.2.2, we present examples that illustrate this *identifiability* problem, and demonstrate that one can overcome this problem by simply estimating the covariance of order flow imbalances.

The presence of causal relations, shown as arrows in Figure 7, may be tested by conditioning on the variable hypothesised to contribute to an observed effect. We use this approach to test for the significance and amplitude of cross-impact channels by conditioning on the common factor in order flow imbalance. In Section 3.3.3, we account for the commonality in order flow by performing a principal component analysis (PCA) on the correlation matrices of returns and OFIs. We adjust for the correlation in order flow using the first principal component of OFI, then test for the statistical significance of any residual cross-impact effect.

We use a two-step procedure. In the first step, we use a one factor model for the order flow imbalance, where the common factor $F^1_{\mathrm{OFI},t}$ is obtained by projecting the order flow vectors onto the order flow first principal component $\boldsymbol{V}^1_{\mathbf{OFI}}$, scaled by its corresponding eigenvalue $\lambda^1_{\mathbf{OFI}}$: $F^1_{\mathrm{OFI},t} = \boldsymbol{V}^1_{\mathbf{OFI}}{}^T(\mathbf{ofi}_t - \alpha_{OFI})/\sqrt{\lambda^1_{OFI}}$:

$$\mathbf{ofi}_t = \alpha + \boldsymbol{V}^1_{\mathbf{OFI}} F^1_{\mathrm{OFI},t} + \boldsymbol{\eta}_t. \tag{34}$$

The residual term $\boldsymbol{\eta}_t$ represents the order flow not accounted for by the common factor $F^1_{\mathrm{OFI,t}}$. We will refer to $\boldsymbol{\eta}_t$ as the 'idiosyncratic' component of order flow.

In a second step, we regress stock returns on the OFI's common factor $F^1_{\mathrm{OFI},t}$, the residuals from the regression in step 1 for stock i ($\eta_{i,t}$) and the residuals from the regression in step 1 for all other stocks $j \neq i$ ($\eta_{j,t}$):

$$\boldsymbol{r}_t = \mu + \gamma F^1_{\mathrm{OFI},t} + \Phi \boldsymbol{\eta}_t + \boldsymbol{\xi}_t, \tag{35}$$

where γ is a vector of impact coefficients due to the common component in order flow, Φ represents the impact of the idiosyncratic component of order flow $\boldsymbol{\eta}_t$. In this way we are able to disentangle the impact on the returns of stock *i* due to common order flow, the stock's idiosyncractic order flow, and the idiosyncratic order flow of all other stocks.

Our results, in Section 3.3.4, show that once a stock's own order flow and the commonality in order flow has been taken into account, the additional explanatory power of all other stocks' order flow is less than 0.5 percentage points.

In section 3.3.5 we also show that, cross-impact coefficients are not only small in magnitude but they are also unstable and change sign randomly over time, which poses a problem for their interpretation and use.

Our results show the main determinants of impact to be idiosyncratic order flow imbalance as well as a market order flow factor that is common across stocks. This leads to a parsimonious approach for *causal* modelling of multi-asset market impact, which does not require introducing any concept of cross-impact.

Our findings in Chapter 3 have important implications for optimal trade execution and transaction costs analysis. They show that accounting for commonality in order flow seems to be a more effective way to build multi-asset market impact models than introducing cross-impact coefficients whose statistical significance and interpretation are not clear.

In section 3.4, we show how our framework may be used for the estimation of portfolio execution costs. First, in section 3.4.1, we show that the presence of common components in order flow leads to *unconditional* estimates of portfolio execution costs which are different from single-asset TCA models.

Letting $\boldsymbol{q}$ denote an $N \times 1$ vector of position sizes (in AV units), and decomposing it into its projection along the first principal component of order flow and an idiosyncratic component:

$$\boldsymbol{q} = F^1_{OFI}(q)\boldsymbol{V}^1_{\mathrm{OFI}} + \eta(\boldsymbol{q}) \qquad \text{where} \qquad F^1_{OFI}(\boldsymbol{q}) = \boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}}, \tag{36}$$

we obtain the *unconditional* estimate of transaction costs:

$$\boldsymbol{q}^T\boldsymbol{r}_t = \underbrace{\boldsymbol{q}^T\mu}_{\text{Trend}} + \underbrace{(\boldsymbol{q}^T\gamma)(\boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}})}_{\text{Common}} + \underbrace{\boldsymbol{q}^T\Phi(\boldsymbol{q} - (\boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}})\boldsymbol{V}^1_{\mathrm{OFI}})}_{\text{Idiosyncratic}}, \tag{37}$$

as a sum of three terms:

- a linear term $\boldsymbol{q}^T\mu$, due to a possible price trend during execution;
- a quadratic cost $(\boldsymbol{q}^T\gamma)(\boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}})$ corresponding to the projection of the portfolio $\boldsymbol{q}$ on the common component of the order flow;
- a quadratic cost corresponding to the idiosyncratic component of order flow.

As we show in section 3.3.3, the first principal component of order flow $\boldsymbol{V}^1_{\mathbf{OFI}}$ has positive weights roughly reflecting a market 'index'. As a result, certain long-short portfolios $\boldsymbol{q}$ may satisfy $\boldsymbol{q}^T\boldsymbol{V}^1_{\mathbf{OFI}} = 0$ in which case the second term in (3.32) is zero and all of the impact is due to the idiosyncratic component. Conversely, if $\boldsymbol{q}$ is aligned with $\boldsymbol{V}^1_{\mathbf{OFI}}$ (which may be the case for an index portfolio), the last term in (3.32) is zero, and the common component in order flow is the only driver of execution costs.

Second, in section 3.4.2, we show how the model allows to use information on the direction of market order flow to obtain *conditional* estimates of portfolio execution costs. This last concept embodies the notion of 'market timing', which is well known to traders.

Assume we are able to arrive at a forecast $\widehat{F}_t^1 = E_t[F^1_{\text{OFI,t}}]$ for the common component of market order flow imbalance. Conditional on this information, the expected order flow imbalance, including the agent's own trade, is $\boldsymbol{q} + \widehat{F}_t^1 \boldsymbol{V}^1_{\mathbf{OFI}}$. This changes the second term in (37) and leads to the following *conditional* estimate of execution cost:

$$\boldsymbol{q}^T \boldsymbol{r_t} = \boldsymbol{q}^T \mu + (\boldsymbol{q}^T \gamma)(\boldsymbol{q}^T \boldsymbol{V}^1_{\text{OFI}} + \widehat{F}_t^1) + \boldsymbol{q}^T \Phi(\boldsymbol{q} - (\boldsymbol{q}^T \boldsymbol{V}^1_{\text{OFI}}) \boldsymbol{V}^1_{\text{OFI}}). \tag{38}$$

We thus observe that the conditional estimate for the second term is affected by the 'market' order flow: in line with traders experience, going 'against' the market order flow or in the same direction does not lead to the same impact. In particular, the second term can be zero if

$$\boldsymbol{q}^T \boldsymbol{V}^1_{\text{OFI}} + \widehat{F}_t^1 = 0, \tag{39}$$

which corresponds to an agent trading in the opposite direction to the aggregate order flow imbalance. This observation is consistent with traders experience of the importance of 'market timing' of trades. For example, it implies that although on average a buy trade will increase the price, a buy trade executed during a market sell-off may end up having zero, or even negative impact.

1

Latent liquidity and market impact

1.1 Introduction

Market impact is arguably one of the most important concepts in market microstructure, and refers to the fact that a trader's order might move the price in an unfavourable direction (Foucault et al., 2013). Market impact is a major concern for traders, who see it as an unwanted cost they wish to minimise, and has received a lot of attention in the literature starting from the seminal papers of Grossman and Stiglitz (1980), Glosten and Milgrom (1985) and Kyle (1985). In these papers, market impact is viewed not merely as a friction, but as a mechanism that maintains price efficiency by allowing prices to adjust to new information. An alternative interpretation is provided by Gode and Sunder (1993) and Farmer et al. (2005) who see impact as a completely mechanical/statistical phenomenon due to order flow fluctuations.

A large strand of literature has investigated the question of how market impact should be modelled, and the prevalent approach is to model it as an observable quantity, by running, for instance, so-called 'price impact' regressions, where the the change in mid price is proportional to the buying or selling pressure (Foucault et al., 2013, p. 56). But, there are clear reasons why we should consider market impact as a *latent* variable, which may be revealed by observing the dynamics of the order book. The resting (visible) liquidity in the order book is rarely, if ever, equivalent to the whole available liquidity even for 'highly liquid' stocks. This is related to the hypotheses of *selective liquidity taking* (Weber and Rosenow, 2005; Bonart and Gould, 2017) and of *exploratory trading* (Clark-Joseph, 2013).

Motivated by this observation, we empirically investigate a microstructure model for price dynamics in which price impact depends on the aggregate activity of the order book as summarised by the *order flow imbalance* (Cont et al., 2014), and is a latent state variable of the order book. This is an extension of the work of Cont et al. (2014) in which liquidity is viewed as constant and observable. Methodologically, we estimate a linear state-space model using the Kalman filter, as linear state space models seem a natural choice to describe a dynamical system in which some variables are latent. Our work is similar in spirit to the work of Mertens (2014), but differs from it in a number of ways. First, we take into account the intraday seasonal component of the the market-impact coefficient. We model the latent market-impact coefficient as being composed of a deterministic and a stochastic component and estimate it on thirteen half-hours buckets. Second, we estimate the model on stocks with different tick sizes and investigate three different measures of order flow imbalance. Finally, we compare the smoothed coefficients from our model with the visible depth.

Our empirical findings show that the model captures market-impact dynamics rather well. We find that the filtered estimated of the market-impact coefficients deliver better out-of-sample forecasts than historical estimates. We find that deviations around the seasonal component of the market-impact coefficient are not very persistent. This is in contrast with the findings of Mertens (2014) who reported market-impact coefficients very close to being a unit root. Finally, we find that for large-tick stocks and ETFs, the smoothed market-impact estimates are strongly related to the average order book depth. The situation is a bit different for the remaining two stocks (AAPL and FB), indicating that the model might not be fully capturing the dynamics of medium-tick stocks. We believe that the ability of our model to capture latent liquidity and track the order book depth may be useful for the design of optimal execution strategies. In particular, optimal execution strategies could dynamically adapt to real-time estimation of market impact so that trades are executed at times where market impact is lower.

1.2 Literature review

The concept of market impact of a trade refers to the notion that a market participant's order may move the price in an unfavourable direction, thereby resulting in a higher execution cost. The seminal papers of Glosten and Milgrom (1985) and Kyle (1985) represent the starting point for the analysis of market impact. The simplest assumption is that market impact is both linear in the traded volume and permanent in time; this can be justified within the framework of the model proposed in Kyle (1985), where an insider trader and 'noise' traders submit orders that are cleared by a market maker at every time step. The model assumes that the price adjustment rule (Δp) of the market maker is linear in the total signed volume (εv):

$$\Delta p = \lambda \varepsilon v. \tag{1.1}$$

In (1.1), λ is a measure of market impact, defined as price change over signed traded volume, and is usually referred to as 'Kyle's lambda'. Indeed, Kyle was the first to formalise the idea that market impact and *market depth* are two sides of the same coin observing that "market depth is measured by the reciprocal of the liquidity parameter $\lambda(t)$" (Kyle, 1985). Moreover, the price adjustment in the Kyle model is *permanent* as the impact $\lambda \varepsilon_n v_n$ of trades in the n-th time interval continues to have its full impact up to time T.[1]

Following Kyle (1985), many authors have studied price impact and the informativeness of limit order books. In particular, Blume et al. (1989) and Hasbrouck (1991) were among the first to study the relationship between price movements and the aggregate order flow imbalance. Blume et al. (1989) found a strong relationship between the order flow imbalance and aggregate returns by analysing 500 NYSE stocks on October 19 and 20, 1987. Hasbrouck (1991) was the first to study the interaction between trades and stock returns more systematically by using data from the New York and American Stock Exchanges over the 62 trading days in the first quarter of 1989. The author analyses quote revisions and trades through a

[1] In particular, the price change between time $t = 0$ and t=T is: $p_T = p_0 + \sum_{n=0}^{N-1} \Delta p_n = p_0 + \lambda \sum_{n=0}^{N-1} \varepsilon_n v_n$.

vector time series model and finds that the full impact of a trade on the security price is not felt instantaneously but with a protracted lag. He also finds that impact is a nonlinear, concave function of the trade innovation size, and that there is Granger causality running from quote revisions to trades.

Following Hasbrouck (1991) several studies have investigated the relationship between trades or order flow imbalance and stock returns. For example, using two years of data for the 20 stocks most actively traded on the Australian market, Brown et al. (1997) finds bi-directional causality between imbalance and return, though not beyond a single day. Hasbrouck and Seppi (2001) find that commonality in the order flow explains roughly two-thirds of the commonality in returns, while Chordia et al. (2002) focus on the aggregate daily order imbalance of trades and find that ordered imbalances reduce liquidity and that they are strongly related to contemporaneous absolute returns after controlling for market volume and market liquidity. Finally, Cont et al. (2014)(further discussed below) demonstrate that price changes are well explained by a linear model driven by the *aggregate order flow imbalance*.

In the *optimal execution* literature, much of the research on market impact has long been focused on the framework pioneered by Bertsimas and Lo (1998) and Almgren and Chriss (2001). Researchers and practitioners adopted the modelling framework proposed in these papers and subsequently developed variants of it to account for stochastic volatility and liquidity (Almgren, 2012; Cheridito and Sepin, 2014). The core idea of these models is to decompose market impact into two separate components: a *permanent* component, that pushes prices up (down) over the course of a buy (sell) order, and that permanently modifies the price process, and a *temporary* or *instantaneous* component, which is the execution cost paid by the trader to take liquidity from the market.

1.2.1 Outline

The remainder of the Chapter is organised as follows. Section 1.3 introduces the market-impact model and its estimation procedure. Section 1.4 presents the model estimation results and discusses the relation between the estimated market-impact coefficients and limit order book depth. Section 1.5 concludes.

1.3 Market impact model

In this section, we propose a microstructure model of market impact which can be seen an extension of the model introduced by Cont et al. (2014). We extend the model of Cont et al. (2014) by considering market impact as an *internal state variable* that affects the dynamics of the limit order book, by examining three alternative measures of order flow imbalance, and by looking not only at 'large-tick' stocks, but also at 'medium-tick' stocks and ETFs, for which the discreteness of price changes is less of an issue. Here we present the model introduced by Cont et al. (2014), our extension, and the estimation procedure by introducing the Kalman filter.

In their model, Cont et al. (2014) show that, over short time intervals, price changes are mainly driven by the *aggregate order flow imbalance*, defined as the imbalance between supply and demand at the best bid and ask prices. By considering price impact as a function of a single aggregate variable, instead of the individual characteristics of the order, the authors therefore depart from the standard modelling approaches proposed, for instance, in Almgren and Chriss (2001). In particular, using NYSE TAQ data for 50 stocks, the authors show that there is a linear relation between order flow imbalance and price changes, with a slope inversely proportional to the market depth.

Cont et al. (2014) consider a stylised example of the order book where the instantaneous effect of order book events can be explicitly computed. Order arrivals and cancellations arrive only at the best bid and ask, and the number of shares (depth) at each price level beyond the best bid and ask is equal to D. When bid (ask) size reaches D, the next passive order arrives one tick above (below) the best quote, initialising a new best level (Cont et al., 2014).

Consider a time interval $[t_{k-1}, t_k]$ and denote by L_k^b, C_k^b the volume (total number of shares) of limit buy orders arrived to and cancelled from the current best bid during that time interval, respectively. Let M_k^b the volume of the marketable sell orders that arrived at the current best bid, and by p_k^b the bid price at time t_k (Cont et al., 2014). Similarly, for the ask side we have the analogous quantities L_k^a, C_k^a, M_k^a, and the ask price at time t_k is p_k^a.

In this stylised order book model, Cont et al. (2014) show that there is a linear relation between order flows L_k^i, C_k^i, M_k^i $(i = b, a)$ and price changes $\Delta p_k^i = (p_k^i - p_{k-1}^i)$ $(i = b, a)$:

$$\Delta p_k^b = \delta \Big[\frac{L_k^b - C_k^b - M_k^b}{D} \Big], \tag{1.2}$$

$$\Delta p_k^a = -\delta \Big[\frac{L_k^a - C_k^a - M_k^a}{D} \Big], \tag{1.3}$$

where δ is the tick size. We can see from (1.2) and (1.3) that the impact of all order book events is additive and only depends on their net imbalance. Consider the mid-price changes normalised by the tick size: $p_k = \frac{p_k^b + p_k^a}{2\delta}$, then:

$$\Delta p_k = \frac{OFI_k}{2D} + \varepsilon_k, \tag{1.4}$$

where

$$OFI_k \equiv (L_k^b - C_k^b - M_k^b) - (L_k^a - C_k^a - M_k^a), \tag{1.5}$$

is the order flow imbalance, or net order flow, and ε_k is a truncation error (Cont et al., 2014, eqn. 3-4). The authors assume a noisy relation between price changes and *OFI*, that holds locally for short intervals of time $[t_{k-1,i}, t_{k,i}] \subset [T_{i-1}, T_i]$, where $[T_{i-1}, T_i]$ are longer intervals:

$$\Delta p_{k,i} = \beta_i OFI_{k,i} + \varepsilon_{k,i}. \tag{1.6}$$

Here, β_i is the *market-impact coefficient* for an i-th time interval and $\varepsilon_{k,i}$ is a noise term summarising influences of other factors, such as, deeper levels of the order book (Cont et al., 2014, eqn. 7). As we can see from (1.6), Cont et al. (2014) allow the market-impact coefficient and the distribution of the noise term to vary with index i, to account for the well-known intraday seasonality. Comparing (1.4) and (1.6), we have:

$$D_i = \frac{1}{2\beta_i}, \tag{1.7}$$

and so we can interpret $\frac{1}{2\beta_i}$ as the *implied* order book depth. The specification in (1.6) can be seen as a model for the instantaneous market impact of order book events.

We extend the model of Cont et al. (2014) by allowing the depth to vary over time. Thus, the price impact of an order now depends on the aggregate activity of the order book, and the level of liquidity of the book. The first variable is observable, the second is latent. For instance, given the same dynamics of the order flow, a smaller depth would lead to a larger price change. Our formulation of depth, or impact, as a latent variable is connected to two theories. The first is the theory of *selective liquidity taking* (Weber and Rosenow, 2005; Bonart and Gould, 2017), according to which traders condition the market order volume to be smaller or equal to the volume at the opposite best in order to avoid market impact. The second is the theory of *exploratory trading* of Clark-Joseph (2013), according to which HFTs use small aggressive orders in order to gather private information about the elasticity of supply and therefore about the market impact of the aggressive orders that follow. HFTs can then trade ahead of predictable demand at only those times when it is profitable to do so (Clark-Joseph, 2013).

HFTs aggressive orders generate valuable private information, specifically, information about the market impact of the aggressive orders that follow. When an HFT places an exploratory order and observes a large market impact, he learns that supply is temporarily inelastic. If the HFT knows that there is going to be more demand soon thereafter, he can place a larger order (even with a big market impact) knowing that the market impact from the coming demand will drive prices up further and ultimately enable him to sell at a premium that exceeds the market impact of his unwinding order. When an HFT knows that supply is temporarily inelastic, he follows a routine demand-anticipation strategy. The purpose of exploratory trading is not to learn about future demand, but rather to identify the times at which trading in front of future demand will be profitable.

1.3.1 Model specification

We start from equation (1.6) and define an econometric model in which the relation between price changes $\Delta p_{k,i}$ and the order flow imbalance $OFI_{k,i}$ is governed by a latent *market-impact process* $\beta_{k,i}$, which is modelled as the sum of a deterministic ($\mu_{k,i}$) and a stochastic component ($\eta_{k,i}$). We include the deterministic component $\mu_{k,i}$ in order to model the diurnal effects of the market-impact coefficient as found in Cont et al. (2014):

$$
\begin{aligned}
\Delta p_{k,i} &= \beta_{k,i} OFI_{k,i} + \varepsilon_{k,i} \qquad \varepsilon_{k,i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon,i}), && (1.8)\\
\beta_{k,i} &= \mu_{k,i} + \eta_{k,i} \quad , && (1.9)\\
\eta_{k,i} &= \rho_i \eta_{k-1,i} + \xi_{k,i} \qquad \xi_{k,i} \sim \mathcal{N}(0, \sigma^2_{\xi,i}). && (1.10)
\end{aligned}
$$

and we assume $\mathrm{cor}(\varepsilon_{s,i}, \xi_{t,i}) = 0$.

Equation (1.10) describes an autoregressive process of order one, so that, if $|\rho_i| < 1$, the process is stationary, and the *average* deviation of the market- impact coefficient from its deterministic component is zero. After a liquidity shock, the market impact reverts to its equilibrium level $E[\beta_{k,i}] = \mu_{k,i}$ at a speed regulated by the parameter ρ_i. Kyle (1985) defines resiliency as "the speed with which prices recover from a random, uninformative shock", and so we can interpret ρ_i as the order book *resiliency coefficient*. The instantaneous market-impact process $\beta_{k,i}$ is related to the order book depth by (1.7) which captures the intuitive concept that, other things equal, market impact is higher (lower) the 'shallower' ('deeper') the order book.

The model proposed in (1.8) - (1.10) is similar in spirit to the model suggested by Mertens (2014), but with some important differences. First, we take into account the intraday seasonal component of the market-impact coefficient by estimating the model for 13 half-hourly intervals, as in Cont et al. (2014) and by modelling the market-impact coefficient as the sum of a deterministic and a stochastic component. Mertens (2014) estimates the model on daily intervals and only considers the stochastic component of price impact leading to the findings of very persistent price impact coefficients. Second, we compare the smoothed market-impact coefficients obtained from the model to the average observable market depth (see Section 1.4.4). Third, we explore three different definitions of the order flow imbalance (see Section 1.4.2).

1.3.2 Model estimation: the Kalman filter

The linear Gaussian *state space model* (SSM) can be written as:

$$y_k = c_k + H_k x_k + r_k \qquad r_k \sim \mathcal{NID}(0, R_k), \tag{1.11}$$

$$x_k = A_{k-1} x_{k-1} + q_{k-1} \qquad q_{k-1} \sim \mathcal{NID}(0, Q_{k-1}), \tag{1.12}$$

where y_k is the *measurement* at time k, and x_k is the *state* of the system at time k. H_k is the *measurement matrix* at time k, A_{k-1} is the *state transition matrix* at time $k-1$ and d_k is a constant at time k. The *measurement noise* r_k and the *process noise* q_k are assumed to be Gaussian and independent of each other.[2]

It is straightforward to see that, in the interval $[T_{i-1}, T_i]$, the state-space representation

[2] For an in-depth treatment of the Kalman filter see Durbin and Koopman (2012) and Hamilton (1994).

of the system in (1.8)-(1.10), is given by:

$$
\begin{aligned}
y_k &\equiv \Delta p_k, \quad c_k \equiv \mu_k OFI_k, \quad H_k \equiv OFI_k, \quad R_k \equiv \sigma^2_{\varepsilon_i}, && (1.13)\\
x_k &\equiv \eta_k, \quad A_k \equiv \rho_i, \quad Q_k \equiv \sigma^2_{\xi_i}. && (1.14)
\end{aligned}
$$

Since we take the contemporaneous order flow imbalance OFI_k as exogenous, the vector of parameters describing the system is given by: $\theta_i = (\rho_i, \sigma^2_{\varepsilon,i}, \sigma^2_{\xi,i})$. We first reformulate the problem by scaling the measurement Δp_k by the tick size τ ($\Delta \tilde{p}_k \equiv \Delta p_k/\tau$), and the order flow imbalance OFI_k by its (estimated) standard deviation $\hat{\sigma}_{\zeta,i}$ ($\tilde{OFI}_k \equiv OFI_k/\hat{\sigma}_{\zeta,i}$). By defining all the other variables accordingly, we get:

$$
\begin{aligned}
\Delta \tilde{p}_{k,i} &= \tilde{\beta}_{k,i} \tilde{OFI}_{k,i} + \tilde{\varepsilon}_{k,i} && \tilde{\varepsilon}_{k,i} \sim \mathcal{NID}(0, \tilde{\sigma}^2_{\varepsilon,i}), && (1.15)\\
\tilde{\beta}_{k,i} &= \tilde{\mu}_{k,i} + \tilde{\eta}_{k,i} \quad , && && (1.16)\\
\tilde{\eta}_{k,i} &= \tilde{\rho}_i \tilde{\eta}_{k-1,i} + \tilde{\xi}_{k,i} && \tilde{\xi}_{k,i} \sim \mathcal{N}(0, \tilde{\sigma}^2_{\xi,i}). && (1.17)
\end{aligned}
$$

Initial values for the optimisation problem are obtained via a two-step process. In the first step, we run a least-square regression as in (1.15) for each interval $[T_{i-1}, T_i]$ of each day in the training set, and obtain the initial values for the state prior and the measurement noise covariance matrix. In particular, for each interval $[T_{i-1}, T_i]$, we set the deterministic component $\tilde{\mu}_{k,i}$ as the mean $\tilde{\beta}_{k,i}$ over the 30 preceding days, the prior mean m_0 as the median of the estimated $\tilde{\eta}_{k,i}$, the initial value for the prior covariance P_0 as the median MSE,[3] and the measurement noise covariance matrix $\tilde{\sigma}^2_{\varepsilon,i}$ as the median of the standardised residuals covariance. In the second step, we take the estimated $\tilde{\beta}$s from step one and fit an AR(1) model as

[3] We compute the median RMSE from the least square regression and square the result.

in (1.17). From this estimation, we obtain the initial value for the parameter $\tilde{\rho}_i$ as the AR(1) coefficient and the initial value for the process noise $\tilde{\sigma}^2_{\xi,i}$ as the estimated variance of the model innovations.

Before going ahead with the optimisation, we follow the suggestion of Durbin and Koopman (2012) and reformulate the problem into an unconstrained optimisation problem[4] by making the following transformations: $\psi_{\varepsilon,i} = \frac{1}{2}log(\tilde{\sigma}_{\varepsilon,i})$, and $\psi_{\xi,i} = \frac{1}{2}log(\tilde{\sigma}_{\xi,i})$. The log-likelihood is easily obtained from the Kalman filter output and is given by:

$$\log\mathcal{L}(Y_T) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(\log|S_t| + v_t^T S_t^{-1} v_t), \tag{1.18}$$

where v_t and S_t are the residual and the system uncertainty, respectively. For each thirty-minute interval of each day in the training sample, we estimate the parameters by maximum likelihood using the quasi-Newton method. Finally, for the same sample, we run the Kalman smoother using the optimised parameters for the same day. Since in the smoothing process we use all available information, the estimates from this step can be interpreted as the best *ex-post* estimates of price impact (Mertens, 2014).

1.4 Estimation results

1.4.1 Data

We use data from the LOBSTER database[5] for six of the most-highly traded stocks and ETFs on the NASDAQ. We select two large tick stocks: Cisco Systems, Inc. (CSCO) and Intel Corp. (INTC), two medium tick stocks: Apple, Inc. (AAPL) and Facebook, Inc. (FB) and

[4]Since $\tilde{\sigma}^2_\varepsilon$ and $\tilde{\sigma}^2_\xi$ need to be positive, the original problem is that of a constrained maximisation.
[5]https://lobsterdata.com.

two ETFs: the PowerShares QQQ Trust, Series 1 ETF (QQQ), which tracks the Nasdaq-100 and the SPDR S&P 500 ETF (SPY), which tracks the S&P 500. The data spans the period: 01-Jan-2017: 31-May-2017, and is organised, each day, in two tables: a *message* table where we can find the timestamp (with millisecond precision), the order type, the order ID, the size and the price direction (buy/sell), and the *orderbook* table where we can find the best bid and ask prices and quantities. We work with the so-called *level I* order book since we are only interested in what happens at the best bid and ask prices.

The dataset is divided into a training set of 110 days and a testing set of 15 days. We use two uniform time grids $\{T_0, ..., T_I\}$ and $\{t_{0,0}, ..., t_{I,K}\}$ with time steps $T_i - T_{i-1} = 30$ minutes and $t_{k,i} - t_{k-1,i} = 10$ seconds. This leads to 13 long time intervals $[T_{i-1}, T_i]$ for each day and 180 observations for each of these intervals.

1.4.2 Variables

We explore three different definitions of order flow imbalance. The first one is the definition introduced by Cont et al. (2014), and summarised in equation (1.5). In the second definition, we make use of the full information available in our database (LOBSTER) and also include executions against hidden liquidity.[6] The order flow imbalance then becomes:

$$OFI_k \equiv V_k^b - V_k^a, \tag{1.19}$$

[6] Hidden limit orders placements and cancellations are not available in the LOBSTER database.

with the volume posted at the best bid during the interval $[t_{k-1}, t_k]$ V_k^b and the volume posted at the best ask V_k^a during the same time interval, defined as:

$$V_k^b \equiv L_k^b - C_k^b - M_k^{b,v} - M_k^{b,h}, \tag{1.20}$$

$$V_k^a \equiv L_k^a - C_k^a - M_k^{a,v} - M_k^{a,h}, \tag{1.21}$$

where the v and h superscripts on market orders represent the visible and hidden components, respectively. In the third definition of order flow imbalance, we go back to the standard practice in the literature of only looking at (visible) trades, in order to compare this with the previous two definitions. This is the definition adopted, among others, by Hasbrouck (1991), Kempf and Korn (1999), Hasbrouck and Seppi (2001) and Chordia et al. (2008). In this case, the order flow imbalance becomes:

$$OFI_k \equiv M_k^{b,v} - M_k^{s,v}. \tag{1.22}$$

1.4.3 Results

Figure 1.1 shows an example of the least-square regression we used in the first step of this process, for a large tick stock (CSCO), a medium tick stock (FB) and an ETF (QQQ) on January 20, 2017 between 12:00 pm and 12:30 pm. In the first row, we show the results using the definition of order flow imbalance in (1.5); in the second row we used the definition of order flow imbalance which also includes hidden market orders (1.19), while in the final row we show the regressions where the order flow imbalance only considers trades (1.22).

If we focus on the first row, we can see how medium tick stocks, such as FB, and ETFs,

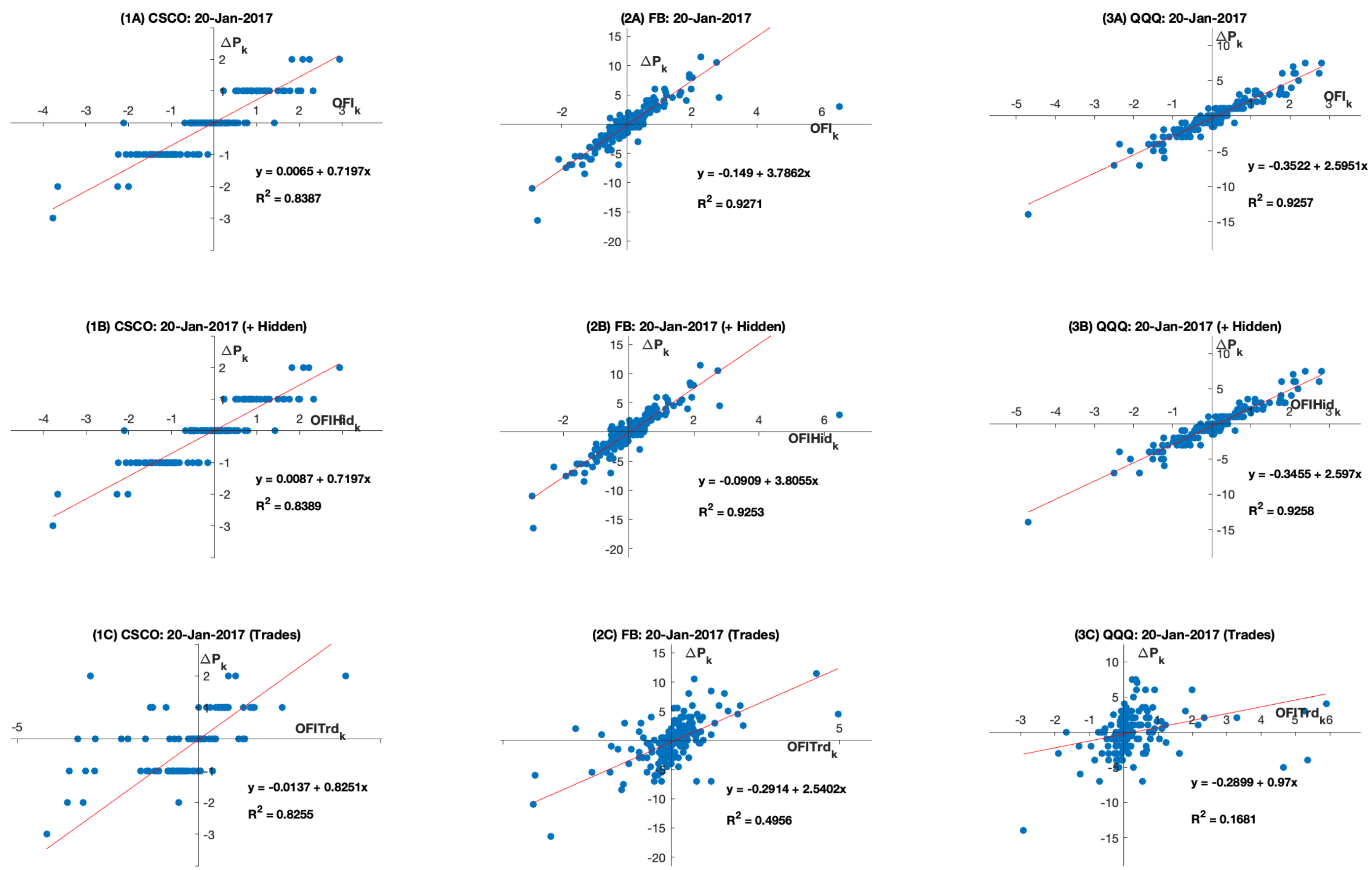

Figure 1.1: Scatter plots and linear regression lines of $\Delta\tilde{p}_k$ on $\tilde{OFI}_k$ for Cisco Systems, Inc. (CSCO), Facebook, Inc. (FB) and PowerShares QQQ Trust, Series 1 ETF (QQQ) on 20-Jan-2017 12:00-12:30pm. In the first row (1A-3A), the order flow imbalance includes all events (limit orders, market orders and cancellations) of the visible limit order book; in the second row (1B-3B), the order flow imbalance includes all events (limit orders, market orders and cancellations) of the visible limit order book plus market orders executed against hidden liquidity; in the third row (1C-3C), the order flow imbalance includes all trades in the visible limit order book.

such as QQQ, display different characteristics compared to a typical large tick stock, such as CSCO, which is the type of stock considered in Cont et al. (2014). In particular, for large tick stocks the order book is densely populated, the spread is often 'locked' at one tick, and prices move little around a narrow grid;[7] as a result, available depth is the dynamic variable that adjusts to uncertainty and risk. This is not the case for other types of stocks or ETFs, where the order book is less densely populated allowing the price to vary more and become the variable that adjusts to uncertainty and risk.[8] From Figure 1.1 we can also see how the linear regression in (1.8) provides a remarkably good fit for the first two definitions of order flow imbalance, with coefficients of determination ranging between 83.9% and 92.7%, confirming the linear relationship introduced by Cont et al. (2014). However, when we consider only trades, this linear relationship no longer seems to hold. In particular, for medium-tick stocks and ETFs, the coefficients of determinations are only 49.6% and 16.8%, respectively. We therefore drop this definition of order imbalance in the rest of our analysis.

In Table 1.1 we show the mean and interquartile range of each of the three parameters $\hat{\theta} = (\hat{\rho}_i, \hat{\sigma}^2_{\varepsilon,i}, \hat{\sigma}^2_{\xi,i})$ for the six stocks under consideration over the whole day. The empirical distributions of the parameters tend to be fairly concentrated, particularly the autoregressive coefficient $\hat{\rho}$. Only the estimates of the process noise, $\hat{\sigma}^2_{\varepsilon}$, for AAPL is quite wide, which is to be expected given the higher dispersion of price changes compared to large-tick stocks and the presence of several outliers. We can see from Table 1.1 how the mean estimates of

[7] This issue has been widely investigated, and alternative approaches that address price discreteness and excess zeros have been proposed in the literature, such as the ordered-probit model, first introduced by Hausman et al. (1992), the ADS model of Rydberg and Shephard (2003), and count models such as the INAR/ACP model of Gross-Klussmann and Hautsch (2013) and the zero-inflated poisson model of Hautsch and Pohlmeier (2001).

[8] Another important difference between large- and small-tick stocks is the use of hidden liquidity; as one moves from large to small tick stocks, hidden liquidity increases quite significantly. This is due to the diminishing role of time priority in smaller tick stocks.

	$\hat{\rho}$			$\hat{\sigma}^2_{\varepsilon}$			$\hat{\sigma}^2_{\xi}$		
	mean	IQR		mean	IQR		mean	IQR	
CSCO	0.126	0.037	0.153	0.989	0.871	1.019	1.486	0.944	2.338
INTC	0.206	0.117	0.227	1.139	0.989	1.217	1.173	0.799	1.803
AAPL	0.619	0.594	0.640	2.713	2.068	3.627	4.011	3.098	5.254
FB	0.670	0.660	0.682	3.324	3.062	4.149	1.013	0.826	1.376
QQQ	0.638	0.552	0.697	2.110	1.330	2.158	1.781	1.305	1.902
SPY	0.680	0.627	0.727	3.726	2.408	4.420	2.467	2.237	2.777

Table 1.1: Daily parameters empirical distributions: mean and IQR for Cisco Systems, Inc. (CSCO), Intel Corp. (INTC), Apple, Inc. (AAPL), Facebook, Inc. (FB), PowerShares QQQ Trust, Series 1 ETF (QQQ) and SPDR S&P 500 ETF (SPY), on 01-Jan-2017: 30-Jun-2017.

the resiliency coefficient $\hat{\rho}$ for the six stocks are quite low for CSCO and INTC, meaning that the market-impact coefficient $\tilde{\beta}_{k,i}$ returns to its mean estimate $\tilde{\mu}_{k,i}$ very quickly. The estimates of $\hat{\rho}$ for the other four stocks are higher and all in a fairly tight range of 0.619-0.670. Nonetheless, these estimates are far from one, implying that although more slowly than for CSCO and INTC, the market-impact coefficients return to their average levels. This is in sharp contrast with the findings of Mertens (2014), where market-impact coefficients are very close to one with the implication of being a near unit root. Our findings suggests that this persistency is mainly due to the intraday seasonal component, and that once this is accounted for in the model, impact coefficients become much less persistent.

