Master's Thesis

Empirical study of a bid-ask model for liquid markets

Douglas M Vieira

June 2014

Advisor: Luca Mertens, PhD

impa

I dedicate this thesis to my father Indio, my mother Rita and my brother Diego, who have always supported me with pleasure in their hearts throughout my life.

*"At the end of the day,
let there be no excuses,
no explanations, no regrets."*

Steve Maraboli

Acknowledgements

I want to express my gratitude to my advisor Luca Mertens for his immense help and for believing in my work. I would also like to thank Jorge Zubelli for his great help throughout my graduation, my friends in Rio de Janeiro who have supported me and made me feel home in this beautiful city.

Furthermore, I also want to thank all those who brought insights to this work in the occasion of the poster presented in Research in Options 2013 conference in Buzios. In special, my friends Adrien Nguyen and Lakshithe Wagalath.

This entire thesis is produced with open-source software. Figures, tables and numerical analysis were produced with `Python` and presented in LaTeX. I would like to acknowledge the efforts of the many authors and maintainers of these languages and their several packages, in special `pandas` and `statsmodels`. The code used in this thesis will be available at my personal website.

Finally, I would like to thank NYSE for the free NBBO sample dataset, which served as database for this thesis.

Abstract

The present thesis is an exploratory data analysis of the high frequency dataset of NBBO quotes (national best bid and offer) for the NYSE market. The focus of the analysis is on the assumptions and results of the Markovian micro structure model introduced in Cont and de Larrard (2013). The model provides distributions for the durations between mid price changes, probabilities of mid price changes conditioned to the state of the order book and a formula describing price intraday volatility in terms of pure microstructure statistics. We verify model assumptions and compare the theoretical results with empirical data for the stocks from the Dow Jones Industrial Average Index, one of the most liquid markets in the world. In this thesis, we conclude that the model gives a good description of the market dynamics at the high frequency level, and produces consistent results with the data, including the aforementioned volatility relationship.

Resumo

Esta dissertação tem como objetivo estudar empiricamente o modelo Markoviano de microestrutura de mercado investigado por Cont and de Larrard (2013), o qual é direcionado a mercados líquidos. Entre outros resultados, o modelo é capaz de prover distribuições para as durações entre mudanças no *mid price*, probabilidades com respeito a mudanças no *mid price* condicionados ao estado do livro de ofertas e relacionar estatísticas de pura micro estrutura com a volatilidade. Utilizando uma base de dados de alta-frequência gratuitamente disponível pela NYSE, investigamos extensivamente os pressupostos do modelo e alguns dos seus resultados para as ações do Índice Dow Jones, um dos mercados mais líquidos do mundo. Nessa dissertação, concluímos que o modelo é capaz de absorver o cerne do funcionamento do mercado em alta frequência e produzir resultados consistentes, incluindo a supracitada relação com a volatilidade.

Contents

List of Figures

1 Literature Review

1.1 Introduction

A large stake of the financial markets are operated through double auction systems, informally known as order books. For instance, the average daily traded volume operated on order books at NYSE (New York Stock Exchange) is of tens of billions of dollars[1]. Due to its relevance in the trading process, the limit order book has been extensively studied in the literature and numerous models have been proposed to address the mechanics of the so-called market microstructure.

Market microstructure models are of key importance to understand price formation. Numerous microstructure characteristics are known to affect price dynamics on a broader scale Madhavan et al. (1997). Furthermore, their study has several applications, such as optimal execution Alfonsi and Schied (2010); Obizhaeva and Wang (2012); Tsoukalas et al. (2013), market making strategies Glosten and Milgrom (1985), liquidity risk Biais and Weill (2009) and transaction costs modelling Glosten and Milgrom (1985); Parlour (1998); Roll (1984).

In this thesis, we consider a Markovian model for market microstructure. This model was first introduced by Stoikov et al. (2010) and was intended to study hidden liquidity. Afterwards, it was further extended by Cont and de Larrard (2013), where various interesting theoretical results were found. In the present thesis, we study to which extent the model assumptions and theoretical results are supported by real data.

The model takes into account only the best bid and ask quantities and prices dynamics. In Cont and de Larrard (2013), the authors present two justifications for not considering the deeper levels of the order book. First, it is shown in Biais et al. (1995) that most of the order flow is directed at best bid and ask prices. Moreover, the best bid and ask dynamics are found in Cont (2001) to be the main driver for price oscillations. This simplification is very welcome both mathematically and technically, since best bid and ask data is cheaper, easier to find and usually easier to process computationally. Throughout this document, we shall refer to the best bid and ask prices and quantities simply as bid and ask prices and quantities.

The model is intended to describe highly liquid markets, where the bid-ask spread is a single tick most of the time. As in Cont and de Larrard (2013), we analyse the stocks that compound the Dow Jones Industrial Average Index, one of the most liquid markets of the world . In Subsection 2.4, it is verified whether these stocks attain this condition. As a matter of illustration, the codes and assets of the Dow Jones Index are enumerated in Table 1.1.

[1]Data from http://www.nyxdata.com/.

	Name
Code	
AA	Alcoa Inc.
AXP	American Express Co.
BA	Boeing Co.
BAC	Bank of America Corp.
CAT	Caterpillar Inc.
CSCO	Cisco Systems Inc.
CVX	Chevron
DD	E.I. DuPont de Nemours & Co.
DIS	Walt Disney Co.
GE	General Electric Co.
HD	Home Depot Inc.
HPQ	Hewlett-Packard Co.
IBM	International Business Machines Corp.
INTC	Intel Corp.
JNJ	Johnson & Johnson
JPM	JPMorgan Chase
KO	Coca-Cola Co.
MCD	McDonald's Corp.
MMM	3M Co.
MRK	Merck & Co. Inc.
MSFT	Microsoft Corp.
PFE	Pfizer Inc.
PG	Procter & Gamble Co.
T	AT&T
TRV	Travelers Cos.
UNH	UnitedHealth Group Inc.
UTX	United Technologies Corp.
VZ	Verizon Communications
WMT	Wal-Mart Stores Inc.
XOM	Exxon Mobil

Table 1.1: Dow Jones Industrial Average Index components from September 24, 2012 until September 22, 2013.

The thesis is organized as follows:

Chapter 1. Literature Review Contextualizes the Markovian model and recall its theoretical formulation and features.

Chapter 2. Bid and Ask Dynamics Study Studies the bid and ask dynamics in a model-free fashion. It investigates the dependence between bid and ask quantities, the behavior of the bid and ask quantities when the mid price is constant and, finally, whether the bid-ask spread is equal to one, as the model demands.

Chapter 3. Parameter Study Studies the parameters of the model focusing on checking its parameter assumptions.

Chapter 4. Analysis of Model Results Finally, after confronting the model assumptions with the empirical data, we investigate the theoretical results and their adherence to their empirical counterpart.

1.2 Model Specification

1.2.1 Model Formulation

Model Assumptions

The model makes the following assumptions:

1. The bid-ask spread is exactly one tick,
2. All order sizes are equal,
3. The bid-ask quantities follow a 2D independent birth and death process while the mid price is constant,
4. When the bid or ask quantity gets fully depleted, random quantities for bid and ask queues are observed and the prices move one tick in the direction of the depletion.

The first assumption is consistent with the idea of liquid markets. The second assumption can be interpreted as a choice of modelling the incoming orders as events[2]. Either for the bid or the ask quantities, the birth and death process in assumption number three is a difference of Poisson processes, which are all assumed to be independent of each other. Finally, the fourth assumption describes the events that generate a mid price change. Section 2.4 is devoted to study the first assumption, Section 2.1 studies the independence feature in the third assumption and Session 2.2 investigates the choice of the birth-death process.

[2] An event is a change in state. This change can be in the price or quantity of the bid or ask queue. Thus, the model do not worry about the order size, only its time and signal. In Section 3.2, we study a potential bias arising from this assumption.

Model Parameters

Given the assumptions, we shall make the model parameters precise. They are

- δ — tick size,
- λ — limit orders rate (birth rate),
- $\mu + \theta$ — market orders and cancellations rate (death rate),
- f and $\tilde{f}$ — conditional probability density functions for bid and ask quantities after a queue gets depleted (f is conditioned to a increase in mid price and $\tilde{f}$ to a decrease in mid price).

Note that the birth and death rates parameters are the same for each dimension. However, sometimes it is interesting to distinguish those rates for each side, so we may introduce λ_{bid}, λ_{ask}, $\mu_{\text{bid}} + \theta_{\text{bid}}$ and $\mu_{\text{ask}} + \theta_{\text{ask}}$.

Still regarding the parameters λ and $\mu + \theta$, it is shown in Cont and de Larrard (2013) that if $\mu + \theta = \lambda$, the expected time for the next price movement is infinity. Thus, it is desired to have $\mu + \theta > \lambda$, as in this case the expected time is finite. In Session 3.2 we study this property.

Session 3.3 is devoted to study the parameters f and $\tilde{f}$ and a measure of market depth[3] derived from these parameters. From the model assumptions, we can note that the model does not rely on an explicit form of the distribution for f and $\tilde{f}$.

Understanding the dynamics of the model

In Figure 1.1, we can visualize the stylized dynamics of the model. The trajectory starts with dot number 1. The parameter $\mu + \theta$ 'pushes' the quantities downwards and leftwards, while the parameter λ 'pushes' the quantities upwards and rightwards.

At dot number 9, the bid queue gets depleted. This makes the mid-price decrease δ and new random quantities are observed at dot number 10. These new random quantities are generated by the distribution $\tilde{f}$. Note that dot number 9 is shown only for illustrative purposes. Since it would imply a spread of two ticks, its existence would violate assumption 1.

1.2.2 Model Features

Negative Autocorrelation in Returns

The existence of negative first autocorrelation is empirically observed in high-frequency data for the log returns (cf., for instance, Tsay (2005)). As pointed out in Cont and de Larrard (2013), this is also true for the mid price process (cf. Cont (2001)). The present model accounts for this feature, which is going to be detailed in Subsection 4.1.2.

[3] Market depth is an abstract concept regarding how much the deeper layers of order book can absorb a market order with a size larger than the best bid and ask quantities.

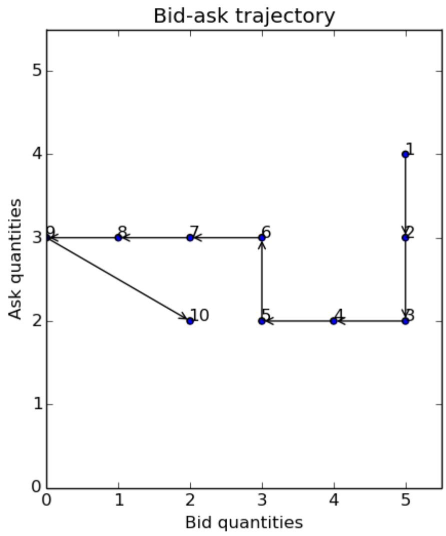

Figure 1.1: Example of a bid-ask trajectory.

Martingale Property of "Efficient Prices"

In Cont and de Larrard (2013), it is also pointed out that various authors, such as Robert and Rosenbaum (2011), consider the so-called "efficient price" and advocates for its martingale property. The "efficient price" is a non-observable process which coincides with the observable price when a trade occurs. Alternatively, it can also be seen as a noiseless version of price process (this noise can be, for instance, caused by the a rounding error to the nearest tick).

The model in Cont and de Larrard (2013) is able to construct such a process. It is defined by the expectation of the price for the next trade. It is shown in Cont and de Larrard (2013) that this process is a martingale if and only if there is no autocorrelation in the returns.

1.3 Theoretical Results of the Model

The main results of the model developed in Cont and de Larrard (2013) are:

1. Distribution of duration until next price movement conditioned to the state of order book, and their tail indices when $\mu + \theta > \lambda$ and when $\mu + \theta = \lambda$;

2. Probability of a mid price increase conditioned to the state of the order book
   a) for the next mid price change, when $\mu + \theta = \lambda$,
   b) for the next mid price change, when $\mu + \theta > \lambda$,
   c) for the next mid price change, when $\mu_{\text{ask}} + \theta_{\text{ask}} \neq \mu_{\text{bid}} + \theta_{\text{bid}}$ and $\lambda_{\text{ask}} \neq \lambda_{\text{bid}}$,
   d) for the n-th mid price change ahead;

3. Probability of consecutive mid price changes in the same direction;
4. Standard deviation of the mid price returns
   a) when $\mu + \theta = \lambda$,
   b) when $\mu + \theta > \lambda$.

In 4, we study 1, 2.a, 3 and 4.a. In order to prepare for the empirical results, this section is intended to give the theoretical formulation of these four selected results. However, before that, it is worth to mention that all the theoretical results in Cont and de Larrard (2013) are given in explicit formulas, so that one does not need to resort to simulations nor numerically solve any equation.

1.3.1 Conditional Distribution for Durations Between Mid Price Changes

Another useful result for the model described in Cont and de Larrard (2013) is related to the conditional distribution of the duration between mid price movements. This distribution is explicitly obtained by

$$\mathbb{P}[\tau > t | q_{\text{bid}} = x, q_{\text{ask}} = y] = \left(\frac{\mu + \theta}{\lambda}\right)^{\frac{x+y}{2}} \psi_x(t)\psi_y(t) \tag{1.1}$$

where $q_\bullet$ is the quantity of the respective queue,

$$\psi_n(t) = \int_t^{+\infty} \frac{n}{u} I_n\left(2\sqrt{\lambda(\mu+\theta)}\right) e^{-u(\lambda+\mu+\theta)} du$$

and $I_\bullet$ is the modified Bessel function of the first kind.

Moreover, Cont and de Larrard (2013) shows that the tail index for this distribution is 1 when $\lambda = \mu + \theta$ — which implies the infinite expectation for the duration between mid price changes — and 2 when $\lambda < \mu + \theta$.

Notice that (1.1) does not involve the parameter f nor $\tilde{f}$. This is expected, since this distribution shall not consider bid and ask quantities either before or after the mid price change.

The formula in (1.1) can be retrieved by considering the bid and ask quantities separately. Let $\tau_\bullet$ be the time spent for a queue (either ask or bid) process to get fully depleted. Since $\tau = \tau_{\text{bid}} \wedge \tau_{\text{ask}}$ and since the processes are independent, we have that

$$\begin{aligned}
\mathbb{P}[\tau < t | q_{\text{bid}} = x, q_{\text{ask}} = y] &= \mathbb{P}[\tau_{\text{bid}} \wedge \tau_{\text{ask}} < t | q_{\text{bid}} = x, q_{\text{ask}} = y] \\
&= \mathbb{P}\left([\tau_{\text{bid}} < t | q_{\text{bid}} = x] \cap [\tau_{\text{ask}} < t | q_{\text{ask}} = y]\right) \\
&= \mathbb{P}[\tau_{\text{bid}} < t | q_{\text{bid}} = x]\mathbb{P}[\tau_{\text{ask}} < t | q_{\text{ask}} = y].
\end{aligned}$$

Thus, the task of finding the conditional distribution for τ is reduced to find the conditional distribution for $\tau_\bullet$. Now, let us consider a random walk that takes increments equal to 1 with probability p and -1 with probability $1-p$ and that the time intervals between those increments are exponentially distributed with mean $1/c$. Then, define the probability $P_k(t)$ to be the probability of a position k at the time t for this random walk. We can notice that

$$P_x(t) = \mathbb{P}[\tau_{\text{bid}} < t | q_{\text{bid}} = x],$$

when

$$p = \frac{\lambda}{\lambda + \mu + \theta}, \quad c = \lambda + \mu + \theta,$$

and an analogous setting for the ask side.

Since the random walk started at the origin, the position $k \neq 0$ at time t can only be attained if there was a jump before t. When that jump occurred, say $t-s$, the position was either $k-1$ or $k+1$ and then it increased or decreased, respectively. Thus, we have that

$$P_k(t) = \int_0^t ce^{-c(t-s)} \left(pP_{k-1}(s) + (1-p)P_{k+1}(s) \right) ds.$$

Differentiating, we find the following ODE

$$P_k'(t) = -cP_k(t) + cpP_{k-1}(t) + c(1-p)P_{k+1}(t) \tag{1.2}$$

with initial conditions $P_0(0) = 1$ and $P_k(0) = 0$ for all $k \neq 0$.

Using Laplace transforms, we can solve (1.2) with its initial conditions, and get the Laplace transform of the density of the distribution for τ_{bid} and τ_{ask}, which is

$$\mathcal{L}_k(u) = \frac{k}{t} \sqrt{\left(\frac{\mu + \theta)}{\lambda} \right)^k} I_k(2\sqrt{\lambda(\mu + \theta)} t) e^{-t(\lambda + \mu + \theta}.$$

With $\mathcal{L}_k(u)$, we can do two things: do the inverse Laplace transform to get the explicit densities of $\tau_\bullet$ — which leads to the explicit distribution of τ, and thus providing (1.1) — and also use the Karamata's Tauberian theorem (cf. Feller (1971)) to get the tail distribution of $\tau_\bullet$ and then compose the tail distribution of τ by multiplying τ_{bid} with τ_{ask}, where we easily extract the tail index.

For further reference, when $\lambda = \mu + \theta$, the tail conditional distribution for τ is

$$\mathbb{P}[\tau > t | q_{\text{bid}} = x, q_{\text{ask}} = y] \overset{t \to \infty}{\sim} \frac{xy}{\pi \lambda} \frac{1}{t},$$

and its inconditional form is

$$\mathbb{P}[\tau > t] \overset{t \to \infty}{\sim} \sum_{i,j=1}^{\infty} \frac{ij}{\pi \lambda} \frac{1}{t} = \frac{D(f)}{\pi \lambda} \frac{1}{t}. \tag{1.3}$$

1.3.2 Probability of an Increase in Mid Price Conditioned to an Order Book State

Another useful result in Cont and de Larrard (2013) is the probability of an increase for the next mid price change. When $\lambda = \mu + \theta$, this probability is explicitly

$$p_{\text{up}} = \frac{1}{\pi} \int_0^{\pi} \left(2 - \cos(t) - \sqrt{(2-\cos(t))^2 - 1}\right)^p \frac{\sin(nt)\cos(t/2)}{\sin(t/2)} dt \tag{1.4}$$

Note that there are no parameters involved in (1.4), which is a very convenient feature of this formula.

To see why this is true, let us first call $\{q_t\}_{t \geq 0}$ the bid and ask quantities process while the mid price is still. Assuming $\lambda = \mu + \theta$, it is easy to see that there exists a symmetrical two-dimensional random walk $\{M_t\}_{t \geq 0}$ and a Poisson process $\{N_t\}_{t \geq 0}$ with mean 2λ such that

$$q_t = M_{N_t}, \forall t \geq 0.$$

This is true because, when we assumed that $\lambda = \mu + \theta$, the probability of an increase or decrease in either bid or ask quantity is equal to each other and since the processes for the bid and ask quantities are independent, it implies that it is a two-dimensional random walk with respect to some time rescale. Since the time interval between two changes in bid or ask quantities are exponentially distributed with mean $1/\lambda = 1/(\mu + \theta)$, then the time interval between changes between events considering both of them indifferently is exponentially distributed with mean $1/(2\lambda)$. Thus, the proper rescale of time for the random walk is a Poisson process with mean 2λ.

However, since we do not consider the time, the time scale is not important, we just want to know the probability that $\{q_t\}_{t \geq 0}$ hits one axis instead of the other. In that case, there is a result in Lawler and Limic (2010) that the probability that $\{M_t\}_{t \geq 0}$ starting at (n, p) hits $(x, 0)$ is

$$\frac{2}{\pi} \int_0^{\pi} e^{-r(t)p} \sin(nt) sin(tx) dt,$$

where

$$r(t) = \cosh^{-1}(2 - \cos(t)).$$

. Thus, we may integrate in $x \geq 1$ to get the probability that the bid and ask quantities hit the ask axis, leading to an increase in mid price, i.e.

$$p_{\text{up}} = \sum_{k=1}^{\infty} \frac{2}{\pi} \int_0^{\pi} e^{-r(t)p} \sin(nt) sin(tx) dt.$$

With some proper arithmetical manipulation we get rid of the sum and get the desired result.

1.3.3 Probability of Consecutive Movements in Mid Price

As pointed out in Subsection 1.2.2, the model accounts for the property of negative first autocorrelation for bid and ask returns time series. Moreover, since the model assumes that the bid-ask spread is always one tick, this means that it would be equivalent to consider either the bid, ask or mid price returns.

For this particular model, an useful result in Cont and de Larrard (2013) is that the first autocorrelation of the mid price returns time series can be determined by computing the probability of consecutive directions for the movement of the mid price. More precisely, the autocorrelation is negative if and only if that probability, defined as p_{cont}, is strictly less than one half. This is intuitively simple, because if p_{cont} is less than one half, it means that it is more likely that, for each movement, the next will be in the opposite direction – which is exactly what negative autocorrelation means.

A very simple expression for p_{cont} can be found if we assume that $\lambda = \mu + \theta$ and $f(x, y) = \tilde{f}(y, x)$. In Cont and de Larrard (2013) this expression is formulated as

$$p_{\text{cont}} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i,j) p^{\text{up}}(i,j), \tag{1.5}$$

where p^{up} is given by (1.4).

To see this, let X_1 and X_2 be consecutive mid price increments. Since we have that f is the distribution of bid and ask quantities conditioned to an increase in mid price, then we have that

$$\mathbb{P}[X_2 = \delta | X_1 = \delta] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i,j) p^{\text{up}}(i,j).$$

On the other hand, we have analogously

$$\mathbb{P}[X_2 = -\delta | X_1 = -\delta] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tilde{f}(i,j)(1 - p^{\text{up}}(i,j)).$$

Since the birth and death rates are the same for bid and ask quantities, we have that

$$p^{\text{up}}(i,j) = 1 - p^{\text{up}}(j,i), \quad \forall (i,j) \in \mathbb{N}^2.$$

Moreover, since we assumed that $f(x, y) = \tilde{f}(y, x)$, we have that

$$\begin{aligned}
\mathbb{P}[X_2 = \delta | X_1 = \delta] &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i,j) p^{\text{up}}(i,j) \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tilde{f}(j,i)(1 - p^{\text{up}}(j,i) \\
&= \mathbb{P}[X_2 = -\delta | X_1 = -\delta]
\end{aligned}$$

Since $\mathbb{P}[X_2 = \delta | X_1 = \delta] = \mathbb{P}[X_2 = -\delta | X_1 = -\delta]$, we conclude that

$$\begin{aligned}
p_{\text{cont}} &= \mathbb{P}[X_2 = \delta | X_1 = \delta]\mathbb{P}[X_1 = \delta] + \mathbb{P}[X_2 = -\delta | X_1 = -\delta]\mathbb{P}[X_1 = -\delta] \\
&= \mathbb{P}[X_2 = \delta | X_1 = \delta] \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i,j) p^{\text{up}}(i,j)
\end{aligned}$$

1.3.4 Volatility Formula

Under the hypotheses that $\lambda = \mu + \theta$ and that $f(i, j) = \tilde{f}(j, i)$, Cont and de Larrard (2013) shows that the volatility of the stock, measured in variance of the mid-price, can be easily computed as

$$\sigma^2 = \delta^2 \frac{\pi\lambda}{D(f)}, \tag{1.6}$$

where $D(f)$ was already defined and discussed in Subsection 3.3.3. This equation is particularly interesting because it links the volatility to pure microstructure parameters, without any (even indirect) dependence on the values of the prices. This result comes from a functional central limit and, for that reason, it should be regarded as an asymptotic result.

The equation (1.6) is obtained by a lengthy demonstration. Because of that, we outline the demonstration and direct the reader to Cont and de Larrard (2013) for the complete proof. The idea is to use a proper rescale of the time such that we can use a functional central limit theorem to obtain a Brownian motion for the mid price dynamics. The $\lambda/D(f)$ ratio in (1.6), in particular, comes from the tail of the durations distribution (1.3). We are interested only in the tail of the durations distribution (1.3) because the result uses the limit for $t \to \infty$.

1.4 Methodology

1.4.1 Overview

The purpose of this section is to describe the methodology which collectively addresses various empirical results in this thesis. Specific methodological issues are described inside the appropriate section where they arise.

1.4.2 Data set

The New York Stock Exchange offers a free sample[4] of the National Best Bid and Offer data. The National Best Bid and Offer consists of the data for the consolidated best bid and ask prices and quantities of all the regulated stock exchanges registered by the Consolidated Tape Association that trade the same assets in the United States at the best bid and ask prices, including the quantities in each side. The sample data is composed of thousands of tickers for the April, 3rd and 4th, 2013 trading sessions totalling 125,949,035 lines for the first day and 117,258,854 for the other.

This data is structured in single table form, where each line represents a change in any of its columns fields. There are 33 fields, but the ones of interest for the purpose of this work are the timestamp, asset code, best bid order price, best bid order quantity, best ask order price and best ask order quantity. The timestamps describe the moment the orders arrive up to its milliseconds.

[4] Which can be downloaded from http://www.nyxdata.com/data-products/daily-taq.

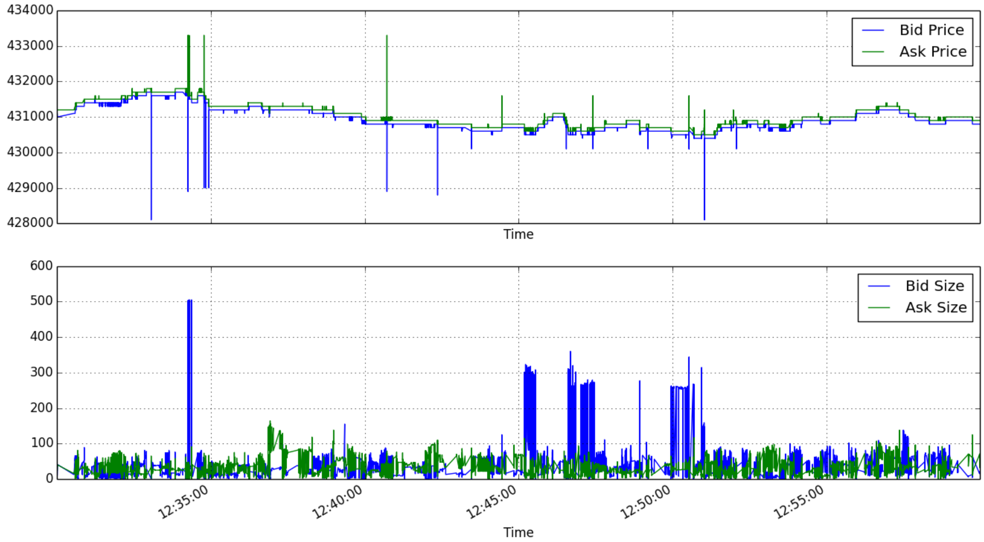

Figure 1.2: Citigroup Inc bid and ask prices and quantities from 12:30 PM to 13:00 PM in April 3rd, 2013.

The model investigated in Cont and de Larrard (2013) does not need any information about orders deeper in the order book nor the information about the trades (although the latter could also be downloaded from the same source). Thus, all of the numerical results contained in this thesis are obtained from this dataset.

In Figure 1.2, we illustrate our dataset with one close-up of the bid and ask prices and quantities. We shall be aware that the prices in the original dataset were multiplied by a ten thousand factor in order to accommodate 4 decimal digits onto an integer format. However, since the tick size for equities in NYSE is 1 cent, the tick size is 100 for our dataset format. Also, we note that the unit of the order sizes are units of trade — i.e., the minimum 'size' is 1 —, which means bulks of 100 shares in our case.

1.4.3 Parameter estimation

The article Cont and de Larrard (2013), where the model is presented, provides no information about parameter estimation. However, a previous article Cont et al. (2010), that presents a similar model, provides some clues for the parameter estimation, which will be used here. All the parameters for the model can be estimated directly from the market.