In Figure 1.2, we also show the diurnal effects of the estimated coefficient $\hat{\rho}_i$ for each of the six stocks. The most striking feature is that the market-impact coefficients of CSCO and INTC are much less persistent than those of the other four stocks/ETFs. Also, these two stocks display stronger diurnal effects than the other four stocks and ETFs, which are not completely captured by our model. QQQ and SPY also shows some diurnal variability, though less than CSCO and INTC, while the persistency of AAPL's and FB's market-impact

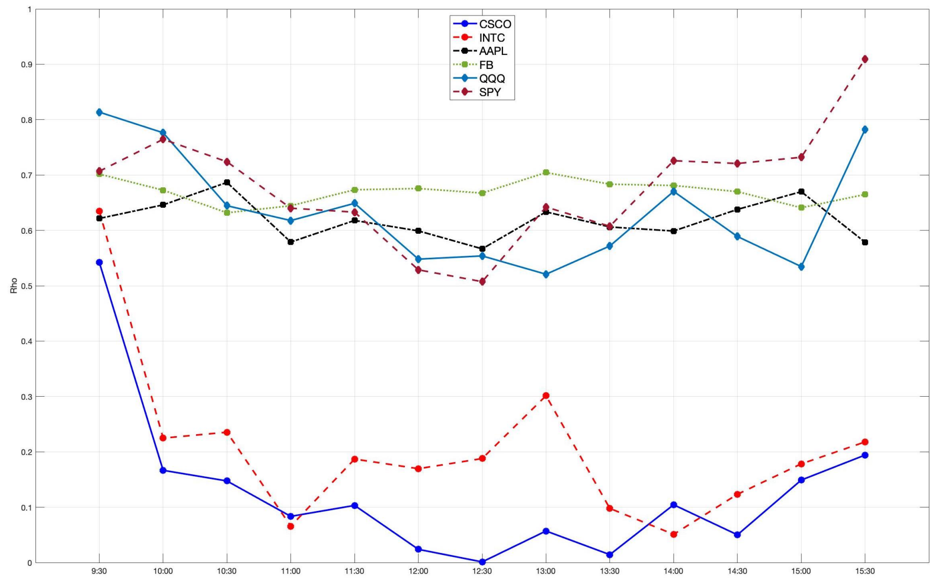

Figure 1.2: Diurnal effects of the estimated coefficient $\tilde{\rho}_i$ (eqn. 1.17) for Cisco Systems, Inc. (CSCO), Intel Corp. (INTC), Apple, Inc. (AAPL), Facebook, Inc. (FB), PowerShares QQQ Trust, Series 1 ETF (QQQ) and SPDR S&P 500 ETF (SPY), on 01-Jan-2017: 30-Jun-2017.

coefficients seems to be fairly constant throughout the day.

Once we have obtained the empirical distributions for the parameters, we can use the model to perform the Kalman smoothing. Figure 1.3 shows the smoothed estimates of the stochastic component of the market-impact coefficient $\tilde{\eta}_k$ (upper panel) and mid price (lower panel) for AAPL on March, 15 2017. This day was chosen to display the real-time evolution of the price impact coefficient in the presence of some external news. Indeed, on March, 15 2017, the Fed's Federal Open Market Committee (FOMC) issued its policy statement for its March, 14-15 meeting, in which it decided to raise its policy rate from 0.75% to 1.00%. Its policy statement showed a more hawkish monetary policy stance than earlier anticipated and the possibility of further rate hikes for the rest of the year. We can see from Figure 1.3 how our measure of market impact spiked very suddenly from a level close to zero to a level

Figure 1.3: In-sample Kalman Smoothed Estimates of the stochastic component of market impact $\tilde{\eta}_k$ (equation 1.17) (upper panel) and mid price (lower panel) for AAPL on March, 15 2017. In the upper panel, the solid blue line is the estimate using only visible quantities in the order book, while the dashed red line is the estimate using also hidden liquidity.

around six in the half hour 12:30-1:00pm, before the release of the decision indicating some re-positioning from traders before the announcement.[9] This coincided with the start of a downward trend in the mid price which lasted until around 2:00 pm.

This is the time the FOMC issued its decision and its policy statement. We can see that following this release, our measure of price impact spiked again from zero to nearly 5 and then reverted to zero. The mid price started an upward trend that lasted from 2:00 pm until approximately 3:30 pm. The market-impact coefficient remained more volatile during the rest of the day. It is also quite clear from the top panel of Figure 1.3 that there seems to be

[9]No other significant economic news was released during this time.

no additional information in using the measure of OFI which also contains hidden liquidity (dashed red line) rather than the measure of OFI containing only visible quantities (solid blue line). These two measures are nearly indistinguishable, with the exception of the first half hour. And even in that initial period, the measure of OFI containing hidden liquidity seems to be proportional to the the measure of OFI containing only visible quantities.[10]

Figure 1.4 shows the performance of the model out-of-sample for CSCO, FB and SPY. For the stochastic component of market impact $\tilde{\eta}_k$, we show three measures: the *historical* market impact (dashed black line), the filtered *real time* estimate of market impact (solid blue line) and the *ex-post* estimate of market impact (dotted red line). The first is defined as the mean of the smoothed estimates from the training set. The second is obtained by running the filter with the parameters set at the median of the estimated in-sample parameters, up to the previous time step, and by setting the prior mean and covariance for the state equal to the median $\tilde{\eta}_k$ and MSE obtained from the first step in the optimisation problem discussed above; again, up to the previous time step. Finally, the third measure of market impact is obtained by Kalman smoothing. The filtered estimate seems to capture the dynamics of market impact quite well, while the historical estimates seem to be very close to zero, with the exception of the beginning of the day for CSCO and FB, and end of the day for SPY. To compare the performance of the real-time and historical estimates of market impact out of sample, we follow Mertens (2014) and compute their mean squared errors (MSE) with respect to the smoothed estimates. From Table 1.2, we can see how, for the 10 days in our testing set, the MSE for the real time estimate (MSE_{rt}) is almost always much smaller than the MSE of the historical estimate (MSE_h), the only exceptions being two days for INTC (26 and 30 June

[10] Although not shown here for brevity, company-related news also have a distinctive effect on latent liquidity and therefore on our price-impact measure.

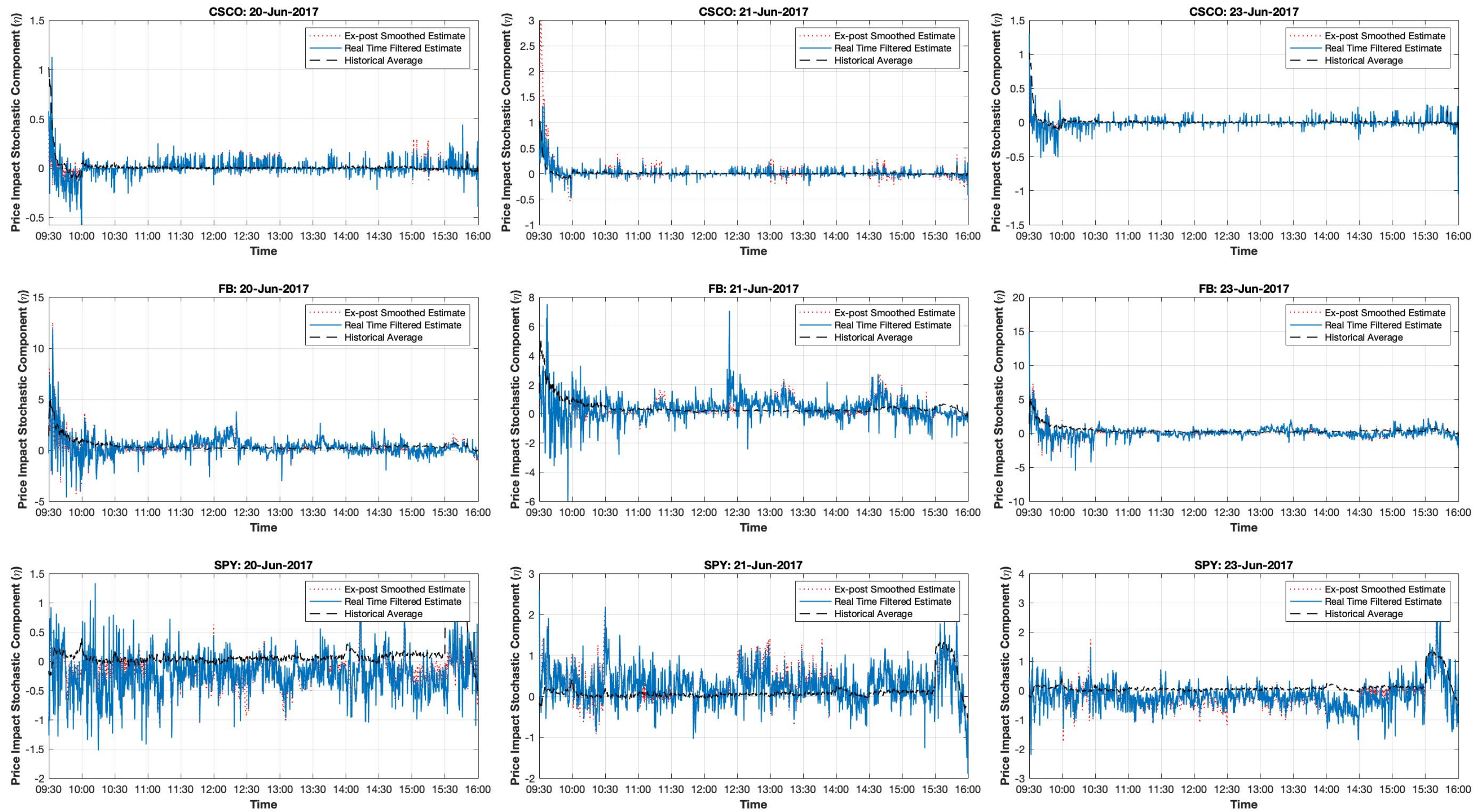

Figure 1.4: Out-of-Sample Kalman Smoothed Estimates for Cisco Systems, Inc. (CSCO), Facebook, Inc. (FB) and PowerShares QQQ Trust, Series 1 ETF (QQQ), on 20-Jun-2017, 21-Jun-2017 and 23-Jun-2017. For the stochastic component of market impact $\tilde{\eta}_k$ (equation 1.17), we show the ex-post Kalman smoothed estimate (dotted red line), the *real-time* Kalman filtered estimate (solid blue line), and the *historical* average market impact (dashed black line), defined as the mean from the in-sample smoothed estimates.

	CSCO			INTC		
date	MSE_h	MSE_{rt}	Gain (%)	MSE_h	MSE_{rt}	Gain (%)
19-Jun-2017	0.019	0.014	23.166	0.025	0.022	10.302
20-Jun-2017	0.008	0.001	82.983	0.012	0.003	77.906
21-Jun-2017	0.036	0.033	9.447	0.044	0.008	82.804
22-Jun-2017	0.012	0.010	14.988	0.019	0.004	79.289
23-Jun-2017	0.007	0.002	78.618	0.017	0.005	68.394
26-Jun-2017	0.009	0.008	10.838	0.029	0.034	-17.097
27-Jun-2017	0.010	0.002	77.307	0.036	0.021	40.827
28-Jun-2017	0.009	0.002	79.498	0.030	0.004	85.058
29-Jun-2017	0.078	0.058	25.400	0.044	0.017	61.173
30-Jun-2017	0.025	0.013	47.029	0.033	0.065	-100.487
median			**36.215**			**64.783**
	AAPL			**FB**		
date	MSE_h	MSE_{rt}	Gain (%)	MSE_h	MSE_{rt}	Gain (%)
19-Jun-2017	0.707	0.128	81.956	0.744	0.200	73.168
20-Jun-2017	0.821	0.102	87.626	0.666	0.194	70.913
21-Jun-2017	0.776	0.112	85.609	0.764	0.255	66.643
22-Jun-2017	1.062	0.126	88.112	1.440	0.490	65.988
23-Jun-2017	1.007	0.235	76.689	0.838	0.111	86.713
26-Jun-2017	1.366	0.377	72.376	1.425	0.407	71.446
27-Jun-2017	1.141	0.443	61.188	2.870	0.738	74.279
28-Jun-2017	0.464	0.258	44.408	0.774	0.285	63.120
29-Jun-2017	2.786	0.805	71.120	5.447	1.082	80.127
30-Jun-2017	0.604	0.327	45.849	1.067	0.345	67.659
median			**74.533**			**71.180**
	QQQ			**SPY**		
date	MSE_h	MSE_{rt}	Gain (%)	MSE_h	MSE_{rt}	Gain (%)
19-Jun-2017	0.086	0.042	50.833	0.187	0.063	66.505
20-Jun-2017	0.207	0.064	69.227	0.135	0.072	46.405
21-Jun-2017	0.171	0.075	56.289	0.215	0.062	70.889
22-Jun-2017	0.111	0.047	57.186	0.132	0.049	62.766
23-Jun-2017	0.184	0.044	76.226	0.228	0.070	69.173
26-Jun-2017	0.473	0.055	88.487	0.237	0.057	76.124
27-Jun-2017	1.128	0.302	73.257	0.777	0.123	84.129
28-Jun-2017	0.251	0.056	77.640	0.247	0.051	79.506
29-Jun-2017	2.648	0.509	80.784	2.440	0.311	87.271
30-Jun-2017	0.447	0.067	85.078	0.263	0.092	64.883
median			**74.742**			**70.030**

Table 1.2: Out-of-sample mean squared error of real time filtered estimates (MSE_{rt}) versus historical estimates (MSE_h) and percentage gain in signal extraction ($Gain = 100 * (MSE_h - MSE_{rt})/MSE_h$) for Cisco Systems, Inc. (CSCO), Intel Corp. (INTC), Apple, Inc. (AAPL), Facebook, Inc. (FB), PowerShares QQQ Trust, Series 1 ETF (QQQ) and SPDR S&P 500 ETF (SPY), on 01-Jan-2017 : 30-Jun-2017.

2017). In the fourth and seventh column of Table 1.2, we show the percentage gain in signal extraction, defined as $100 * (MSE_h - MSE_{rt})/MSE_h$. The median percentage gain in signal extraction ranges from around 36.2% for CSCO to 74.7% for SPY, indicating how real-time monitoring of market impact leads to superior out-of-sample results compared to historical measures.[11]

1.4.4 Order flow imbalance and depth

In this section, we test how our smoothed estimate of the market-impact coefficient $\tilde{\beta}_i$ relates to the order book depth D_i. Our model states that

$$D_i = \frac{1}{2\beta_i}, \tag{1.23}$$

which, as previously mentioned, captures the intuitive concept that, other things equal, market impact is higher (lower) the 'shallower' ('deeper') the order book. We therefore run the following least-square regression:

$$\tilde{\beta}_i = \alpha_i + \lambda_i \frac{1}{2D_i} + \varepsilon_i, \tag{1.24}$$

where $\tilde{\beta}_i$ is the Kalman smoothed estimate and D_i is the average order book depth at the best.[12] These are Mincer-Zarnowitz regressions (Mincer and Zarnowitz, 1969), and so we expect α_i to be equal to zero, and both λ_i and the R^2 to be equal to one. In Table 1.3 we show the median estimates of this regression for the six stocks and ETFs under consideration. The 'Visible-Visible' results show the median estimates when $\tilde{\beta}_i$ is the smoothed estimate with

[11] We also carried a similar out-of-sample analysis for the order flow measure that includes hidden order executions with very similar results.

[12] We take an arithmetic average of the bid and ask queues. Geometric average and weighted average with weights given by the number of events on each side of the queue give similar results.

	Visible-Visible			Total-Total		
	$\hat{\alpha}_i$	$\hat{\lambda}_i$	$R^2(\%)$	$\hat{\alpha}_i$	$\hat{\lambda}_i$	$R^2(\%)$
CSCO	0.458	0.814	57.16	0.425	0.820	60.84
INTC	0.306	0.842	43.59	0.312	0.822	43.83
AAPL	0.713	0.281	21.93	0.687	0.254	18.74
FB	0.681	0.276	21.33	0.634	0.269	19.90
QQQ	0.838	0.978	13.12	0.872	0.941	13.52
SPY	0.479	0.877	25.98	0.518	0.834	23.26

Table 1.3: Median least-square estimates of the regression: $\tilde{\beta}_i = \alpha_i + \lambda_i \frac{1}{2D_i} + \varepsilon_i$ for Cisco Systems, Inc. (CSCO), Intel Corp. (INTC), Apple, Inc. (AAPL), Facebook, Inc. (FB), PowerShares QQQ Trust, Series 1 ETF (QQQ) and SPDR S&P 500 ETF (SPY) over the period 01-Jan-2017 : 30-Jun-2017; see eqn. (1.24). $\tilde{\beta}_i$ is the Kalman smoothed estimate, and D_i the depth (average queue size at the best). The 'Visible-Visible' results show the median estimates when $\tilde{\beta}_i$ is the smoothed estimate with order flow imbalance and depth only containing visible quantities (no hidden market orders). The 'Total-Total' results show the median estimates when $\tilde{\beta}_i$ is the smoothed estimate with order flow imbalance and depth also containing hidden market orders.

order flow imbalance and depth only containing visible quantities (no hidden market orders). The 'Total-Total' results show the median estimates when $\tilde{\beta}_i$ is the smoothed estimate with order flow imbalance and depth also containing hidden market orders.

We can see that the estimated coefficients, $\hat{\lambda}_i$ are close to one for CSCO, INTC, QQQ and SPY, and we cannot reject the hypothesis of $\lambda_i = 1$ at the 1% significance level in around $17 - 22\%$ of the cases for CSCO and INTC, and in around $31 - 47\%$ of the cases for QQQ and SPY. For AAPL and FB, however $\hat{\lambda}_i$ is significantly different from one in almost all cases, which is possibly due to the fact that for medium-tick stocks deeper levels of the order book exert a stronger influence on price dynamics than for large-tick stocks. The estimated constant coefficient $\hat{\alpha}_i$ is statistically different from zero at the 1% significance level in all regressions with the exception of around $1 - 4\%$ of regressions for CSCO and INTC. Similarly, a Wald test for the joint hypothesis of $\alpha_i = 0$ and $\lambda_i = 1$ is rejected in all regressions with the exception of $1-3\%$ of regressions for CSCO and INTC. The coefficient of determination (R^2) ranges from around 13% of QQQ to 60.84% for INTC. For AAPL and FB, the situation is a bit different and the estimated coefficients are not close to one, indicating that the model might not be fully capturing the dynamics of medium tick stocks. These results are quite similar to those of Cont et al. (2014): although their price impact-depth regressions have greater explanatory power than ours in terms of R^2 for large tick stocks, they also find that, for medium- and small-tick stocks, the estimated coefficients $\hat{\lambda}_i$ are much smaller than one, and that the coefficients of determination are lower than for large-tick stocks.

1.5 Conclusion

Markets operate in a state of high *latent* liquidity and low revealed liquidity. Since the outstanding liquidity is so small, traders try to conceal their intentions and try to avoid market impact by executing large orders only incrementally. Motivated by this observation, we have proposed a market microstructure model in which the market impact of an individual order depends on the aggregate activity of the order book as summarised by a single aggregate variable: the *order flow imbalance*. We have empirically tested this model on large and medium tick stocks, as well as ETFs using three different measures of order flow imbalance. Our empirical findings show that the model captures rather well the market-impact dynamics. We find that filtered estimates of market-impact coefficients deliver better out-of-sample forecasts than historical estimates. We find that once the seasonality of the market-impact coefficient is accounted for, the deviations around this seasonal component are not very persistent. This is in contrast with previous findings in the literature that reported estimates of market-impact coefficients very close to one. Finally, we find that for large-tick stocks and ETFs, the smoothed market-impact estimates are strongly related to the average order book depth. The situation is a bit different for the remaining two stocks (AAPL and FB), indicating that the model might not be fully capturing the dynamics of medium-tick stocks. We believe that the ability of our model to capture latent liquidity and track the order book depth may be useful for the design of optimal execution strategies. In particular, optimal execution strategies could use real-time estimation of market impact to dynamically adapt their execution schedule so that trades are executed at times when market impact is lower.

2

Trade duration, volatility and market impact

This chapter is based on the paper "Trade duration, volatility and market impact" which is a joint work with Professor Rama Cont.

2.1 Introduction

2.1.1 The determinants of execution costs

The analysis of trade execution costs is a key ingredient in the design and performance measurement of trading strategies. A common approach to the quantification of these costs has been to separate one's own demand or supply contribution from the aggregate contribution of the rest of the market, and, consequently, to decompose execution costs into two components: *market impact* and *volatility risk* (Collins and Fabozzi, 1991). Market impact refers to the fact that a trader's order might move the price in an unfavourable direction. Volatility risk refers to possibility of adverse moves in market prices during execution, regardless of the trader's action.

Starting with the seminal paper of Almgren and Chriss (2001), the problem of minimising execution costs has been studied through the angle of a trade-off between market impact and volatility risk, as if they were two separate, independent components. With a few notable exceptions (Almgren et al., 2005), the prevailing view in the literature is that the market impact of a trade is the main determinant of execution costs, while the contribution of price volatility averages to zero if one examines a large number of trades.

Most of the theoretical models on optimal execution have therefore focused on optimising the market impact component, and the empirical studies on market impact have tried to 'explain' execution costs only in terms of trade characteristics, usually through some correlation analysis or regression studies. In particular, the empirical literature on the *square-root law* of market impact (Almgren et al., 2005; Moro et al., 2009; Toth et al., 2011; Bacry et al., 2015; Donier and Bonart, 2015; Gomes and Waelbroeck, 2015; Zarinelli et al., 2015) points to the *size* of the trade as the key variable determining market impact, while the (mostly theoretical) literature on optimal execution, pioneered by Almgren and Chriss (2001), considers the *speed* at which the trade is executed, and the related concept of *participation rate*, as the key determinants of execution costs.

Despite the wide use of such models, the theoretical and empirical evidence underlying their justification is not entirely convincing for a number of reasons. First, most of the empirical work that looks at price changes with trade size (Torre, 1999; Moro et al., 2009; Gomes and Waelbroeck, 2015; Bacry et al., 2015; Toth et al., 2016) or participation rate (Almgren and Chriss, 2001; Bacry et al., 2015) is based on the analysis of *unconditional* relations, and ignores the influence of the aggregate net order flow on trades' execution costs (Hasbrouck, 1991, 2007; Cont et al., 2014). But unconditional correlation analysis, being symmetric, has little to say with respect to *causal relations* and may yield spurious 'relations' between quantities even when there is none. In order to uncover meaningful *causal* relations, one needs to *condition* on relevant variables (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009).

Second, many such empirical studies focus on 'average market impact' by averaging over many trades with similar size, implicitly treating the market impact and volatility components as independent. The former, due to one's own demand, is 'explained' in terms of trade characteristics; the latter, due to the market's aggregate demand is, with a few exceptions (Bacry et al., 2015; Almgren et al., 2005), averaged out under the assumption of independence across time. This notion of average impact may be relevant for market participants executing a large number of trades sequentially across 'independent' market scenarios and independently of the market order flow, but these assumptions may not hold in the case of institutional investors, for which the observed execution costs may deviate significantly from such an average. Many market participants execute trades based on market signals, and their order flow is often correlated with the market order flow. As a result, the two components of execution costs are *not* independent in general, so the contribution of market volatility may fail to average to zero.

2.1.2 Contributions

In the present study, we provide a different perspective on the concept of market impact and the determinants of execution costs based on a *causal* analysis of the relation between execution costs, volatility, trade duration, trade size and order flow imbalance. Using a data set of millions of institutional trade executions, we apply *conditional* regression analysis (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009) to uncover causal relations between these variables.

Our *causal* analysis shows that the dominant component of execution costs for orderly trade executions is not the 'market impact' of trades but price volatility. We show that price changes accompanying trade executions are mainly driven by the *aggregate (net) order flow imbalance* rather then characteristics of individual trades, even when the trade size is large. In particular, we show that a buy (sell) trade may be accompanied by a negative (positive) price change with high probability ($\simeq 40\%$). Once we condition on market order flow imbalance, features of individual trades lose their explanatory power.

A second finding is that the amplitude of price changes during execution is proportional to the square-root of trade *duration*, rather than trade size, a relation which has a simple explanation in term of the scaling of volatility as a function of the time horizon (Bachelier, 1900). Moreover, we observe that, once we condition on trade duration, the dependence of both absolute and signed execution costs on trade size vanishes almost entirely.

Our study also provides evidence that, conditional on trade duration, the *participation rate* has little or no influence on the amplitude of price changes during execution. In fact, we find that, for trades of a given size, *slower* execution leads to higher execution costs, which is explainable by higher exposure to volatility risk during execution. In particular, we find that for day-long VWAP trades, a large increase in the traded quantity (and in the participation rate) from $\simeq 18.5\%$ to $\simeq 50\%$ of daily volume leaves absolute price changes almost unchanged, and leads to a *decline* in signed price changes, contradicting the idea that the participation rate is the main determinant of execution costs.

Although these results fly in the face of some assumptions commonly used in the literature on optimal trade execution, they are consistent with previous studies of Engle et al. (2012); Bershova and Rakhlin (2013) on trade executions in equity markets and Salem et al. (2018) on trade execution in the US Treasury market, which similarly underline volatility risk as the main determinant of execution costs.

Our findings have implications for optimal trade execution and transaction cost analysis. They question in particular the received notion of 'price impact of a trade' and challenge the view that 'market impact' should be treated as the main determinant of execution costs. Our findings also question the use of participation rate as the key determinant of execution costs, a feature common to many optimal execution models. Finally, they suggest that instead of focusing on optimising market 'impact' at a trade-by-trade level, investors should focus on modelling the aggregate dynamics of net order flow and on managing the main source of risk: price volatility. In particular, this means that for *orderly trade executions* (Avellaneda and Cont, 2013), where trade size and the participation rate are within reasonable bounds, the focus should be on modelling and forecasting order flow dynamics, since execution costs will be different depending on whether a trade leans with or against the market order flow (Bechler and Ludkovski, 2015). This is related to the concept of *trade timing*, which is further analysed in section 3.4.

2.1.3 Relation with previous literature

The concept of market impact of a trade refers to the notion that a market participant's order may move the price in an unfavourable direction, thereby resulting in a higher execution cost. An early reference is Kyle (1985) who introduced the concept that trades by a market participant may have an impact on the market price. In Kyle's model, this impact was modeled as a linear function of the trade size. Huberman and Stanzl (2004) reinforced this assumption of linearity by showing that, provided liquidity is constant, permanent impact must be linear to prevent arbitrage.

Following these insights, a considerable body of empirical research has investigated the relation between trade size and price variations during trade execution (Chan and Lakonishok, 1993, 1995; Torre, 1997, 1999; Lillo et al., 2003; Lillo and Farmer, 2004; Bouchaud et al., 2009; Toth et al., 2011; Bershova and Rakhlin, 2013; Cont et al., 2014; Bacry et al., 2015; Gomes and Waelbroeck, 2015; Toth et al., 2016). In particular, several studies point to a concave dependence of the amplitude of price variations on trade size, sometimes represented as a square root function. These findings are sometimes referred to as the 'square-root law of market impact' and may be summarised by the following relation between the size Q of a trade, as a percentage of the daily market volume of the traded asset from open to close V_D, and its expected 'market impact' $\mathcal{I}(\frac{Q}{V_D})$ on the asset price (Toth et al., 2011):

$$\mathcal{I}\Big(\frac{Q}{V_D}\Big) = Y\sigma_D\sqrt{\frac{Q}{V_D}}, \tag{2.1}$$

where $\mathcal{I}(\frac{Q}{V_D})$ is the average execution cost conditioned on trade size as a percentage of daily volume, where the cost is usually defined as the signed percentage difference between the arrival price at the start of the trade and the (volume-weighted) average execution price ('trade VWAP'), Y is a constant, and σ_D is the daily asset volatility. Pohl et al. (2017) give arguments based on dimensional analysis as to why one should expect a square root dependence.

In fact, there is considerable variation in empirical findings: many studies have pointed to deviations from this square root dependence (Lillo et al., 2003) and propose variations of this model where the square root is replaced by a power law with some exponent $\alpha < 1$. Some of these findings are summarised in Table 2.1.

However, a closer inspection leads to many questions regarding the methodology and the conclusions of these studies. In Figure 2.1 we present a graphical model summarising standard 'market impact' models' causal dynamics. We follow the graphical model literature and use plates to represent iteration over N trades or 'metaorders' (Jordan, 2004).

As shown in the causal graph in Figure 2.1, most of the aforementioned studies attempt to explain price changes during execution by using trade features, such as size or speed, together with volatility and volume, as sole explanatory variables. This contrasts with the more traditional view which relates market price fluctuations to fluctuations in *aggregate* supply

Authors	Exponent α	Asset	Region
Lillo *et al.* (2003)	0.10-0.50	Equities	US
Almgren *et al.* (2005)	0.60	Equities	US
Moro *et al.* (2009)	0.64-0.72	Equities	UK
	0.44-0.48	Equities	Spain
Toth *et al.* (2011)	0.50-0.60	Futures	NA
Bacry *et al.* (2015)	0.40-0.45	Equities	Europe
Bershova and Rakhlin (2015)	0.47	Equities	US
Donier and Bonart (2015)	0.50	Bitcoin	BTC/USD
Gomes and Walbroeck (2015)	0.50	Equities	NA
Zarinelli *et al.* (2015)	0.47	Equities	US
Toth *et al.* (2016)	0.50	Options	NA

Table 2.1: Summary of empirical studies on the (power-law) dependence of 'market impact' on trade size. An exponent of 0.5 corresponds to a square-root dependence.

and demand, represented by the net order flow (see e.g. Blume et al. (1989); Hasbrouck (1991); Brown et al. (1997); Hasbrouck and Seppi (2001); Chordia et al. (2002); Cont et al. (2014)).

Additionally, most of the empirical work relating price changes and trade size (Torre, 1997, 1999; Moro et al., 2009; Gomes and Waelbroeck, 2015; Bacry et al., 2015; Toth et al., 2011, 2016) has been based on the analysis, typically using logarithmic regressions, of *unconditional* relations; i.e. without conditioning on relevant variables such as the execution horizon or *duration* of the trade, or the aggregate net order flow. An exception is Zarinelli et al. (2015), who study market impact of large institutional trades using the same database as the present work (ANcerno); these authors study the joint dependence of price changes on the participation rate and market trading volume and investigate the effect of conditioning on market capitalisation, participation rate and market trading volume (but not execution horizon or order flow imbalance). Bacry et al. (2015) present a joint analysis of the trade size and duration, but do not go as far as examining conditional relations.

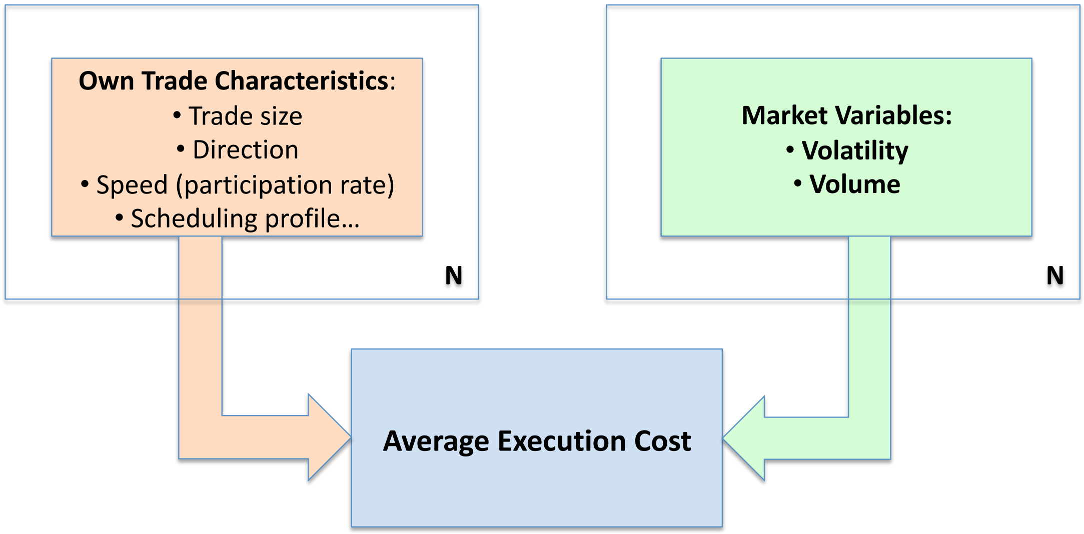

Figure 2.1: Causal representation of standard market impact models. The *average* execution cost (over N trades or 'meta-orders') is solely explained in terms of trade characteristics, volatility and trading volume, without any reference to *aggregate* demand and supply. Volatility is treated as *independent* from the market participant's trade and its characteristics.

Unfortunately, unconditional correlation analysis is symmetric, and has little to say with respect to causal relations; in fact, it may yield spurious 'relations' between quantities even when there is none. In order to uncover meaningful causal relations, one needs to condition on relevant variables (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009). A key aspect of our analysis is therefore to focus on such *conditional* relations, in particular conditioning on trade duration, and aggregate order flow imbalance, to elicit the role of different variables.

Starting with Almgren and Chriss (2001), a common feature of *optimal execution* models, summarised in Figure 2.1, has been to focus on *average* price changes during trade execution, where averaging is done across N trades of similar size, and to cast the optimal execution problem in terms of a trade-off between 'market impact' and volatility risk (Almgren and Chriss, 2001; Alfonsi et al., 2010; Gatheral, 2010; Predoiu et al., 2011; Curato et al., 2017). These models consider the *speed of execution* as the key variable driving the 'market impact' term, whilst the 'volatility term' arises from price variations during execution, *independently* of the execution strategy. The speed of execution is often parametrised through the *participation rate*, defined as volume executed per unit time as a fraction of total market volume during the same period, and it is generally assumed that slower execution reduces the impact component, whilst increasing exposure to volatility risk.

In contrast, the square-root law formula (2.1) focuses on the influence of trade *size* but does not consider the rate of trading or how the trade is executed. In fact, if taken at face value, the square root law implies that the impact of a trade is *invariant* with respect to the execution horizon, or duration of the execution i.e. that speeding up the execution by a factor 2 has no influence on its market impact (Bouchaud et al., 2018).

The causal approach proposed in this Chapter differs from the standard paradigm used in many 'market impact' models. We start from the premise that price movements are determined by *aggregate* supply and demand and thus, in markets where orders are anonymised, the order flow of a market participant influences price movements *through* the aggregate order flow. Figure 2.2 shows the causal chain of relations: a market participant's order merge into the aggregate order flow, which, together with the market depth provided by standing (passive) orders in the limit order book, determines the market price and the execution cost. This approach emphasises the relation between each participants order flow and the market order flow imbalance in the determination of the resulting price move, which is consistent with traders' experience of the importance of 'market timing' in trade execution. Moreover, in contrast with standard 'market impact' models, in the present study we examine the influ-

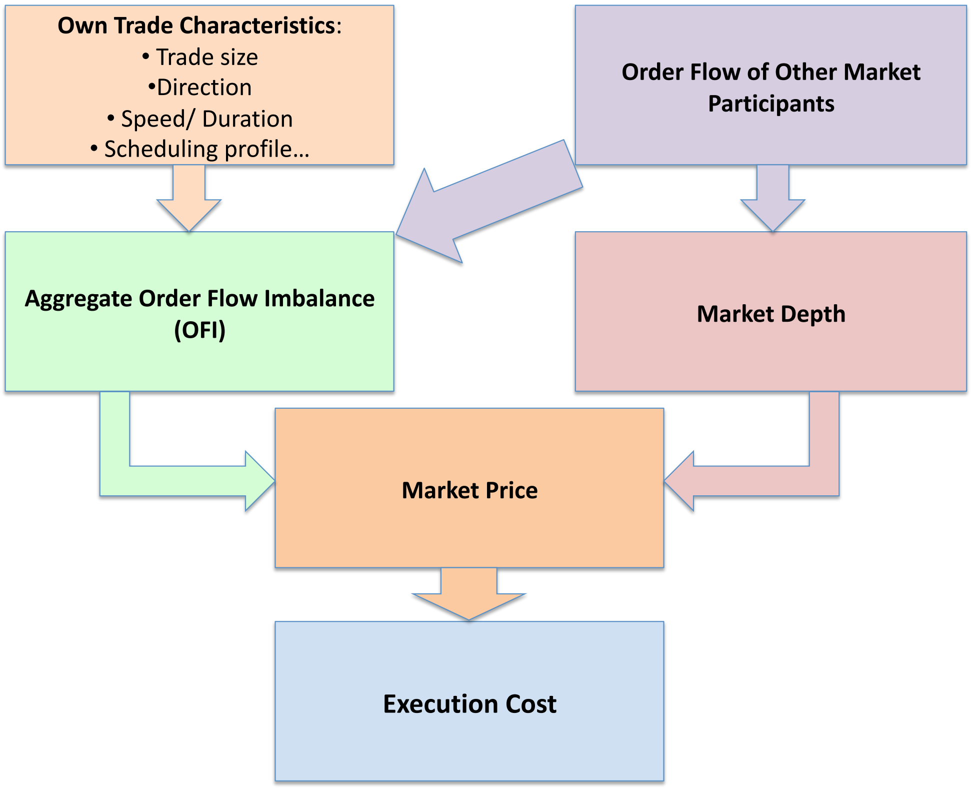

Figure 2.2: Causal representation of execution costs proposed in this paper. A market participant's order merge into the *aggregate order flow*, which interacts with the limit order book to determine the market depth and, consequently, the market price and the execution cost.

ence of these variables on price variations on a trade-by-trade basis, not just on average.