<table>
<tr><th>Queue</th><th>Price movement</th><th>Order type</th><th>Order size</th></tr>
<tr><td rowspan="2">Bid</td><td>Up</td><td>Limit order</td><td>Bid size at price movement</td></tr>
<tr><td>Down</td><td>Cancellation or market order</td><td>Bid size just before price movement</td></tr>
<tr><td rowspan="2">Ask</td><td>Up</td><td>Cancellation of market order</td><td>Ask size just before price movement</td></tr>
<tr><td>Down</td><td>Limit order</td><td>Ask size at price movement</td></tr>
<tr><td>Both</td><td>Still</td><td colspan="2">The consecutive positive differences between bid or ask sizes are limit orders and the consecutive negative differences between bid or ask sizes are cancellations or market orders</td></tr>
</table>

Table 1.2: Summary of the distinction between order types and sizes.

Distinguishing Causes for Bid and Ask Queues Changes

A simple, yet important task when analysing the time series of bid and as quantities of high-frequency databases is to distinguish the cases of the changes of bid or ask quantities. For instance, let us suppose that the bid size at some timestamp is 1000 and then increases to 1200. This increase may be due to a limit order, but it also may be due to a market order that consumed the whole queue and exposed a deeper bid queue with total quantity 1200. In order to do this distinction, the price time series must be taken into account.

Let us say that we are looking at the bid price and quantity and we start at a fixed point. While the bid price is still the same, the rise of the quantities is due to the placing of new limit orders – for the model, it would be driven by λ_{bid} — while its depletion is due to the cancellation or market orders — for the model, it would be driven by $\mu_{\text{bid}} + \theta_{\text{bid}}$. However, on the exact moment when the price goes down in the bid offer, we have the bid queue depletion. This means that the quantity immediately before this event is the quantity of the last cancellation or market order for that queue and the quantity observed right on the time of the event is the new quantity randomly generated by $\tilde{f}$. On the other case, when the price goes up, we have that a limit order with size equal to the whole new queue observed precisely at the change in the bid price. And this quantity was generated by f.

Note that the reasoning is the same for the ask queue when the prices are still and opposite when the price changes. Table 1.2 summarizes this reasoning.

Empirical distributions

The need to estimate distributions empirically occurs in Sections 2.1, 3.3, 4.1 and 4.2. We shall make a distinction between two groups of distributions that arise in this study: distributions with respect to events and distributions with respect to bid and/or ask quantities states. Since events are discrete, the estimation of those empirical distributions are simply frequency tables of occurrences. This is the case for the estimation of the conditional transitional distributions f and $\tilde{f}$.

However, when the probability of a state is determined by its persistence in time, we have to treat them differently. Thus, if we want to estimate a distribution with respect to an arbitrary point in time, this continuum feature must also be addressed. In this case, the frequency must be thought as the sum of time intervals in which the state is present relative to the studied time scope. This is the case for the estimation of the bid and ask quantities joint distribution in Section 2.1, probabilities of mid price increase conditioned to the order book state in Section 4.1 and the conditional distribution for the time duration between mid price changes in Section 4.2.

2 Bid and Ask Dynamics Study

2.1 Bid-Ask Quantities Interdependence

In this section we analyse the empirical joint density[1] of the bid and ask queues. In Figure 2.1, we can observe an empirical distribution for the bid and ask sizes. It can be seen that, while the main density seems quite uncorrelated, the extreme values, which are mainly located near the axes, are producing negative correlation. In addition to this particular case, Figure 2.2 suggests that the negative correlation between bid and ask sizes are persistent throughout our the dataset. Although we conclude that the assumption is not accurate, it should be a reasonable approximation since this negative correlation is small.

A possible interpretation of this fact can be made by the assumption of a 'fair price' of the asset that lies between the bid and ask prices. If, for instance, that price is closer to the bid price, there is more pressure to lower the mid-price, thus forcing the bid quantity to withdraw and the ask quantity to get bigger.

Note that the persistence of this negative correlation do not affect directly the assumption of the model that the bid quantities process is independent of the ask process. The assumption takes into account the relationship between increments, but not the total volume of the queues.

2.2 Bid-Ask Trajectories for Fixed Mid-prices

2.2.1 Methodology

The purpose of this section is to analyse the mean behavior of the bid and ask quantities between mid-price changes. For each mid-price state, we have a trajectory for bid and ask quantities. The analysis follows from gathering all these trajectories and translating them to a common starting point at the origin. Thus, each unit of time in the horizontal axis mean a unit of time past the mid-price change. Since those trajectories have different lengths, it is difficult to address what should we characterize by mean behaviour. In event time[2], the trajectories length are integers, thus we could formulate the following methods:

1. Consider the mean for each event time step (after the mid-price change) only for the values available for that event time step;

[1] For the methodology regarding this estimation, see Section 1.4

[2] Event time is the time counted in microstructure events. In our case, these events are the change in bid or ask quantities or prices.

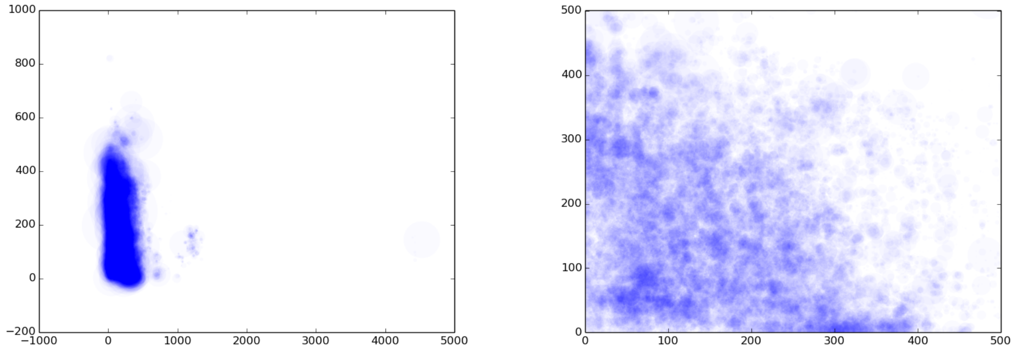

Figure 2.1: Full and zoomed empirical joint density for the bid and ask quantities — horizontal and vertical axes, respectively — for Citigroup Inc. in April 3rd, 2013. The larger the circle, the more frequent the bid and ask quantities. The computed correlation for this distribution is -0.17.

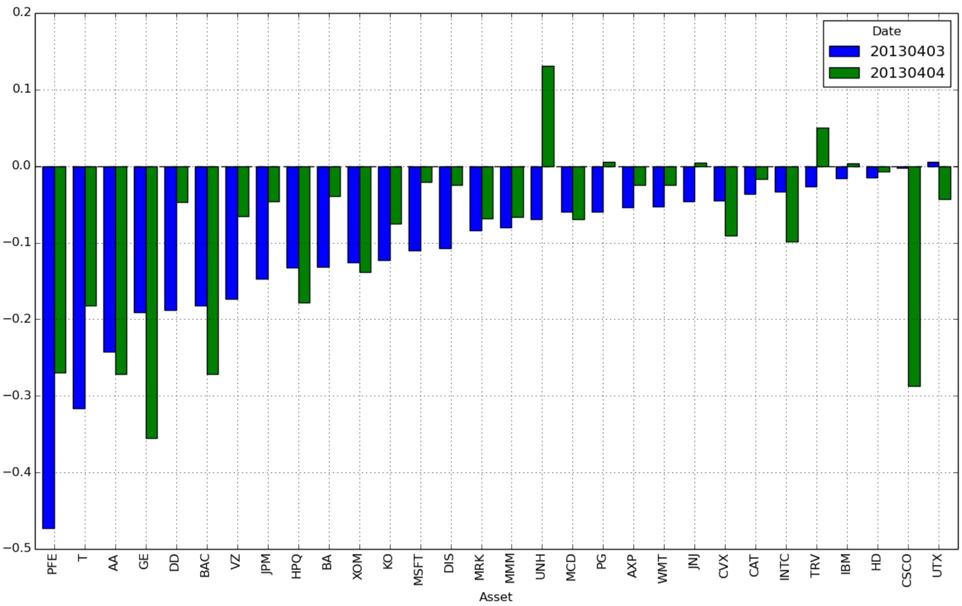

Figure 2.2: Correlations between bid and ask sizes computed from the empirical distributions.

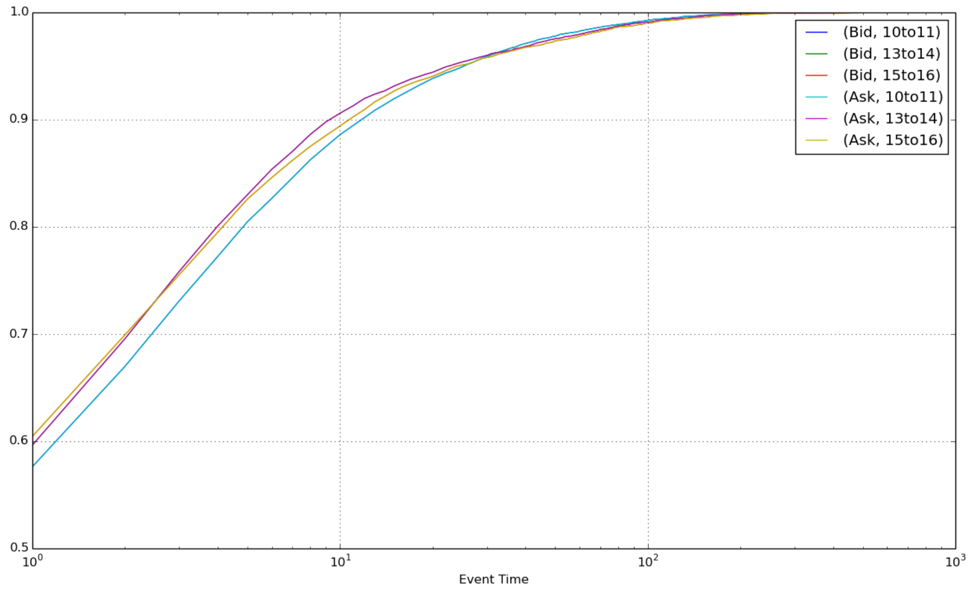

Figure 2.3: Accumulated frequency of time durations for each queue states.

2. Extend the trajectories with zeroes in order to have them all with the same length and then take the mean for each event time step (after the mid-price change);

3. Extend the trajectories by repeating their last value in order to have them all with the same length and then take the mean for each event time step;

4. Take the mean for each event time step for each group of trajectories with the same length.

Methods 1, 2, 3 all present biases. The farther from the origin, each method presents, respectively,

1. More chaotic values, since less sample is given for the mean as the trajectories are ending;

2. Lower values, since the filling zeroes pushes the mean downwards, as the trajectories are ending;

3. Flat values, since the mean gets computed more from filling constant values than from fluctuating trajectories, as the trajectories are ending.

Thus, among these methodologies, we choose the 4th method. However, it comes at the cost that we then have to look at several mean trajectories instead of only one. For our analysis, we consider the trajectories for the Citigroup Inc. stock in April 3rd, 2013 in three distinct time frames of the day — 10:00 to 11:00, 13:00 to 14:00 and 15:00 to

16:00. This is done because intraday patterns in market dynamics generate different market microstructure regimes (cf. Gourieroux et al. (1999)). Moreover, we restrict the analysis to the trajectories with a maximum of 10 event time length, since they represent roughly 90% of the data, as it can be seen in Figure 2.3.

An alternative methodology to the one we apply would be to perform the analysis in physical time. We plan to address this in future work.

In our analysis, the mean behavior of the bid and ask quantities are also presented in terms of the mean of the z-scores of the trajectories. In other words, for each trajectory, we filter its mean and scale, then we group those with the same length and take the mean for each event time step for each group. With this technique, we avoid that trajectories with high values dominate the mean, so that we can concentrate on the shape of the trajectory rather than its level and scale.

2.2.2 Trajectories Conditioned to Mid-Price Changes

In this section, we study the mean behavior of the bid and ask trajectories conditioned to an increase or decrease in mid price prior or posterior to the trajectory. According to the model, since we have Markovian birth and death processes, then the average trajectory should be monotonic regardless of the prior condition. Moreover, if we have the desired condition that $\mu + \theta > \lambda$[3] then we should have only decreasing average trajectories in any combination of prior and posterior conditions. We see clearly in Figures 2.4 and 2.5 that this is not the case.

Let us first condition to an increase or decrease in mid price prior to the bid and ask trajectories. As depicted in Figure 2.4, the starting bid quantities are mostly lower when we had a prior increase in mid price. To understand why this should hold, let us first consider the case when the mid price increases, and analyse the bid quantities. If the cause for the mid price increase was an ask price increase, then the bid queue is expected to remain the same. If the cause was a bid price increase, then it means that a completely new queue appeared by a sole limit order, which should grow with other limit orders placed at that queue. On the other case, if the mid price decreases, then, conversely, either the bid queue remains the same or the bid queue turns to be an existing queue deeper in the book. An analogous reasoning explains the opposite effect for the ask quantities.

Now, let us focus on the condition of increase or decrease in mid price posterior to the bid and ask trajectories. Figure 2.5 tells us that the ending bid quantities get lower when the mid price is going to decrease. This is an interesting fact, since it shows that the queue gets depleted gradually, and not abruptly. An analogous effect is also observable to the ask queue.

Furthermore, we can also observe for the mentioned cases — starting bid lower when there is prior price increase, ending bid lower when there is posterior price decrease and the analogous cases for ask prices — that concave trajectories are predominant in Figures 2.4 and 2.5. Since we have the fact that there is a negative autocorrelation in

[3] In 3.2, we show that this condition is satisfied

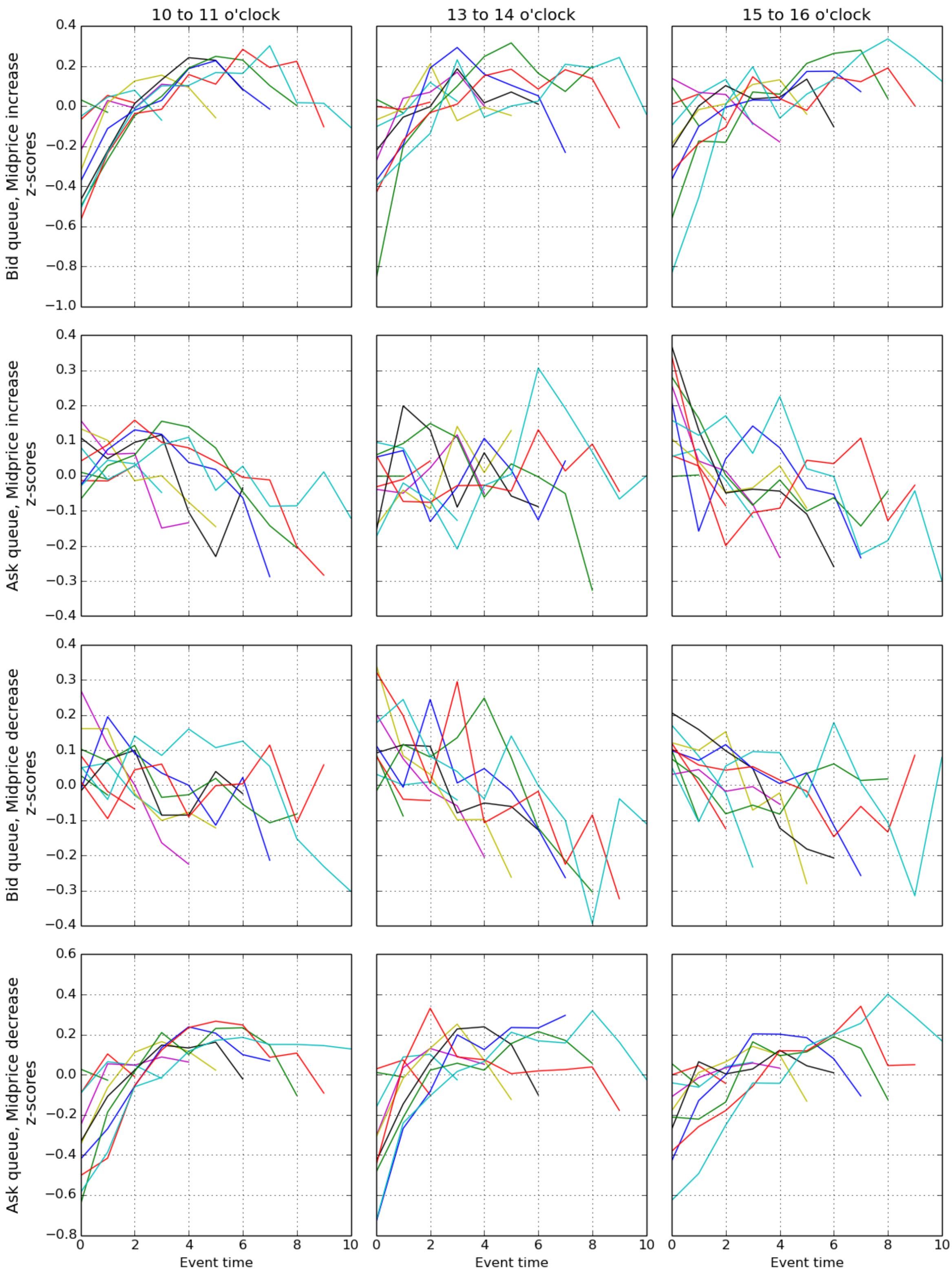

Figure 2.4: Mean normalized bid and ask sizes trajectories conditioned to prior mid price movements

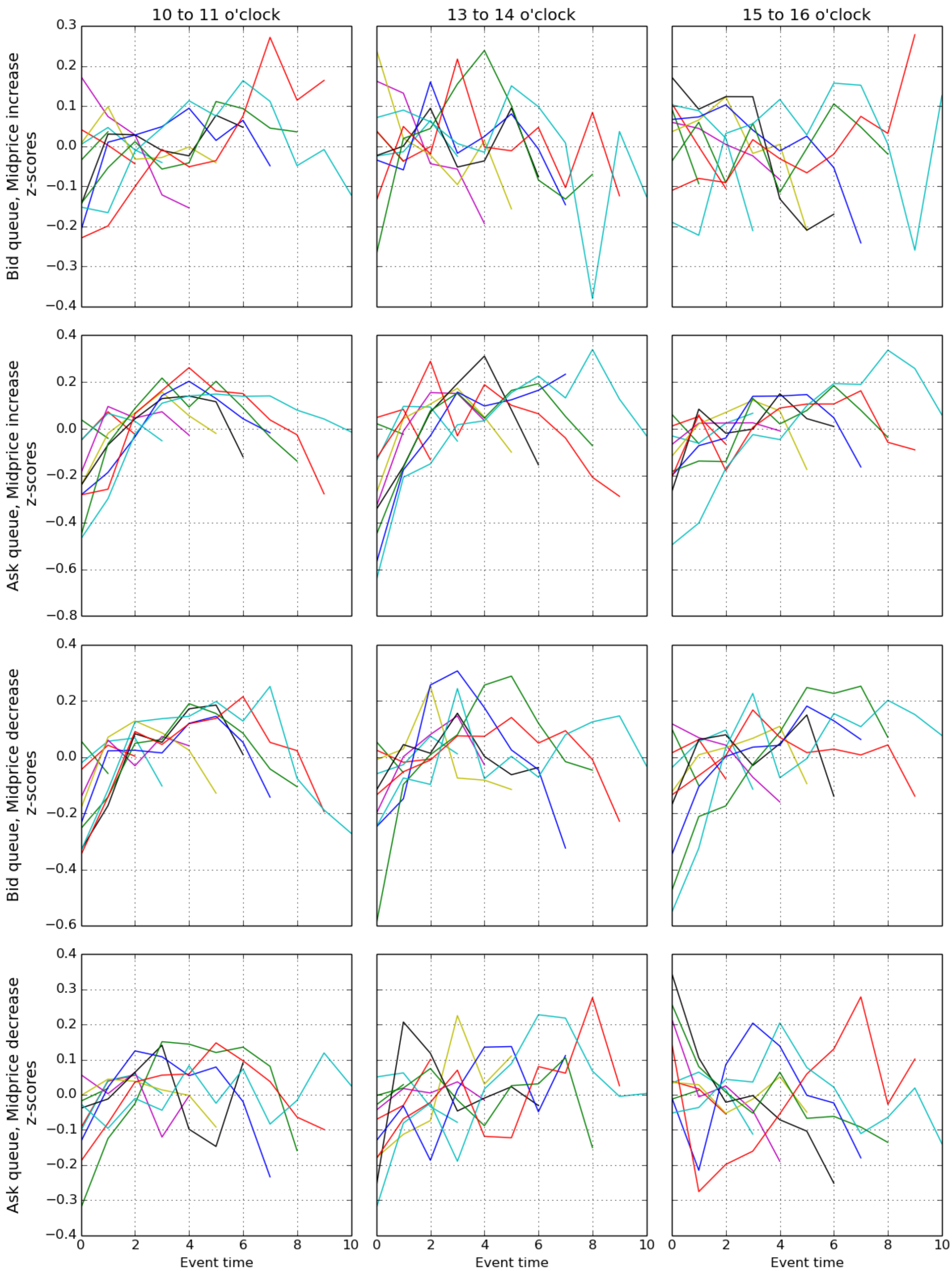

Figure 2.5: Mean normalized bid and ask sizes trajectories conditioned to posterior mid price movements

mid prices (cf. Cont and de Larrard (2013) and Subsection 4.1.2), it is expected that we have a mid price decrease when there is a prior mid price increase and vice-versa. Thus, this concave trajectories are simply a blend of the cited effects.

2.3 Mean-Reverting Behavior in Bid-Ask Quantities

This section is devoted to check whether there exists a long-run level of liquidity. Its existence would introduce a mean reversion behavior to the bid and ask quantities processes. In order to find this long-term level of liquidity for bid and ask quantities, we simply compute the means for the marginals of the distributions estimated in 2.1. Then, we can separate the bid and ask trajectories which start either above or below this level. If the mean-reversion effect is present, then the trajectories that start above the mean should converge downwards to the mean and the trajectories below should converge upwards to it. Figure 2.6 was produced with this methodology, and it suggests the existence of this long-run level of liquidity.

Moreover, it should be interesting to see whether the change in mid-price are disturbances that push the trajectories away from their long-run level of liquidity. This verification can be attained by computing the mean square deviations of the bid and ask quantities to their long-run level of liquidity for each event time step. Thus, Figure 2.7 tells us that there is no significant relation of the starting and ending points — close to the mid-price changes — that affects the long-run level of liquidity. It seems, however, that the length of the trajectories are related to the deviation to the mean level of liquidity, and this relation is not monotonic.

2.4 Bid-Ask Spread in Ticks

The first assumption of the model states that the bid-ask spread is always one tick. In this section we verify the fraction of time in which the bid-ask spread is exactly one tick for stocks in the Dow Jones Index. In Figure 2.8, eighteen among the thirty stocks are at least 80% of the time in both days with a bid-ask spread of only one-tick.

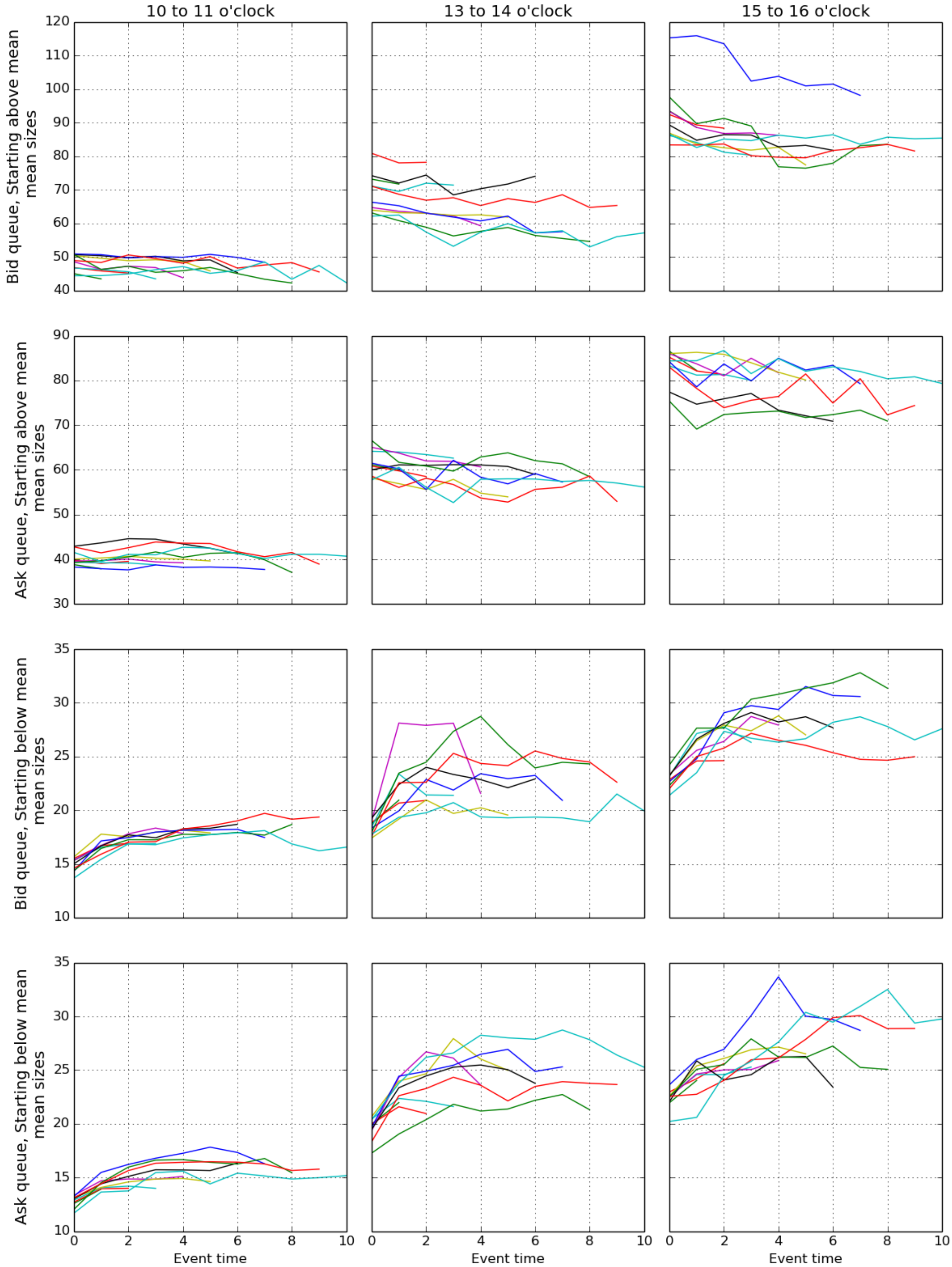

Figure 2.6: Mean bid and ask sizes trajectories conditioned to the starting value being above or below the mean level of bid or ask sizes.