Our analysis also emphasises the role of trade duration, or execution horizon, as a key conditioning variable for the causal explanation of execution costs. The role of duration as a key variable for the explanation of price variations during trade execution has also been noted in (Dufour and Engle, 2000; Engle et al., 2012; Bershova and Rakhlin, 2013; Easley et al., 2015; Bechler and Ludkovski, 2015; Salem et al., 2018). Engle et al. (2012) emphasised the temporal dimension of execution costs and the contribution of volatility. Bershova and

Rakhlin (2013) analysed US equity trades executed by Alliance Bernstein from January 2009 to June 2011 and find that impact is a square-root function of trade duration and report an R^2 of 29% (Bershova and Rakhlin, 2013, Table 5). Salem et al. (2018) report order flow imbalance to be the most important feature for predicting price impact in the US Treasury market and find that *slower* the execution leads to *higher* price impact (Salem et al., 2018, p. 9). These findings are consistent with ours and provide independent corroboration of our conclusions in other markets and data sets.

2.1.4 Outline

The remainder of the Chapter is organised as follows. Section 2.2 describes our data sets and defines the quantities of interest. Section2.3 examines the determinants of the direction of price variations during trade execution, Section 2.4 examines execution costs and their relation to trade size and duration. Section 2.5 discusses the role of trading speed and participation rate. Section 2.6 concludes.

2.2 Data sources and quantities of interest

2.2.1 The ANcerno database

We use institutional trading data provided by ANcerno Ltd., a consulting firm that monitors trade execution costs for institutional clients.[1] The dataset contains institutional trading data on US equities, non-US equities and fixed-income by pension funds, such as CalPERS, the Commonwealth of Virginia, and the YMCA retirement fund; and money managers, such as

[1] The ANcerno database has been used in several academic papers, such as Anand et al. (2012, 2013); Chakrabarty et al. (2017); Chemmanur et al. (2009); Chicayachantana et al. (2017); Hu (2009); Zarinelli et al. (2015). For a complete list of papers, description and history of the dataset see Hu et al. (2018).

MFS, Putnam Investments, Lazard Asset Management and Fidelity. Our empirical analysis focuses on US equity trades between 01-Jan-2008 and 31-Dec-2012.

Hu et al. (2018) and Anand et al. (2012) provide a detailed description of the ANcerno dataset. According to Hu et al. (2018), ANcerno Ltd. provides consultancy and advisory services to 1,000+ institutional clients representing over 12% of CRSP volume, higher than the previous estimate of $8-10\%$ provided by Puckett and Yan (2011). Concerning the quality of the data, Anand et al. (2012) argues that survivorship bias is not an issue for two reasons. First, ANcerno Ltd. representatives have stated that there is no such bias. Second, not all institutions are present at the end of their sample period, suggesting that the ANcerno data does not contain only surviving institutions (Anand et al., 2012). Selection bias is also investigated, where ANcerno Ltd. clients might systematically differ from the typical institution, and rejected as there are no explicit client requirements (e.g. dollar size of fund managed, type of institution or number of trades executed) and stocks traded by ANcerno institutions do not statistically differ from stocks traded by the average 13F institution. The only notable difference is that ANcerno institutions trade in larger sizes (Anand et al., 2012).

In this Chapter, we refer to a collection of transactions that implements the decision to buy or sell a certain quantity of shares as a *trade*.[2] The ANcerno database contains a variable which allows us to link together the sequence of transactions involved in the execution of a trade.[3] We focus on a subset of trades with the following criteria:

1. We focus on stocks in the Russell 3000 Index (i.e. liquid stocks), for which intraday data is available for estimating volatility and order book depth;

[2] Referred to as 'parents orders' or 'meta-orders' in some papers (Bouchaud et al., 2009).

[3] Earlier versions of the database also contained a client identifier; see for example Puckett and Yan (2011), Anand et al. (2012) and Hu et al. (2018).

2. We retain trades with duration lasting at least two minutes (since in our dataset a large proportion of price changes is zero for trades lasting only one minute) and composed of at least five transactions (since we want to analyse the execution cost of a trade or 'parent order' rather than the execution cost of a single execution);

3. We do not consider overnight trades to avoid overnight risk: only trades executed with start and end time within the same trading day are considered;

4. We only consider transactions with execution time before 4:01 PM to avoid estimation biases associated with closing trades;

For each trade we observe, among other variables, the ticker of the stock traded, the trade direction ($\varepsilon = 1$ for buy trades, and $\varepsilon = -1$ for sell trades), the arrival time (t_{start}), the execution time of the last transaction (t_{end}), the arrival price (p_{start}), defined as the average of the bid and ask prices at the time (t_{start}) the trade is initiated, the volume-weighted average execution price (p_{vwap}), the volume of the trade (Q) and the volume of each transaction i (v_i) that constitutes such trade, the daily volume of the traded stock from open to close (V_D), the volume of the traded stock during execution (V_P), the daily highest (p_h) and lowest (p_l) prices. These quantities are reported at one-minute frequency. We define the following variables:[4]

[4] In the Appendix, we report some results were the percentage execution cost is computed by taking as benchmark the market VWAP, rather than the arrival price.

$$
\begin{aligned}
\textit{Trade Duration} \quad T \quad &\equiv \quad t_{\text{end}} - t_{\text{start}}, && (2.2)\\
\textit{Percentage Execution Cost} \quad C \quad &\equiv \quad \varepsilon \frac{(p_{vwap} - p_{\text{start}})}{p_{start}}, && (2.3)\\
\textit{Unsigned Percentage Execution Cost} \quad R \quad &\equiv \quad \frac{(p_{vwap} - p_{\text{start}})}{p_{start}}, && (2.4)\\
\textit{Daily Volume Fraction} \quad &\equiv \quad \frac{Q}{V_D}, && (2.5)\\
\textit{Average Participation Rate} \quad \pi \quad &\equiv \quad \frac{Q}{V_P}, && (2.6)\\
\textit{Daily Volatility} \quad &\equiv \quad \sigma_D && (2.7)
\end{aligned}
$$

Different estimators may be used for the daily volatility σ_D:

- the normalised daily range:

$$\sigma_D = \frac{p_h - p_l}{\frac{1}{2}(p_h + p_l)};$$

- the GARCH(1,1) estimator (Bollerslev, 1986) based on daily returns;
- a realised volatility estimator based on 5-minute intraday returns (Andersen et al., 2003).

2.2.2 Other data sources

To match transaction data with data on market prices and order flow, we supplement the ANcerno data with intraday order book data from LOBSTER, [5] which provides detailed tick by tick limit order book data for NASDAQ stocks. LOBSTER contains all limit order submissions, cancellations and executions on each trading day and all events are time-stamped

[5] https://lobsterdata.com.

with millisecond precision (Bibinger et al., 2017), reconstructed from NASDAQ's historical TotalView-ITCH data (Huang and Polak, 2011). The LOBSTER *message* files contain the order type, order ID, size and the price direction (buy/sell). The LOBSTER *orderbook* files contain the bid and ask prices and quantities. In our study, we use "Level I" data from LOBSTER; these are all order-book events at the best bid and ask prices (Huang and Polak, 2011).

From the LOBSTER database, we compute the *order flow imbalance* (OFI) for the time interval $[t - \Delta t, t]$ as the volume posted at the best bid V_t^b, minus the volume posted at the best ask V_k^a during that time interval, divided by the total volume posted at the best prices $(V_t^b + V_t^a)$ (Cont et al., 2014):

$$OFI_t \equiv \frac{V_t^b - V_t^a}{V_t^b + V_t^a}, \tag{2.8}$$

where:

$$V_t^b \equiv L_t^b - C_t^b - M_t^b, \qquad V_t^a \equiv L_t^a - C_t^a - M_t^a. \tag{2.9}$$

L_t^b and C_t^b denote the total volume of limit buy orders arrived to and cancelled from the current best bid during the time interval $[t - \Delta t, t]$, respectively. M_t^b denotes the total size of the marketable sell orders that arrived to the current best bid. Similarly, for the ask side we have the analogous quantities L_t^a, C_t^a, M_t^a.

The LOBSTER data can also be used to compute intraday indicators of volatility (Andersen et al., 2003).

2.2.3 Descriptive statistics

After applying the filters mentioned in section 2.2.1, we end up with 410,792 trades over 1,764 different stocks, corresponding to 10,026,055 transactions. Figure 2.3 shows the histograms of trade duration in minutes (left panel), (log) daily volume fraction in percentage (central panel), and (log) participation rate in percentage (right panel) for all trades in our dataset. These trades have 389 different duration values, with an average trade duration of 118.94 minutes and standard deviation of 126.69. We can see from the left panel of Figure 2.3 that most trades have either short durations or durations corresponding to one trading day. There are 16,710 trades (4.07%) with a duration of 2 minutes, 13,236 trades (3.22%) with a duration of 3 minutes, 10,620 trades (2.59%) with a duration of 4 minutes, 9,519 trades (2.32%) with a duration of 5 minutes, and 10,018 trades (2.44%) with a duration corresponding to a full trading day (389-390 minutes).

From the central panel of Figure 2.3 we can see that although many trades are small in size, there is a considerable number of trades that exceed 5% of daily volume. In particular, the average daily volume fraction is 3.27%, with a standard deviation of 5.78 percentage points, a minimum of $3.5^{-6}\%$ and a maximum of 88.82%. The number of trades with a daily volume fraction equal to or greater than 5% is 78,036 (19.00%); those with a daily volume fraction equal to or greater than 20% is 10,634 (2.59%), and those with a daily volume fraction equal to or greater than 50% is 364 (0.09%).

The right panel of Figure 2.3, shows a similar story for the participation rate. The average participation rate is 12.39%, with a standard deviation of 14.70 percentage points, a minimum of $3.8^{-5}\%$ and a maximum of 98.90%. The number of trades with a participation rate equal to or greater than 5% is 242,855 (59.12%); those with a participation rate

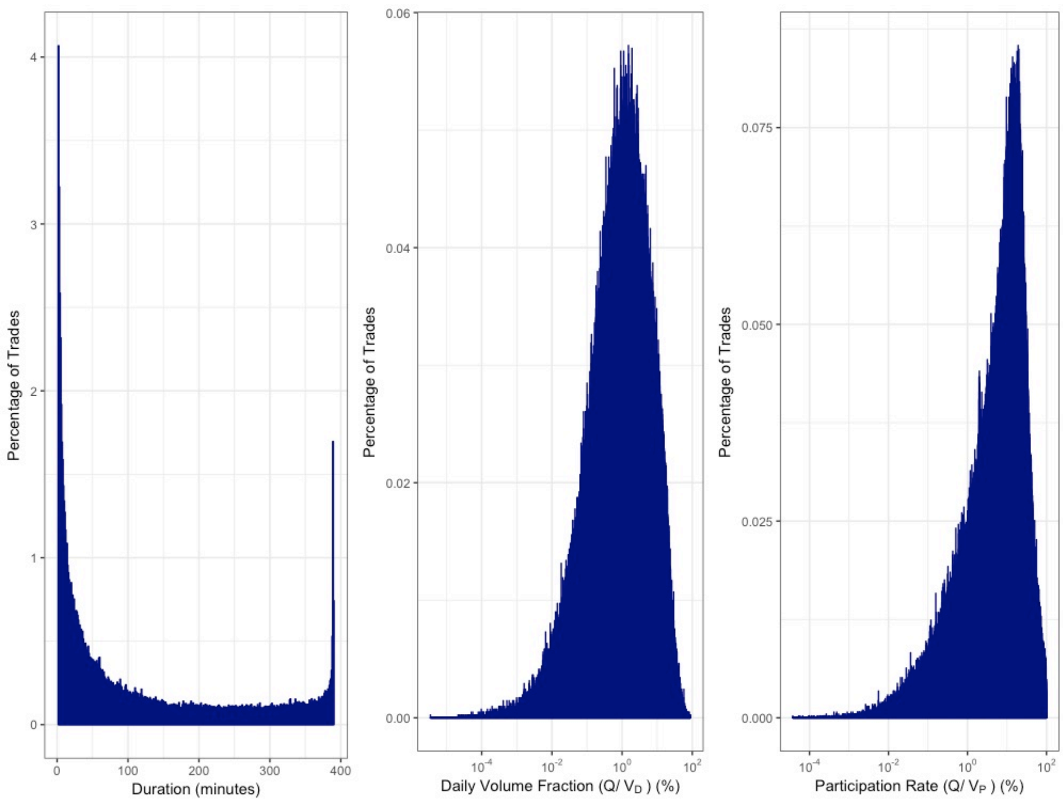

Figure 2.3: Histograms of trade duration in minutes (left), (log) daily volume fraction (Q/V_D) in percentage (centre), and (log) participation rate (Q/V_P) in percentage (right), for all trades in ANcerno (2008-2012).

equal to or greater than 20% is 85,506 (20.81%), and those with a participation rate equal to or greater than 50% is 12,998 (3.16%). We will devote special attention to large trades, whose participation rates fall in the 4th quartile range, with participation rates of 17.6% or more.

2.3 Direction of price variations during trade execution

2.3.1 Do buy trades increase the market price?

2.3.1.1 Why averaging is not relevant for institutional investors

One of the key assumptions in market impact modelling is that, on average, price changes during execution are in the same direction as the trade: buy trades are assumed to push the price up and sell trades to push the price down. That a buy (sell) trade should push the price up (down) *on average* seems intuitive, but what many studies fail to report in detail is the level of noise around such average estimates. Most of the studies mentioned in section 2.1.3 attempt to 'average out' this effect by averaging price variations over a large number of trades with sizes in a given range (see e.g. Zarinelli et al. (2015)). Almgren et al. (2005), Bershova and Rakhlin (2013) and Bacry et al. (2015) are among the few studies which attempt to quantify the contribution to execution costs of this 'noise' term –which is none other than market volatility– and avoiding taking averages.

The averaging argument only holds in very specific circumstances. If we think of the price change contemporaneous to a trade as being composed of two terms, a *systematic* term due to one's own execution and a random term due to market volatility, then averaging across trades only makes sense when analysing a large number of consecutive executions over non-overlapping or 'independent' market scenarios, where trades are placed independently of the market order flow. In this situation it may be argued that the volatility term may average to zero, leaving the systematic 'impact' as the main contribution to the 'average impact'. However, in reality many market participants execute trades based on market signals, and their order flow is often correlated with the market order flow. As a result, the two components of execution costs are not independent in general, so the contribution of market volatility may fail to average to zero.

This is even more the case for institutional market participants, such as those that make up the ANcerno database. Most of these market participants execute infrequent but large trades – only a few trades a day – so the notion of 'average impact' is not really relevant or meaningful for such market participants. The average execution cost over just a few trades would tell the investor very little about the potential loss and about the worst case scenario. Indeed, as observed in Almgren et al. (2005), in absence of such averaging, volatility is the main contribution to execution costs for such market participants. It is therefore necessary to understand how execution costs behave when this term is not averaged to zero, and how this term behaves with respect to other trade characteristics such as size, participation rate, and speed.

Figures 2.4-2.5 show *average* execution costs as a function of trade size, as they are usually presented in the square-root law literature (Zarinelli et al., 2015): one sorts the data into 50 *equally populated* bins $[q_k, q_{k+1})$ for the daily volume fraction Q/V_D, the k-th bin corresponding to observations with $Q/V_D \in [q_k, q_{k+1})$. The conditional expectation of the price change given the trade size Q/V_D is estimated by averaging over each bin.

Figure 2.4 shows the average *absolute* execution cost, normalised by daily volatility, estimated as:

$$\widehat{E}\Big(\Big|\frac{C}{\sigma_D}\Big| \quad \Big|\frac{Q}{V_D} \in [q_k, q_{k+1})\Big) = \frac{1}{N}\sum_{j=1}^{N}\Big|\frac{C_j}{\sigma_{D,j}}\Big| 1_{[q_k,q_{k+1})}\Big(\frac{Q}{V_D}\Big), \tag{2.10}$$

and Figure 2.5 shows the average *signed* execution cost, normalised by daily volatility, esti-

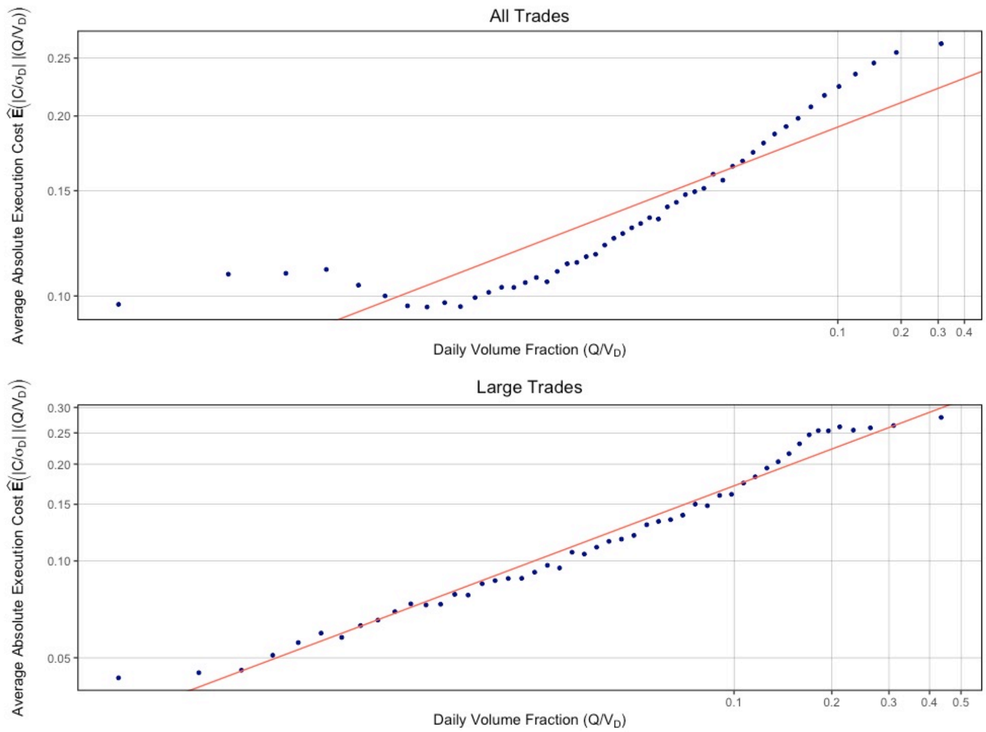

Figure 2.4: *Average* absolute execution cost standardised by daily volatility versus *average* daily volume fraction $\overline{q_k}$ for all trades (upper panel) and large trades only (lower panel) (2008-2012). Average values are computed across 50 *equally-populated* bins for the daily volume fraction. The logarithmic regression (2.12) yields $\hat{\beta}_{s,k} = 0.136$ for all trades ($R^2 = 0.79$) and $\hat{\beta}_{s,k} = 0.380$ for large trades ($R^2 = 0.98$).

mated as:

$$\widehat{E}\Big(\frac{C}{\sigma_D}\quad\Big|\frac{Q}{V_D} \in [q_k, q_{k+1})\Big) = \frac{1}{N}\sum_{j=1}^{N} \frac{C_j}{\sigma_{D,j}} 1_{[q_k,q_{k+1})}\Big(\frac{Q}{V_D}\Big), \tag{2.11}$$

where $\widehat{E}$ denotes the empirical average over the data set. The upper panels in Figures 2.4-2.5 correspond to results for all trades and the lower panels correspond to results restricted to the largest trades (upper quartile for Q/V_D). Denoting by $\overline{|c_k|}$ and by $\overline{c_k}$ the average absolute and signed execution cost standardised by daily volatility in the bin $[q_k, q_{k+1})$ corresponding to observations with $Q/V_D \in [q_k, q_{k+1})$, respectively, and by $\overline{q_k}$ the average daily volume

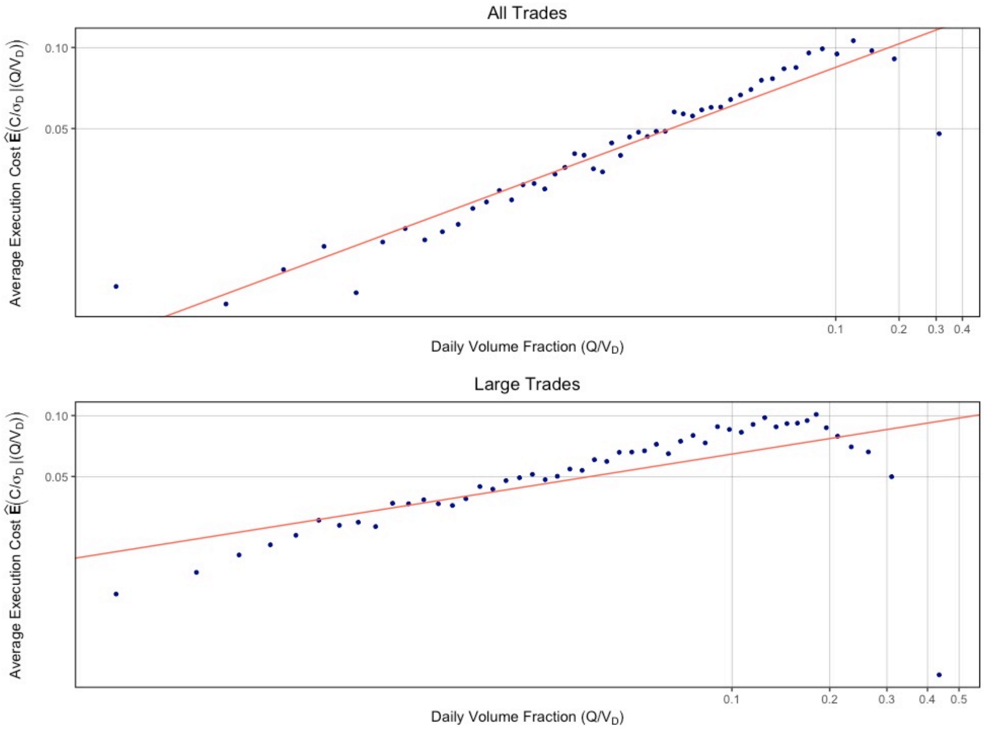

Figure 2.5: *Average* execution cost standardised by daily volatility versus *average* daily volume fraction $\overline{q_k}$ for all trades (upper panel) and large trades only (lower panel) (2008-2012). Average values are computed across 50 equally-populated bins for the daily volume fraction. The logarithmic regression (2.13) yields $\hat{\beta}_{s,k} = 0.290$ for all trades ($R^2 = 0.91$) and $\hat{\beta}_{s,k} = 0.254$ for large trades ($R^2 = 0.38$).

fraction in the bin $[q_k, q_{k+1})$, we run the following logarithmic regressions:

$$\log \overline{|c_k|} = \alpha_k + \beta_{s,k} \log \overline{q_k} + \xi_k, \tag{2.12}$$

$$\log \overline{c_k} = \alpha_k + \beta_{s,k} \log \overline{q_k} + \xi_k, \tag{2.13}$$

where $\overline{q_k}$ is the average daily volume fraction in the bin $[q_k, q_{k+1})$, we find that the square-root model is not satisfied in our dataset.

For all trades the estimated regression coefficient $\hat{\beta}_{s,k}$ is only 0.136 when looking at the absolute value of execution costs and only 0.290 when looking at signed execution costs, which is very far from the 0.5 indicating a square-root relationship. For large trades, we find an estimated regression coefficient $\hat{\beta}_{s,n}$ of 0.380 when looking at the absolute value of execution costs and only 0.254 when looking at the signed execution costs. Further, it is clear from the upper panels of Figures 2.4-2.5, that both small and very large trades deviate from the square-root law. In particular, the lower panel of Figure 2.4 shows that the average absolute execution cost is essentially the same for all trades with a daily volume fraction of 18.5% and above, while the lower panel of Figure 2.5 shows that the average signed execution cost is actually declining for these. This strongly contradicts the idea that the size and participation rate are the main drivers of execution costs.

In Figure 2.6, we show the histograms of the duration values and (log) participation rate for these trades. It is clear from the left panel of Figure 2.6 that these are very large trades which are mostly executed over the duration of the whole trading day. It is then not surprising to see these that the absolute execution cost associated with these VWAP trades is the same regardless of the quantity traded, because this is nothing else than the absolute daily return.

A final point worth mentioning in relation to *average* execution costs, is that the estimation of the power-law exponent is very sensitive to the way the data is handled. In the square-root literature, the data set is usually divided into *equally-populated* bins for the daily volume fraction Q/V_D and then the average execution costs for each bin is estimated as in (2.11) (Zarinelli et al., 2015). However, it is clear that in so doing, the interval width for Q/V_D increases almost exponentially. Most of the trades in ANcerno have small daily vol-

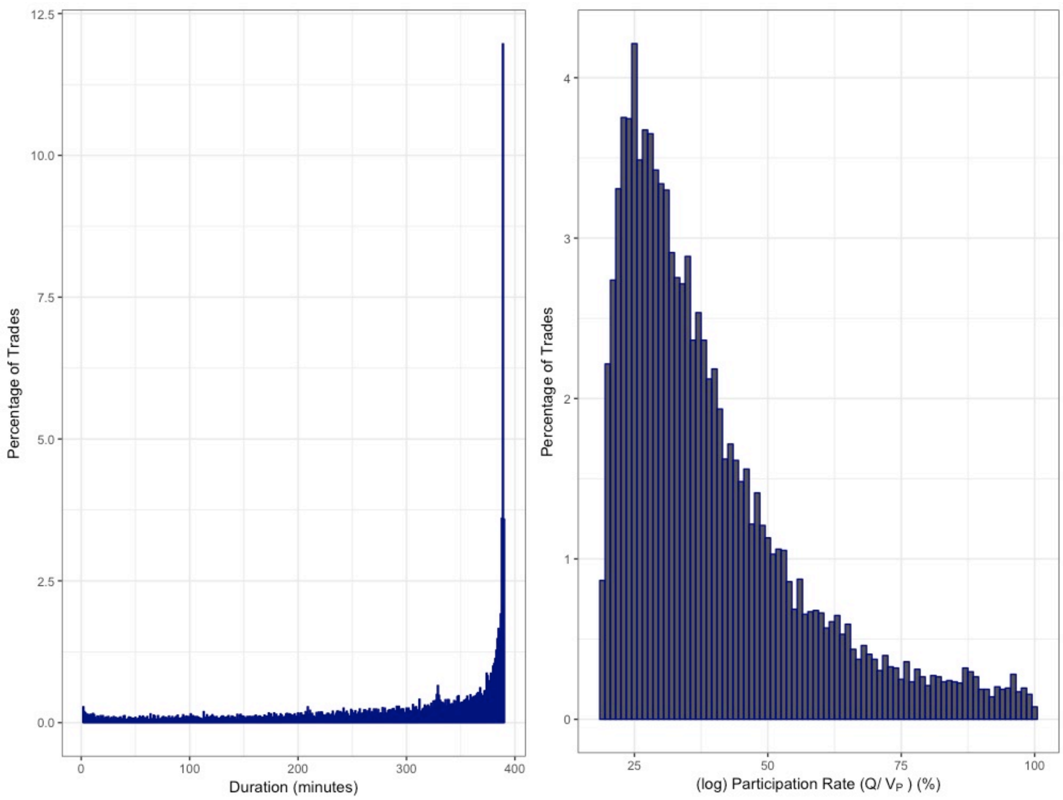

Figure 2.6: Histogram of trade duration values (left) and (log) participation rate for all trades with a daily volume fraction Q/V_D above 18.5% (2008-2012).

ume fractions, and relatively few trades have large daily volume fractions. Thus, if one divides the data into *equally-populated* bins, most of the bins will capture the small trades, while few, very wide bins, will capture the large trades. Further, as the interval width increases, each bin will be increasingly dominated by the smaller trades within each bin.

In other disciplines, where power-law exponents arise naturally, these exponents are estimated by averaging the data over *equally-spaced* intervals (Pickering et al., 1995; White et al., 2008).[6] Figure 2.7 shows what happens if we divide the data set into 50 *equally-spaced* in-

[6]An alternative method is to divide the data into bins having constant *logarithmic* width. This means that

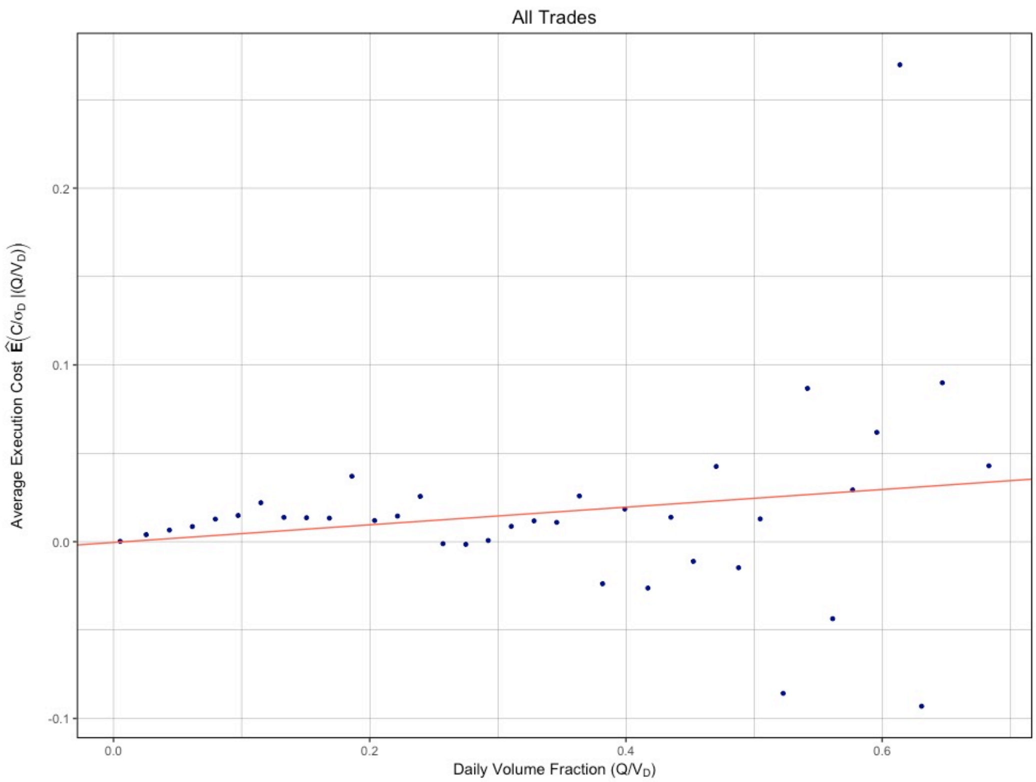

Figure 2.7: *Average* execution cost standardised by daily volatility $\overline{c_k}$ versus *average* daily volume fraction $\overline{q_k}$ for all trades (2008-2012). Average values are computed across 50 *equally-spaced* bins for the daily volume fraction, and bins with fewer than 50 trades are discarded. The estimated regression coefficient $\hat{\beta}_{s,k}$, from a linear regression of the average cost on the average daily volume fraction is 0.05 ($R^2 = 0.03$).

tervals for the daily volume fraction, and take the average execution cost as in (2.11), with $N \geq 50$. The picture is completely different from Figures 2.4-2.5, and we can see how there is no clear relationship between the average execution cost and the daily volume fraction ($R^2 = 3.26\%$). Indeed, the estimated regression coefficient $\hat{\beta}_{s,k}$ is only 0.05, and for large trades, the average execution cost is arbitrarily large and positive or large and negative. This shows that the way the exponent is estimated in the square-root or any power law model is

the data is spread more evenly among the intervals (Pickering et al., 1995).

extremely dependent on the way the data is averaged.

2.3.1.2 Price changes and trade direction of individual trades

For the reasons mentioned above, we focus in this study on all trades in our dataset, rather than their average values. We start by looking at the *sign* of price changes during execution and by computing the conditional probabilities that these price changes are in the same direction of trades. Then, we look at the standard square-root law model of market impact, and test its explanatory power on the observed data (again without any averaging). Finally, we analyse and quantify the noise term surrounding the square-root model estimates and draw conclusions about the model performance, the notion of 'price impact of a trade' and the main determinants of execution costs. Figure 2.8 presents a scatter-plot of the *unsigned* execution costs R (standardised by their daily volatility) against the *signed* daily volume fraction (positive for buy trades and negative for sell trades). If the assumption that prices moved in the same direction of trades were true, we would only see points in the quadrants where $R > 0, Q > 0$ and $R < 0, Q < 0$ and no points in the quadrants defined by $R > 0, Q < 0$ and $R < 0, Q > 0$. The scatter-plot clearly shows that this is not the case. Points are distributed across all four quadrants. In particular, the joint probabilities for each quadrant are as follows:

$$\mathcal{P}(R > 0, Q > 0) = 33.0\%, \tag{2.14}$$

$$\mathcal{P}(R < 0, Q > 0) = 19.0\%, \tag{2.15}$$

$$\mathcal{P}(R > 0, Q < 0) = 17.7\%, \tag{2.16}$$

$$\mathcal{P}(R < 0, Q < 0) = 30.3\%. \tag{2.17}$$

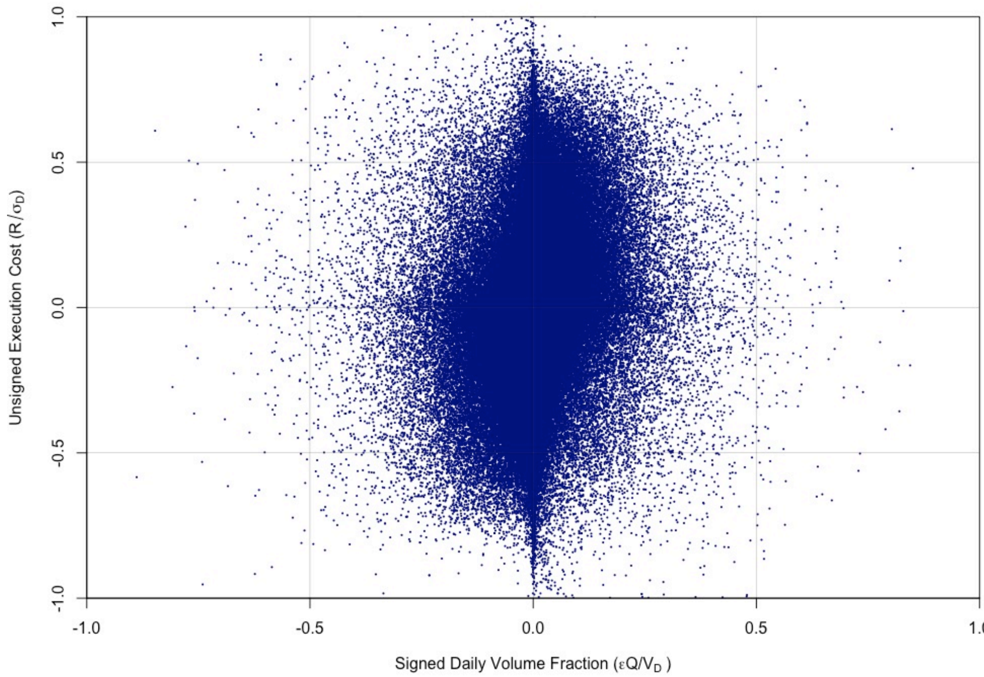

Figure 2.8: Scatter-plot of the *unsigned* execution cost standardised by the daily volatility R/σ_D versus *signed* daily volume fraction $\varepsilon Q/V_D$ for all trades in our dataset (2008-2012).

The conditional probabilities for a buy trade are:

$$\mathcal{P}(R > 0|Q > 0) = 63.5\%, \qquad \mathcal{P}(R < 0|Q > 0) = 36.5\%, \tag{2.18}$$

and for a sell trade:

$$\mathcal{P}(R < 0|Q < 0) = 63.1\%, \qquad \mathcal{P}(R > 0|Q < 0) = 36.9\%. \tag{2.19}$$

The conditional probabilities of equations (2.18) and (2.19) clearly show that the primary assumption of traditional impact models does not hold in our dataset. The sign of price changes does not seem to be determined by the direction of the individual trade. In fact, we observe a *negative impact* in nearly 40% of the cases. This finding brings into question the received notion of 'price impact of a trade', that is, the idea that the price change contemporaneous to a trade is driven by that trade. [7]

In market impact models that assume prices change with the direction of the trade, only noise can explain why prices move in the opposite direction of the trade. For instance, according to the square-root model of market impact, R is a square-root function of the daily volume fraction:

$$R = Y\left(\varepsilon\sigma_D\sqrt{\frac{Q}{V_D}}\right) + \xi \tag{2.20}$$

where $\varepsilon = 1$ for buy trades and $\varepsilon = -1$ for sell trades, and Y is a constant. A regression yields the estimate $\hat{Y} = 6.54$ with an estimated standard error of 0.002, a t-statistic of 180 and significance at the 1% level. However, the coefficient of determination is only $R^2 = 5.5\%$, which shows that the square-root model of market impact provides a poor explanation for the direction and magnitude of the unsigned execution costs. 94.5% of the variation across trades is given by the residual, whose standard deviation $\sigma_\xi \sim 0.01$ is of the same order as the daily volatility of the stock return, showing that volatility is by far the main determinant of the direction and magnitude of price variations during trade execution.