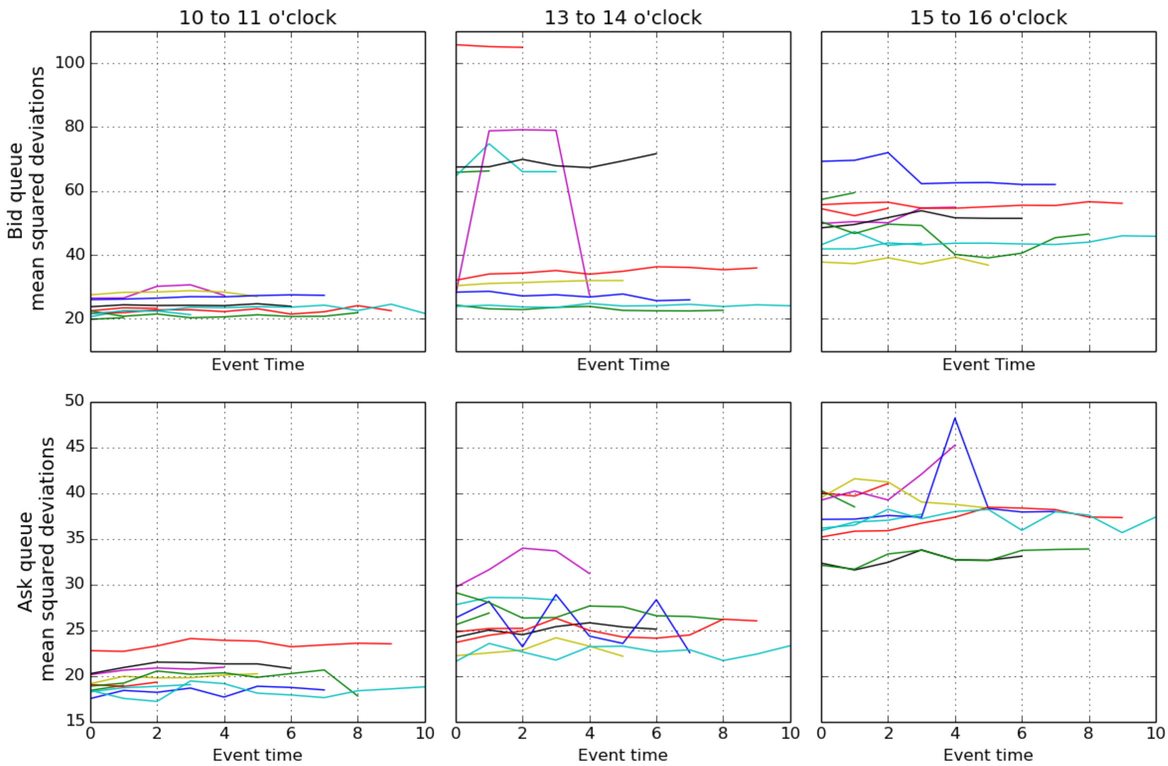

Figure 2.7: Trajectories for the mean squared bid and ask sizes deviations from the bid or ask mean level.

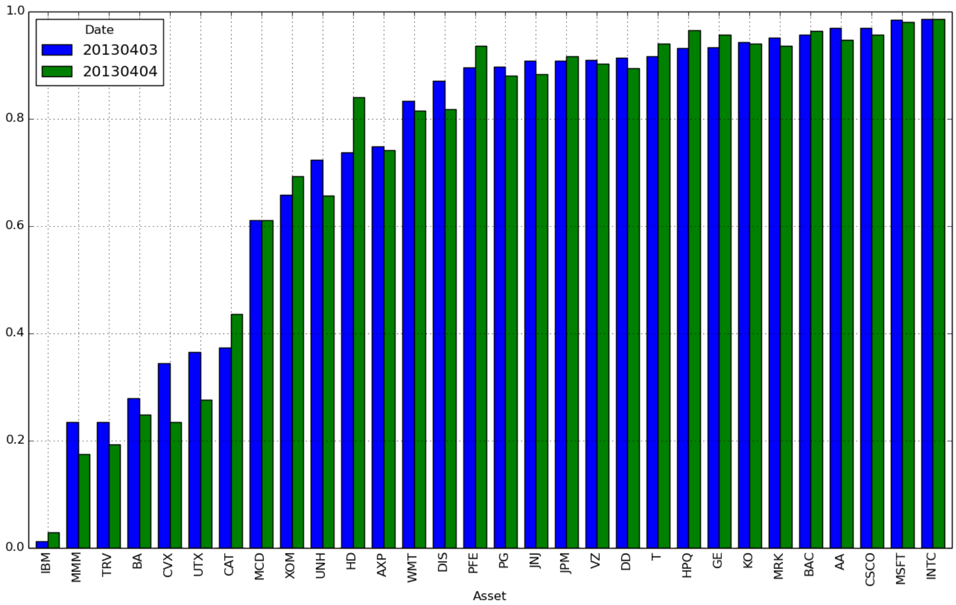

Figure 2.8: Fraction of the day in which the bid-ask spread is exactly one tick for April 3rd and 4th, 2013 from 10:00 AM until 4:00 PM.

3 Parameter Study

3.1 Overview

As outlined in Subsection 1.2.1, the parameters of the model presented in Cont and de Larrard (2013) are tick size δ, birth rate λ, death rate $\mu + \theta$ and the conditional transitional distributions f and $\tilde{f}$. The tick size δ is given by our dataset and is the same for all studied assets. The other parameters are studied in this chapter as follows:

Section 3.2 Studies the parameters λ and $\mu + \theta$. This section investigates whether we have the desired property that $\mu + \theta > \lambda$, the potential bias caused by the assumption that all order sizes are equal, and whether it is safe to assume that $\lambda = \lambda_{\text{bid}} = \lambda_{\text{ask}}$ and $\mu + \theta = \mu_{\text{bid}} + \theta_{\text{bid}} = \mu_{\text{ask}} + \theta_{\text{ask}}$.

Section 3.3 Studies the parameters f and $\tilde{f}$. This section investigates the issues in estimating the distributions f and $\tilde{f}$, describes those distributions qualitatively and discuss whether the assumption $f(x, y) = \tilde{f}(y, x)$ is reasonable. Finally, it also introduces the derived parameter $D(f)$, discusses some estimation methods and the estimates for our dataset.

3.2 Order Flow Parameters

3.2.1 Methodology

Estimation of the Order Flow Parameters

The order flow parameters — λ and $\mu+\theta$ — are parameters for the mean of exponential distributions; thus, they can be estimated simply by taking a sample mean. Let us take λ for instance. One approach would be to take a simple arithmetic mean of limit orders either on the ask side or the bid side (since λ drives both the bid and ask sides) for a fixed time length interval. Or, in order to use more data, we could take the simple mean of all limit orders for a fixed time length interval and divide by two. More specifically, let N be the total number of limit orders for all intervals, T be the total time length of our sample time series and τ be the length of our fixed length interval. Therefore, the number of intervals would be T/τ, and we would have an estimation $\hat{\lambda}$ numerically defined as

$$\hat{\lambda} = \frac{1}{2}\frac{N}{T/\tau} = N\frac{\tau}{2T}. \tag{3.1}$$

However, this approach may introduce some bias if we consider our model assumptions. The model states that all orders sizes are equal. Thus, if we see that the mean order size for the limit orders are smaller than for market orders and cancellations, the model would speed the time it takes for the queue to be depleted compared to reality. In reality, we would have small orders but in a high flow that go in the direction of depletion, but the model does not consider those sizes, and this would introduce bias. In order to prevent this bias, we introduce correction weights w_λ and $w_{\mu+\theta}$, such that our estimates are now

$$\hat{\lambda} = w_\lambda \hat{N}_\lambda \frac{\tau}{2T}, \quad \hat{\mu} + \hat{\theta} = w_{\mu+\theta} \hat{N}_{\mu+\theta} \frac{\tau}{2T},$$

where T is the timespan of our data, τ is the time unit of λ and $\mu+\theta$, $N_\bullet$ is the number of orders observed for a certain order type and

$$w_\lambda = \bar{Q}_\lambda / \bar{Q}, \quad w_{\mu+\theta} = \bar{Q}_{\mu+\theta} / \bar{Q},$$

where $\bar{Q}$ is the mean order size of all orders and $\bar{Q}_x$ is the mean order size for the orders related to x. This methodology to counter the mentioned potential bias is hinted in Cont et al. (2010).

For a more detailed analysis, however, we may distinguish λ and $\mu+\theta$ for bid and ask, as mentioned in Subsection 1.2.1. Thus, analogously we will have

$$\begin{gathered}
\hat{\lambda}_{\text{bid}} = w_{\lambda,\text{bid}} \hat{N}_{\lambda,\text{bid}} \frac{\tau}{T}, \quad \hat{\mu}_{\text{bid}} + \hat{\theta}_{\text{bid}} = w_{\mu+\theta,\text{bid}} \hat{N}_{\mu+\theta,\text{bid}} \frac{\tau}{T}, \\
\hat{\lambda}_{\text{ask}} = w_{\lambda,\text{ask}} \hat{N}_{\lambda,\text{ask}} \frac{\tau}{T}, \quad \hat{\mu}_{\text{ask}} + \hat{\theta}_{\text{ask}} = w_{\mu+\theta,\text{ask}} \hat{N}_{\mu+\theta,\text{ask}} \frac{\tau}{T}, \\
w_{\lambda,\text{bid}} = \bar{Q}_{\lambda,\text{bid}} / \bar{Q}_{\text{bid}}, \quad w_{\mu+\theta,\text{bid}} = \bar{Q}_{\mu+\theta,\text{bid}} / \bar{Q}_{\text{bid}}, \\
w_{\lambda,\text{ask}} = \bar{Q}_{\lambda,\text{ask}} / \bar{Q}_{\text{ask}}, \quad w_{\mu+\theta,\text{ask}} = \bar{Q}_{\mu+\theta,\text{ask}} / \bar{Q}_{\text{ask}}.
\end{gathered}$$

This time, without the $1/2$ term, since we have split the parameters that affect the ask queue and the parameters that affect the bid queues. Clearly, when $\lambda_{\text{bid}} = \lambda_{\text{ask}}$, then

$$\lambda = \lambda_{\text{bid}} = \lambda_{\text{ask}},$$

and the same goes with the $\mu+\theta$ parameter.

In order to compute the correct estimates in a dataset with only the best bid and ask quantities and prices, we should take into consideration the distinction between the causes of bid and ask quantities changes, which are detailed in Subsection 1.4.3. Furthermore, we shall be aware that the limit orders that cause the change in bid or ask prices should be discarded when computing $N_\bullet$ and $\bar{Q}_\bullet$. To explain that, let us consider the event where there is a mid price change. There is a market order or cancellation that caused this mid-price change. This market order or cancellation is considered for the estimation, since it was the last 'push' caused by the parameter $\mu+\theta$. Then, there are new quantities which are generated by the conditional distributions f and $\tilde{f}$. On one side of the book, we see a queue that was already there one tick deeper in the book. On the other side, there is a new queue. Although the creation of this new queue consists in the placement of a limit order, it is a starting quantity for the bid and ask trajectories until the next queue depletion. Thus, we should not consider this limit order to estimate $N_\bullet$ and $\bar{Q}_\bullet$. Note that this automatically indicates that there is a natural bias for $\hat{\lambda}$ to be smaller than $\widehat{\mu+\theta}$, since there are limit orders that are not considered for the estimation.

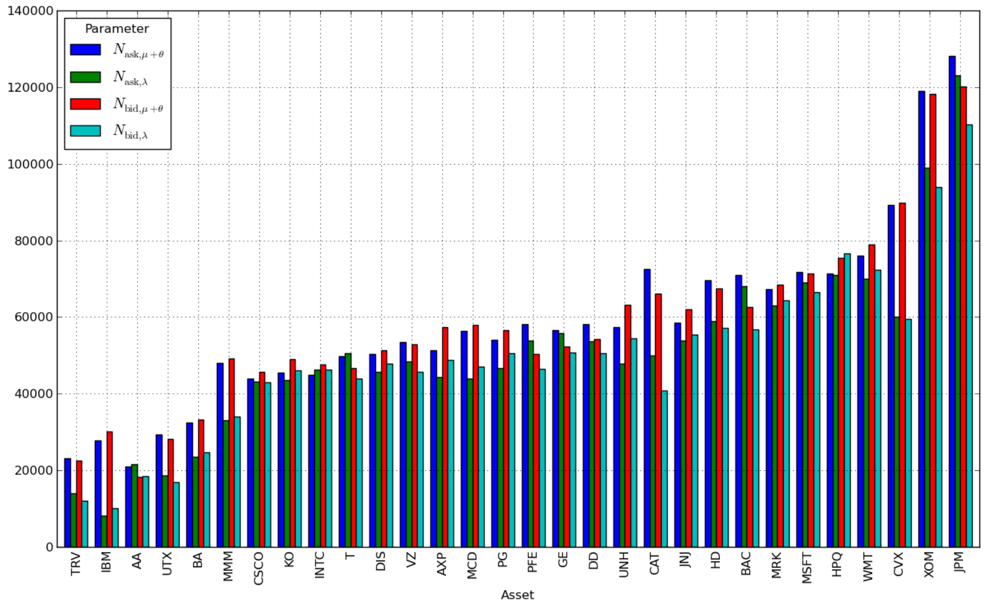

Figure 3.1: Number of orders in April 3rd, 2013.

3.2.2 Analysis of the Parameters

3.2.3 Number of Orders in a Trading Session

The number of orders in a day is one possible interpretation of the liquidity of an asset. In Figure 3.1, we may have an idea of which ones are more liquid than the others.

In this figure, it is also possible to observe that the cancellations and market orders — $N_{\bullet,\mu+\theta}$ — are usually[1] more numerous than their respective limit orders — $N_{\bullet,\lambda}$ —, but the number of bid and ask orders varies at the same level of magnitude. Both facts are favorable indications for the desired property that $\mu + \theta$ dominates λ, and for the assumption that $\lambda = \lambda_{\text{bid}} = \lambda_{\text{ask}}$ and $\mu + \theta = \mu_{\text{bid}} + \theta_{\text{bid}} = \mu_{\text{ask}} + \theta_{\text{ask}}$, respectively.

3.2.4 Mean Sizes of Orders

In Figure 3.2, we notice that the mean size for the cancellations, market and limit orders on the bid side — $\bar{Q}_{\text{bid},\bullet}$ — is virtually equal for each stock, thus the correction in this side is not necessary.

Looking at the ask side perspective, the mean sizes for cancellation and market orders — $\bar{Q}_{\text{ask},\mu+\theta}$ — clearly dominated their respective mean sizes for the limit orders — $\bar{Q}_{\text{ask},\lambda}$. Note that most cases where $N_{\bullet,\hat{\lambda}}$ was larger than $N_{\bullet,\hat{\mu}+\hat{\theta}}$ were on the bid side. Thus, this bias correction did not interfere too much to reverse the dominance between the

[1]The exceptions in April 3rd, 2013 on the bid side were Alcoa's, Intel's, Travelers' and Hewlett Packard's stocks.

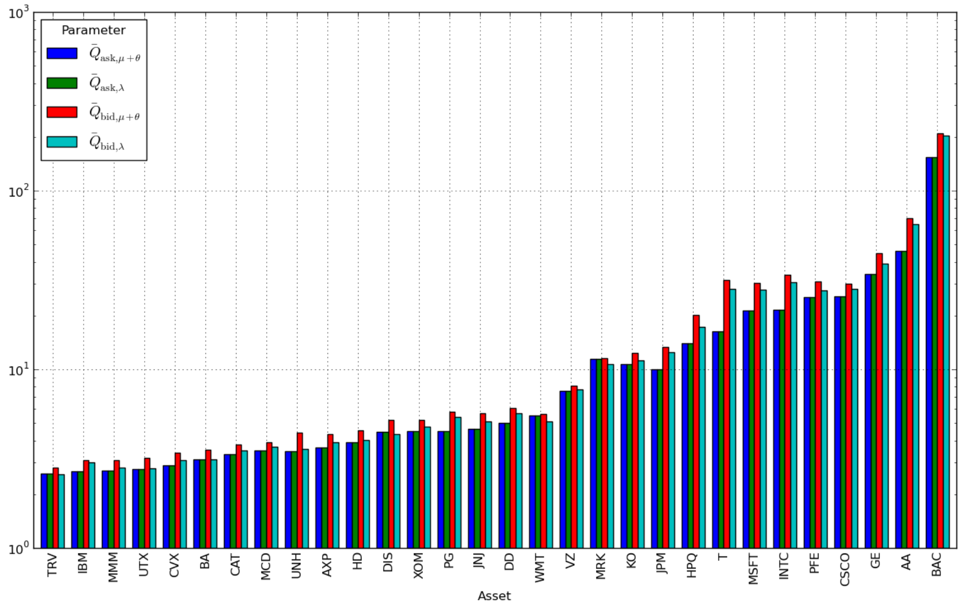

Figure 3.2: Mean order sizes in April 3rd, 2013.

order flow parameters. Nevertheless, the results were in favor of the model assumption of $\mu + \theta$ dominance over λ.

Moreover, we should note that the vertical axis is in logarithmic scale, which means that the mean order size varies heavily for each stock.

3.2.5 Birth and Death Parameters

As expected from the previous statistics, we see in Figure 3.3 that most of our estimates $\hat{\mu} + \hat{\theta}$ and $\hat{\mu}_\bullet + \hat{\theta}_\bullet$ surpass $\hat{\lambda}$ and $\hat{\lambda}_\bullet$, as desired.

Also, we note that $\hat{\lambda}_{\text{bid}}$ and $\hat{\mu}_{\text{bid}} + \hat{\theta}_{\text{bid}}$ are usually respectively approximately $\hat{\lambda}_{\text{ask}}$ and $\hat{\mu}_{\text{ask}} + \hat{\theta}_{\text{ask}}$. Since this is true, we should consider the remark in 1.4.3 that if $\lambda_{\text{bid}} = \lambda_{\text{ask}}$, we shall have $\lambda = \lambda_{\text{bid}} = \lambda_{\text{ask}}$, and this occurs analogously for the parameter $\mu + \theta$.

We should also note in those figures that $\hat{\lambda}$ and $\hat{\mu}+\hat{\theta}$ have high correlation. This means that the stocks differ from others on the rate of the overall incoming limit and market orders (and cancellations) but not on the resilience of the quantities in the queues — i.e. the more positive the difference between the death and birth rates, the less resilient the queues are to the depletion.

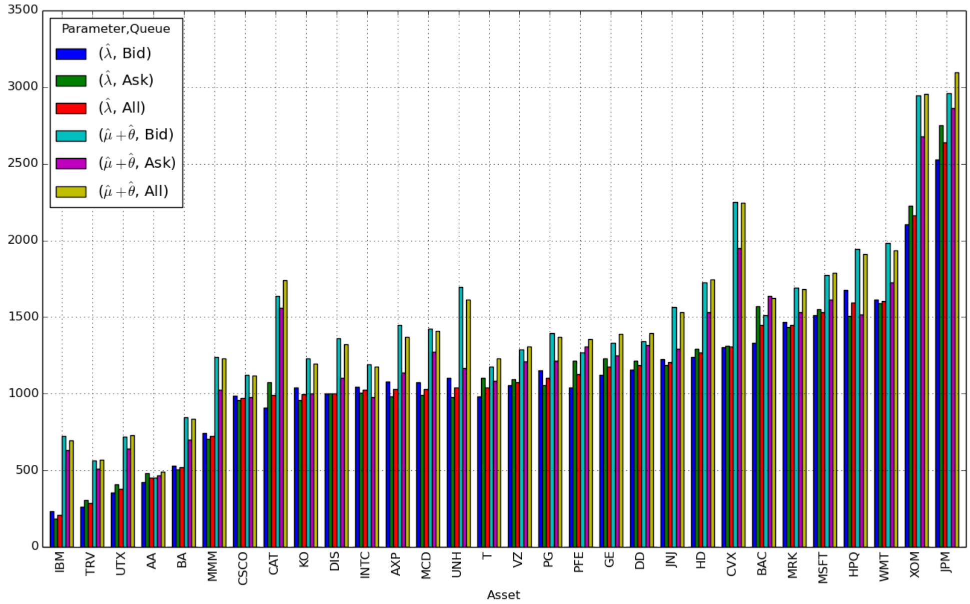

Figure 3.3: $\hat{\lambda}$ and $\hat{\mu} + \hat{\theta}$ in April 3rd, 2013.

3.3 Transitional Quantities Distribution

3.3.1 Methodology

The model assumption that the bid-ask spread is always one poses a special difficulty to estimate the empirical distributions for f and $\tilde{f}$. By intuition, when there is a queue depletion, we would expect that the bid-ask spread is increased by at least one tick, which is not possible by the model. However, when a queue gets depleted in real data, it is seldom the case that the opposite queue immediately builds as the model imposes. These cases occur in real data only when there is a aggressive limit order that is larger than size of the best bid or ask queue so that it executes all the bid or ask queue and the remaining quantity of the order turns to be the new queue exactly at the same price. Thus, when estimating f, the problem is to decide when and how one should measure the resting depth of the bid and ask queues. If it was the bid (resp. ask) that has just risen, then it is clear to see what is the new quantity for the bid (resp. ask) side, but not for the ask (resp. bid) if it did not rise at the same time. In order to account for this issue, we propose two different methodologies.

The first approach is to make another assumption to the model, which is the independence between the random bid and ask quantities after a queue depletion. This is a special case of the model proposed by Cont and de Larrard (2013). In this method, when estimating f, we will only account for the bid quantities when the bid price has risen, and only account for the ask quantities when the ask price has risen. This composes the marginals for f and, analogously, for $\tilde{f}$ in an event time framework as described in

Section 1.4. Then, we use that assumption to compute the joint distribution, where each joint probability can be computed by the product of the respective marginal probabilities.

The second approach consists in collecting events that are similar to the depletion event described by the model and taking the distribution of the bid and ask quantities for these events. More precisely, we look for the cases where there was an increase in bid price (to estimate f) or decrease in ask price (to estimate $\tilde{f}$) such that the bid-ask spread is exactly one tick. Contrary to the previous case where we only keep the bid (resp. ask) quantities when the bid (resp. ask) price has moved, here we keep both quantities for each event, and also put the additional requirement that the bid ask spread is one tick.

Note that this approach also lets the ideal situation of simultaneous increase in bid and ask prices — for f — or the opposite movement — for $\tilde{f}$ — with one-tick bid-ask spread, to be taken into consideration. While this methodology does not need to assume the independence of bid and ask quantities for f and $\tilde{f}$ as the other method, it comes with two disadvantages. First, we do not consider all the mid price movements. Secondly, let us take the case to estimate f: when the bid price increases, if the ask price has not increased together, the dynamics of the ask queue is not considered from the beginning, i.e. we have already let λ or $\mu + \theta$ affect the queue before we take the quantity into consideration.

As a final remark, in order to estimate the distributions f and $\tilde{f}$, we should use the normalized quantities in order to fulfil the assumption that all orders have size 1. However, in this chapter, since our analysis concentrates on the shape of the distribution, the quantities are not normalized so that we can preserve our intuition of small and big order sizes.

3.3.2 Distribution

With the independence assumption made in Subsection 3.3.1, the marginals — which we denote by f_{bid}, f_{ask}, $\tilde{f}_{\text{bid}}$ and $\tilde{f}_{\text{ask}}$ — are enough to explain all the distribution for f and $\tilde{f}$. By looking at Figure 3.4, f_{bid} and f_{ask} tell us that, after a mid-price increase, it is more likely to observe a bigger queue at the ask price than at the bid price. This is due to the fact that, after the ask queue has been depleted, the new ask quantity was already there (it was one tick deeper in the book), while the new bid quantity is being constructed by the incoming limit orders. For $\tilde{f}$, the effect is exactly opposite, so we would expect that $f(x, y)$ is roughly $\tilde{f}(y, x)$ for all x and y, which can also be verified in Figure 3.4 as f_{bid} and $\tilde{f}_{\text{bid}}$ are respectively approximately $\tilde{f}_{\text{ask}}$ and f_{ask}. These observations are not only true for the displayed, but also for all the other empirical distributions taken from the Dow Jones Index stocks.

We can also notice in Figure 3.4 that the shape of the marginals for the Microsoft Corp. stock is different from those of the JPMorgan Chase stock, although both stocks are among the most liquid of the Dow Jones Index. There are patterns that repeat from a stock to another and, for each individual stock, the patterns usually repeat from a day to the other.

Now, regarding the second method, we can see in Figure 3.5 the differences in the

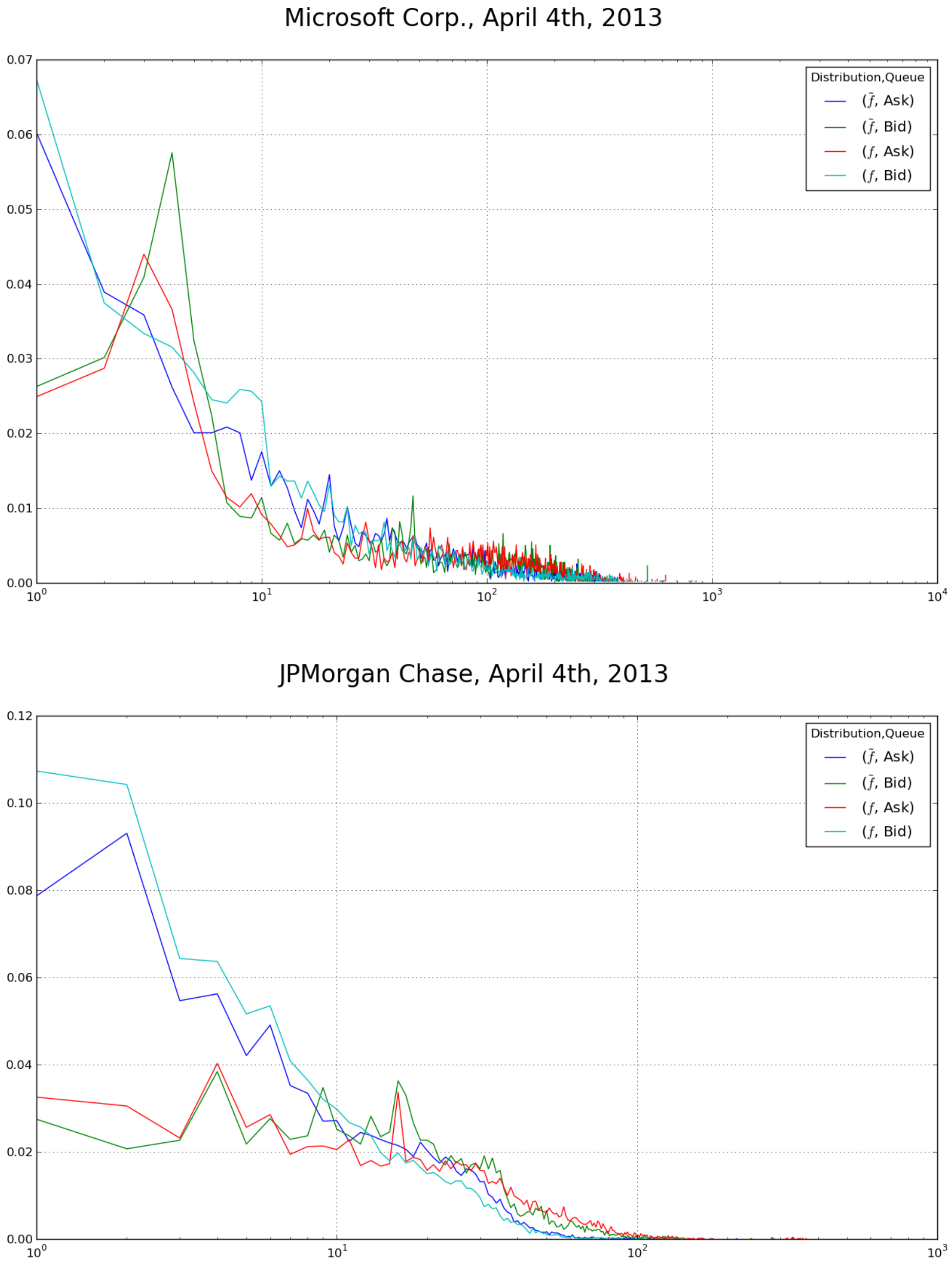

Figure 3.4: f and $\tilde{f}$ distribution marginals for Microsoft Corp and JPMorgan Chase, respectively, in April 4th, 2013, estimated with the independence assumption.

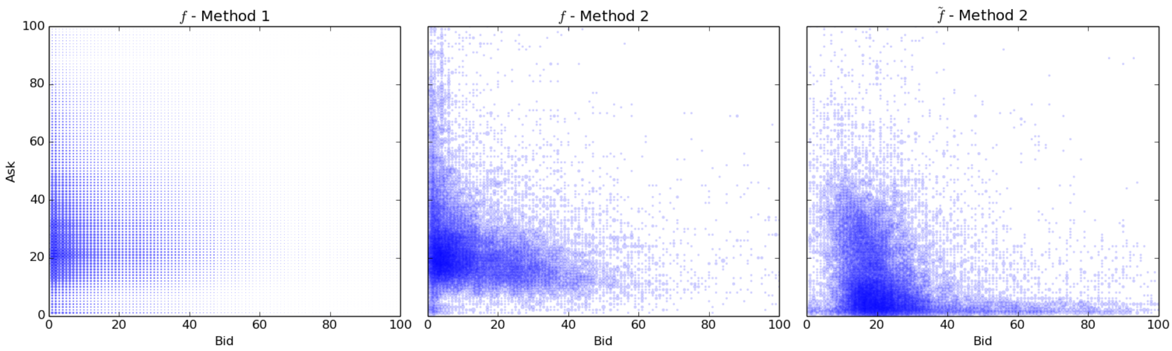

Figure 3.5: f and $\tilde{f}$ joint distributions for Citigroup Inc in April 4th, 2013, using both methods.

joint distribution for both methods. It is easy to see that all the observations made for the first method also holds for the distributions estimated by the second method. However, there is a feature that in the joint distribution of the second method that does not appear on the first one. For the distribution f estimated with the second method, we can see a dependence between the marginal correlations, which infers that for large bid queue quantities, it is more likely to see smaller ask queue quantities. This clearly shows the slight negative correlation for bid and ask dynamics when the mid price has just changed, as it was expected from our study in 2.1.