We can also use the square-root model to compute the conditional probabilities of points falling in the 'wrong' quadrants (where price change and trade direction have opposite signs)

[7] Similar results, where execution costs are computed using the market VWAP as benchmark, are presented in Appendix A.1.1.

by using models for binary choice. Assuming a certain distribution $F_{\sigma_\xi}(x)$ for the residual in equation (2.20), we can compute the probability of a positive price change given a buy trade ($Q > 0$) from (2.20) as:

$$
\begin{aligned}
\mathcal{P}\Big(R > 0 \Big| Q > 0\Big) &= \mathcal{P}\Big(Y\sigma_D\sqrt{\frac{Q}{V_D}} + \xi \geq 0\Big), \\
&= \mathcal{P}\Big(\xi \geq -Y\sigma_D\sqrt{\frac{Q}{V_D}}\Big), \\
&= 1 - F_{\sigma_\xi}\Big(- Y\sigma_D\sqrt{\frac{Q}{V_D}}\Big), \\
&= F_{\sigma_\xi}\Big(Y\sigma_D\sqrt{\frac{Q}{V_D}}\Big), \qquad (2.21)
\end{aligned}
$$

and the probability of a negative price change given a buy trade as:

$$
\mathcal{P}\Big(R < 0 \Big| Q > 0\Big) = 1 - F_{\sigma_\xi}\Big(Y\sigma_D\sqrt{\frac{Q}{V_D}}\Big). \qquad (2.22)
$$

We can do the same for sell trades ($Q < 0$) and compare these probabilities with the empirical probabilities from our dataset. In this process, we neglect the probability of a price change exactly equal to zero because this occurs in only 0.27% of our observations.

Table 2.2 shows the empirical conditional probabilities for the sign of a price change (positive/negative) given the trade direction (buy/sell) and compares them with the probabilities of the square-root model (2.20) computed as in (2.21)-(2.22), where the error term ξ is assumed to follow either a logistic (logit model) or normal (probit model) distribution. The conditional probability of a negative price change from a buy trade in our data is 36.5%, while the estimated conditional probability from the logit model is only 6.1%, with a 99%

Sell ($Q < 0$)	Buy ($Q > 0$)	
Empirical: 36.9%	Empirical: 63.5%	**Positive change**
Logit: 6.3%	Logit: 93.9%	($R > 0$)
Probit: 3.2%	Probit: 97.1%	
Empirical: 63.1%	Empirical: 36.5%	**Negative change**
Logit: 93.7%	Logit: 6.1%	($R < 0$)
Probit: 96.8%	Probit: 2.9%	

Table 2.2: Conditional probabilities for the sign of a price change (positive/negative) *given* trade direction (buy/sell). The Table reports the empirical probabilities from our dataset versus the probabilities implied by the square-root model of equation (2.20) under the assumption of a logistic (logit) and normal (probit) distribution for the error term ξ.

confidence interval of $[5.7\% - 6.5\%]$, and from the probit model is only 2.9%, with a 99% confidence interval of $[2.5\% - 3.3\%]$. The empirical probability of a negative price change given a buy trade is therefore 152 standard errors away from what is predicted by the square-root model assuming a logistic distribution for the error term, and 168 standard errors when the error term follows a normal distribution. Similarly, the conditional probability of a positive price change from a sell trade is 36.9%, while the estimated conditional probability from the logit model is just 6.3%, with a 99% confidence interval of $[5.8\% - 6.8\%]$, and from the probit model is only 3.2%, with a 99% confidence interval of $[2.7\% - 3.7\%]$. The probability of a positive change given a sell trade is therefore 153 standard errors away from what is predicted by the square-root model assuming a logistic distribution for the error term, and 168 standard errors when the error term follows a normal distribution. These findings further support the view that the square-root model is inconsistent with the observed data and that the direction of price moves is not determined by the buy/sell nature of individual trades.

2.3.2 Price changes and order flow imbalance

The results in subsection 2.3.1 show that the sign of the price change does not seem to be determined by the direction of the individual trade. Indeed, nearly 40% of trades have a *negative impact*, and therefore the claim that individual trades *cause* prices to move in the same direction is highly tenable. This causality claim comes from the fact that in standard 'market impact' models individual trade characteristics *cause* price movements without any regard for the *aggregate* contribution to demand and supply from the rest of the market (see Figure 2.1).

In contrast, we start from the premise that price movements are determined by *aggregate* supply and demand and thus, in markets where orders are anonymised, the order flow of a market participant only influences price movements through the aggregate order flow (see Figure 2.2). This claim is testable, because it implies that *conditional* on the aggregate order flow imbalance, the size of the individual trade has no significant explanatory power on execution costs.

We evaluate the relationship between trade size and price changes *conditional* on the aggregate order flow imbalance on a subset of our ANcerno sample. In particular, we select all trades in 77 stocks that overlap the Nasdaq 100 index. This is because we can only compute *OFI* for Nasdaq stocks, for which the LOBSTER database records every single order book event at, at least, millisecond precision (Bibinger et al., 2017).

We sort the trades into 50 *equally-populated* bins for the *OFI*. We then run, for each bin k, the following conditional regression of the unsigned execution cost R, standardised by the

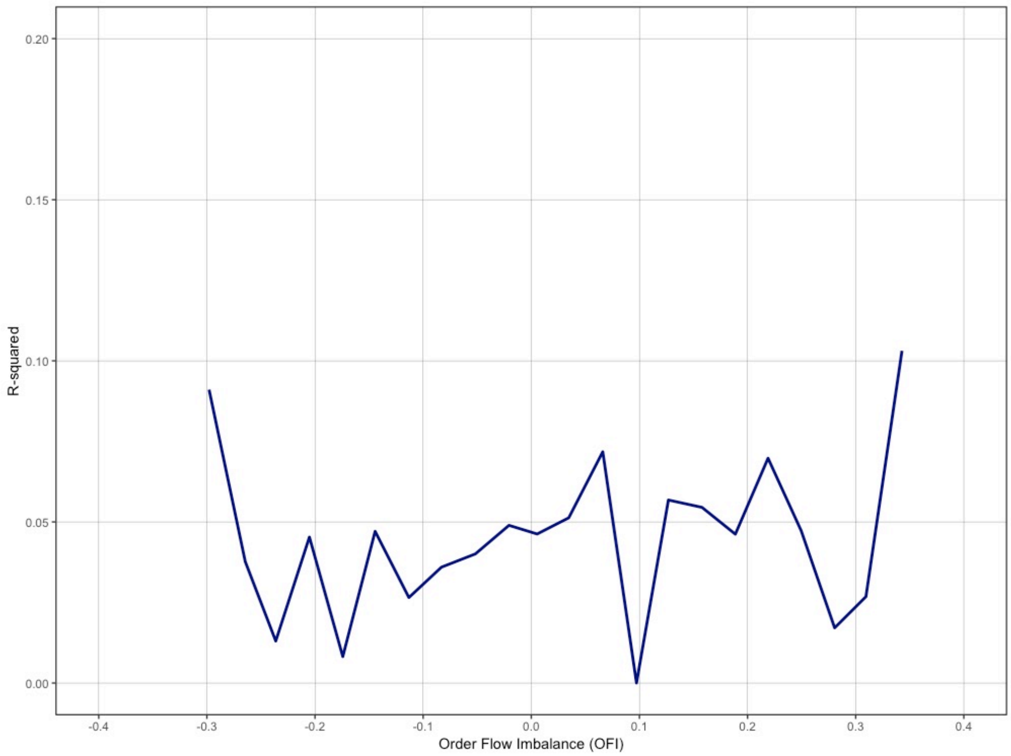

Figure 2.9: R-squared for the conditional regressions of the unsigned execution cost, standardised by daily volatility, on the signed daily volume fraction, *given* the *OFI* (2.23).

daily volatility, on the daily volume fraction Q/V_D, given the OFI:

$$\left(\frac{R}{\sigma_D}\right)_k = \alpha_k + \beta_{s,k}\left(\varepsilon\frac{Q}{V_D}\right)_k + \xi_k. \tag{2.23}$$

Figure 2.9 shows the coefficient of determination of these regressions against the *OFI* value on which they have been conditioned.[8] We can see how with exception of large imbalances ($|OFI| > 0.30$), the R^2 is very low, with an average value of only 3.9% for $OFI \in [-0.30, 0.30]$. These results once again show that price moves during execution are not determined by the

[8] Similar results, where execution costs are computed using the market VWAP as benchmark, are presented in Appendix A.1.2.

characteristics of individual trades, and that the claim that individual trades *cause* the market price to move, and increase execution costs, might be spurious. Instead, as shown in Figure 2.2, a trade only affects the market price, and execution costs, through the *aggregate* order flow imbalance. These findings are consistent with a recent work on execution costs in the US Treasury market, which finds that the order flow imbalance is the most important feature in predicting price changes (Salem et al., 2018, exhibit 5, p.7). They are also consistent with the findings of Easley et al. (2015), which shows that the trade side and order imbalance are key variables in modelling execution costs, and with the work of Bechler and Ludkovski (2015) which proposes a dynamic model combining the Almgren-Chriss framework, the order flow and the model of Easley et al. (2015).

Overall, the results in this section invalidate the standard notion of the 'price impact of a trade' and suggest that volatility is the main determinant of execution costs. They imply that investors should focus on aggregate order flow modelling, rather than trade-induced price moves.

2.4 Execution costs and the role of trade duration

2.4.1 Where does the square-root law come from?

The previous section showed that it is the aggregate (net) order flow imbalance, rather than the individual trade, that determines the sign of price changes during execution and therefore that the idea that the 'impact' of a trade is the main determinant of execution costs needs to be rethought. So, why do empirical studies consistently find a square-root dependence of price changes on trade size? The reason is that these empirical studies, like ours, are based on datasets composed of *orderly* institutional trade executions, defined as executions in which the trade *duration* T is modulated according to the trade size Q in order to maintain a participation rate below some reasonable bound (Avellaneda and Cont, 2013). In fact, the two variables are *proportional* if trading occurs at a fixed participation rate throughout the day. If π is the average participation rate,

$$Q = \pi\, V_D T(Q). \tag{2.24}$$

When returns are uncorrelated, the amplitude of a price move over a horizon τ scales like $\sigma_D\sqrt{\tau}$. So, over the trade duration T, the typical amplitude of a price change will be:

$$\sigma_D\sqrt{T} = \sigma_D\sqrt{\frac{Q}{\pi V_D}} = Y\sigma_D\sqrt{\frac{Q}{V_D}}, \tag{2.25}$$

where $Y = 1/\sqrt{\pi}$. In other words, the 'square-root law' of market impact turns out to be nothing more than the (well known) square-root scaling of volatility with the execution horizon (Bachelier, 1900), which explains the ubiquity of this 'law'.

This explanation is consistent with our findings in section 2.3.1. If we look at the average over all trades, the constant Y, should be close to the average of $1/\sqrt{\pi}$ over the whole sample. The regression in (2.20) gave us an estimated value for Y of 6.54, implying an average participation rate over the whole sample of 2.33% which is quite close to the value of 1.51% implied by the dataset.

Our explanation is also consistent with previous findings on the influence of trade duration on execution costs. Engle et al. (2012) and Easley et al. (2015) emphasised the temporal dimension of transaction cost analysis and the importance of the execution horizon in the analysis of execution costs. Bacry et al. (2015) present a market impact model where the return is a function of the daily volume fraction and duration (Bacry et al., 2015, eqn. 4, and Table 2) with power-law dependence with exponents 0.54-0.59 for the daily volume fraction and between -0.35 and -0.23 for the duration but do not perform any analysis of impact conditional on duration. Bershova and Rakhlin (2013) analyse trades executed by Alliance Bernstein's trading desk and run a non-linear regression of execution costs on trade duration for *all* observations, without averaging. They find execution costs are a square-root function of *duration* and report a coefficient of determination of 29% (Bershova and Rakhlin, 2013, eqn. 7, Table 5 and Fig. 4). Finally, Salem et al. (2018) examine a dataset of trade executions in the US Treasury market and report that execution costs increase with duration (Salem et al., 2018, p. 9).

2.4.2 Execution costs and trade size conditional on trade duration

The explanation for the 'square-root law' findings in the literature provided in (2.25) is empirically testable. If the square-root law arises from the square-root scaling of volatility with trade duration, we should find that execution costs do not depend on trade size when conditioned on duration. If we look at all trades, rather than just the averages, and condition on trade duration, we find that the relationship between execution costs and trade size indeed disappears. This is clear from Figure 2.10, which shows four scatter-plots of (log) absolute execution costs $|C/\sigma_D|$ versus (log) daily volume fraction Q/V_D conditioned on representative short- (2 minutes), medium- (5 minutes), and high-duration (389 minutes) values. The cloud of points in each of these four scatter-plots is essentially flat, showing no relationship between execution costs and trade size. Similarly, Figure 2.11, which shows the scatter-plot of signed execution costs C/σ_D versus (log) daily volume fraction Q/V_D conditioned on trade

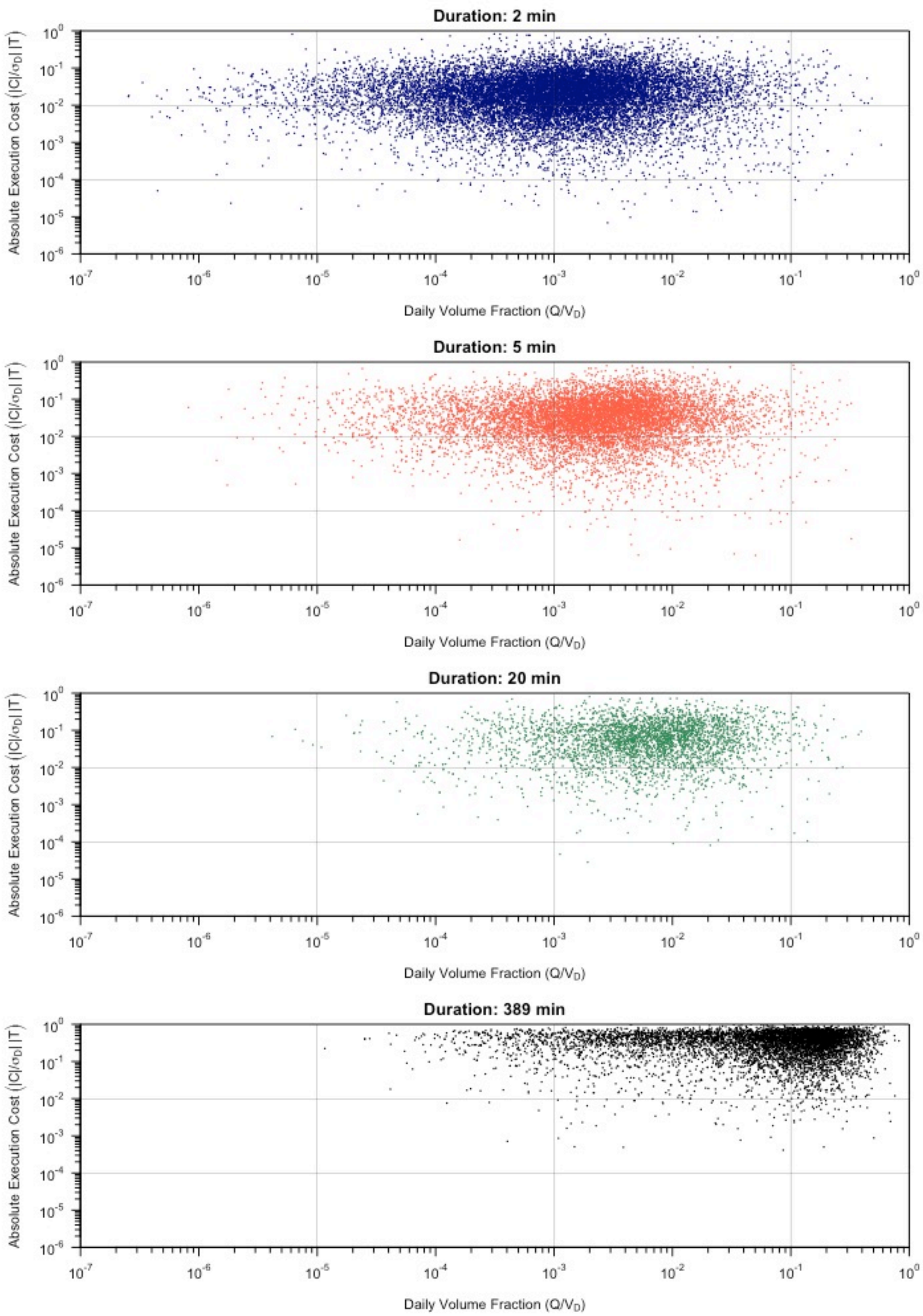

Figure 2.10: Scatter-plots of (log) absolute execution cost $|C/\sigma_D|$ versus (log) daily volume fraction Q/V_D conditional on four representative trade duration values: $T = 2, 5, 20$ and 389 minutes (2008-2012).

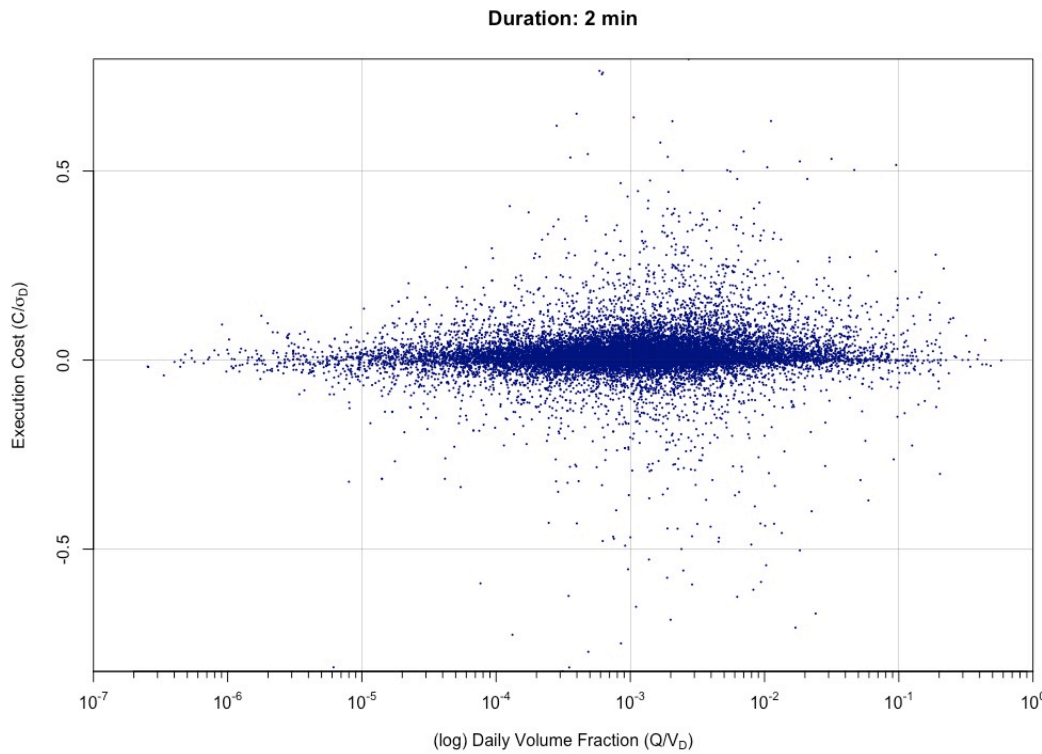

Figure 2.11: Scatter-plot of execution cost C/σ_D versus (log) daily volume fraction Q/V_D conditional on trade duration $T = 2$ minutes (2008-2012).

duration $T = 2$ minutes, reveals no relationship between execution costs and trade size. To test this formally, we sort the data into 389 bins for the trade duration T, where T is expressed in minutes.

The conditional logarithmic regression of the absolute execution cost $|C|$, standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T, is estimated, for each bin k as:

$$\log\left(\left|\frac{C}{\sigma_D}\right|\right)_k = \alpha_k + \beta_{s,k} \log\left(\frac{Q}{V_D}\right)_k + \xi_k, \qquad (2.26)$$

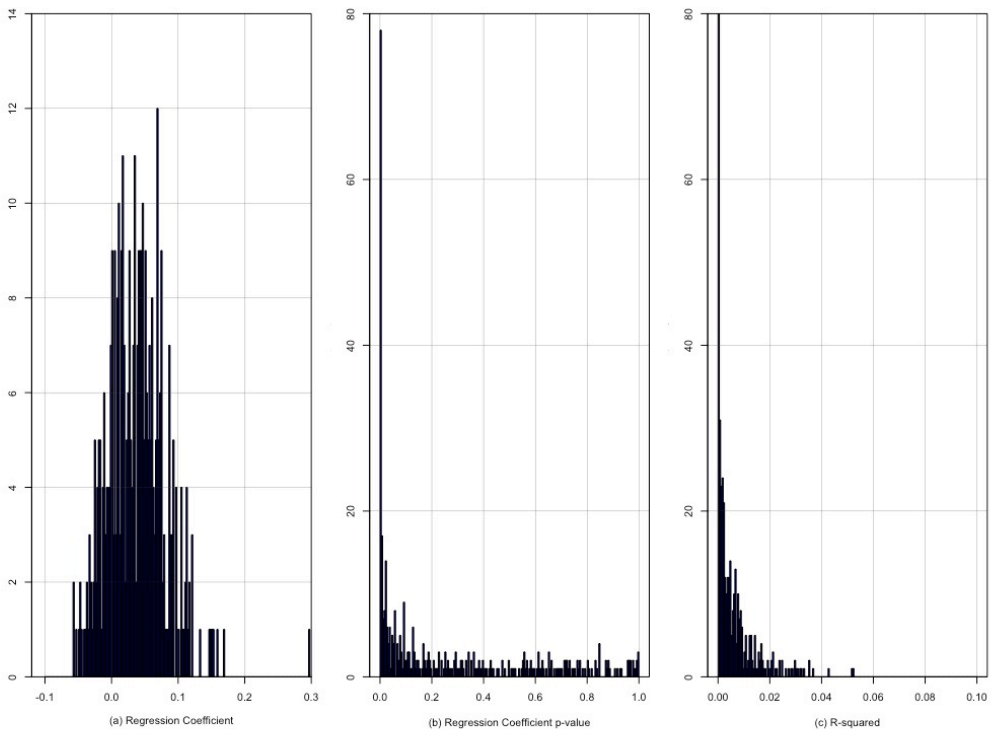

Figure 2.12: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional regressions of absolute execution costs, standardised by the daily volatility, on the daily volume fraction, given trade duration (2.26).

and the conditional non-linear least square (NLS) regression for the execution cost C, standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T, is estimated as:

$$\left(\frac{C}{\sigma_D}\right)_k = Y_k \left(\frac{Q}{V_D}\right)_k^{\beta_{s,k}} + \xi_k. \tag{2.27}$$

Figures 2.12-2.13 show the histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$, their p-values and the coefficients of determination for the log regressions (2.26) and the NLS regressions (2.27), respectively. When we look at the absolute value of execution costs, $\hat{\beta}_{s,k}$

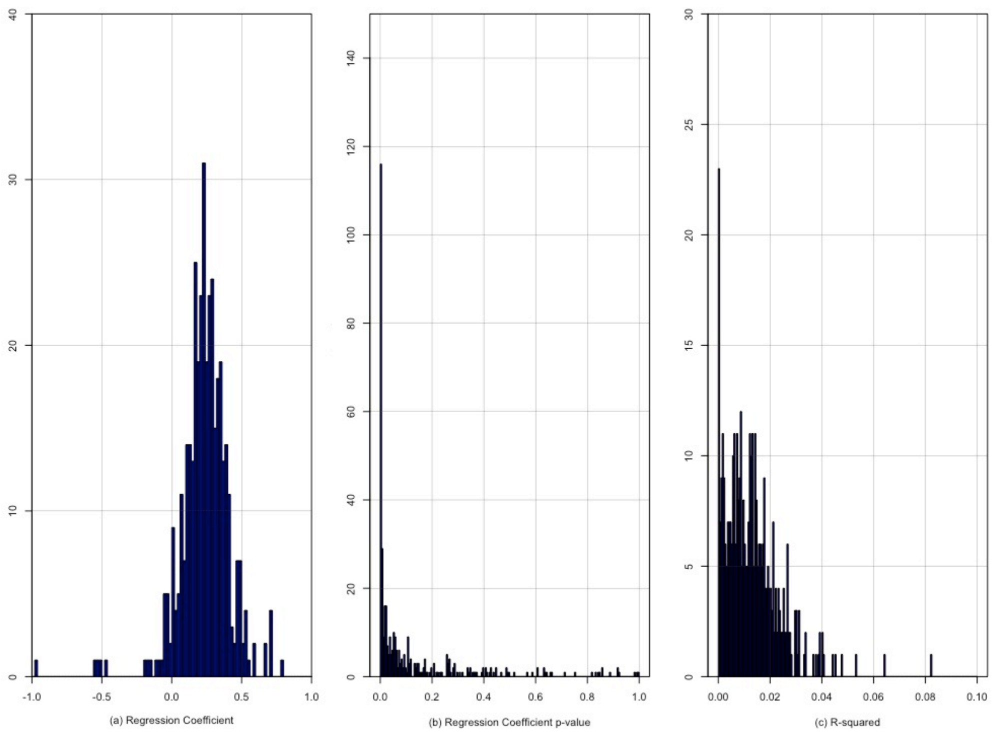

Figure 2.13: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional non-linear least square regressions of signed execution costs, standardised by the daily volatility, on the daily volume fraction, given trade duration (2.27).

is statistically significant at the 5% level in 146 of the 389 regressions (37.5%) (Figure 2.12, panel (b)). However, in the vast majority of cases, its estimate is very close to zero (Figure 2.12, panel (a)). On average, $\hat{\beta}_{s,k}$ is 0.04, with an average coefficient of determination of only 0.7%, and 358 out of 389 (92%) regressions having an R^2 smaller than 2% (Figure 2.12, panel (c)). In other words, once conditioned on duration, trade size has nearly zero explanatory power on the absolute value of execution costs.

The analysis of signed execution costs paints a similar picture. $\hat{\beta}_{s,k}$ is statistically significant at the 5% level in 218 of the 389 regressions (56.0%) (Figure 2.13, panel (b)), but the value is, once again, very close to zero in most cases, and in any case far from 0.5 (Figure 2.13, panel (a)). On average, $\hat{\beta}_{s,k}$ is 0.23, with an average coefficient of determination of only 1.2%,[9] and 314 out of 389 (81%) regressions having an R^2 smaller than 2% (Figure 2.13, panel (c)).

These results strongly support the explanation provided above for the 'square root law of market impact' as arising from the square-root scaling of volatility with trade duration (Bachelier, 1900). The implication is that the two components of execution costs –'market impact' and 'volatility risk'– are in fact two sides of the same coin: market volatility.

2.4.3 Joint analysis of execution costs, trade size and trade duration

A step-wise multivariate log-regression of the absolute value of execution costs on trade size, duration and volatility further supports our results. We proceed as follows. Firstly, we perform a log-log regression of the absolute value of execution costs on the daily volume fraction Q/V_D. Secondly, a log-log regression of absolute value of execution costs on trade duration T. Thirdly, a regression with both explanatory variables. In each case, the daily volatility σ_D is included as an additional explanatory variable. The three regressions are as follows:

$$\log|C| = \alpha + \beta_s \log\frac{Q}{V_D} + \beta_v \log\sigma_D + \xi, \tag{2.28}$$

$$\log|C| = \alpha + \beta_d \log T + \beta_v \log\sigma_D + \xi, \tag{2.29}$$

$$\log|C| = \alpha + \beta_s \log\frac{Q}{V_D} + \beta_d \log T + \beta_v \log\sigma_D + \xi. \tag{2.30}$$

[9] The coefficient of determination is not defined for non-linear least square regressions. We compute a pseudo R^2 as follows: $R^2 = 1 - [(\frac{C}{\sigma_D} - \widehat{\frac{C}{\sigma_D}})^2/(N-1)\sigma^2_{C/\sigma_D}]$, where $\widehat{\frac{C}{\sigma_D}}$ is fitted value for the execution cost , σ^2_{C/σ_D} is the variance of the execution cost, and N is the number of observations in the bin $[t_k, t_{k+1})$.

	Dependent variable: Absolute Execution Cost $\|C\|$					
	All Trades			Large Trades		
Model	(2.28)	(2.29)	(2.30)	(2.28)	(2.29)	(2.30)
Constant α	-2.071^{***}	-2.070^{***}	-1.954^{***}	-1.675^{***}	-2.151^{***}	-2.033^{***}
	(0.013)	(0.012)	(0.014)	(0.027)	(0.026)	(0.040)
Trade size Q/V_D	0.201^{***}		0.038^{***}	0.451^{***}		0.071^{***}
	(0.001)		(0.001)	(0.003)		(0.005)
Duration T		0.437^{***}	0.414^{***}		0.464^{***}	0.412^{***}
		(0.001)	(0.001)		(0.003)	(0.005)
Volatility σ_D	0.895^{***}	0.911^{***}	0.905^{***}	0.888^{***}	0.874^{***}	0.874^{***}
	(0.003)	(0.003)	(0.003)	(0.007)	(0.006)	(0.007)
Observations	410,792	410,792	410,792	102,698	102,698	102,698
R^2	0.212	0.343	0.345	0.284	0.335	0.346
Adjusted R^2	0.212	0.343	0.345	0.284	0.335	0.336
Residual Std. Error	1.382	1.262	1.260	1.350	1.301	1.300
F Statistic	$55{,}307^{***}$	$107{,}191^{***}$	$72{,}017^{***}$	$20{,}361^{***}$	$25{,}892^{***}$	$17{,}357^{***}$
AIC	1,431,635	1,357,075	1,355,981	353,163	345,533	345,343
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2.3: Estimated logarithmic regressions of absolute execution costs on daily volume fraction, duration and daily volatility for all trades (left) and for large trades only (right) (2008-2012), equations (2.28 - 2.30). In parentheses: heteroskedasticty-consistent Newey-West standard errors.

Table 2.3 presents the results of these regressions for all trades (columns 2-4) and for large trades only (columns 5-7). In the first regression for all trades (column 2), the regression coefficient on the daily volume fraction, $\hat{\beta}_s$, is 0.201, and is statistically significant at the 1% level. This is very far from 1/2, so we do not find any evidence to support the 'square-root law' in terms of trade size. In the second regression for all trades (column 3), the regression coefficient on trade duration, $\hat{\beta}_d$, is 0.437 and is statistically significant at the 1% level. This is much closer to 1/2, and supports our previous finding that price changes during execution are a square-root function of trade *duration* rather than trade size. Also, as expected, we find a regression coefficient close to one on the daily volatility ($\hat{\beta}_v = 0.911$), and we notice that the coefficient of determination for this regression is much higher than the first regression

(34.3% versus 21.2%).

In the third regression, where the daily volume fraction and trade duration are included as regressors (column 4), the coefficient on trade duration remains close to 1/2 (specifically 0.414) while the regression coefficient on daily volume fraction drops to only 0.038. Both regression coefficients are statistically significant at the 1% level and collinearity between the two variables is rejected based on the variance inflation factor test.[10] The inclusion of the daily volume fraction as a statistically significant explanatory variable is confirmed when we compare the nested models in equations (13) and (14), with an ANOVA test (*F*- statistic: 1097.5, *p*-value: $< 2.2\,10^{-16}$) or using the Akaike Information Criterion (1,355,981 versus 1,357,075). However, despite being statistically significant, the estimated regression coefficient on the daily volume fraction is very small and it implies that a 1 percentage point increase in daily volume fraction leads to only a 0.038% increase in the absolute value of execution costs. Moreover, the increase in the adjusted coefficient of determination compared to the second regression is minimal: from 34.3% to 34.5%.

The findings are similar if we focus only on large trades. In the first regression (column 5), the regression coefficient on the daily volume fraction, $\hat{\beta}_s$, is 0.451, and is statistically significant at the 1% level. This is in line with the stylised fact of a 'square-root law' in terms of trade size reported in the literature. However, if we look at the results of the second regression (column 6), the estimated regression coefficient on trade duration, $\hat{\beta}_d$, is also close to 1/2 (0.464) and statistically significant at the 1% level, and the coefficient of determination for this regression is higher (33.5% versus 28.4%).

[10] The square root of the variance inflation factor indicates the degree to which the confidence interval for the estimated regression coefficient is expanded relative to a model with uncorrelated predictors (Fox and Monette, 1992). In this case we find values of 1.15 for both variables, suggesting that multicollinearity is not a problem.

Finally, when daily volume fraction and duration are included in the regression (column 7), the coefficient on trade duration remains close to 1/2 (specifically 0.412) while the regression coefficient on daily volume fraction drops to only 0.071. Both regression coefficients are statistically significant at the 1% level, and the inclusion of the daily volume fraction as a statistically significant explanatory variable is confirmed when we compare the nested models in equations (13) and (14), with an ANOVA test (F- statistic: 191.4, p-value: $< 2.2\ 10^{-16}$) or using the Akaike Information Criterion (345,343 versus 345,533). However, despite being statistically significant, the estimated regression coefficient on the daily volume fraction is very small and it implies that a 1 percentage point increase in daily volume fraction leads to only a 0.071% increase in the absolute value of execution costs.

For large trades, we also notice that that the sum of the estimated coefficients on duration and daily volume fraction when both variables are included is similar to the coefficient on trade duration alone, when only this variable and daily volatility are considered; that the standard error on $\hat{\beta}_d$ doubles when we include daily volume fraction; and that the increase in the adjusted coefficient of determination is small: from 33.5% to 33.6%. Suspecting a multicolinearity problem in equation (2.30), we compute the variance inflation factor, whose square root indicates the degree to which the confidence interval for the estimated regression coefficient is expanded relative to a model with uncorrelated predictors. We find a value of around 1.779 on both daily volume fraction and duration, and one on daily volatility. These results confirm that the two variables are collinear, and suggest that for large trades the dependence between the absolute value of execution costs and the daily volume fraction mainly comes from the correlation between trade size and duration, which is 0.78. This is not surprising, since we are looking at very large trades, whose participation rate is greater than 16%.

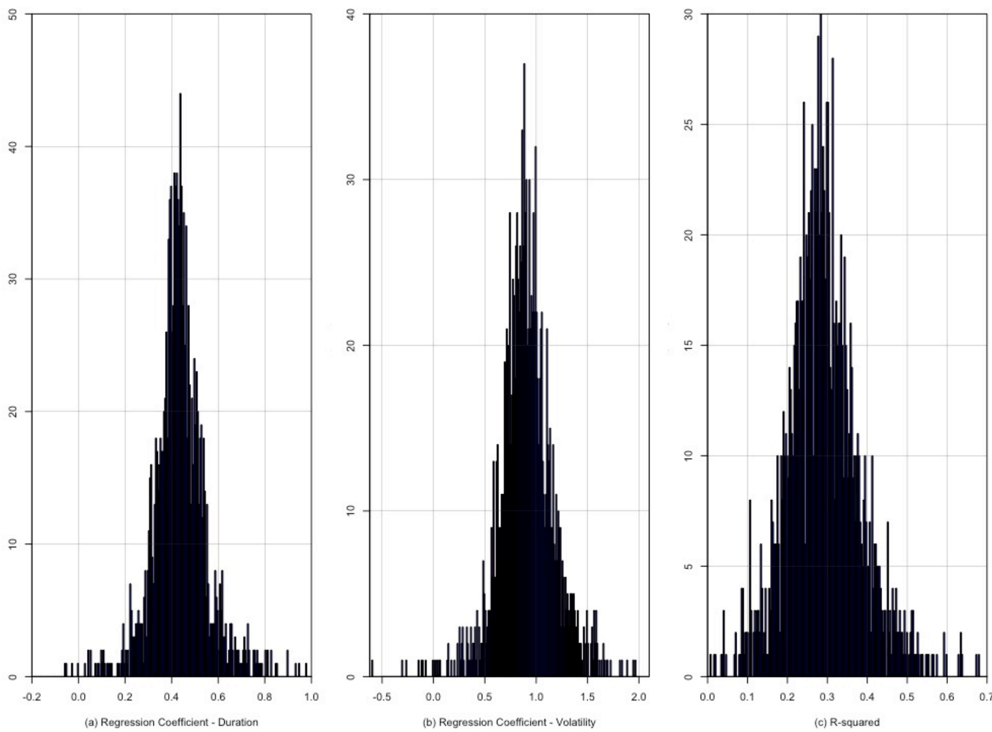

Figure 2.14: Histograms of estimated regression coefficients for duration $\hat{\beta}_d$ (a), and volatility $\hat{\beta}_v$ (b), and R-squared (c) for the logarithmic regressions of absolute execution costs on trade duration and volatility (2.29), estimated independently for each stock in our database.

These trades are likely to have been executed over similarly long durations. Instead, collinearity between these two variables does not arise for all trades, where the correlation between trade size and duration is only 0.45.

The model in equation (2.29) is also independently estimated for each stock. Figure 2.14 shows the histograms of the estimated regression coefficients and the coefficients of determination. The mean value of $\hat{\beta}_d$ is 0.448, and we can see how most of the estimated $\hat{\beta}_d$ coefficients are indeed close to 1/2. In particular, 95.51% of $\hat{\beta}_d$ coefficients are statistically significant at the 5% level. The mean value of $\hat{\beta}_v$ is 0.914, and we can see how most of the estimated

$\hat{\beta}_v$ coefficients are indeed close to one, as we would expect. In particular, 94.07% of $\hat{\beta}_d$ coefficients are statistically significant at the 5% level. The average coefficient of determination is 28.98% and its maximum value is 68.68%.

The results from these multiple-regressions show very clearly that the variation of the amplitude of price changes accompanying trade executions is much better explained in terms of a square-root dependence on trade *duration* rather than trade size, with volatility being the proportionality coefficient. As explained in section 2.4.1, this square-root relationship arises naturally from well known square root scaling of volatility with horizon and does not require to invoke any concept of 'market impact'.