3.3.3 A Market Depth Parameter

The parameter f appears in two results that are subject to study in Chapter 4. One is with respect to computing volatility and the other is with respect to computing the probability of consecutive movements in mid price — i.e., an increase in mid price given that it had previously increased or a decrease given a prior decrease. For the former case, it appears inside the derived parameter

$$D(f) := \sum_{i,j} ij f(i,j).$$

Actually, the parameter $D(f)$ does not really depend on the condition that the mid price has risen, only that it has changed. This is due to the assumption that $f(i,j) = \tilde{f}(j,i)$, which we have shown in Subsection 3.3.2 to be a very reasonable assumption. Then, we have that

$$D(f) = \sum_{i,j} ij f(i,j) = \sum_{i,j} ij \tilde{f}(j,i) = \sum_{i,j} ji \tilde{f}(j,i) = D(\tilde{f}).$$

This observation is given in Cont and de Larrard (2013) when this parameter first appears into the computation of the equation for the volatility. Thus, it is convenient to introduce the distribution g, which is simply the new bid and ask quantities immediately after a mid price change. Thus, if we let U be the event of an increase in mid price, D be the decrease in mid price and C be the change of mid price, we have the relation

$$g(x,y) = f(x,y)\mathbb{P}(U|C) + \tilde{f}(x,y)\mathbb{P}(D|C).$$

Therefore, since $\mathbb{P}(U|C) + \mathbb{P}(D|C) = 1$,

$$\begin{aligned}
D(f) &= \sum_{i,j} ij f(i,j) = && (3.2)\\
&= \mathbb{P}(U|C)\sum_{i,j} ij f(i,j) + \mathbb{P}(D|C)\sum_{i,j} ij f(i,j) = && (3.3)\\
&= \sum_{i,j} ij\mathbb{P}(U|C) f(i,j) + \sum_{k,l} lk\mathbb{P}(D|C)\tilde{f}(k,l) = && (3.4)\\
&= \sum_{i,j} ij\left(\mathbb{P}(U|C) f(i,j) + \mathbb{P}(D|C)\tilde{f}(i,j)\right) = && (3.5)\\
&= D(g). && (3.6)
\end{aligned}$$

Now, the intuition behind $D(f)$ is clearer. As mentioned in Cont and de Larrard (2013), it is a measure of market depth and can be precisely described as the square of the geometric mean between bid and ask quantities after a mid price change. The idea of depth of the order book comes from the interpretation that higher $D(f)$ indicates that the new queues are filled with a greater stack of orders, which better absorbs large rates of market order and cancellations.

In this study, we propose three methodologies to compute $D(f)$. Two methodologies are simply computing $D(g)$ using the two methods we proposed in Subsection 3.3.2 adapted to find g instead of f.

The third method is taking the unconditional mean of the bid and ask quantities. This is motivated by the study in Chapter 2. We have shown in Section 2.2 that the bid-ask trajectories start increasing under some conditions of prior mid price movement, which is not what should be expected for the birth-death process with a dominant death rate — i.e., with $\mu + \theta > \lambda$. Thus, the starting quantities for the bid-ask trajectories should be bigger in order to counter this assumption. Moreover, we have also seen, in Section 2.3, evidence of a long-run level of liquidity, that is not disturbed specifically for the mid price change event. Thus, it suggests that we can use the unconditional joint distribution for bid and ask queues in order to estimate $D(f)$, which would then be simply the expectation of the product of the bid and ask queue sizes.

In Figure 3.6 we can see the distributions involved in each of the three methods to estimate $D(f)$. The comparison between the first two distributions is directly related to the equivalent comparison in Subsection 3.3.2. The third one was already presented in Section 2.1, and we can verify that it concentrates less probability for the small order sizes than in the other two distribution, just as we observed in the last paragraph.

Finally, Figure 3.7 shows the estimated $D(f)$ for all stocks of the Dow Jones Index. As it is expected, the third method resulted in larger estimations than the other two methods. And, due to the similar approach of the first two methods, we had close

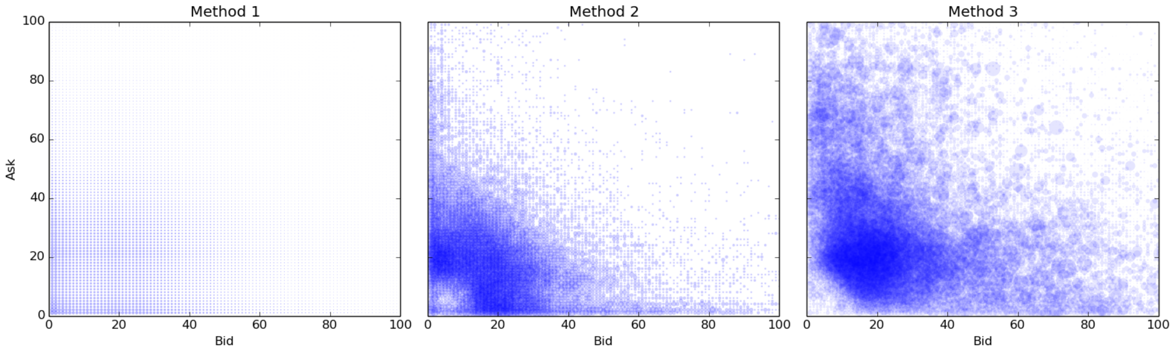

Figure 3.6: Joint distributions g in two methods and the unconditional joint distribution for bid and ask queues sizes for Citigroup Inc's stock in April 3rd, 2013.

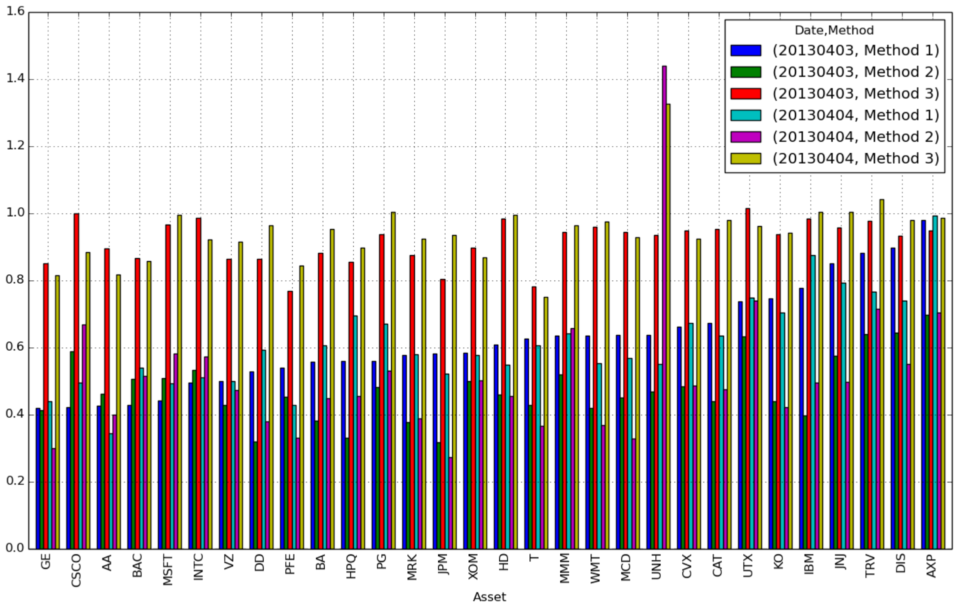

Figure 3.7: $D(f)$

estimates between the two for each stock. It is interesting to notice that the third method produced estimates that are quite similar across different stocks.

4 Analysis of Model Results

4.1 Probabilities for Mid Price Movements

4.1.1 Conditional Probability of a Mid-Price Increase

In Figure 4.1 we confront the empirical and theoretical probabilities of an increase in mid price conditioned to the bid and ask quantities. The theoretical distributions are computed according to (1.4). Similarly to what we have done in Section 3.3, when we studied f and $\tilde{f}$, we condition the probability distributions to the real bid and ask quantities instead of the modeled normalized bid and ask quantities.

Before we analyze Figure 4.1, we recall that, as one can see in Figure 2.1, most of the bid and ask quantities are concentrated within the area below the descending diagonal of each picture. As we can see in the graph of the differences of probabilities in Figure 4.1, the mixed blue and red disks in the upper right corner show the poor estimation of the probabilities due to the lack of samples.

As it should be expected, Figure 4.1 shows that the larger the ask quantities compared to the bid quantities, the more likely is an increase in mid price for both empirical and theoretical distributions. In particular, in the empirical distribution, we see mainly three regions: bid quantities very close to zero, ask quantities very close to zero, and the rest of the intermediate cases. The probabilities for each region is, respectively, high, low and quite uniform. For the theoretical distribution, however, these regions are less distinguishable and the probabilities vary very smoothly throughout the plane. These differences are clearly noted in the third graph of Figure 4.1 where, for the area very close to the axes and for an intersecting ascending diagonal, we have very small discrepancies and, outside this area, two main regions of discrepancy. Moreover, in this ascending diagonal we have the cases where the bid and ask prices are very close, which means that we should expect a 0.5 probability of a mid price increase. Therefore, except for the cases when the bid or ask quantities are very low or when they are very close to each other, the empirical data suggests that the bid and ask quantities are less informative to the prediction of an increase in mid-price than what the model describes.

Still in Figure 4.1, a particular region of interest, which is when the bid and ask quantities are less than 20 and 30, respectively. We can see a better adherence of the model to the border of this region compared to the general picture. In the interior of this region, however, we have a uniform domination of the empirical probability with respect to the theoretical one. The border of this region is actually very close to where the most frequent bid and ask quantities lie. This shows that for small orders, the dynamics of these probabilities are different. Further research is required to better understand this

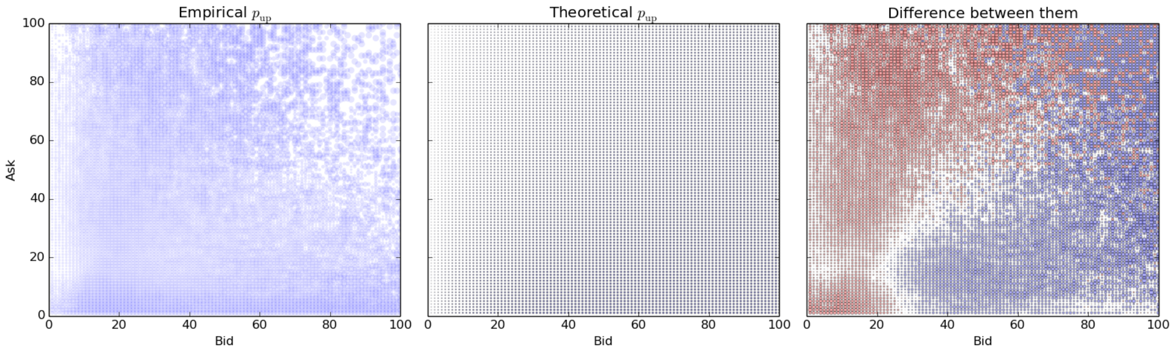

Figure 4.1: Empirical and theoretical conditional probabilities for the bid and ask quantities for Citigroup Inc stock in April 3rd, 2013 and their difference. For the first two graphs, the larger the circle, the more frequent the bid and ask quantities. For the third graph, red and blue disks represent the cases where empirical distribution were higher and lower to the theoretical, respectively, and the size of the disk represents the magnitude of the difference.

dynamics.

4.1.2 Probability of Consecutive Mid-Price Movements

Furthermore, we have computed the probability of consecutive mid price movements using the formula mentioned in Subsection 1.3.3. These computations are represented in Figure 4.2. As discussed in Subsections 1.2.2 and 1.3.3, we would expect to see that p_{cont} is below 0.5 because, since this would characterize the negative correlation in mid price returns, which is a stylized fact in market microstructure. And indeed that is the case for all stocks except Alcoa in April 4th, 2013. Thus, the negative autocorrelation of the first difference in the mid price is effectively captured by the model, although varying in magnitude among stocks.

4.2 Durations Distribution

4.2.1 Conditional distribution

Since it is difficult to produce a reasonable graphical representation of a conditional distribution that takes values in $\mathbb{N}^2 \times [0, +\infty)$ and maps them to the $(0, 1)$ interval — i.e. that attributes a probability to inputs of bid and ask quantities and time values —, we shall use only one pair of bid and ask quantities to illustrate the distribution. Because of this limitation, little sample is available to estimate the conditional empirical distribution effectively and thus purpose of this section is to give an intuition of the general case.