We also analyse how these relationships hold when we look at signed execution costs, rather than their absolute values. We run the following NLS regressions:

$$
\begin{aligned}
C &= Y\left(\frac{Q}{V_D}\right)^{\beta_s} \sigma_D^{\beta_v} + \xi, && (2.31)\\
C &= Y T^{\beta_d} \sigma_D^{\beta_v} + \xi, && (2.32)\\
C &= Y\left(\frac{Q}{V_D}\right)^{\beta_s} T^{\beta_d} \sigma_D^{\beta_v} + \xi, && (2.33)
\end{aligned}
$$

The results of these regressions are presented in Table 2.4 and are similar for all trades (columns 2-4) and for large trades only (columns 5-7). In the first regression (columns 2 and 5), the estimated regression coefficient on the daily volume fraction, $\hat{\beta}_s$, is 0.318 for all trades and 0.373 for large trades, and both coefficients are statistically significant at the 1% level. For all trades, this is closer to 0.5 than what we found for the log regression on absolute values (0.201), but the explanatory power is much lower ($R^2 = 4.9\%$ versus 21.2%), and quite similar to what

	Dependent variable: Execution Cost C					
	All Trades			Large Trades		
Model	(2.31)	(2.32)	(2.33)	(2.31)	(2.32)	(2.33)
Constant Y	6.73***	0.359***	1.410***	2.863***	0.186***	0.441 ***
	(0.006)	(0.006)	(0.072)	(0.210)	(0.014)	(0.055)
Trade size Q/V_D	0.318***		0.187***	0.373***		0.126***
	(0.004)		(0.004)	(0.009)		(0.015)
Duration T		0.359***	0.305***		0.379***	0.275***
		(0.015)	(0.006)		(0.009)	(0.014)
Volatility σ_D	1.094***	1.163***	1.153***	0.968***	0.959***	0.967***
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)
Observations	410,792	410,792	410,792	102,698	102,698	102,698
R^2	0.049	0.049	0.055	0.050	0.054	0.054
Note:				*p<0.1; **p<0.05; ***p<0.01		

Table 2.4: Estimated NLS regressions of signed execution costs on daily volume fraction, duration and daily volatility for all trades (left) and for large trades only (right) (2008-2012), equations (2.31 - 2.33). Standard errors in parentheses.

we found when testing the square-root model in (2.20).

In the second regression for all trades (columns 3 and 6), the estimated regression coefficient on trade duration, $\hat{\beta}_d$, is 0.359 for all trades and 0.379 for large trades, and both coefficients are statistically significant at the 1% level. This is closer to 1/2, though the difference is less striking than when we look at the absolute value of execution costs and the coefficients of determination for these regressions are quite similar to those from model (2.31).

In the third regression, where the daily volume fraction and duration are included as regressors (columns 4 and 7), the estimated regression coefficient on trade duration remains around the 0.30 mark while the regression coefficient on daily volume fraction drops to only 0.13-0.19. Despite being statistically significant, the estimated regression coefficients on the daily volume fraction are very far from 0.5. Moreover, for all trades, the increase in the coefficient of determination compared to the second regression is minimal (5.5% versus 4.9%)

while for large trades there is virtually no increase in the coefficient of determination.

These findings are in line with Bershova and Rakhlin (2013) who also find a coefficient close to 0.5 when running a NLS of execution costs on trade duration. More importantly, they seem to confirm the results in section 2.3.1.2 that, even if there is a 'market impact' term, this is very small compared to the volatility term and it contributes much less to the total execution costs. The implication is that, rather than trying to optimise 'impact' at a trade-by-trade level, investors should focus on managing price volatility.

2.5 The role of trading speed and participation rate

2.5.1 Execution costs and participation rate conditional on trade duration

Most optimal execution models consider the participation rate, which is a measure of the speed of trading, as the key driver of execution costs. The assumption made by these models is that faster trading leads to higher market impact and therefore to higher execution costs.

However, as discussed in Section 2.4.2, the analysis of large trades shows no such dependence of execution costs on the participation rate or the trading speed . As shown in Figure 2.4 (lower panel), for trades executed over the entire trading day with size larger then 18.5% of daily volume, an increase in the size, and therefore in the participation rate has essentially no influence on the amplitude of the execution cost during the trade. And Figure 2.5 (lower panel) shows that for these trades the signed execution cost actually *decreases*.

Driven by these observations, we take a closer look at the influence of the (average) participation rate on both absolute and signed execution costs, conditional on trade duration. Figure 2.15, shows four scatter-plots of (log) absolute execution costs (standardised by daily

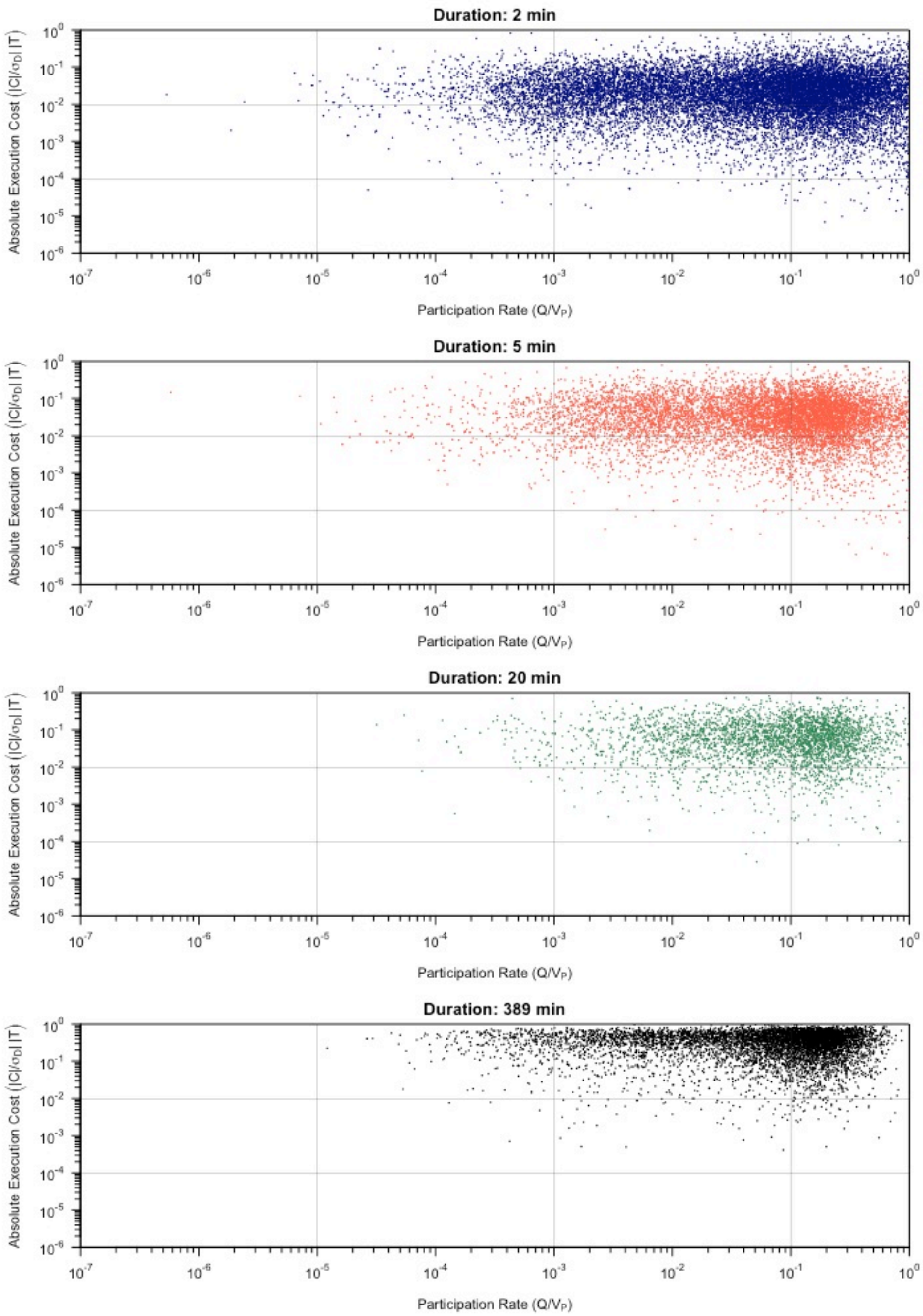

Figure 2.15: Scatter-plots of (log) absolute execution cost $|C/\sigma_D|$ versus (log) participation rate Q/V_P conditional on four representative trade duration values: $T = 2, 5, 20$ and 389 minutes (2008-2012).

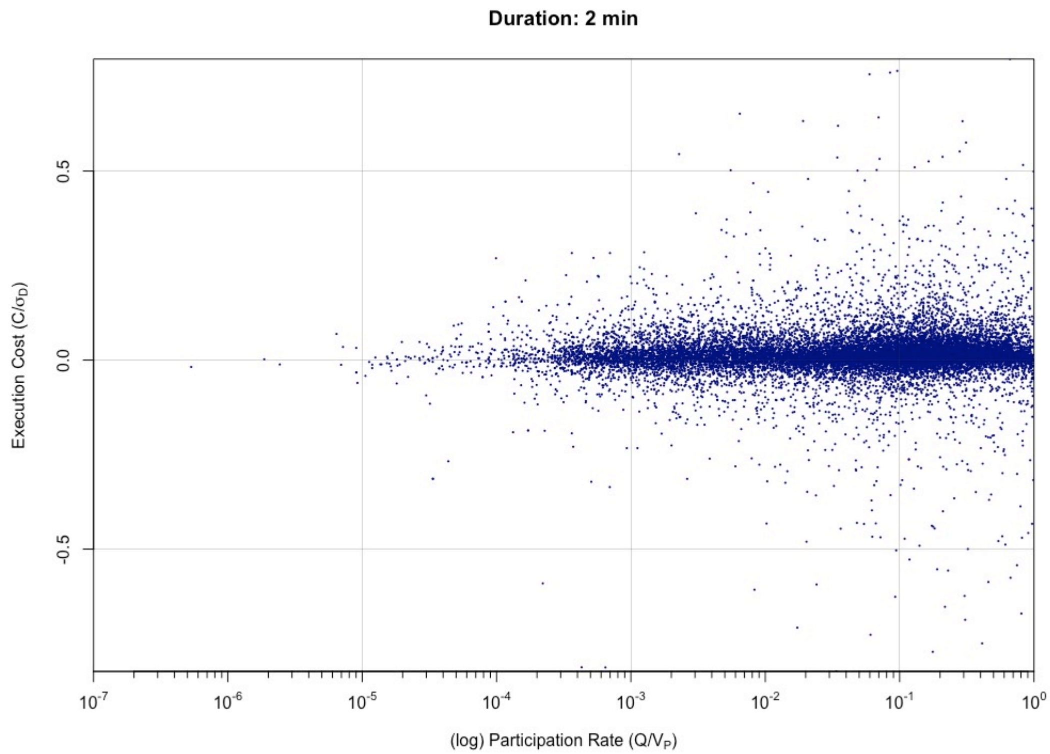

Figure 2.16: Scatter-plot of execution cost $|C/\sigma_D|$ versus (log) participation rate Q/V_P conditional on trade duration $T = 2$ minutes (2008-2012).

volatility) $|C/\sigma_D|$ versus (log) participation rate Q/V_P conditioned on representative short- (2 minutes), medium- (5 minutes), and high-duration (389 minutes) values. The cloud of points in each of these four scatter-plots is essentially flat, showing no relationship between execution costs and the participation rate. Similarly, Figure 2.16 , which shows the scatter-plot of signed execution costs C/σ_D versus (log) participation rate Q/V_P conditioned on trade duration $T = 2$ minutes, reveals no relationship between execution costs and the participation rate. To test this formally, we sort the data into 389 bins $[t_k, t_{k+1})$ for the trade duration T, the k-th bin corresponding to observations with $T \in [t_k, t_{k+1})$, and T is expressed in minutes.

The conditional logarithmic regression of the absolute execution cost $|C|$, standardised by the daily volatility, on the participation rate Q/V_P, given the trade duration T, is estimated, for each bin k, as:

$$\log\left(\left|\frac{C}{\sigma_D}\right|\right)_k = \alpha_k + \beta_{p,k}\log\left(\frac{Q}{V_P}\right)_k + \xi_k, \tag{2.34}$$

and the conditional non-linear least square (NLS) regression for the execution cost C, standardised by the daily volatility, on the participation rate Q/V_P, given the trade duration T, is estimated as:

$$\left(\frac{C}{\sigma_D}\right)_k = Y_k\left(\frac{Q}{V_P}\right)_k^{\beta_{p,k}} + \xi_k, \tag{2.35}$$

Histograms of the estimated regression coefficients $\hat{\beta}_{p,k}$, their p-value and the coefficient of determination for all 389 regressions are presented in Figure 2.17 for the log regressions and in Figure 2.18 for the NLS regressions. When we look at the absolute value of execution costs, $\hat{\beta}_{p,k}$ is statistically significant at the 5% level in 46 of the 389 regressions (11.8%, Figure 2.17 panel (b)). However, its average estimate is only 0.02 (Figure 2.17, panel (a)), and the average coefficient of determination for the regression in (2.34) is only 0.5%, with 372 out of 389 (96%) regressions having an R^2 smaller than 2% (Figure 2.17, panel (c)). In other words, once conditioned on trade duration, the participation rate has essentially zero explanatory power on the absolute value of execution costs. More importantly, we also find *negative* regression coefficients, implying that an increase in the participation rate during the same trade horizon, and therefore an increase in the trading speed, leads to *lower* absolute execution costs, as also reported in Salem et al. (2018).

The analysis of signed execution costs paints a similar picture. $\hat{\beta}_{p,k}$ is statistically signif-

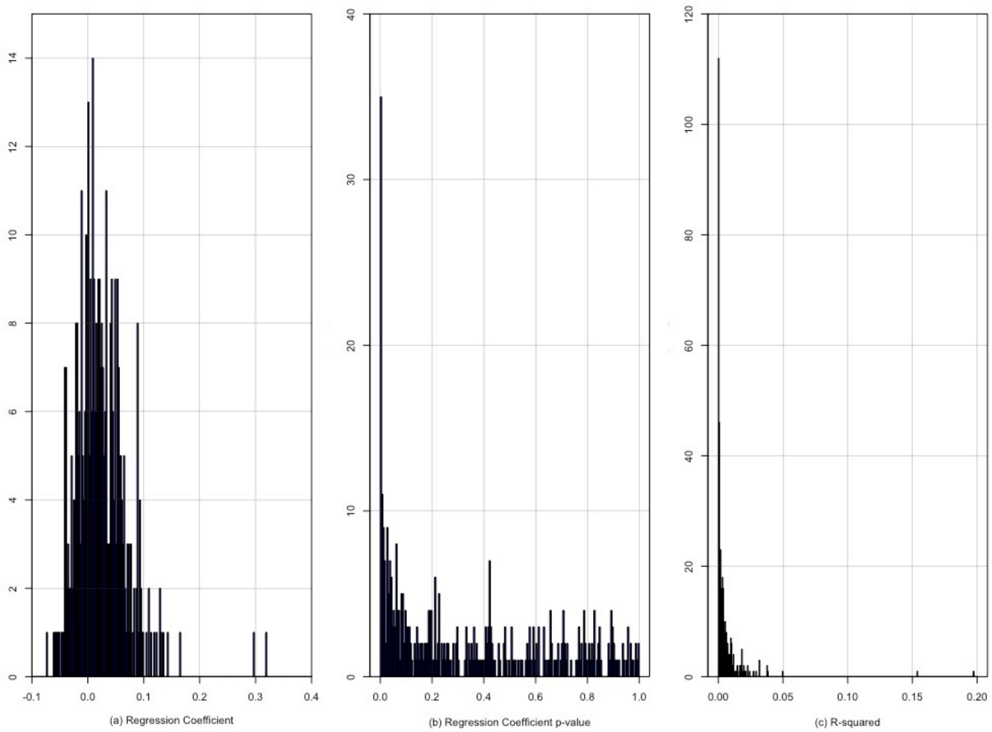

Figure 2.17: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional logarithmic regressions of absolute execution costs, standardised by daily volatility, on the participation rate, given trade duration (2.34).

icant at the 5% level in 184 of the 389 durations (47.3%) (Figure 2.18, panel (b)), but the value is again often very close to zero, and in any case far from 0.5 (Figure 2.18, panel (a)). On average, $\hat{\beta}_{p,k}$ is 0.20, the average coefficient of determination for the regressions is only 1.1%, and 333 out of 389 (85.6%) regressions having an R^2 smaller than 2% (Figure 2.18, panel (c)).

Thus, these results strongly contradict the notion that the participation rate is *the* main driver of execution costs, as emphasised in much of the current literature on optimal execution.

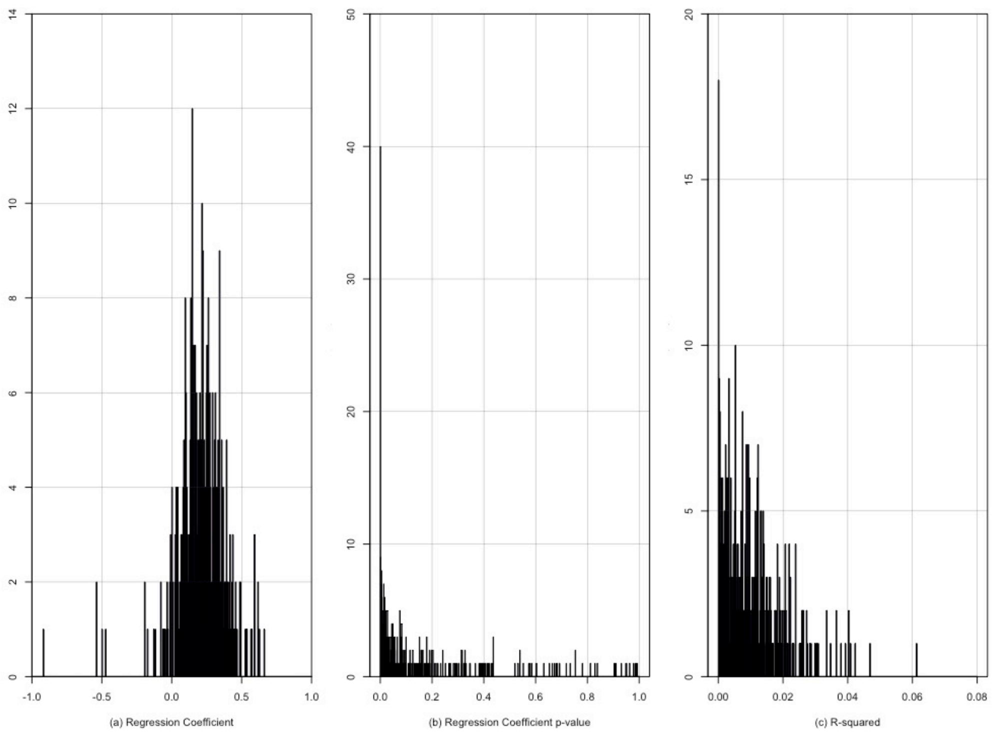

Figure 2.18: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional non-linear least square regressions of execution costs, standardised by daily volatility, on the participation rate, given trade duration (2.35).

2.5.2 Joint analysis of execution costs, participation rate and trade duration

We also run a step-wise multivariate log-regression of the absolute value of execution costs on the participation rate, duration and volatility to further investigate the role of the participation rate on execution costs, once trade duration is controlled for. We estimate the following log-log regressions:

$$
\begin{aligned}
\log|C| &= \alpha + \beta_p \log\frac{Q}{V_P} + \beta_v \log\sigma_D + \xi, && (2.36)\\
\log|C| &= \alpha + \beta_d \log T + \beta_v \log\sigma_D + \xi, && (2.37)\\
\log|C| &= \alpha + \beta_p \log\frac{Q}{V_P} + \beta_d \log T + \beta_v \log\sigma_D + \xi. && (2.38)
\end{aligned}
$$

Table 2.5 presents the results of these regressions for all trades (columns 2-4) and for large trades only (columns 5-7). In the first regression for all trades (column 2), the estimated regression coefficient on the participation rate, $\hat{\beta}_p$, is -0.014 and is statistically significant at the 1% level. The *negative* regression coefficient is clearly at odds with the notion that higher participation rates lead to higher execution costs.

In the second regression for all trades (column 3), the estimated regression coefficient on trade duration, $\hat{\beta}_d$, is 0.437, and is statistically significant at the 1% level, as already found in section 2.4.3. When both the participation rate and trade duration are considered as explanatory variables (column 4), the coefficient on trade duration remains close to 1/2 (specifically 0.438) while the regression coefficient on the participation rate becomes positive. Both regression coefficients are statistically significant at the 1% level. However, despite being statistically significant, the estimated regression coefficient on the participation rate is only 0.011, implying that a 1 percentage point increase in participation rate to only a 0.011% increase in the absolute value of execution costs. Moreover, the adjusted coefficient of determination is the same (34.3%) as in the second regression.

The findings are even more striking if we focus only on large trades. The estimated regression coefficient $\hat{\beta}_p$ from the first regression is -0.485 (column 5), and is statistically signif-

	Dependent variable: Absolute Execution Cost $\lvert C \rvert$					
	All Trades			Large Trades		
Model	(2.36)	(2.37)	(2.38)	(2.36)	(2.37)	(2.38)
Constant α	-2.949^{***}	-2.071^{***}	-2.039^{***}	-3.698^{***}	-2.151^{***}	-2.312^{***}
	(0.013)	(0.012)	(0.013)	(0.031)	(0.026)	(0.028)
Participation Rate Q/V_P	-0.014^{***}		0.011^{***}	-0.485^{***}		-0.119^{***}
	(0.001)		(0.001)	(0.011)		(0.010)
Duration T		0.437^{***}	0.438^{***}		0.464^{***}	0.457^{***}
		(0.001)	(0.001)		(0.003)	(0.003)
Daily Volatility σ_D	0.930^{***}	0.911^{***}	0.910^{***}	0.908^{***}	0.874^{***}	0.874^{***}
	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)	(0.007)
Observations	410,792	410,792	410,792	102,698	102,698	102,698
R^2	0.148	0.343	0.343	0.139	0.335	0.336
Adjusted R^2	0.148	0.343	0.343	0.139	0.335	0.336
Residual Std. Error	1.437	1.262	1.262	1.481	1.301	1.300
F Statistic	$35,724^{***}$	$107,191^{***}$	$71,508^{***}$	$8,316^{***}$	$25,892^{***}$	$17,328^{***}$
AIC	1,463,712	1,357,075	1,356,983	372,049.6	345,533.1	345,400.1
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2.5: Estimated logarithmic regressions of absolute execution cost on participation rate, duration and daily volatility for all trades (left) and only for large trades (right) equations (2.36 - 2.38) (2008-2012). In parentheses: heteroskedasticty-consistent Newey-West standard errors.

icant at the 1% level, confirming a negative relationship between the participation rate and absolute execution costs. When both the average participation rate and trade duration are included in the regression (column 7), the coefficient on trade duration remains close to 1/2 (specifically 0.457) and the regression coefficient on the participation rate remains negative at -0.119, implying that a 1 percentage point increase in the participation rate leads to a 0.119 *decrease* in the absolute value of execution costs.

These results confirm that, in orderly executions, an increase in the average participation rate at constant trade duration leads to a *decrease* in the absolute value of execution costs. Or, in other words, an *increase* in the trading speed leads to a *decrease* in the amplitude of price changes during execution, similar to the findings of Salem et al. (2018) for the US Treasury market. This is simply because a faster execution entails a shorter duration, and therefore a lower exposure to volatility risk during execution. Note that this is the opposite effect to what

	Dependent variable: Execution Cost C					
	All Trades			Large Trades		
Model	(2.39)	(2.40)	(2.41)	(2.39)	(2.40)	(2.41)
Constant Y	1.738***	0.359***	0.499***	0.214***	0.186***	0.090 ***
	(0.070)	(0.006)	(0.021)	(0.034)	(0.014)	(0.008)
Participation Rate Q/V_P	0.112***		0.150***	−0.699***		− 0.558***
	(0.004)		(0.004)	(0.034)		(0.031)
Duration T		0.359***	0.456***		0.379***	0.367***
		(0.015)	(0.006)		(0.009)	(0.009)
Volatility σ_D	1.038***	1.163***	1.155***	0.872***	0.959***	0.951***
	(0.007)	(0.006)	(0.006)	(0.013)	(0.012)	(0.011)
Observations	410,792	410,792	410,792	102,698	102,698	102,698
R^2	0.032	0.049	0.053	0.033	0.054	0.057
Note:				*p<0.1; **p<0.05; ***p<0.01		

Table 2.6: Estimated NLS regressions of execution cost on participation rate, duration and daily volatility for all trades (left) and only for large trades (right) (2008-2012), equations (2.31 - 2.33). Standard errors in parentheses.

is usually assumed in many optimal execution models.

We also analyse how these relationships hold when we look at signed execution costs rather than their absolute value. We run the following NLS regressions:

$$C = Y\Big(\frac{Q}{V_P}\Big)^{\beta_p}\sigma_D^{\beta_v} + \xi, \tag{2.39}$$

$$C = YT^{\beta_d}\sigma_D^{\beta_v} + \xi, \tag{2.40}$$

$$C = Y\Big(\frac{Q}{V_P}\Big)^{\beta_p}T^{\beta_d}\sigma_D^{\beta_v} + \xi, \tag{2.41}$$

The results of these regressions are presented in Table 2.6. In the first regression (columns 2 and 5), the regression coefficient on the participation rate, $\hat{\beta}_p$, is 0.112 for all trades and -0.699 for large trades, which, for large trades, confirms the negative relationship between execution costs and participation rate.

In the second regression (columns 3 and 6), the regression coefficient on trade duration, $\hat{\beta}_d$, is 0.359 for all trades and 0.379 for large trades, and both coefficients are statistically significant at the 1% level.

In the third regression, where both the participation rate and trade duration are included as regressors (columns 4 and 7), the coefficient on trade duration remains close to 0.5 while the regression coefficient on the participation rate drops to 0.15 for all trades and remains negative, at -0.558, for large trades. The coefficients are statistically significant at the 1% level, but once again, the inclusion of the participation rate as an additional explanatory variable only leads to a very small increase in the coefficient of determination compared to the second regression (5.3% versus 4.9% for all trades, and 5.7% versus 5.4% for large trades).

These results confirm our previous findings that in orderly executions, an increase in the average participation rate at constant trade duration leads to a *decrease* in execution costs.

2.6 Conclusion

A generation of models has emphasised the importance of market 'impact' as the main determinant of execution costs and defined optimal execution in terms of a trade-off between two independent components: 'market impact' and price volatility. Our empirical analysis suggests that in fact 'market impact' and price volatility are two sides of the same coin. We have shown that execution costs are mainly driven by the *aggregate (net) order flow imbalance* rather than the 'impact' of trades and that, once conditioned on trade duration, trade size has very little explanatory power for price changes during execution. We have found that execution costs are explained by the square-root of *duration* and that this relationship arises from the well-known square-root scaling of volatility as a function of time. When looking at the role of the participation rate in determining execution costs, we have found that for trades with high participation rates, mainly executed on a VWAP schedule, a large increase in the trade size (and in the participation rate) leaves execution costs almost unchanged, contradicting one of the key assumptions of optimal execution models. Finally, our analysis of the relationship between execution costs and the average participation rate *conditional* on trade duration shows that the trading speed affects costs in the opposite direction than what has been assumed in optimal execution models: the *slower* the execution the *higher* the price variation during the trade.

Our findings bring into question the notion of 'market impact' defined at the trade level and highlight the need to rethink what it means to optimally execute a trade. Our results suggest that it is more meaningful to focus on the modelling of aggregate order flow dynamics and the management of portfolio volatility during execution rather than the optimisation of 'impact' at a trade-by-trade level.

3

Multi-asset market impact and order flow commonality

This chapter is based on the paper "Multi-asset market impact and order flow commonality" which is a joint work with Professor Rama Cont.

3.1 Introduction

3.1.1 Market impact and cross-impact

The notion of *cross-impact* refers to the fact that the price of a given asset may be influenced not only by its own order flow, but also by the order flow of *other* assets, thus resulting in higher execution costs when more than one security is traded at the same time. The empirical finding of positive covariance between the returns of one asset and the order flow imbalance of other assets has been claimed as evidence of cross-impact, leading to multi-asset market-impact models with off-diagonal *cross-impact coefficients* (Alfonsi et al., 2016; Benzaquen et al., 2017; Mastromatteo et al., 2017; Min et al., 2018).

In their simplest form, such models may be seen as multi-asset extensions of the Kyle (1985) model, parameterised by a matrix of impact coefficients whose off-diagonal elements are meant to capture how trades in one asset influence the price of other assets (Caballé and Krishnan, 1994). Such multi-asset impact models have been the subject of several recent theoretical and empirical studies (Andrade et al., 2008; Schied et al., 2010; Boulatov et al., 2013; Pasquariello and Vega, 2015; Alfonsi et al., 2016; Schöneborn, 2016; Benzaquen et al., 2017; Mastromatteo et al., 2017; Tsoukalas et al., 2017; Min et al., 2018; Schneider and Lillo, 2019; Tomas et al., 2020). While some authors *assume* cross-impact without detailing the underlying causal mechanism that might give rise to this phenomenon (e.g. Schied et al., 2010; Schöneborn, 2016), other authors present models of sequential trades in which market makers' quote revisions give rise to cross-impact (Pasquariello and Vega, 2015).

Despite the wide use of such models, the theoretical and empirical evidence underlying their justification is not entirely convincing for a number of reasons. The mere presence of positive covariation between the returns of an asset and the order flow of another asset is not sufficient to provide a causal justification for the existence of cross impact as a separate phenomenon from, for example, correlation between the order flow of the two assets. Common components in order flow - which are well documented empirically (Hasbrouck and Seppi, 2001) - naturally arise from cross-asset trading (Bernhardt and Taub, 2008), index trading (Harford and Kaul, 2005) and portfolio rebalancing (Cont and Wagalath, 2016), and may lead to similar effects without the need to invoke cross-impact.

In this Chapter we investigate these questions using a *causal* modelling approach (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009), carefully distinguishing measures of association - such as correlations - from causal relations, to arrive at a causal model for market impact in a multi-asset setting.

In particular, we investigate whether there is any empirical evidence for the concept of cross-impact or whether a simpler and more parsimonious explanation may be given for the observed correlations among returns and order flows across assets without invoking this notion.

3.1.2 Contribution: causal analysis and modelling of multi-asset market impact

In this Chapter, we reexamine the empirical evidence from equity markets and provide a different perspective on the notion of cross-impact.

Starting from a linear model for market impact, we show that cross-impact is not needed to explain the positive covariation between the returns of an asset and the order flow of other

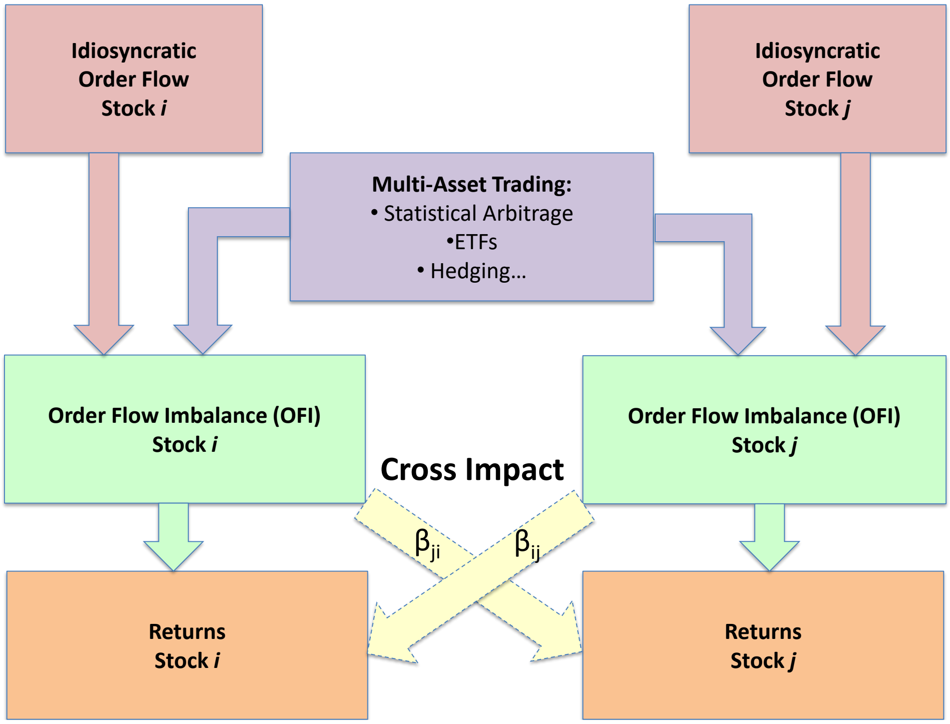

Figure 3.1: Causal representation of multi-asset impact models. The returns of each asset are influenced by the asset's own order flow imbalance, which is the superposition of an idiosyncratic order flow and of a *common component* across assets. This common component leads to correlation in order flow imbalance across assets. Cross-impact models posit the existence of an additional direct influence of the order flow imbalance of an asset on the returns of other assets, represented by the diagonal arrows in the causal diagram.

assets. Instead, we provide evidence that this covariation simply arises from the correlation of order flow imbalances across assets. A significant proportion of the demand and supply for assets originates in portfolio rebalancing strategies or trading strategies -index or ETF trading, pairs trading, statistical arbitrage strategies and hedging strategies- which involve positions in several assets, thereby generating correlations in order flow.

We investigate these questions through the angle of causal modelling (Haavelmo, 1943, 1944; Heckman and Pinto, 2015; Pearl, 2009), which emphasises causal relations among variables, testable through conditioning.

Figure 3.1 shows the causal chain of relations for two such assets i and j: market participants trade in stocks i and j at the same time, which leads to correlation between the OFIs of the two stocks. The return of each asset is determined by the order flow imbalance, which is the superposition of a common component and the idiosyncratic order flow in each security. The presence of this common component leads to correlations in order flow imbalance across assets, and results in correlations between the returns of an asset and order flow of other assets.

Cross-impact models insert an additional mechanism, represented in Figure 3.1 as a pair of diagonal arrows, through which the order flow of one asset would somehow influence the prices of another asset, bypassing the supply/demand mechanism usually assumed to underpin price formation.

The causal graph in Figure 3.1 suggests that without conditioning on the common component in the order flow, one cannot distinguish between cross-impact and order flow correlation. In Section 3.2.2, we present examples that illustrate this *identifiability* problem, and demonstrate that one can overcome this problem by simply estimating the covariance of order flow imbalances.

The presence of causal relations, shown as arrows in Figure 3.1, may be tested by conditioning on the variable hypothesised to contribute to an observed effect. We use this approach to test for the significance and amplitude of cross-impact channels by conditioning on the common factor in order flow imbalance. Our results, in Section 3.3.4, show that once a stock's own order flow and the commonality in order flow has been taken into account, the additional explanatory power of all other stocks' order flow is less than 0.5 percentage points.

In section 3.3.5 we show that, in addition to being small in magnitude, cross-impact coefficients are unstable and change sign randomly over time, which poses a problem for their interpretation and use.

Our results show the main determinants of impact to be idiosyncratic order flow imbalance as well as a market order flow factor common across stocks. This leads to a parsimonious approach for *causal* modelling of multi-asset market impact, which does not require introducing any concept of cross-impact.

Our findings have important implications for optimal trade execution and transaction costs analysis. They show that accounting for commonality in order flow seems to be a more effective way to build multi-asset market impact models than introducing cross-impact coefficients whose statistical significance and interpretation are not clear.

In section 3.4, we show how our framework may be used for the estimation of portfolio execution costs. First, we show that the presence of common components in order flow leads to *unconditional* estimates of portfolio execution costs which are different from single-asset TCA models. Second, we show how the model allows to use information on the direction of market order flow to obtain *conditional* estimates of portfolio execution costs.

3.1.3 Relation with previous literature

The extension of the model in Kyle (1985) to multiple assets led, almost inevitably, to trying to understand and model how prices in an asset may be influenced by trades in other assets. Caballé and Krishnan (1994) developed a model that relates prices and order flows of different assets and derive the consequences for covariance structure of these variables. Pasquariello and Vega (2015) present a multi-asset sequential trade model in which cross-impact arises, in equilibrium, from the strategic cross-trading of informed speculators and the market makers' subsequent cross inference based on the speculators' order flow and present empirical evidence that cross-impact is economically and statistically significant, and often negative.

Boulatov et al. (2013) present a multi-period model of strategic trading in which cross-impact arises as a results of market makers updating their quotes on an asset based on the order flow of other assets. Using institutional investor orders on the NYSE, these authors find statistical evidence of a lead-lag correlation between the returns of a stock and the informed order flow on another stock.

Andrade et al. (2008) propose a model in which the magnitude of cross-impact depends on the correlation among stocks' underlying cash flows, and their relative attractiveness as hedging instruments for liquidity providers, and provide empirical evidence from the Taiwan Stock Exchange. Our results, although at intraday frequencies and using different data sources, are consistent with these findings.

Cross-impact has also been considered in the context of optimal execution in multivariate extensions of Almgren and Chriss (2001). The focus in these studies is the formulation of strategies for optimal portfolio liquidation in the presence of volatility risk. For example, Schied et al. (2010) and Schöneborn (2016) use this framework to study optimal execution strategies for investors with constant absolute risk aversion, and for a more general absolute risk aversion setting, respectively. However, the authors *assume* cross-impact, rather than detailing the mechanism through which it might arise. Tsoukalas et al. (2017) develop a tractable order book model to capture liquidity supply and demand dynamics in a multi-asset setting and show that even when the goal is to trade a single asset, optimal execution may involve transitory trades in other assets, leading to cross-impact. However, no empirical evidence is provided to support this model, which does not distinguish cross-impact from order flow correlation, an important point which we address in this paper.

Other models of cross-impact in trade execution include Mastromatteo et al. (2017), Min et al. (2018) and Tomas et al. (2020). Finally, the notion of cross-impact has also been studied from the perspective of absence of dynamic arbitrage by Alfonsi et al. (2016) and Schneider and Lillo (2019). In these papers, cross-impact is, once again, assumed rather than derived.

Commonalities in liquidity and order flow have been widely documented in the literature (Chordia et al., 2000; Hasbrouck and Seppi, 2001; Heston et al., 2010; Pástor and Stambaugh, 2004) and cannot be dismissed as one of the main sources of positive covariation between the returns of a stock and the order flow of other stocks. Harford and Kaul (2005) document common factors in order flow arising from index trading; Bernhardt and Taub (2008) study order flow correlations arise from speculative cross-asset trading. Cont and Wagalath (2016) study endogenous correlations in order flow and returns arising from portfolio rebalancing of institutional investors. Cont and Wagalath (2013) show how liquidation of large positions by institutional investors may lead to correlated order flow and endogenous correlations in price movements. All these mechanisms lead to significant correlations between asset returns and order flow in other assets without introducing the concept of cross-impact. We build on these ideas and show that they lead to a plausible and parsimonious explanation for the positive covariance of returns and order flow imbalance across different assets.

3.1.4 Outline

The remainder of the Chapter is organised as follows. Section 3.2 describes a linear model of market impact from trading and discusses the issue of identifiability of the cross-impact coefficients through an example. Section 3.3 presents empirical evidence for our causal model, obtained from the analysis of order flow, trades and quotes for NASDAQ stocks. Section 3.4 discusses the application of our model to the analysis of portfolio execution costs. Section 3.5 discusses the implications of our findings for the modelling of market impact.

3.2 Order flow correlation and cross impact

3.2.1 A linear model of market impact

Linear models of market impact, going back to Kyle (1985), are based on a linear relation between price movements (or returns) and aggregate net order flow for an asset.

Cont et al. (2014) provide empirical evidence for such a linear model of market impact at intraday frequencies using limit order book data, using the concept of *order flow imbalance* (OFI):

$$R_{i,t} = \mu_i + \beta_i \, \mathrm{OFI}_{i,t} + \varepsilon_{i,t}, \tag{3.1}$$

where $R_{i,t}$ is the return of asset i, $\mathrm{OFI}_{i,t}$ is the order flow imbalance for asset i, defined as the net order flow during the time interval $[t - \Delta t, t]$, and $\varepsilon_{i,t}$ is a noise term summarising influences of other factors (e.g., deeper levels of the order book, etc.).

The multi-asset extension of (3.1) to N assets may be written in matrix form as:

$$\boldsymbol{R}_t = \mu + \beta \, \mathbf{OFI}_{\boldsymbol{t}} + \boldsymbol{\varepsilon}_{\boldsymbol{t}}, \tag{3.2}$$

where μ, $\boldsymbol{R_t}$, and $\mathbf{OFI}_{\boldsymbol{t}}$ are respectively $N \times 1$ vectors of constants, returns and order flow imbalances:

$$\mu \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}, \quad \boldsymbol{R_t} \equiv \begin{bmatrix} R_{1,t} \\ R_{2,t} \\ \vdots \\ R_{N,t} \end{bmatrix}, \quad \mathbf{OFI}_{\boldsymbol{t}} \equiv \begin{bmatrix} \mathrm{OFI}_{1,t} \\ \mathrm{OFI}_{2,t} \\ \vdots \\ \mathrm{OFI}_{N,t} \end{bmatrix},$$

$\boldsymbol{\varepsilon_t}$ is an $N \times 1$ vector of uncorrelated error terms:

$$E(\boldsymbol{\varepsilon_t}) = 0, \qquad \mathrm{cov}(\boldsymbol{\varepsilon_t}) = \Sigma, \qquad \mathrm{cov}(\boldsymbol{\varepsilon_t}, \mathbf{OFI}_{\boldsymbol{t}}) = 0, \tag{3.3}$$

and β is an $N \times N$ matrix of market impact coefficients:

$$\beta \equiv \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots \\ \vdots & \ddots & \\ \beta_{N1} & & \beta_{NN} \end{bmatrix}, \tag{3.4}$$

whose off-diagonal terms represent 'cross-impact'. Let $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{K}}$ be the covariance matrices of returns and OFI:

$$\begin{aligned} \boldsymbol{\mathcal{C}} &\equiv E\Big[(\boldsymbol{R_t} - \mu_{\boldsymbol{R}})(\boldsymbol{R_t} - \mu_{\boldsymbol{R}})^{\boldsymbol{T}}\Big], && (3.5) \\ \boldsymbol{\mathcal{K}} &\equiv E\Big[(\mathbf{OFI}_{\boldsymbol{t}} - \mu_{\mathbf{OFI}})(\mathbf{OFI}_{\boldsymbol{t}} - \mu_{\mathbf{OFI}})^{\boldsymbol{T}}\Big], && (3.6) \end{aligned}$$

where $\mu_{\boldsymbol{R}}$ is an $N \times 1$ vector of mean returns and $\mu_{\mathbf{OFI}}$ is an $N \times 1$ vector of mean OFIs. We can also define covariance matrix between returns and OFIs as:

$$\boldsymbol{\mathcal{M}} = \mathrm{cov}(\boldsymbol{R_t}, \mathbf{OFI}_t) = E\Big[(\boldsymbol{R_t} - \mu_{\boldsymbol{R}})(\mathbf{OFI}_t - \mu_{\mathbf{OFI}})^T\Big] = \beta\boldsymbol{\mathcal{K}} \tag{3.7}$$

In (3.7), the diagonal terms represent the covariance between the returns of asset i and its own order flow imbalance, while the off-diagonal terms represent the covariance between the returns of asset i and the order flow imbalance of assets $j \neq i$.

3.2.2 Are cross-impact coefficients identifiable?

The covariance (3.7) of order flow imbalance with asset returns has been the subject of several empirical studies for different asset classes (Mastromatteo et al., 2017; Tomas et al., 2020). These studies find positive off-diagonal elements in $\boldsymbol{\mathcal{M}}$, and have used this as justification for the extension of univariate impact models to multiple assets with off-diagonal, *cross-impact coefficients* in (3.4). But the presence of off-diagonal terms in the matrix $\boldsymbol{\mathcal{M}}$ may arise from a much simpler and more intuitive mechanism, without invoking the notion of cross-impact: namely, the fact that OFIs of different assets are correlated.

That order flow across different assets may be correlated is an intuitive consequence of the widespread use of multi-asset trading strategies. Market participants often adopt trading strategies (such as pairs trading, index arbitrage, and other statistical arbitrage strategies) that involve more than one asset. This is more apparent for some asset classes, such as fixed-income and commodities, where there are clear theoretical reasons that underpin the simultaneous trading of assets with different maturities. As shown in the causal diagram depicted in Figure 3.1, this leads to correlation among returns and OFIs of different assets, without any need to invoke cross-impact.

The following simple example shows that off-diagonal terms in $\boldsymbol{\mathcal{M}}$ may arise from two different sources: correlation in order flow *or* off-diagonal elements in the cross-impact matrix β, leading to an identifiability problem for such cross-impact coefficients. Consider the model (3.2) with two assets ($N = 2$):

$$
\begin{aligned}
R_{1,t} &= \beta_{11}\mathrm{OFI}_{1,t} + \beta_{12}\mathrm{OFI}_{2,t} + \varepsilon_{1,t}, && (3.8)\\
R_{2,t} &= \beta_{22}\mathrm{OFI}_{2,t} + \beta_{21}\mathrm{OFI}_{1,t} + \varepsilon_{2,t}, && (3.9)
\end{aligned}
$$

where we have dropped the constant terms μ_1 and μ_2 for simplicity. Assume that the covariance matrices of returns, OFIs, and errors terms are given, respectively, by:

$$
\boldsymbol{\mathcal{C}} = \begin{bmatrix} c_1^2 & \rho_c c_1 c_2 \\ \rho_c c_1 c_2 & c_2^2 \end{bmatrix}, \quad \boldsymbol{\mathcal{K}} = \begin{bmatrix} k_1^2 & \rho_k k_1 k_2 \\ \rho_k k_1 k_2 & k_2^2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \qquad (3.10)
$$

so that the returns are correlated, with correlation coefficient ρ_c, OFIs are correlated with correlation coefficient ρ_k, and the error terms are uncorrelated. Under these assumptions, (3.7) yields:

$$
\boldsymbol{\mathcal{M}} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \begin{bmatrix} \beta_{11}k_1^2 + \beta_{12}\rho_k k_1 k_2 & \beta_{11}\rho_k k_1 k_2 + \beta_{12}k_2^2 \\ \beta_{22}\rho_k k_1 k_2 + \beta_{21}k_1^2 & \beta_{22}k_2^2 + \beta_{21}\rho_k k_1 k_2 \end{bmatrix}. \qquad (3.11)
$$

We can see from (3.11) how the off-diagonal elements, m_{12} and m_{12}, are functions of β_{12} and β_{21}, *and* $\rho_k = \mathrm{corr}(\mathrm{OFI}_1, \mathrm{OFI}_2)$. So from the knowledge of $\boldsymbol{\mathcal{M}}$, one cannot draw any conclusions with respect to cross-impact. Indeed, different combinations of ρ_k and β_{ij} lead to the same off-diagonal elements in $\boldsymbol{\mathcal{M}}$, including combinations where β_{12} and β_{21} are equal to zero.

The identification problem becomes clear when we solve (3.11) for β_{12} and β_{21}:

$$\beta_{12} = m_{12}\left[\frac{1}{k_2^2(1-\rho_k^2)}\right] - m_{11}\left[\frac{\rho_k}{k_1k_2(1-\rho_k^2)}\right], \tag{3.12}$$

$$\beta_{21} = m_{21}\left[\frac{1}{k_1^2(1-\rho_k^2)}\right] - m_{22}\left[\frac{\rho_k}{k_1k_2(1-\rho_k^2)}\right]. \tag{3.13}$$

The coefficients β_{12} and β_{21} can only be identified by estimating or assuming a value for the OFI correlation ρ_k. If one assumes the OFI correlation to zero, as in Tomas et al. (2020), then (3.12)-(3.13) collapses to:

$$\beta_{12} = \frac{m_{12}}{k_2^2}, \quad \beta_{21} = \frac{m_{21}}{k_1^2}. \tag{3.14}$$

In this case, β_{12} and β_{21} are identifiable and the off-diagonal elements of $\boldsymbol{\mathcal{M}}$ are solely due to cross-impact.

Two numerical examples help illustrate this identification problem:

Example 1: Consider the case where order flows are uncorrelated ($\rho_k = 0$) and normalised to have unit variance, with market impact coefficients given by:

$$\boldsymbol{\mathcal{K}}_{\mathbf{0}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta}_{\mathbf{0}} = \begin{bmatrix} 0.6 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}. \tag{3.15}$$

Assuming uncorrelated error terms, the covariance matrix of returns and OFIs is given by:

$$\mathcal{M}_0 = \begin{bmatrix} 0.6 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}. \tag{3.16}$$

Now, consider the case where OFIs are correlated, with correlation coefficient $\rho_k = 0.5$*, and the matrix of market impact coefficients is the same as in (3.15), but with the off-diagonal elements set to zero:*

$$\mathcal{K}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad \beta_1 = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \tag{3.17}$$

which leads to the following covariance matrix between returns and OFIs:

$$\mathcal{M}_1 = \begin{bmatrix} 0.6 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}. \tag{3.18}$$

We observe that $\mathcal{M}_1 = \mathcal{M}_0$*, yet they arise from two very different mechanisms: in the first case, order flows are uncorrelated across assets and cross-asset correlations between order flow and returns arise from cross-impact; in the second case, order flows are correlated, and there is no cross-impact mechanism. This examples shows that the mere observation of positive or non-zero off-diagonal terms in (3.16) is not sufficient to draw any conclusions regarding cross-impact.*

Example 2: Consider now the case where $\rho_k = 0.2$*, with OFIs normalised to have unit variance, with impact coefficients given by:*

$$\mathcal{K}_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}, \quad \beta_2 = \begin{bmatrix} 0.6 & 0.5 \\ 0.5 & 0.6 \end{bmatrix}. \tag{3.19}$$

The covariance between of returns and OFI is then given by:

$$\mathcal{M}_2 = \begin{bmatrix} 0.7 & 0.62 \\ 0.62 & 0.7 \end{bmatrix}. \tag{3.20}$$

Now, consider the case where OFIs are correlated, with correlation coefficient $\rho_k = 0.6$, *and the matrix of market impact coefficients market impact coefficients are:* $\beta_{11} = \beta_{22} = 0.5125$, *and* $\beta_{12} = \beta_{21} = 0.3125$. *That is to say:*

$$\mathcal{K}_3 = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}, \quad \beta_3 = \begin{bmatrix} 0.5125 & 0.3125 \\ 0.3125 & 0.5125 \end{bmatrix}. \tag{3.21}$$

Assuming, once again, that the error terms are uncorrelated, the covariance matrix of returns and OFIs is given by:

$$\mathcal{M}_3 = \begin{bmatrix} 0.7 & 0.62 \\ 0.62 & 0.7 \end{bmatrix}. \tag{3.22}$$

Again, we observe $\mathcal{M}_3 = \mathcal{M}_2$ *yet these covariances result from two very different mechanisms: in the first case, there was low correlation in the order flow and high cross-impact; in the second case, there is high correlation in the order flow, but low cross-impact. As in Example 3.2.2, this example shows that cross-impact coefficients may not be identified solely from the covariance of returns with order flow.*

3.3 Empirical analysis of returns and order flow imbalances

3.3.1 Data sources and quantities of interest

We use intraday order book data from LOBSTER, [1] which provides detailed tick-by-tick limit order book data for NASDAQ stocks. LOBSTER contains all limit order submissions, cancellations and executions on each trading day and all events are time-stamped with millisecond precision (Bibinger et al., 2017), reconstructed from NASDAQ's historical TotalView-ITCH data (Huang and Polak, 2011). The LOBSTER *message* files contain the order type, order ID, size and the price direction (buy/sell). The LOBSTER *orderbook* files contain the bid and ask prices and quantities. In our study, we use "Level I" data from LOBSTER; these are all order-book events at the best bid and ask prices (Huang and Polak, 2011).

From the LOBSTER database, we select all stocks belonging to the NASDAQ-100 Index. Our empirical analysis focuses on the period between 01-Jan-2008 and 30-Jun-2010, and after filtering out the stocks for which the dataset does not span the whole sample period, we end up with 67 stocks.

For each stock i, we estimate the daily volatility σ_i as the square root of the daily realised volatility estimator from 5-minute returns (Andersen et al., 2003).

In our analysis we will focus on $\Delta t = 1$ minute intervals. For each one-minute window $[t - \Delta t, t]$, and for each stock i, we compute:

- the return $R_{i,t}$, defined as the percentage change of the mid price between $t - \Delta t$ and t;

[1] https://lobsterdata.com.

- the net (buy) order flow $V^b_{i,t}$ at the best bid during $[t - \Delta t, t]$:

$$V^b_{i,t} \equiv L^b_{i,t} - C^b_{i,t} - M^b_{i,t},$$

  where $L^b_{i,t}$ and $C^b_{i,t}$ denote respectively the volume of new limit buy orders and cancellations at the current best bid and $M^b_{i,t}$ denotes the volume of marketable sell orders flowing to the current best bid during $[t - \Delta t, t]$;

- the net (sell) order flow $V^a_{i,t}$ at the best ask during $[t - \Delta t, t]$:

$$V^a_{i,t} = L^a_{i,t} - C^a_{i,t} - M^a_{i,t},$$

  where $L^a_{i,t}$ and $C^a_{i,t}$ denote respectively the volume of new limit sell orders and cancellations at the current best ask and $M^a_{i,t}$ denotes the volume of marketable buy orders flowing to the current best ask during $[t - \Delta t, t]$.

The *order flow imbalance* $\mathrm{OFI}_{i,t}$ for the interval $[t - \Delta t, t]$ is defined as (Cont et al., 2014):

$$\mathrm{OFI}_{i,t} \equiv V^b_{i,t} - V^a_{i,t}. \tag{3.23}$$

To account for intraday seasonality patterns in volume and order flow, we define $AV_{i,t}$ as the average total order flow $V^b_{i,t} + V^a_{i,t}$ at the best bid and ask over the (one-minute) time bucket $[t - \Delta t, t]$ and use this variable to define an average intraday profile of order flow.

In order to make our results easier to interpret, we normalise $R_{i,t}$ and $\mathrm{OFI}_{i,t}$ as in Obizhaeva

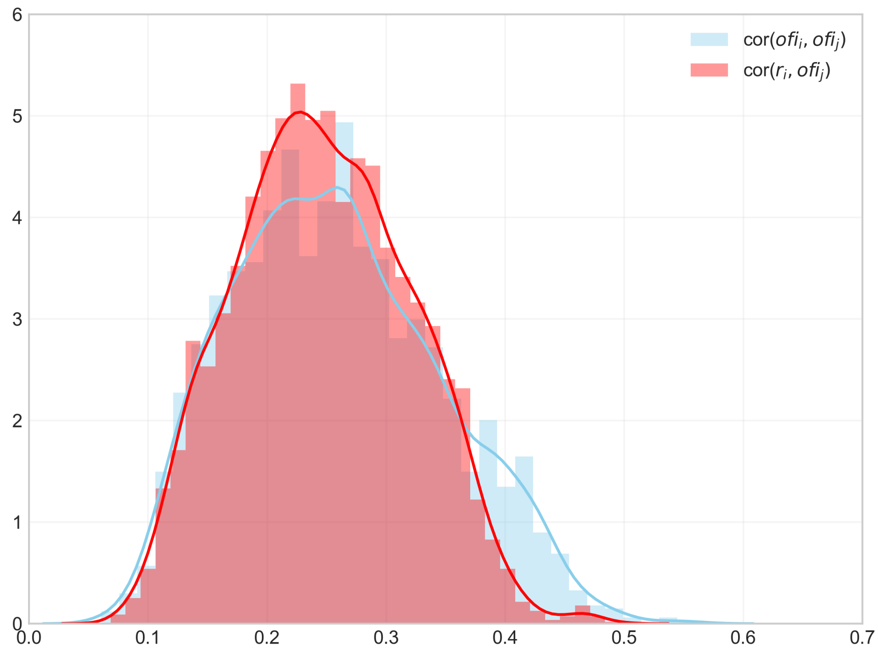

Figure 3.2: Comparison of the distribution of correlations between 1-minute $\mathrm{OFI}_{i,t}$ and $\mathrm{OFI}_{j,t}$, $\forall i \neq j$ (blue) with the distribution of correlations between 1-minute $r_{i,t}$ and $\mathrm{OFI}_{j,t}$, $\forall i \neq j$ (red) (Jan-2008: Jun-2010).

(2012). After denoting by σ_i the volatility of asset i we define:

$$r_{i,t} = \frac{R_{i,t}}{\sigma_i \sqrt{\Delta t}}, \qquad \mathrm{ofi}_{i,t} = \frac{\mathrm{OFI}_{i,t}}{AV_{i,t}}. \tag{3.24}$$

With this normalisation the coefficients of the β matrix in (3.2) can be compared directly and they can be interpreted as the number of standard deviations the price of a stock would move if one traded $x\%$ of average volume in the same stock (diagonal elements), or in a different stock (off-diagonal elements).

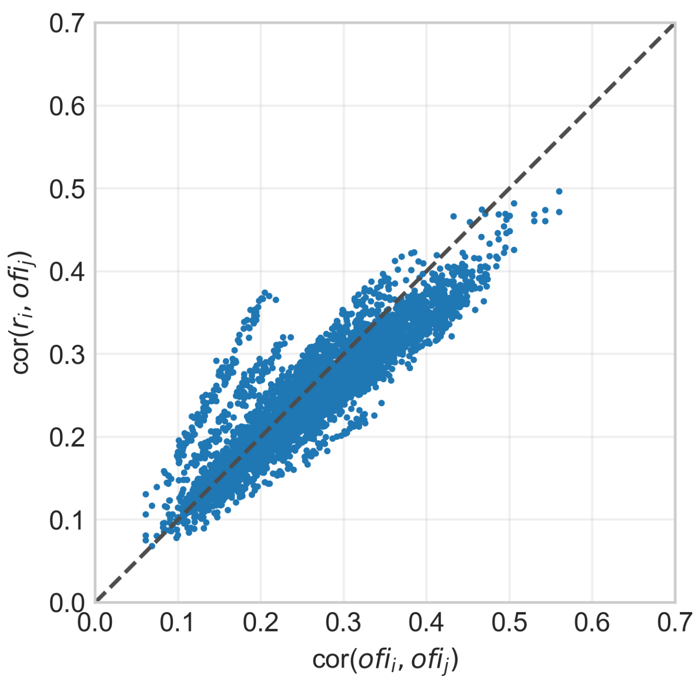

Figure 3.3: Comparison of pairwise correlations between 1-minute $\mathrm{OFI}_{i,t}$ and $\mathrm{OFI}_{j,t}, \forall i \neq j$ with pairwise correlations between 1-minute $r_{i,t}$ and $\mathrm{OFI}_{j,t}, \forall i \neq j$ for the same equity pairs (Jan-2008: Jun-2010). The dashed line corresponds to equality.

Figure 3.2 shows the distribution of the correlation coefficients between $\mathrm{ofi}_{i,t}$ and $\mathrm{ofi}_{j,t} i, j = 1 \ldots N$ (blue), and the distribution of the correlation coefficients between $r_{i,t}$ and $\mathrm{ofi}_{j,t} \quad \forall i, j = 1 \ldots N, i \neq j$ (red). We observe that the correlation coefficients between the returns and OFI of different stocks seem to follow the same distribution of the correlation coefficients between OFIs. The average pairwise correlation of OFI across stocks is 0.26 while the average correlation between returns and OFI is 0.25. The distribution of the correlation between returns and OFI is slightly more concentrated around the mean, with a standard deviation of 0.07 versus 0.09 for OFI correlations.

Figure 3.3 compares the correlation between returns and OFI and the correlation between OFIs for the same pair of stocks. For most pairs of stocks these values are close to being equal, suggesting that the correlation between order flow and returns may arise from correlation of order flows themselves. To test this assumption we now proceed to do a more detailed analysis based on a comparison of nested regression models and principal component analysis.

3.3.2 Regression analysis

We estimate the linear regression model (3.1) using ordinary least squares:

$$\boldsymbol{r}_t = \mu + \beta\, \mathbf{ofi}_t + \boldsymbol{\varepsilon}_t. \tag{3.25}$$

and examine whether off-diagonal terms β_{ij}, $i \neq j$ have significant explanatory power.

Figure 3.4 shows a heat map of the price-impact matrix β, estimated from (3.25). It is immediately clear from this heat map that the off-diagonal terms, representing the impact of other stocks order flows on a stock's returns, are much smaller in magnitude than the diagonal terms, which represent the impact of a stock's own order flow on its returns. Figure 3.5 further illustrates this point by presenting the relative frequency distributions of the diagonal (idiosyncratic) impact coefficients β_{ii} (blue) and of the off-diagonal cross-impact coefficients β_{ij} (red). The mean diagonal coefficient β_{ii} is 2.64 with a standard deviation of 0.78, while the mean off-diagonal cross-impact coefficient β_{ij} is only 0.032 with a standard deviation of 0.06. This means that the off-diagonal terms are only about 1.2% in magnitude when compared with the diagonal terms. Further, Figure 3.5 shows that whilst most off-diagonal terms are positive (76.91%), a number of cross-impact coefficients are negative (23.09%), indicating cross-impact in the opposite direction to what is suggested by many cross-impact models.

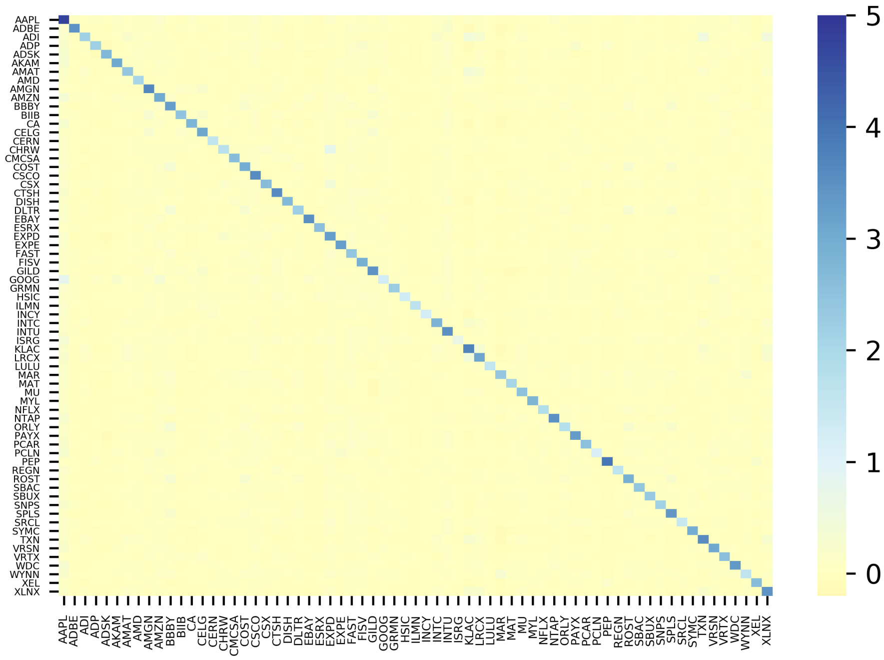

Figure 3.4: Heat map of the market impact matrix β estimated from (3.25) (Jan-2008: Jun-2010).

Given these results, we test the hypothesis that the off-diagonal terms are equal to zero, using Newey-West heteroskedasticty-consistent estimators for standard errors. Out of $4,422$ off-diagonal terms, 1,846 (41.3%) are not found to be statistically significant at the 1% level. The remainder of the coefficients are found to be significant at 1% level. For each stock i, we also run an F test for the hypothesis that all $\beta_{ij}(i \neq j)$ are jointly equal to zero and find that we can reject this hypothesis at the 1% significance level in all 67 regressions.

However, as explained in Section 3.2.2, the mere presence of statistically significant off-diagonal terms in β does not provide a causal justification for the existence of cross-impact.

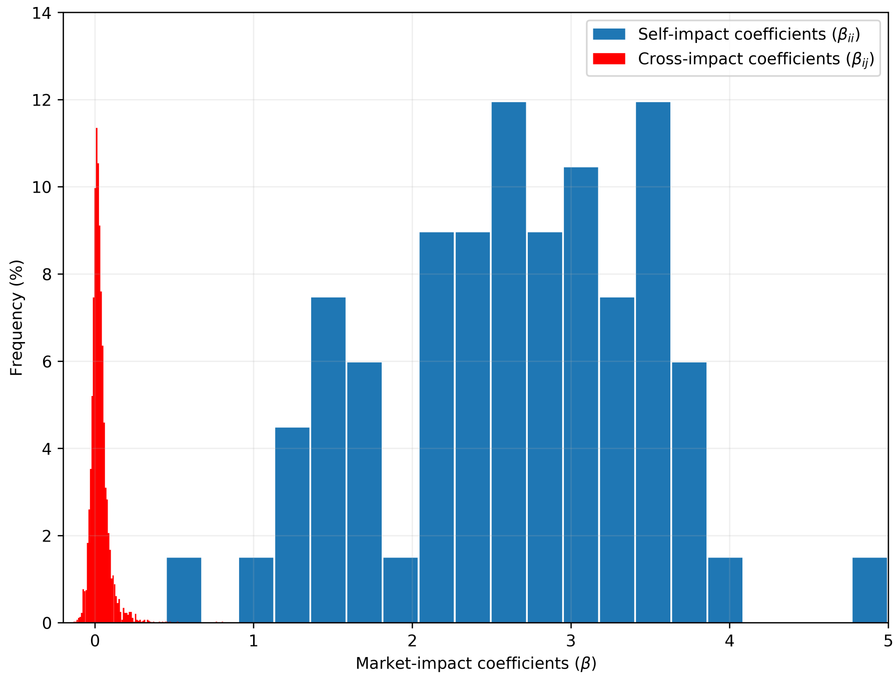

Figure 3.5: Distribution of the diagonal (idiosyncratic-impact) coefficients β_{ii} (blue) and off-diagonal (cross-impact) coefficients β_{ij} (red) estimated from (3.25) (Jan-2008: Jun-2010).

Since the correlation in order flows has not been taken into account, these coefficients cannot be interpreted as evidence of cross-impact. We delve further into this by conducting a stepwise regression analysis and a principal component analysis of order flow imbalance and returns.

3.3.3 Common components in order flow imbalance

As shown in Section 3.2.2, failing to account for the presence of common factors in order flow leads to an identification problem and makes it difficult to draw conclusions regarding the presence of cross-impact as a phenomenon distinct from correlation in order flow.

We account for the commonality in order flow by performing a principal component

	Returns			Order Flow Imbalance		
Eigenvalue	(1)	(2)	(3)	(1)	(2)	(3)
Value	21.44	1.70	1.48	18.94	1.29	1.21
Explained Variance	32.00%	2.54%	2.21%	28.27%	1.92%	1.80%
Cum. Explained Variance	32.00%	34.54%	36.75%	28.27%	30.19%	32.00%

Table 3.1: Principal component analysis of returns and order flow imbalance (Jan-2008: Jun-2010). The table shows the first three eigenvalues, the variance explained by each eigenvalue and the total cumulative explained variance.

analysis (PCA) on the correlation matrices of returns and OFIs. We adjust for the correlation in order flow using the first principal component of OFI, then test for the statistical significance of any residual cross-impact effect.

Table 3.1 reports the results of this PCA for the returns and order flow imbalances of the 67 stocks under consideration. For each of these two variables, we report the eigenvalues, the variance explained by each eigenvalue and the cumulative variance explained by the sum of the current and all previous eigenvalues. The sum of eigenvalues is equal to $N = 67$.

From Table 3.1, we see that the first eigenvalue of returns is 21.44, implying that $21.44/67 = 32.00\%$ of the total variation in 1-minute returns can be explained by a single common factor. These results suggest a stronger commonality in intraday returns than reported in previous studies: Hasbrouck and Seppi (2001) found that 21% of total variation in 15-minute returns on 30 Dow Jones stocks was accounted for by the first principal component, while Benzaquen et al. (2017) report a first principal component accounting for 23% of total variation in 5-minute returns across 275 US stocks (NASDAQ and NYSE). Column 3 in Table 3.1 shows that the second eigenvalue of returns is 1.70, explaining 2.54% of the variance, while column 4 of the same table shows that the third eigenvalue explains 2.21% of the variance. Altogether, the first three principal components of returns account for 36.75% of the total

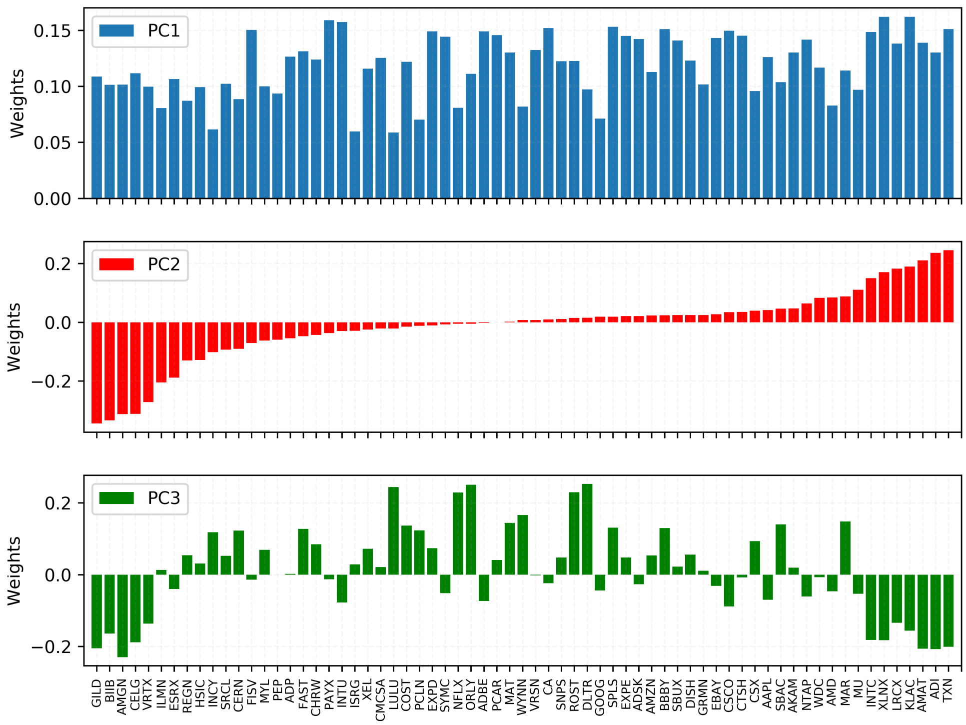

Figure 3.6: First, second, and third principal components (V^1_{OFI}, V^2_{OFI}, and V^3_{OFI}) of 1-minute order flow imbalance (OFI) for NASDAQ-100 (Jan-2008: Jun-2010).

variance, which is higher than the value 27.87% found by Hasbrouck and Seppi (2001).

The first eigenvalue of OFI suggests strong commonality in order flow. The first eigenvalue of OFI shows that 28.27% of the total variance in 1-minute OFI across all stocks can be explained by a single common factor. This is much higher than the 13.5% found by Hasbrouck and Seppi (2001) on a 15-minute interval, although the comparison needs to be taken with caution since the authors used a different variable to summarise the order flow. Column 6 in Table 3.1 shows that the second eigenvalue of OFI is 1.29, explaining 1.92% of the variance, while the third eigenvalue is 1.21, explaining 1.80% of the variance. Altogether, the first three eigenvalues of OFI account for 32.00% of the total variance, which is smaller than

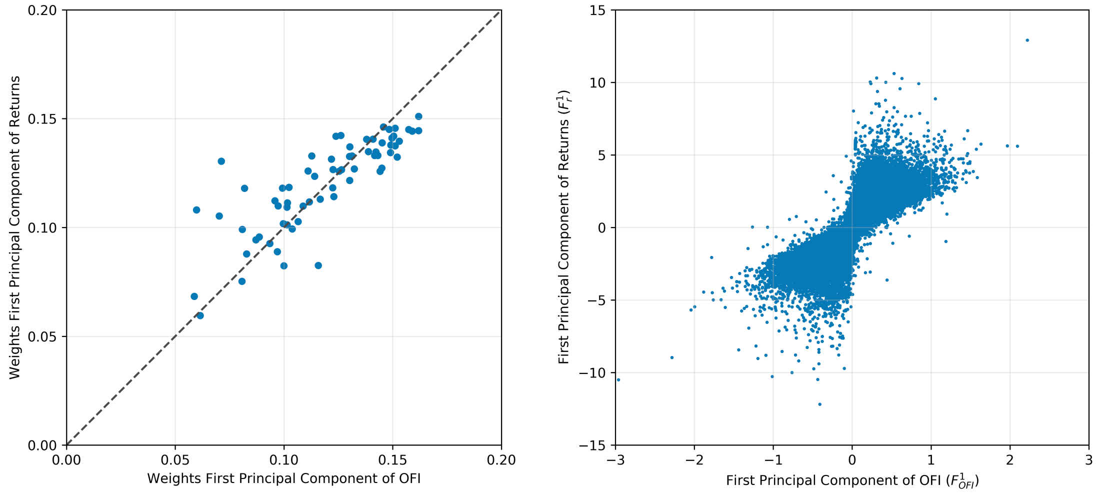

Figure 3.7: First principal component (PC) of 1-minute returns (F_r^1) vs. first principal component of 1-minute order flow imbalance (F_{OFI}^1). Left: loadings of the first PC for returns vs first PC for order flow imbalance. Right: time series of $F_{r,t}^1$ vs $F_{\mathrm{OFI},t}^1$ (Jan-2008: Jun-2010). The correlation coefficient is 87%.

what we found for returns. As a comparison, Hasbrouck and Seppi (2001) found that the first three eigenvalues of their order flow variable accounted for only 20.50% of the total variance.

Figure 3.6 displays the first three principal components of order flow imbalance: $\boldsymbol{V}_{\mathbf{OFI}}^1$, $\boldsymbol{V}_{\mathbf{OFI}}^2$, and $\boldsymbol{V}_{\mathbf{OFI}}^3$. The first principal component, whose loadings across all stocks are positive, clearly represents a 'level' effect. The second principal component is a long/short market neutral portfolio with loadings of opposite sign for pharmaceuticals and semiconductors, while the third principal component also shows clustering of coefficients across sectors.

The left panel in Figure 3.7 shows that the first principal component of order flow imbalance is strongly aligned with that of returns. The corresponding factor processes $F_{r,t}^1$ and $F_{\mathrm{OFI,t}}^1$, are obtained by projecting the returns and order flow vectors onto their respective first

	$r_{QQQ,t}$	$r_{SPY,t}$	$F^1_{r,t}$
$r_{QQQ,t}$	1.00		
$r_{SPY,t}$	0.87	1.00	
$F^1_{r,t}$	0.91	0.88	1.00
	$\text{OFI}_{QQQ,t}$	$\text{OFI}_{SPY,t}$	$F^1_{\text{OFI,t}}$
$\text{OFI}_{QQQ,t}$	1.00		
$\text{OFI}_{SPY,t}$	0.61	1.00	
$F^1_{\text{OFI,t}}$	0.82	0.62	1.00

Table 3.2: Correlation matrices of returns (upper panel) and order flow imbalance (OFI) (lower panel). The upper panel shows the correlation matrix of QQQ ETF returns ($r_{QQQ,t}$), SPY ETF returns ($r_{SPY,t}$), and the first principal component of returns ($F^1_{\text{r,t}}$). The lower panel shows the correlation of the order flow imbalance of the QQQ ($\text{OFI}_{QQQ,t}$) and SPY ETFs ($\text{OFI}_{SPY,t}$) with the first principal component of OFIs ($F^1_{\text{OFI,t}}$) (Jan-2008: Jun-2010).

principal components $\boldsymbol{V}^1_r$ and $\boldsymbol{V}^1_{\textbf{OFI}}$, scaled by their respective first eigenvalues λ^1_r and $\lambda^1_{\textbf{OFI}}$: $F^1_{\text{r},t} = \boldsymbol{V}^{1\,T}_{\mathbf{r}}(\mathbf{r}_t - \alpha_r)/\sqrt{\lambda^1_r}$, $F^1_{\text{OFI},t} = \boldsymbol{V}^1_{\textbf{OFI}}{}^T(\mathbf{ofi}_t - \alpha_{OFI})/\sqrt{\lambda^1_{OFI}}$ and are also very highly correlated, with a correlation coefficient of 87.26%. The right panel in Figure 3.7 shows a scatter plot of these factors. The correlation coefficient of the second principal component of returns and the second principal component of OFIs is 53.88%, while the correlation coefficient of the third principal components is 37.16%.

To better interpret the first principal components of returns and OFIs we examine their correlation with the returns and OFIs of the QQQ and SPY ETFs, which track the NASDAQ-100 and the $S\&P$ 500, respectively. The upper panel of Table 3.2 shows the correlation matrix of returns, including the returns of the QQQ ETF ($r_{QQQ,t}$) the returns of the SPY ETF ($r_{SPY,t}$) and the first principal component of returns ($F^1_{r,t}$). We can see from this table that the first principal component of returns is strongly positively correlated with the returns of the overall NASDAQ-100 and $S\&P$ 500 indices, with correlation coefficients of 91.37% and 88.16%, respectively. Similarly, the lower panel of Table 3.2 shows the correlation of order flow imbalance for two large ETFs, QQQ ($\text{OFI}_{QQQ,t}$) and SPY ($\text{OFI}_{SPY,t}$) with the common factor in order flow ($F^1_{\text{OFI,t}}$). This common factor is strongly correlated with the overall OFI of the NASDAQ-100, with a correlation coefficient of 82.00%, and with the OFI of the $S\&P$ 500, with a correlation coefficient of 61.80%. Therefore, one can interpret this common factor as the 'aggregate order flow imbalance'.

3.3.4 Is there additional explanatory power in cross-impact terms?

In the previous section we found that 32% of the explained variance in returns and 28.27% of the explained variance in OFI is accounted for by their respective first principal components. The question we ask ourselves next is: do we still need cross-impact coefficients once this commonality has been taken into account?

We use a two-step procedure. In the first step, we use a one factor model for the order flow imbalance, where the common factor is given by $F^1_{\text{OFI},t}$:

$$\mathbf{ofi}_{\boldsymbol{t}} = \alpha + \boldsymbol{V}^1_{\mathbf{OFI}} F^1_{\text{OFI},t} + \eta_{\boldsymbol{t}}. \tag{3.26}$$

The residual term $\eta_{\boldsymbol{t}}$ represents the order flow not accounted for by the common factor $F^1_{\text{OFI,t}}$. We will refer to $\eta_{\boldsymbol{t}}$ as the 'idiosyncratic' component of order flow.

In a second step, we regress stock returns on the OFI's common factor $F^1_{\text{OFI},t}$, the residuals from the regression in step 1 for stock i ($\eta_{i,t}$) and the residuals from the regression in step 1 for all other stocks $j \neq i$ ($\eta_{j,t}$):

$$\boldsymbol{r_t} = \mu + \gamma F^1_{\text{OFI},t} + \Phi \eta_{\boldsymbol{t}} + \xi_{\boldsymbol{t}}, \tag{3.27}$$

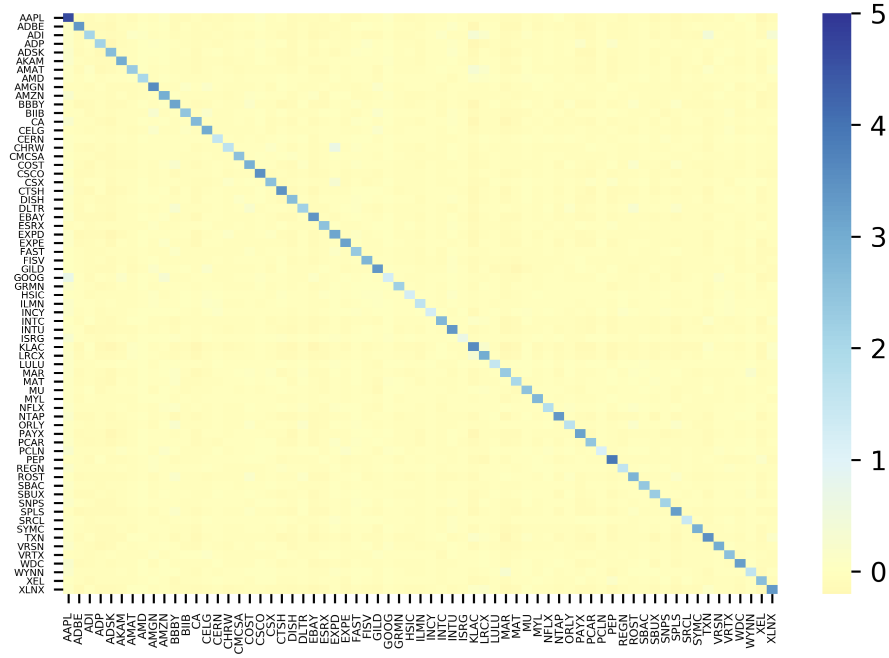

Figure 3.8: Heat map of the matrix Φ estimated from Eq. (3.27) (Jan-2008: Jun-2010).

where γ is a vector of impact coefficients due to the common component in order flow, Φ represents the impact of the idiosyncratic component of order flow η_t. In this way we are able to disentangle the impact on the returns of stock i due to common order flow, the stock's idiosyncractic order flow, and the idiosyncratic order flow of all other stocks.

Presence of cross-impact would translate in this setting into off-diagonal coefficients φ_{ij} which are statistically significant and of an economically significant magnitude when compared to a stock's idiosyncratic coefficient φ_{ii}. Figure 3.8 shows a heat map of the market impact coefficient matrix Φ, estimated from (3.27), using the standard-error correction for generated regressors proposed in Murphy and Topel (1985). This is quite similar to Figure

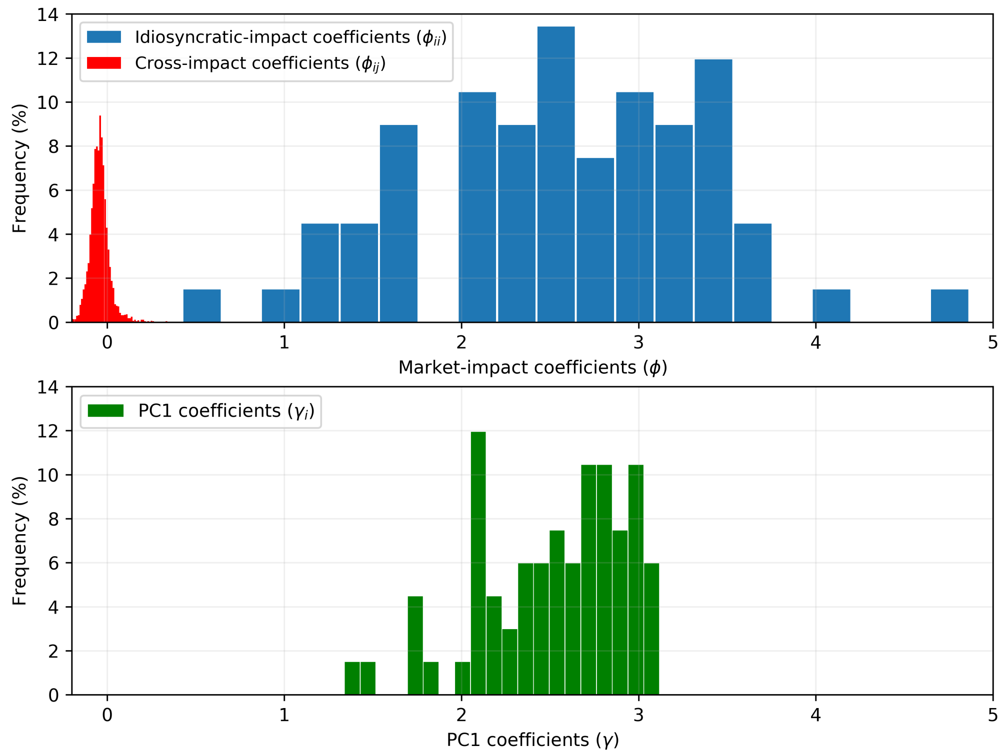

Figure 3.9: Estimation results for Eq. (3.27) (Jan-2008: Jun-2010). Upper panel: distribution of the estimated diagonal (idiosyncratic-impact) coefficients φ_{ii} (blue) and off-diagonal (cross-impact) coefficients φ_{ij} (red). Lower panel: coefficients γ_i on common order flow factor (first principal component of OFI).

3.4, and once again shows that the off-diagonal terms, representing the impact of a stock's order flow on another stock's returns, once the OFI's first common factor has been taken into account, are much smaller in magnitude compared to the diagonal terms, representing a stock's idiosyncratic impact. Figure 3.9 presents the relative frequency distributions of the diagonal "idiosyncratic-impact" coefficients φ_{ii} (upper panel - blue), the off-diagonal " cross-impact" coefficients φ_{ij} (upper panel - red), together with the frequency distribution of the coefficients of the OFI's common factor γ_i (lower panel - green). The mean diagonal idiosyncratic-impact coefficient φ_{ii} is 2.57 with a standard deviation of 0.77, while the mean off-diagonal cross-impact coefficient φ_{ij} is -0.039 with a standard deviation of 0.06, and the average coefficient for the first principal component γ_i is 2.50, with a standard deviation of 0.39. This means that once we account for the first principal component, off-diagonal terms not only remain very small in magnitude when compared with the diagonal terms, but they are also very small when compared with the coefficients on the first principal component of OFIs.

The diagonal coefficients φ_{ii} and all coefficients γ_i on the common factor are statistically significant at the 1% significance level. Also, some off-diagonal coefficients (64.36%) remain statistically significant at the 1% significance level. In order to test whether or not we can reject the null hypothesis that the off-diagonal Φ coefficients are jointly equal to zero for any of the 67 regressions, we compare the model in (3.27), with the nested model in which only a stock's own OFI and the common order flow factor F^1_{OFI} are taken into account:

$$\boldsymbol{r_t} = \mu + \gamma F^1_{\text{OFI},t} + \boldsymbol{diag}(\Phi)\eta_{\boldsymbol{t}} + \xi_{\boldsymbol{t}}, \tag{3.28}$$

where $\boldsymbol{diag}(\Phi)$ is a vector containing only the diagonal elements of the impact matrix Φ. Using an F-test we cannot reject the null hypothesis that the off-diagonal coefficients are jointly equal to zero for any of the 67 regressions. However, the increase in explanatory power that one achieves by using model (3.27) instead of the model (3.28) with cross-impact with thousands of additional parameters is very small. This is shown in Figure 3.10, where the upper panel presents the histogram of adjusted R-squared when a stock's returns are regressed on a stock's idiosyncratic OFI and the common OFI factor (3.28). In this case, the average adjusted R-squared is 43.31%, with a standard deviation of 9.8 percentage points. The lower panel of Figure 3.10 shows the histogram of adjusted R-squared when a stock's returns are

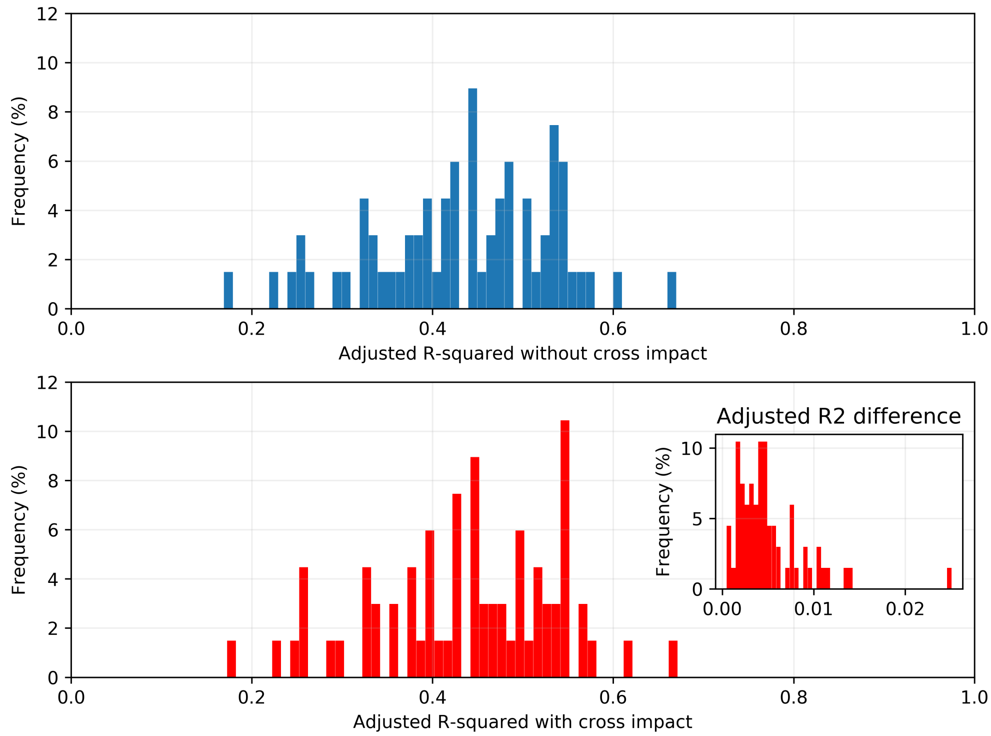

Figure 3.10: Histograms of adjusted R-squared from regressions (3.28) and (3.27) (Jan-2008: Jun-2010). The upper panel shows the R-squared from Eq. (3.28), where a stock's returns are regressed on the stock's own residual OFI and the common factor in OFI (no cross-impact). The lower panel shows the R-squared from Eq. (3.27), where a stock's returns are regressed on the stock's own OFI, the common factor in OFI, and all other stocks' OFIs. The smaller histogram in the lower panel shows the difference in adjusted R-squared between the model with cross-impact in Eq. (3.27) and the model without cross-impact in Eq. (3.28).

regressed on a stock's idiosyncratic OFI, the first principal component of OFIs, and all other stocks' idiosyncratic OFIs (3.27). We can see that the increase in explanatory power given by adding the order flow of all other stocks is minimal: the average adjusted R-squared is 43.84%, with a standard deviation of 9.8 percentage points. The inset graph in the lower panel of Figure 3.10 shows the difference in adjusted R-squared between (3.27) and (3.28). This graph shows that once a stock's own order flow and the first principal component of

OFI have been taken into account, adding the order flow of all other stocks as explanatory variables only increases the adjusted R-squared by 0.5 percentage points, on average.

Thus, despite being jointly statistically significant, the cross-impact coefficients φ_{ij} in (3.27) are not of an economically significant magnitude. The inclusion of all other stocks' OFIs only increases the explanatory power by 0.5 percentage points, on average, but at the same time the number of coefficients to be estimated increases from $67 \times 2 = 134$ to $67^2 = 4,489$. The red histogram in the upper panel of Figure 3.9 also shows that the interpretation and use of these cross-impact terms is problematic. We can see that once the correlation in order flow is taken into account, most cross-impact coefficients change sign, and become *negative*. Indeed, while the percentage of negative off-diagonal terms from regression (3.25) is only 23.09%, this percentage increases to 84.46% when we include the first principal component of OFIs in the regression.

The results of this section clearly suggest that the main determinants of execution costs are a stock's own order flow imbalance and the common component in OFI, and that once these two factors have been accounted for, the additional contribution of all other stocks' OFIs is negligible. Thus, accounting for commonality in the order flow seems to be a more effective way to build multi-asset market impact models than introducing cross-impact terms, whose contribution to total impact is very small and whose interpretation is ambiguous. This leads to a more parsimonious approach for modelling multi-asset market impact, in which the number of parameters to be estimated is drastically reduced.

3.3.5 Instability of cross-impact coefficients

Multivariate market impact models rely on the calibration of the matrix of cross-impact coefficients and its stability over time. Stability of the off-diagonal cross-impact coefficients in Φ is perhaps more important than their statistical significance. Indeed, cross-impact terms whose sign changes over time pose problems for interpretation and use.

In the previous section we found that the vast majority of the off-diagonal terms in Φ become negative once we take into account the commonality in OFIs. To test whether or not these coefficients are stable over time, we divide our sample into 30 monthly sub-samples. For each sub-sample, we perform a PCA on returns and OFIs and repeat the two-step process of (3.26)-(3.27).

Figure 3.11 shows the evolution of the idiosyncratic-impact coefficient φ_{ii} of Apple Inc. (AAPL) (upper panel) and the cross-impact coefficient φ_{ij} of Costco Wholesale Corporation (COST) (lower panel) over a one-month moving window with shaded 99% confidence intervals. We can clearly see from the lower panel how the cross-impact coefficient is unstable and changes sign randomly through time.

Similarly, Figure 3.12 shows the evolution of the idiosyncratic-impact coefficient φ_{ii} of Apple Inc. (AAPL) (upper panel) and the cross-impact coefficient φ_{ij} of O'Reilly Automotive Inc. (ORLY) (lower panel) over a one-month moving window. Once again, the cross-impact coefficient is unstable and changes sign randomly through time.

These results clearly show that cross-impact coefficients are not stable over time, even for liquid 'blue chip' stocks. Thus, a cross-impact model that is based on the calibration of Φ under these circumstances may lead to spurious results. Together with the results from the previous section, this suggests that instead of focusing on cross-impact, one should focus on

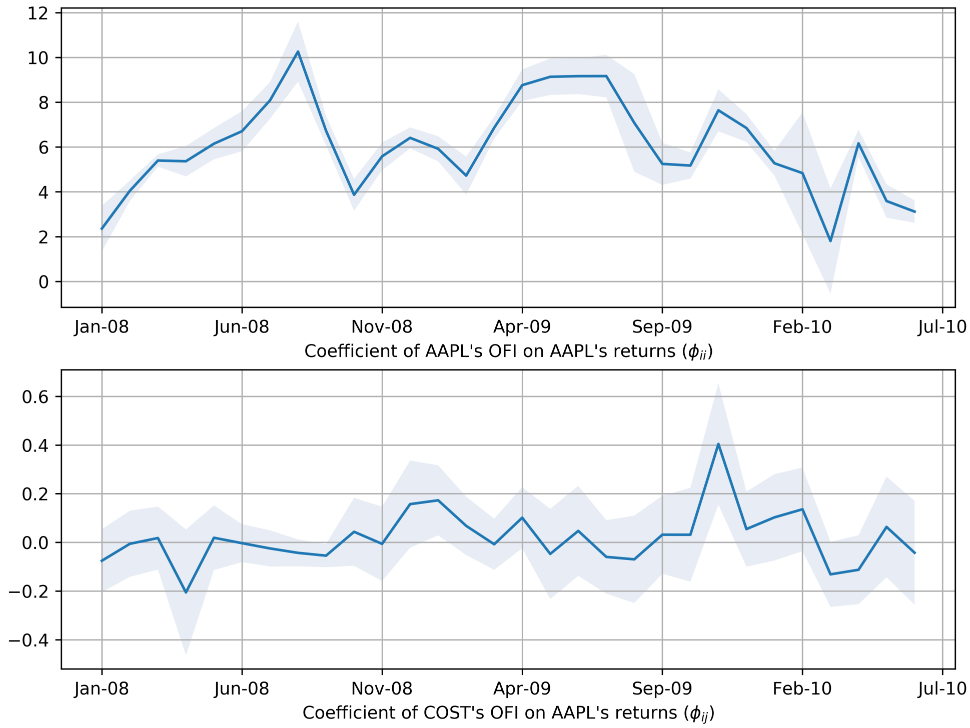

Figure 3.11: Evolution of the AAPL's idiosyncratic-impact coefficient φ_{ii} (upper panel) and cross-impact coefficient of COST's OFI on AAPL's returns φ_{ij} (lower panel) from Eq. (3.25), computed over a one-month moving window (Jan-2008: Jun-2010). Shaded blue bands represent 99% confidence intervals computed using the standard-error correction for generated regressors proposed in Murphy and Topel (1985). The coefficient φ_{ij} is unstable and changes sign randomly.

understanding and modelling the correlation in OFIs. One can then simply use the common OFI factor to extend the univariate market impact model without the need for introducing cross-impact terms.

3.4 Application: analysis of portfolio execution costs

We now illustrate how our multi-asset market impact model may be used to analyse the execution costs of portfolio trades. Most of the literature on transaction cost analysis (TCA)

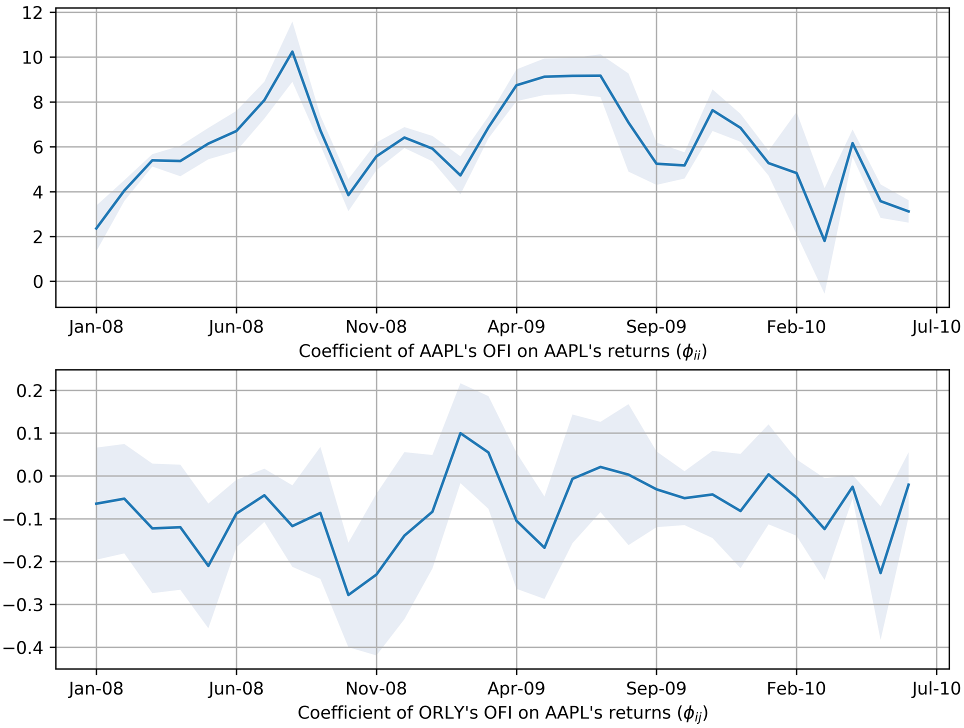

Figure 3.12: Evolution of AAPL's idiosyncratic-impact coefficient φ_{ii} (upper panel) and the cross-impact coefficient of ORLY's OFI on AAPL's returns φ_{ij} (lower panel) from Eq. (3.25), computed over a one-month moving window (Jan-2008: Jun-2010). Shaded bands represent 99% confidence intervals computed using the standard-error correction for generated regressors proposed in Murphy and Topel (1985). The coefficient φ_{ij} is unstable and changes sign randomly.

focuses on 'expected' transaction costs. This corresponds to an *unconditional* analysis of execution costs. However, our model also allows to estimate *conditional* execution costs given a forecast for market order flow. We discuss both concepts and show how their difference embodies the notion of 'market timing', a notion which is well known to traders.

3.4.1 Unconditional analysis of portfolio execution costs

The multi-asset model proposed in the previous section has an immediate application to portfolio transaction cost analysis (TCA). Equation (3.27), and its nested equivalent (3.28), provide the expected execution cost in basis points of each asset i. Let $\boldsymbol{q}$ denote an $N \times 1$ vector of position sizes (in AV units):

$$\boldsymbol{q} \equiv \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{bmatrix}. \tag{3.29}$$

To arrive at an *unconditional* estimate of transaction costs, we note that $E[F^1_{\text{OFI,t}}] = 0$, so we can interpret $\boldsymbol{q}$ as the expected order flow imbalance during execution. We can decompose $\boldsymbol{q}$ into its projection along the first principal component of order flow and an idiosyncratic component:

$$\boldsymbol{q} = F^1_{OFI}(q)\boldsymbol{V}^1_{\text{OFI}} + \eta(\boldsymbol{q}) \qquad \text{where} \qquad F^1_{OFI}(\boldsymbol{q}) = \boldsymbol{q}^T\boldsymbol{V}^1_{\text{OFI}}. \tag{3.30}$$

This leads to an expected impact of:

$$\boldsymbol{r_t} = \mu + \gamma\boldsymbol{q}^T\boldsymbol{V}^1_{\text{OFI}} + \Phi\eta(q). \tag{3.31}$$

Multiplying (3.31) by $\boldsymbol{q}$, we obtain the total execution cost:

$$\boldsymbol{q^T r_t} = \underbrace{\boldsymbol{q^T}\mu}_{\text{Trend}} + \underbrace{(\boldsymbol{q^T}\gamma)(\boldsymbol{q}^T\boldsymbol{V}^1_{\text{OFI}})}_{\text{Common}} + \underbrace{\boldsymbol{q^T}\Phi(\boldsymbol{q} - (\boldsymbol{q}^T\boldsymbol{V}^1_{\text{OFI}})\boldsymbol{V}^1_{\text{OFI}})}_{\text{Idiosyncratic}}, \tag{3.32}$$

as a sum of three terms:

- a linear term $\boldsymbol{q^T}\mu$, due to a possible price trend during execution;

- a quadratic cost $(\boldsymbol{q}^T\gamma)(\boldsymbol{q}^T\boldsymbol{V}^1_{\text{OFI}})$ corresponding to the projection of the portfolio $\boldsymbol{q}$ on the common component of the order flow;

- a quadratic cost corresponding to the idiosyncratic component of order flow.

What is noteworthy is that any of these three terms may be zero. For example, as the first principal component $\boldsymbol{V}^1_{\textbf{OFI}}$ (displayed in Figure 3.6) has positive weights roughly reflecting a market 'index', certain long-short portfolios $\boldsymbol{q}$ may satisfy $\boldsymbol{q}^T\boldsymbol{V}^1_{\textbf{OFI}} = 0$ in which case the second term in (3.32) is zero and all of the impact is due to the idiosyncratic component.

Conversely, if $\boldsymbol{q}$ is aligned with $\boldsymbol{V}^1_{\textbf{OFI}}$ (which may be the case for an index portfolio), the last term in (3.32) is zero, and the common component in order flow is the only driver of execution costs.

The following examples illustrate the difference between estimating portfolio transaction costs using (3.32), with zero off-diagonal elements in Φ, versus a standard single-asset TCA model.

Consider the case of two portfolios:

- a long-only \$5m equally-weighted portfolio of Apple Inc. (AAPL) and Intel Corporation (INTC), corresponding to 10% of AV in each asset; and

- a long-short \$5m portfolio consisting of a \$2.5m long position in AAPL and a \$2.5m short position in INTC.

Note that the positions in the two portfolios have the same size but different signs.

If we use a single-asset impact model such as (3.1), the expected execution cost is the same for both portfolios, and equal to the sum of the expected execution cost of the two positions, leading to an estimated execution cost of 10.58 basis points (bps), or around $\$5,293$.

In the multi-asset model (3.32), execution costs of the two portfolios are different, since they are computed at the *portfolio* level and take into account the contribution of the order flow common factor. In this case, the expected execution cost of the long-only portfolio is around 9.82bps or \$4, 910, and consists of $\sim$ 0.008bps, or \$4 due to the trend component, 5.13bps, or \$2, 566 due to the common component, and 4.68bps, or \$2, 340 due to the idiosyncratic component. The expected portfolio execution cost for the long-short portfolio is smaller and equal to $\sim$ 3.71bps or \$1, 853, and consists of -0.004bps, or $-\$2$ due to the trend component, 0.97bps, or $-\$485$ due to the common component, and 4.68bps, or \$2, 340 due to the idiosyncratic component. The smaller execution costs for the long-short portfolio are due to the fact that the principal components of AAPL and INTC are almost equal in magnitude, and so their impact coefficients γ cancel out at a portfolio level.

3.4.2 Conditional analysis of portfolio execution costs: market timing

More interestingly, we can also use our model to perform a *conditional* execution cost analysis, which is related to the issue of *market timing*.

Assume we are able to arrive at a forecast $\widehat{F}^1_t = E_t[F^1_{\mathrm{OFI,t}}]$ for the common component of market order flow imbalance. Conditional on this information, the expected order flow imbalance, including the agent's own trade, is $\boldsymbol{q} + \widehat{F}^1_t \boldsymbol{V}^1_{\mathbf{OFI}}$. This changes the second term in (3.31) and leads to the following *conditional* estimate of execution cost:

$$\boldsymbol{q}^{\boldsymbol{T}}\boldsymbol{r_t} = \boldsymbol{q}^{\boldsymbol{T}}\mu + (\boldsymbol{q}^{\boldsymbol{T}}\gamma)(\boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}} + \widehat{F}^1_t) + \boldsymbol{q}^{\boldsymbol{T}}\Phi(\boldsymbol{q} - (\boldsymbol{q}^T\boldsymbol{V}^1_{\mathrm{OFI}})\boldsymbol{V}^1_{\mathrm{OFI}}). \qquad (3.33)$$

We thus observe that the conditional estimate for the second term is affected by the 'market' order flow: in line with traders experience, going 'against' the market order flow or in the same direction does not lead to the same impact. In particular, the second term can be zero if

$$\boldsymbol{q}^T \boldsymbol{V}^1_{\mathrm{OFI}} + \widehat{F^1_t} = 0, \tag{3.34}$$

which corresponds to an agent trading in the opposite direction to the aggregate order flow imbalance.

This observation is consistent with traders experience of the importance of 'market timing' of trades. For example, it implies that although on average a buy trade will increase the price, a buy trade executed during a market sell-off may end up having zero, or even negative impact (i.e. yield a better price than the arrival price).

Let us consider the same long-only \$5m equally-weighted portfolio of AAPL and INTC from the previous section, and consider the execution of this portfolio under two different forecasts $\widehat{F^1_t}$ for the common component of market order flow imbalance:

- $\widehat{F^{1,h}_t} = 0.9$ corresponding to the execution of the portfolio during a strong upward trend;
- $\widehat{F^{1,l}_t} = -0.9$, corresponding to the execution of the portfolio during a market sell-off.

In the first case, the buy trade adds to the already strongly positive market order flow imbalance, resulting in a portfolio execution cost of 98.23bps, or \$49, 115. In the second case, the trader executes the buy trade against the strongly negative market order flow imbalance, resulting in a portfolio execution cost of −78.58bps, or −\$39, 290.

3.5 Discussion

Recent studies on market impact have introduced the notion of cross-impact to explain the positive covariation between the returns of an asset and the order flow of other assets. We have first shown, through several examples, that one cannot use the existence of such positive covariance terms to infer the existence of cross-impact. This positive covariance may be explained, without invoking the notion of cross-impact, by simply noticing that order flow imbalances are correlated across stocks, a phenomenon which arises naturally from multi-asset trading strategies.

We have provided empirical evidence to support our explanation: we have shown that returns and order flows are driven by highly correlated common factors, and that this commonality in order flow is the dominant contribution to the covariance between order flow imbalance and return across assets.

Our results demonstrate that the main determinants of execution costs are an asset's own order flow (OFI) and a market order flow factor common across stocks. Once this common factor in order flow is accounted for, introducing cross-impact terms provides little or no additional explanatory power. We are therefore unable to uncover any empirical justification for including cross-impact mechanisms in market impact models. Our results also show that once the commonality in order flow has been taken into account, nearly 85% of cross-impact terms become *negative*, and that the sign of these cross-impact terms changes randomly through time. The fact that cross-impact coefficients are small in magnitude and unstable over time poses serious problems for their interpretation and use.

Our findings imply that a simpler and more parsimonious approach which takes into account the commonality in order flow seems to be a more effective way to build multi-asset market impact models than introducing cross-impact coefficients whose statistical significance and interpretation are not clear.

Finally, we have shown that our framework may be used for the estimation of portfolio execution costs. *Unconditional* estimates using our framework differ from single-asset TCA models' estimates, since they take into account a common component in order flow across assets. *Conditional* estimates allow us to include information on the direction of market order flow, and provide a quantitative estimate of 'market timing' on execution costs.

References

Alfonsi, A., Fruth, A., Schied, A., 2010. Optimal execution strategies in limit order books with general shape functions. Quantitative Finance 10, 143–157.

Alfonsi, A., Klöck, F., Schied, A., 2016. Multivariate transient price impact and matrix-valued positive definite functions. Mathematics of Operations Research 41, 914–934.

Almgren, R., 2012. Optimal trading with stochastic volatility and liquidity. SIAM Journal on Financial Mathematics 3, 163–181.

Almgren, R., Chriss, N., 2001. Optimal execution of portfolio transactions. Journal of Risk 3, 5–40.

Almgren, R., Thum, C., Hauptman, H., Li, H., 2005. Direct estimation of equity market impact. Risk 18, 58–62.

Amihud, Y., Mendelson, H., 1980. Dealership market: market making with inventory. Journal of Financial Economics 8, 31–53.

Anand, A., Irvine, P., Puckett, A., Venkataraman, K., 2012. Performance of institutional trading desks: an analysis of persistence in trading costs. Review of Financial Studies 25, 558–598.

Anand, A., Irvine, P., Puckett, A., Venkataraman, K., 2013. Institutional trading and stock resiliency: evidence from the 200-2009 financial crisis. Journal of Financial Economics 108, 773–797.

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71, 579–625.

Andrade, S. C., Chang, C., Seasholes, M. S., 2008. Trading imbalances, predictable reversals, and cross-stock price pressure. Journal of Financial Economics 88, 406–423.

Avellaneda, M., Cont, R., 2013. Close-Out Risk Evaluation: A new risk management approach for central counterparties. Technical report, Finance Concepts.

Bachelier, L., 1900. Théorie de la spéculation. Annales scientifiques de l'École Normale Superieure 17, 21–86.

Bacry, E., Iuga, A., Lasnier, M., Lehalle, C.-A., 2015. Market impact and the life cycle of investors orders. Market microstructure and liquidity 1, 1550009.

Bechler, K., Ludkovski, M., 2015. Optimal execution with dynamic order flow imbalance. SIAM Journal of Financial Mathematics 6, 1123–1151.

Benzaquen, M., Mastromatteo, I., Eisler, Z., Bouchaud, J.-P., 2017. Dissecting cross-impact on stock markets: an empirical analysis. Journal of Statistical Mechanics: Theory and Experiment 2017, 023406.

Bernhardt, D., Taub, B., 2008. Cross-asset speculation in stock markets. The Journal of Finance 63, 2385–2427.

Bershova, N., Rakhlin, D., 2013. The non-linear market impact of large trades: evidence from buy-side order flow. Quantitative Finance 13, 1759–1778.

Bertsimas, D., Lo, A., 1998. Optimal control of execution costs. Journal of Financial Markets 1, 1–50.

Bibinger, M., Hautsch, N., Malec, P., Reiss, M., 2017. Estimating the spot covariation of asset prices – statistical theory and empirical evidence. Journal of Business & Economic Statistics 0, 1–17.

Blume, M., Mackinlay, A., Terker, B., 1989. Order imbalances and stock price movements on October 19 and 20, 1987. Journal of Finance 44, 827–848.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307 – 327.

Bonart, J., Gould, M. D., 2017. Latency and liquidity provision in a limit order book. Quantitative Finance 17, 1601–1616.

Bouchaud, J.-P., Bonart, J., Donier, J., Gould, M., 2018. Trades, quotes and prices. Cambridge University Press, Cambridge, UK.

Bouchaud, J.-P., Farmer, J., Lillo, F., 2009. How markets slowly digest changes in supply and demand. Handbook of Financial Markets: Dynamics and Evolution.

Boulatov, A., Hendershott, T., Livdan, D., 2013. Informed trading and portfolio returns. Review of Economic Studies 80, 35–72.

Brown, P., Walsh, D., Yuen, A., 1997. The interaction between order imbalance and stock price. Pacific-Basin Finance Journal 5, 539–557.

Caballé, J., Krishnan, M., 1994. Imperfect competition in a multi security market with risk neutrality. Econometrica 62, 695–704.

Chakrabarty, B., Moulton, P., Trzcinka, C., 2017. The performance of short-term institutional trades. Journal of Financial and Quantitative Analysis 52, 1–26.

Chan, L., Lakonishok, J., 1993. Institutional trades and intraday stock price behavior. Journal of Financial Economics 33, 173–199.

Chan, L., Lakonishok, J., 1995. The behavior of stock prices around institutional trades. Journal of Finance 50, 1147–1174.

Chemmanur, T., He, S., Hu, G., 2009. The role of institutional investors in seasoned equity offerings. Journal of Financial Economics 94, 384–411.

Cheridito, P., Sepin, T., 2014. Optimal trade execution under stochastic volatility and liquidity. Applied Mathematical Finance 21, 1–21.

Chicayachantana, C., Jiang, P., Jiang, C., Sharma, V., 2017. Permanent price impact asymmetry of trades with institutional constraints. Journal of Financial Markets 36, 1–16.

Chordia, T., Roll, R., Subrahmanyam, A., 2000. Commonality in liquidity. Journal of Financial Economics 56, 3–28.

Chordia, T., Roll, R., Subrahmanyam, A., 2002. Order imbalance, liquidity, and market returns. Journal of Financial Economics 65, 111–130.

Chordia, T., Roll, R., Subrahmanyam, A., 2008. Liquidity and market efficiency. Journal of Financial Economics 87, 249–268.

Clark-Joseph, A. D., 2013. Three essays on trading behaviour. Doctoral dissertation Harvard University. http://nrs.harvard.edu/urn-3:HUL.InstRepos:11156818.

Collins, B. M., Fabozzi, F. J., 1991. A methodology for measuring transaction costs. Financial Analysts Journal 47, 27–36.

Cont, R., Kukanov, A., Stoikov, S., 2014. The price impact of order book events. Journal of Financial Econometrics 12, 47–88.

Cont, R., Wagalath, L., 2013. Running for the exit: distressed selling and endogenous correlation in financial markets. Mathematical Finance 23, 718–741.

Cont, R., Wagalath, L., 2016. Institutional investors and the dependence structure of asset returns. International Journal of Theoretical and Applied Finance 19, 1650010.

Curato, G., Gatheral, J., Lillo, F., 2017. Optimal execution with non-linear transient market impact. Quantitative Finance 17, 41–54.

Demsetz, H., 1968. The cost of transacting. The Quarterly Journal of Economics 82, 33–53.

Donier, J., Bonart, J., 2015. A million metaorder analysis of market impact on the bitcoin. https://arxiv.org/pdf/1412.4503.pdf.

Dufour, A., Engle, R. F., 2000. Time and the price impact of a trade. The Journal of Finance 55, 2467–2498.

Durbin, J., Koopman, S., 2012. Time series analysis by state space methods. Oxford University Press, Oxford, UK.

Easley, D., Prado, M. L., O'Hara, M., 2015. Optimal execution horizon. Mathematical Finance 25, 640–672.

Engle, R., Ferstenberg, R., Russell, J., 2012. Measuring and modeling execution cost and risk. Journal of Portfolio management 38, 14–28.

Farmer, J., Patelli, P., Zovko, I., 2005. The predictive power of zero intelligence in financial markets. Proceedings of the National Academy of Sciences of the United States of America 102, 2254–2259.

Foucault, T., Pagano, M., Röell, A., 2013. Market liquidity. Theory, evidence and policy. Oxford University Press, Oxford.

Fox, J., Monette, G., 1992. Generalized collinearity diagnostics. Journal of the American Statistical Assosiation 87 (417), 178–183.

Garman, M., 1976. Market microstructure. Journal of Financial Economics 3, 257–275.

Gatheral, J., 2010. No-dynamic-arbitrage and market impact. Quantitative Finance 7, 749–759.

George, T., Kaul, G., Nimalendran, M., 1991. Estimation of the bid–ask spread and its components: a new approach. Review of Financial Studies 4, 623–656.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.

Gode, D. K., Sunder, S., 1993. Allocative efficiency of markets with zero-intelligence traders: market as a partial substitute for individual rationality. Journal of Political Economy 101, 119–137.

Gomes, C., Waelbroeck, H., 2015. Is market impact a measure of the information value of trades? market response to liquidity vs informed metaorders. Quantitative Finance 15, 773–793.

Gross-Klussmann, A., Hautsch, N., 2013. Predicting bid-ask spreads using long memory autoregressive conditional poisson models. Journal of Forecasting 32, 724–742.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. The American Economic Review 70, 393–408.

Haavelmo, T., 1943. The statistical implications of a system of simultaneous equations. Econometrica 11, 1–12.

Haavelmo, T., 1944. The probability approach in econometrics. Econometrica 12 (Supplement), iii–vi and 1–115.

Hamilton, J. D., 1994. Time series analysis. Princeton University Press, Princeton.

Harford, J., Kaul, A., 2005. Correlated order flow: Pervasiveness, sources, and pricing effects. The Journal of Financial and Quantitative Analysis 40, 29–55.

Harris, L., 2003. Trading and Exchanges Market Microstructure for Practitioners. Oxford University Press, Oxford.

Hasbrouck, J., 1991. Measuring the information content of stock trades. Journal of Finance 46, 179–207.

Hasbrouck, J., 2007. Empirical market microstructure. Oxford University Press.

Hasbrouck, J., Seppi, D. J., 2001. Common factors in prices, order flows, and liquidity. Journal of Financial Economics 59, 383–411.

Hasbrouck, J., Sofianos, G., 1993. The trades of market-makers: an analysis of NYSE specialists. Journal of Finance 48, 1565–1595.

Hausman, J. A., Lo, A. W., MacKinlay, A. C., 1992. An ordered probit analysis of transaction stock prices. Journal of Financial Economics 31, 319–379.

Hautsch, N., Pohlmeier, W., 2001. Econometric analysis of financial transaction data: pitfalls and opportunities. Allgemeines Statistisches Archiv 86, 5–30.

Heckman, J., Pinto, R., 2015. Causal analysis after Haavelmo. Econometric Theory 31, 115–151.

Heston, S. L., Korajczyk, R., Sadka, R., 2010. Intraday patterns in the cross-section of stock returns. The Journal of Finance 65, 1369–1407.

Holthausen, R. W., Leftwich, R. W., Mayers, D., 1987. The effect of large block transactions on security prices: a cross-sectional analysis. Journal of Financial Economics 19, 237–267.

Hu, G., 2009. Measures of implicit trading costs and buy-sell asymmetry. Journal of Financial Markets 12, 418–437.

Hu, G., Jo, K., Wang, Y., Xie, J., 2018. Institutional trading and Abel Noser data. Unpublished working paper. Hong Kong Polytechnic University.

Huang, R., Polak, T., 2011. Lobster: Limit order book reconstruction system. Unpublished working paper. Humboldt-Universität zu Berlin.

Huang, R. D., Stoll, H. R., 1994. Market microstructure and stock return predictions. Review of Financial Studies 7, 179–213.

Huberman, G., Stanzl, W., 2004. Price manipulation and quasi arbitrage. Econometrica 72, 1247–1275.

Jordan, M. I., 2004. Graphical models. Statistical Science 19, 140–155.

Keim, D. B., Madhavan, A., 1996. The upstairs market for large-block transactions: analysis and measurement of price effects. Review of Financial Studies 9, 1–36.

Keim, D. B., Madhavan, A., 1997. Transaction costs and investment style: an inter- exchange analysis of institutional equity trades. Journal of Financial Economics 46, 265–292.

Keim, D. B., Madhavan, A., 1998. The cost of institutional equity trades. Financial Analysts Journal 54, 50–69.

Kempf, A., Korn, O., 1999. Market depth and order size. Journal of Financial Markets 2, 29–48.

Kissell, R., 2006. The expanded implementation shortfall: understanding transaction cost components. Journal of Trading Summer, 6–16.

Kissell, R., 2014. The Science of Algorithmic Trading and Portfolio Management. Academic Press, Elsevier.

Kraus, A., Stall, H. R., 1972. Price impacts of block trading on the New York Stock Exchange. Journal of Finance 27, 569–588.

Kyle, A. S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1335.

Lillo, F., Farmer, J., 2004. The long memory of the efficient market. Studies in Nonlinear Dynamics & Econometrics 8, 1–33.

Lillo, F., Farmer, J., Mantegna, R., 2003. Master curve for price impact function. Nature 421, 129–130.

Loeb, T. F., 1983. Trading cost: the critical link between investment information and results. Financial Analysts Journal 19, 39–44.

Madhavan, A., 2000. Market microstructure: a survey. Journal of Financial Markets 3, 205–258.

Madhavan, A., Cheng, M., 1997. In search of liquidity: block trades in the upstairs and downstairs markets. Review of Financial Studies 10, 175–204.

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? a transaction-level analysis of NYSE stocks. Review of Financial Studies 10, 1035–1064.

Madhavan, A., Sofianos, G., 1998. An empirical analysis of NYSE specialist trading. Journal of Financial Economics 48, 189–210.

Mastromatteo, I., Benzaquen, M., Eisler, Z., Bouchaud, J.-P., 2017. Trading lightly: cross-impact and optimal portfolio execution. https://ssrn.com/abstract_id=2949748 .

Mertens, L. P., 2014. Latent price impact. Unpublished working paper. IMPA. SSRN: https://ssrn.com/abstract=2481070.

Min, S., Maglaras, C., Moallemi, C. C., 2018. Cross-sectional variation of intraday liquidity, cross-impact, and their effect on portfolio execution. https://arxiv.org/pdf/1811.05524.pdf .

Mincer, J. A., Zarnowitz, V., 1969. The evaluation of economic forecasts, in J A Mincer (ed.), economic forecasts and expectations. National Bureau of Economic Research. New York. Ch.1, 1–46.

Moro, E., ans L Moyano, J. V., Gerig, A., Farmer, J., Vaglica, G., Lillo, F., Mantegna, R., 2009. Market impact and trading profile of large trading orders in stock markets. Physical Review E 80, 066102.

Murphy, K. M., Topel, R. H., 1985. Estimation and inference in two-step econometric models. Journal of Business and Economic Statistics 3, 370–379.

Obizhaeva, A., 2012. Liquidity estimates and selection bias. Available at SSRN: https://ssrn.com/abstract=
1178722 .

O'Hara, M., 1995. Market Microstructure Theory. Basil Blackwell.

Pasquariello, P., Vega, C., 2015. Strategic cross-trading in the U.S. stock market. Review of Finance 19, 229–282.

Pástor, L., Stambaugh, R. F., 2004. Liquidity risk and expected stock returns. Journal of Political Economy 111, 642–685.

Pearl, J., 2009. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, New York, second ed.

Perold, A. F., 1988. The implementation shortfall: paper versus reality. Journal of Portfolio Management 14, 4–9.

Perold, A. F., Sirri, E., 1993. The implementation shortfall: paper versus reality. Working Paper. Harvard Business School.

Pickering, G., Bull, J. M., Sanderson, D. J., 1995. Sampling power-law distributions. Tectonophysics 248, 1–20.

Pohl, M., Ristig, A., Schachermayer, W., Tangpi, L., 2017. The amazing power of dimensional analysis: Quantifying market impact. Market Microstructure and Liquidity 03, 1850004.

Predoiu, S., Shaikhet, G., Shreve, S., 2011. Optimal execution in a general one-sided limit-order book. SIAM Journal on Financial Mathematics 2, 183–212.

Puckett, A., Yan, S., 2011. The interim trading skills of institutional investors. Journal of Finance 66, 601–633.

Roll, R., 1984. A simple implicit measure of the effective bid– ask spread in an efficient market. Journal of Finance 39, 1127–1139.

Rydberg, T. H., Shephard, N., 2003. Dynamics of trade-by-trade price movements: decomposition and models. Journal of Financial Econometrics 1, 2–25.

Salem, M., Younger, J., Zhao, Z., 2018. Exploring drivers of liquidity in the US treasury market. J.P.Morgan, Liquid lunch, 06 April 2018.

Schied, A., Schöneborn, T., Tehranchi, M., 2010. Optimal basket liquidation for CARA investors is deterministic. Applied Mathematical Finance 17, 471–489.

Schneider, M., Lillo, F., 2019. Cross-impact and no-dynamic arbitrage. Quantitative Finance 19, 137–154.

Schöneborn, T., 2016. Adaptive basket liquidation. Finance and Stochastics 20, 455–493.

Smidt, S., 1971. Which road to an efficient stock market: free competition or regulated monopoly? Financial Analysts Journal 27.

Stoll, H., 1978. The supply of dealer services in securities markets. Journal of Finance 33.

Tomas, M., Mastromatteo, I., Benzaquen, M., 2020. How to build a cross-impact model from first principles: Theoretical requirements and empirical results. arXiv:2004.01624 .

Torre, N. G., 1997. BARRA Market Impact Model Handbook. BARRA Inc., Berkeley, CA.

Torre, N. G., 1999. The Market Impact Model. BARRA Research Insights, BARRA Inc.

Toth, B., Eisler, Z., Bouchaud, J.-P., 2016. The square-root impact law also holds for option markets. Wilmott 85, 70–73.

Toth, B., Lemperiere, Y., Deremble, C., Lataillade, J., Kockelkoren, J., Bouchaud, J.-P., 2011. Anomalous price impact and the critical nature of liquidity in financial markets. Physical Review X 1, 021006.

Tsoukalas, G., Wang, J., Giesecke, K., 2017. Dynamic portfolio execution. Management Science 65, 1949–2443.

Vega, J. D. L., 1688. Confusion de Confusiones: Portions Descriptive of the Amsterdam Stock Exchange. (Translation by H. Kellenbenz, Harvard University, 1957).

Wagner, W. H., Edwards, M., 1993. Best execution. Financial Analysts Journal 49, 65–71.

Walras, L., 1900. Élements d' écoonomie politique pure. Lausanne, F. Rouge, fourth ed.

Wang, S., Schäfer, R., Guhr, T., 2016a. Cross-response in correlated financial markets: individual stocks. The European Physical Journal B 89, 1–16.

Wang, S., Schäfer, R., Guhr, T., 2016b. Price response in correlated financial markets: empirical results. arXiv:1510.03205v2 .

Weber, P., Rosenow, B., 2005. Order book approach to price impact. Quantitative Finance 5, 357–364.

White, E. P., Enquist, B. J., Green, J. L., 2008. On estimating the exponent of power-law frequency distributions. Ecology 89, 905–912.

Zarinelli, E., Treccani, M., Farmer, J., Lillo, F., 2015. Beyond the square root: evidence for logarithmic dependence of market impact on size and participation rate. Market Microstructure and Liquidity 01 (02), 1550004.

A

Appendix

A.1 Execution costs using the market VWAP as benchmark

In this Appendix, we present some additional results for Chapter 2, where both signed and unsigned execution costs are computed by taking the market volume-weighted average price

(VWAP) during execution of a trade as benchmark. We define these two variables as follows:

$$\begin{aligned}
\textit{Percentage Execution Cost} \quad C_{VWAP} &\equiv \varepsilon \frac{(p_{vwap} - p_{\mathrm{VWAP}})}{p_{\mathrm{VWAP}}}, && \text{(A.1)} \\
\textit{Unsigned Percentage Execution Cost} \quad R_{VWAP} &\equiv \frac{(p_{vwap} - p_{\mathrm{VWAP}})}{p_{VWAP}}, && \text{(A.2)} \\
& && \text{(A.3)}
\end{aligned}$$

where, as before, p_{vwap} is the *trade's* volume-weighted average execution price, p_{VWAP} is the *market's* volume-weighted average execution price between the trade's arrival time and the execution time of its last transaction, and ε is the trade direction ($\varepsilon = 1$ for buy trades, and $\varepsilon = -1$ for sell trades).

A.1.1 Price changes and trade direction of individual trades

Figure A.1 presents a scatter-plot of the *unsigned* execution costs R_{WAP} standardised by daily volatility against the *signed* daily volume fraction (positive for buy trades and negative for sell trades). The plot is different from Figure 2.8 in Chapter 2, but it still shows that points are distributed across all four quadrants and that prices during execution do not move in the same direction of trades.

In particular, the joint probabilities for each quadrant are as follows:

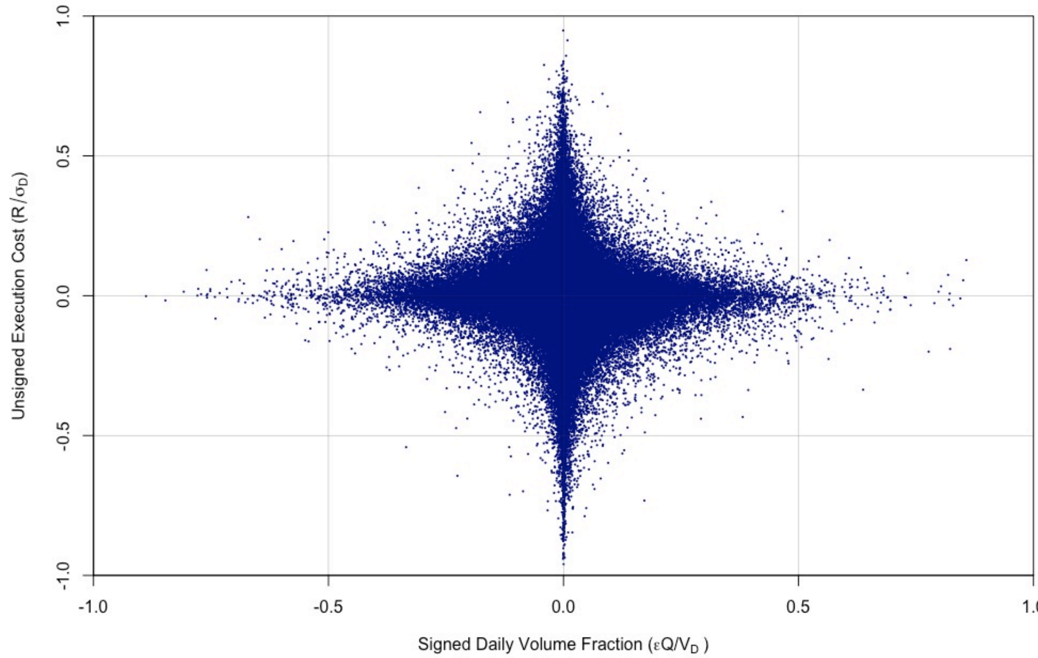

Figure A.1: Scatter-plot of the *unsigned* execution cost (using the market VWAP as benchmark price) standardised by the daily volatility R_{VWAP}/σ_D versus *signed* daily volume fraction $\varepsilon Q/V_D$ for all trades in our dataset (2008-2012).

$$\mathcal{P}(R_{VWAP} > 0, Q > 0) \quad = \quad 27.2\%, \tag{A.4}$$

$$\mathcal{P}(R_{VWAP} < 0, Q > 0) \quad = \quad 25.1\%, \tag{A.5}$$

$$\mathcal{P}(R_{VWAP} > 0, Q < 0) \quad = \quad 23.0\%, \tag{A.6}$$

$$\mathcal{P}(R_{VWAP} < 0, Q < 0) \quad = \quad 24.7\%. \tag{A.7}$$

The conditional probabilities for a buy trade using the market VWAP as benchmark are:

$$\mathcal{P}(R_{VWAP} > 0|Q > 0) = 51.9\%, \qquad \mathcal{P}(R_{VWAP} < 0|Q > 0) = 48.1\%, \qquad \text{(A.8)}$$

and for a sell trade :

$$\mathcal{P}(R_{VWAP} < 0|Q < 0) = 51.8\%, \qquad \mathcal{P}(R_{VWAP} > 0|Q < 0) = 48.2\%. \quad \text{(A.9)}$$

The conditional probabilities of equations (A.8) and (A.9) clearly show that the primary assumption of traditional impact models does not hold in our dataset even if we use the market VWAP, instead of the arrival price, as benchmark. The sign of price changes does not seem to be determined by the direction of the individual trade. In fact, we observe a *negative impact* in nearly 50% of the cases, which is even higher than the $\simeq 40\%$ we observed in Chapter 2 using the arrival price p_{start} as benchmark.

The square-root model of market impact:

$$R_{VWAP} = Y\left(\varepsilon\sigma_D\sqrt{\frac{Q}{V_D}}\right) + \xi \qquad \text{(A.10)}$$

also seem to perform worse, when we take the market VWAP as benchmark. In this case a regression yields the estimate $\hat{Y} = -0.03$ with an estimated standard error of 0.001, a t-statistic of -28.01 and significance at the 1% level. However, the coefficient of determination is only $R^2 = 0.2\%$, which is even lower than the 5.5% reported in Chapter 2. This further shows that the square-root model of market impact provides a poor explanation for the direction and magnitude of the unsigned execution costs. 99.8% of the variation across trades is given by the residual and volatility is by far the main determinant of the direction and magnitude of price variations during trade execution.

A.1.2 Price changes and order flow imbalance

We also reevaluate the relationship between trade size and price changes *conditional* on the aggregate order flow imbalance presented in section 2.3.2.

Once again, we sort the trades into 50 *equally-populated* bins for the *OFI*. We then run, for each bin k, the following conditional regression of the unsigned execution cost R_{VWAP} standardised by the daily volatility, on the daily volume fraction Q/V_D, given the OFI:

$$\left(\frac{R_{VWAP}}{\sigma_D}\right)_k = \alpha_k + \beta_{s,k}\left(\varepsilon\frac{Q}{V_D}\right)_k + \xi_k. \tag{A.11}$$

Figure A.2 shows the coefficient of determination of these regressions against the *OFI* value on which they have been conditioned. The blue line shows the results presented in section 2.3.2, where we used the arrival price as benchmark, while the red line shows the results using the market VWAP as benchmark.

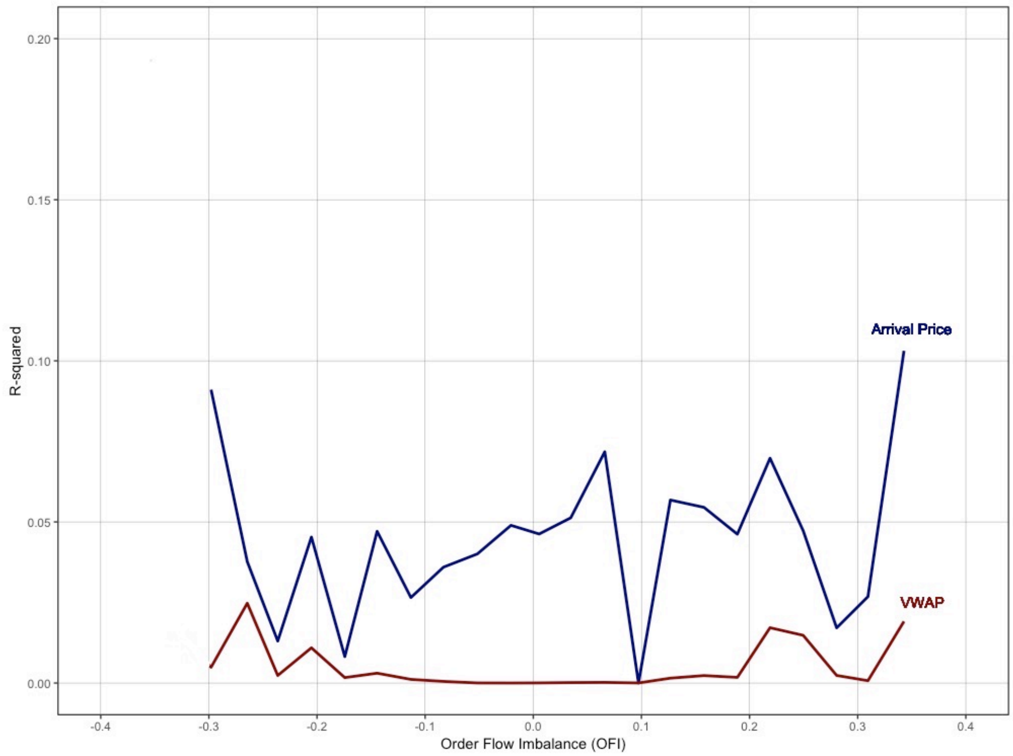

Figure A.2: R-squared for the conditional regressions of the unsigned execution cost, standardised by daily volatility, on the signed daily volume fraction, *given* the *OFI* (A.11). The figure compares the R-squared obtained from using the arrival price as benchmark for execution costs (blue line) versus the market VWAP between placement and execution (red line).

We can see how the R^2 obtained when using the market VWAP as benchmark is much lower that the R^2 obtained using the arrival price. The average R^2, for $OFI \in [-0.30, 0.30]$, is only 0.4% when using the market VWAP versus 3.9% when using the arrival price. These results once again show that price moves during execution are not determined by the characteristics of individual trades, and that the claim that individual trades *cause* the market price to move, and increase execution costs, might be spurious. Instead, as shown in Figure 2.2, a trade only affects the market price, and execution costs, through the *aggregate* order flow imbalance.

A.1.3 Execution costs and trade size conditional on trade duration

The analysis presented in section 2.4.2, showed that the relationship between execution costs and trade size disappears, when conditioning on trade size. This relationship also seem to disappears when using the market VWAP as benchmark. Figure A.3 shows four scatter-plots of (log) absolute execution costs $|C_{VWAP}/\sigma_D|$ versus (log) daily volume fraction Q/V_D conditioned on representative short- (2 minutes), medium- (5 minutes), and high-duration (389 minutes) values. The cloud of points in each of these four scatter-plots is essentially flat, and sometimes downward sloping, showing no relationship between execution costs and trade size.

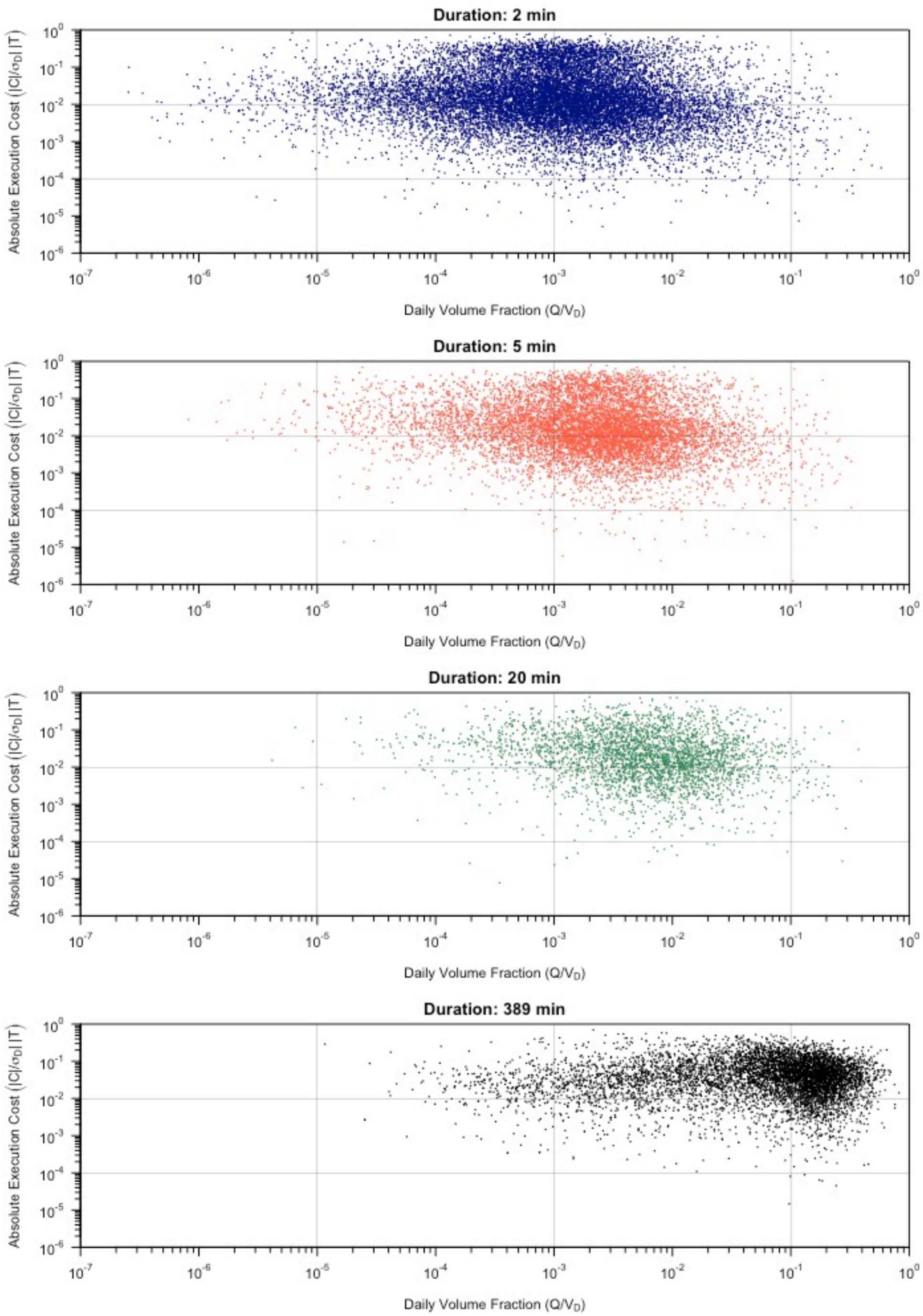

Figure A.3: Scatter-plots of (log) absolute execution cost $|C_{VWAP}/\sigma_D|$ versus (log) daily volume fraction Q/V_D conditional on four representative trade duration values: $T = 2, 5, 20$ and 389 minutes (2008-2012).

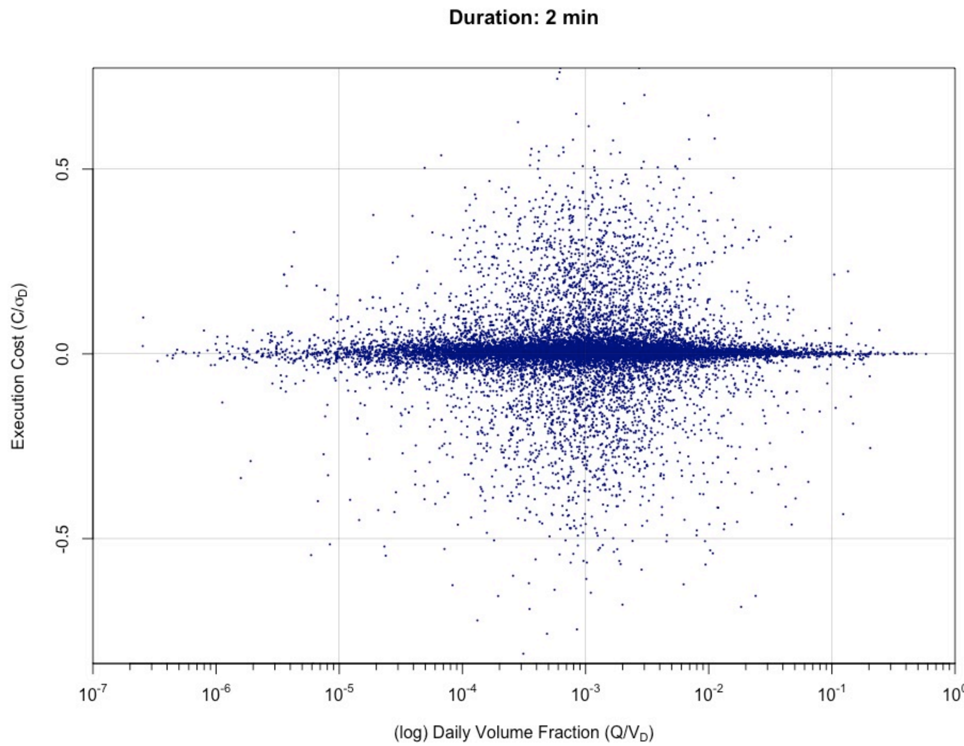

Figure A.4: Scatter-plot of execution cost C_{VWAP}/σ_D versus (log) daily volume fraction Q/V_D conditional on trade duration $T = 2$ minutes (2008-2012).

Similarly, Figure A.4, which shows the scatter-plot of signed execution costs C_{VWAP}/σ_D versus (log) daily volume fraction Q/V_D conditioned on trade duration $T = 2$ minutes, reveals no relationship between execution costs and trade size, and is similar to Figure 2.11.

The conditional logarithmic regression of the absolute execution cost $|C_{VWAP}|$, standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T, has been estimated once again, for each bin k as:

$$\log\left(\left|\frac{C_{VWAP}}{\sigma_D}\right|\right)_k = \alpha_k + \beta_{s,k} \log\left(\frac{Q}{V_D}\right)_k + \xi_k, \tag{A.12}$$

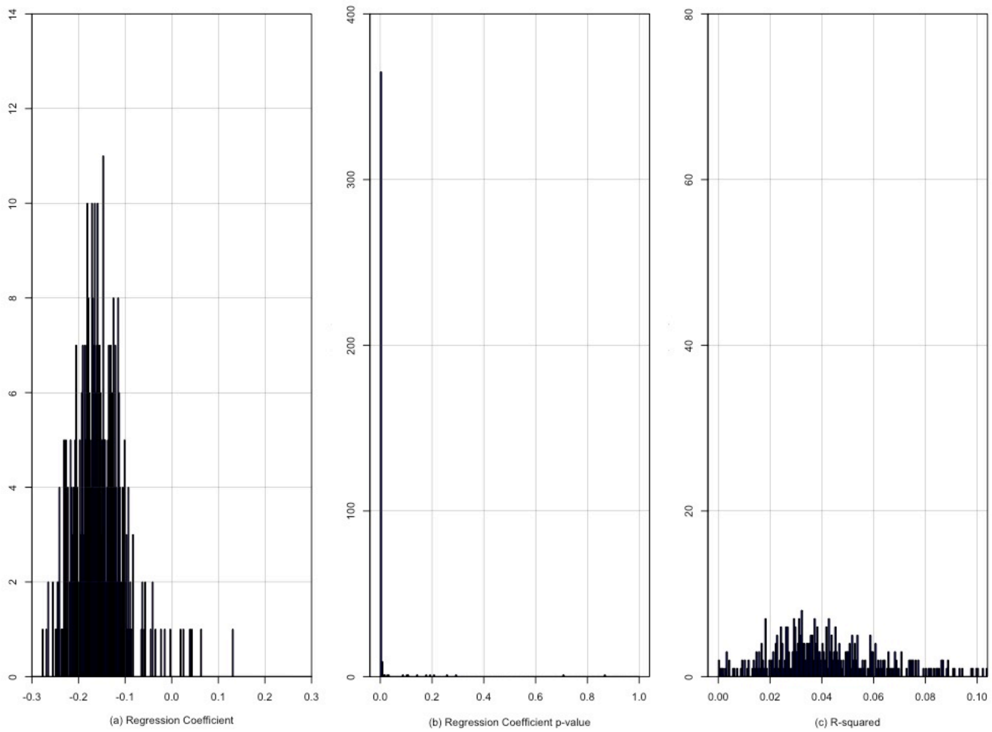

Figure A.5: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional regressions of absolute execution costs, standardised by the daily volatility, on the daily volume fraction, given trade duration (A.12).

and the conditional non-linear least square (NLS) regression for the execution cost C_{VWAP}, standardised by the daily volatility, on the daily volume fraction Q/V_D, given the trade duration T, is estimated as:

$$\left(\frac{C_{VWAP}}{\sigma_D}\right)_k = Y_k\left(\frac{Q}{V_D}\right)_k^{\beta_{s,k}} + \xi_k. \tag{A.13}$$

Figures A.5-A.6 show the histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$, their p-values and the coefficients of determination for the log regressions (A.12) and the NLS regressions (A.13), respectively. When we look at the absolute value of execution costs, $\hat{\beta}_{s,k}$

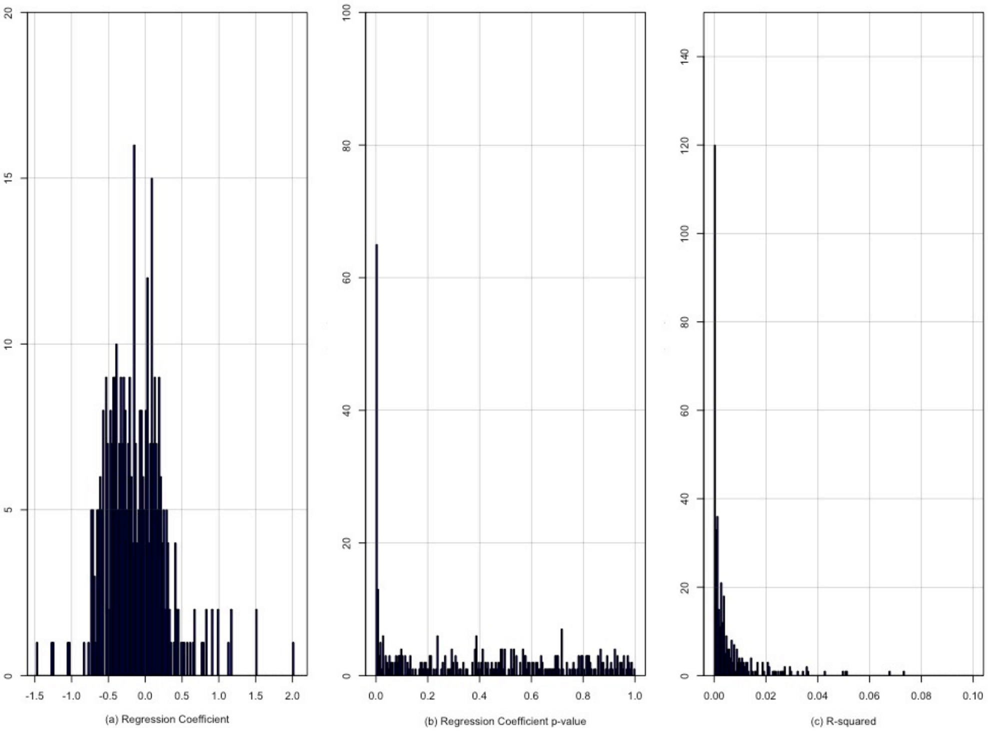

Figure A.6: Histograms of the estimated regression coefficients $\hat{\beta}_{s,k}$ (a), their p-values (b), and R-squared (c) for the conditional non-linear least square regressions of signed execution costs, standardised by the daily volatility, on the daily volume fraction, given trade duration (A.13).

is statistically significant at the 5% level in all cases versus 146 of the 389 regressions (37.5%) when considering the arrival rate as benchmark (Figure A.5, panel (b)). However, in the vast majority of cases, its estimate changes from being very close to zero (when using the arrival price as benchmark) to being negative (Figure A.5, panel (a)). On average, $\hat{\beta}_{s,k}$ is -0.16, with an average coefficient of determination of 4.3% versus 0.7% when considering the arrival price as benchmark.

The analysis of signed execution costs paints a similar picture. $\hat{\beta}_{s,k}$ is statistically significant at the 5% level in all 389 regressions (Figure A.6, panel (b)), but its value changes, once again, from very close to zero (when using the arrival price as benchmark) to negative in most cases (Figure A.6, panel (a)). On average, $\hat{\beta}_{s,k}$ is -0.13, with an average coefficient of determination of only 0.3%,[1] versus 1.2% when considering the arrival rate as benchmark. (Figure A.6, panel (c)).

[1] The coefficient of determination is not defined for non-linear least square regressions. We compute a pseudo R^2 as follows: $R^2 = 1 - [(\frac{C}{\sigma_D} - \widehat{\frac{C}{\sigma_D}})^2/(N-1)\sigma^2_{C/\sigma_D}]$, where $\widehat{\frac{C}{\sigma_D}}$ is fitted value for the execution cost , σ^2_{C/σ_D} is the variance of the execution cost, and N is the number of observations in the bin $[\mathrm{t}_k, t_{k+1})$.