Figure 4.3 illustrates the differences between the empirical and theoretical conditional

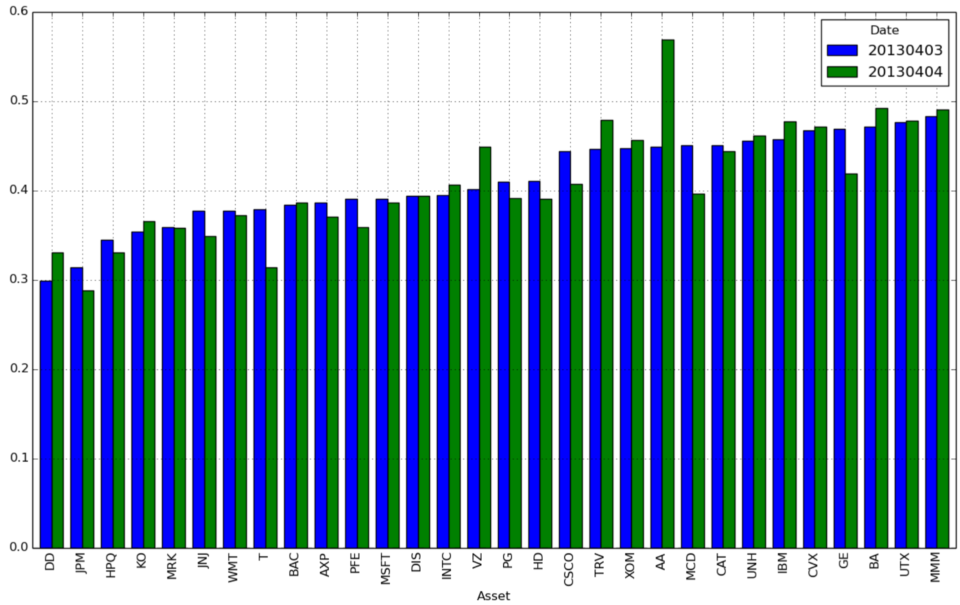

Figure 4.2: Probabilities p_{cont} for the stocks in the Dow Jones Index.

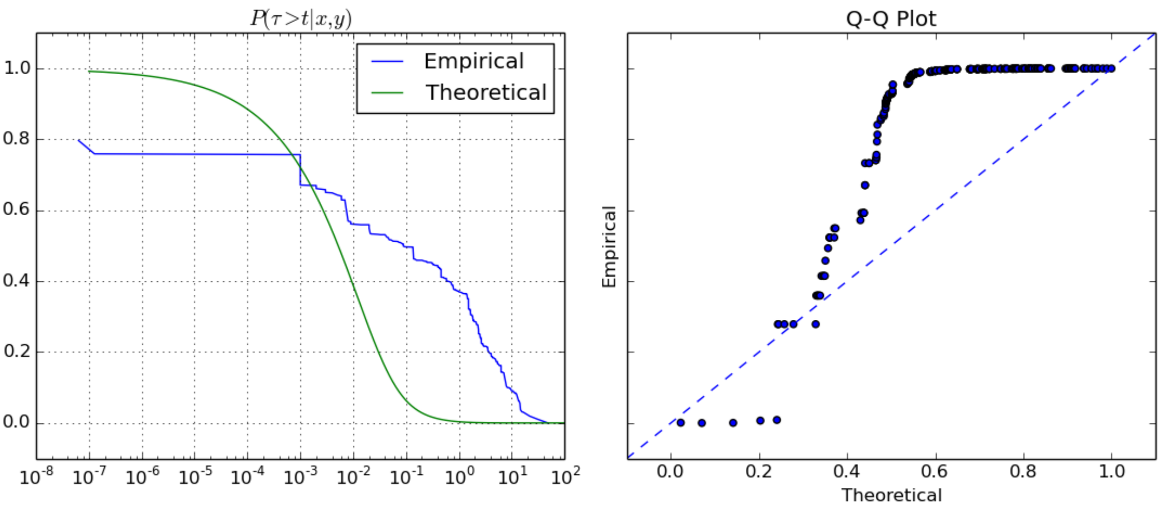

Figure 4.3: Comparison between the theoretical and empirical distributions for the durations between mid price changes conditioned to the most frequent bid and ask quantities, which are 13 and 16, respectively for the Citigroup Inc stock in April 3rd, 2013.

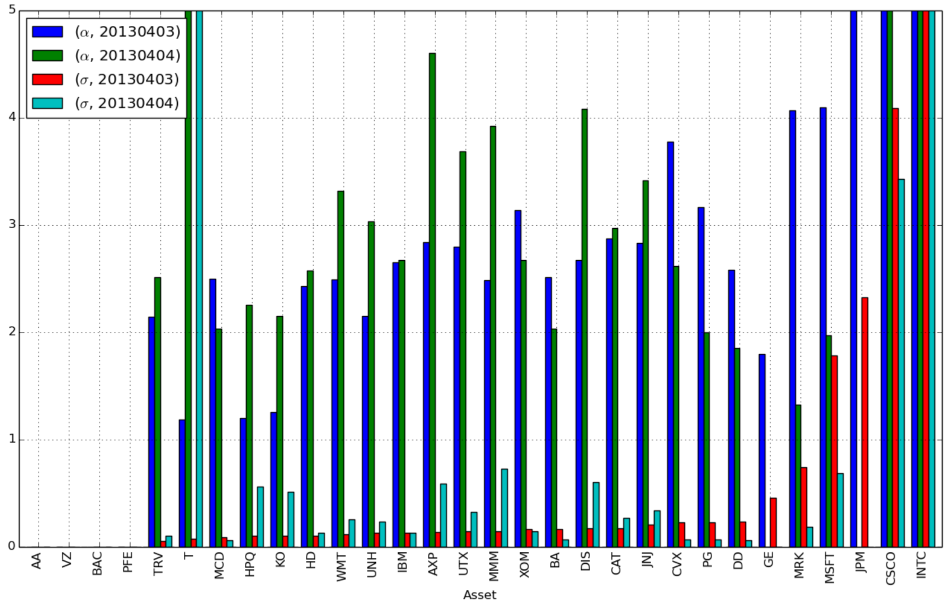

Figure 4.4: Estimated tail indices (represented by α) and the standard deviation of the errors for the Dow Jones Index stocks.

distributions using the most frequent bid and ask quantities pair. The theoretical distribution is computed using (1.1) and is clearly different from the empirical one.

4.2.2 Tail indices

Although it is difficult to verify the theoretical distributions of the durations by confronting to actual empirical probabilities, we can compute the tail index for the distributions. Let us recall, from Section 1.3 that the tail index of the distribution for the durations is either 2 if $\lambda > \mu + \theta$ or 1 if $\lambda = \mu + \theta$. It is worth to remark that the latter explains why the expected duration between mid price changes is infinity when $\lambda = \mu + \theta$. In order to estimate the tail indices of real data, the maximum likelihood estimation technique was employed.

In Figure 4.4, we should note that when σ is very low, the tail indices are indeed close to 1 or close to 2. Since the tail index estimation requires a large sample size, it may not be precise for some stocks, and thus, we can see that tail indices of 1 and 2 are reasonable candidates. This implies that the theoretical distributions for the durations may reflect the reality if a reasonable choice of parameters is made.

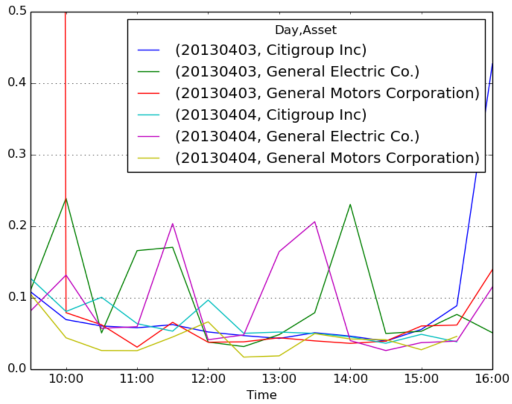

Figure 4.5: Annualized standard deviation for 1-minute log returns throughout a trading session.

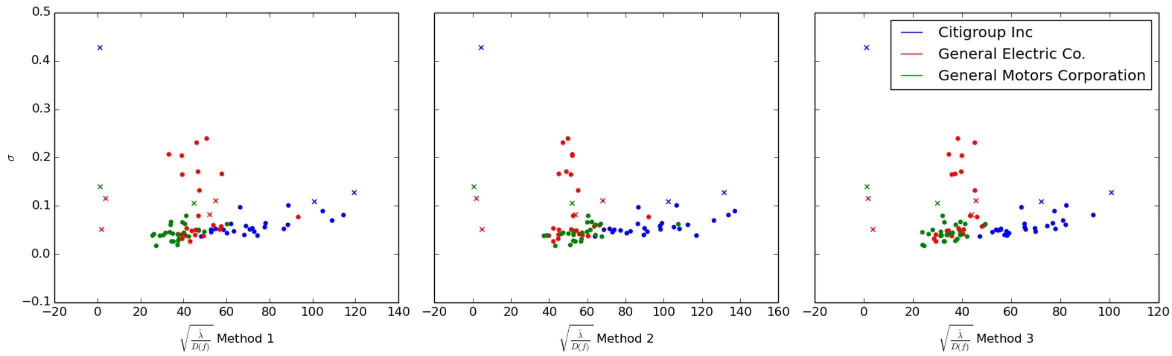

Figure 4.6: Annualized standard deviation for 1-minute log returns versus . Each observation in this Figure is a 30-minute high-frequency data strip for the specified stock and date. Cross markers are data from the first and last halves of the respective trading session.

4.3 Volatility

4.3.1 Theoretical results

4.3.2 Intra-day analysis

In this subsection, we analyze the three assets that were highlighted in Cont and de Larrard (2013). In the period of our dataset, however, only General Electric Co. is in the Dow Jones Index. The volatility of those three assets are depicted in Figure 4.5. It is interesting to notice the tendency of the volatility to be convex throughout the day, a stylized fact of intraday market dynamics. Another observation to be made is that the standard deviations from General Electric Co. are the most unstable among the three stocks in Figure 4.5.

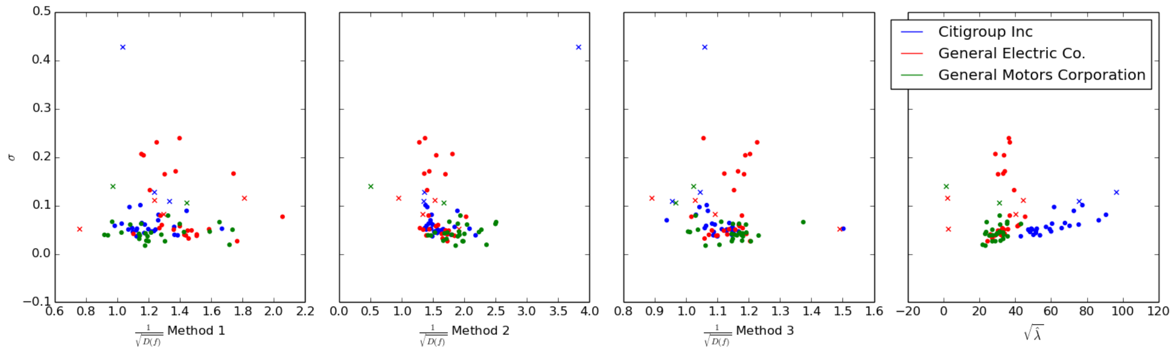

Figure 4.7: Annualized standard deviation for 1-minute log returns. Each observation in this Figure is a 30-minute high-frequency data strip for the specified stock and date. Cross markers are data from the first and last halves of the respective trading session.

Before we start to analyze the figures, notice that, by (1.6), we immediately conclude that

$$\sigma \propto \sqrt{\frac{\lambda}{D(f)}}, \tag{4.1}$$

since $\delta^2\pi$ is constant.

In Figures 4.6 and 4.7, we study both the relation (4.1) and the influence of its components to volatility. The first observation to be made is that both the parameters and the factor $\sqrt{\lambda/D(f)}$ can vary a lot for a particular asset considering different time spans, although the model considers them to be constant in time. Nevertheless, Figure 4.6 shows that indeed the relation (4.1) holds quite well for Citigroup's and General Motors' stocks for all three possible methods of estimating $D(f)$ described in Section 3.3.

On the other hand, Figure 4.7 shows that the $1/\sqrt{D(f)}$ part in the relation (4.1) does not contribute too much for the accuracy of this relation, so that $\sqrt{\lambda}$ alone accounts for most of this correlation.

For all seven cases, the General Electric Co. stock presented the worst fit. This is probably due its estimated variance, which was the most unstable among the other stocks, as we have already noted in Figure 4.5.

In addition, we shall note that there are clusters of observations for each stock, mainly for the λ parameter. This implies that, although the parameters vary in time, we can still use this parameters to describe some characteristics of the stocks in terms of liquidity and market depth.

Still in Figures 4.6 and 4.7, we can observe some outliers. These were all located in the first and last half hours of the trading session. Since the bid-ask spread in those times are often wider than on the rest of the trading session, we take them out from our data.

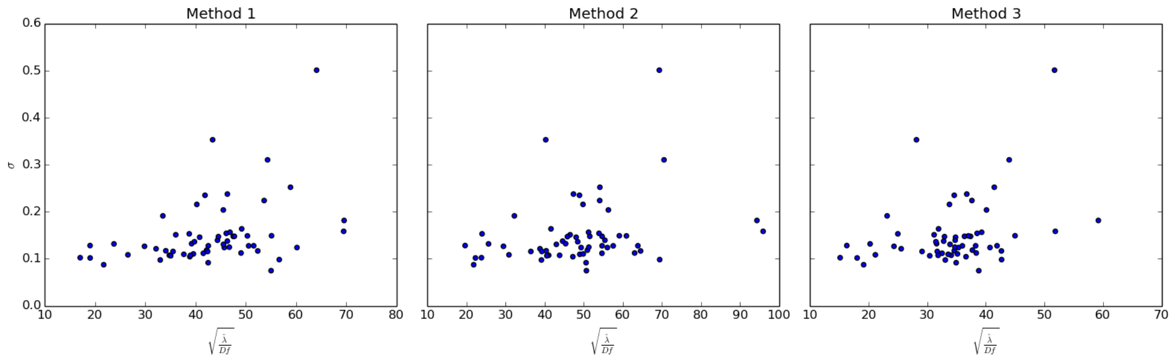

Figure 4.8: Scatter plot of $\sqrt{\frac{\hat{\lambda}}{D(f)}}$ versus the computed standard deviation of the mid price log returns for samples taken in 10-minutes time intervals. Each graph shows the plot for different methods for the $D(f)$ estimation, and each point in each graph is a stock from the Dow Jones Index in April 3rd, 2013 or April 4th, 2013.

4.3.3 Cross-Section analysis

In the previous subsection the relation (4.1) indeed holds for some stocks if we fix a particular stock and analyze it throughout different time periods. Another possibility is to fix the time period and analyze the relation for different stocks. Figure 4.8 shows relation (4.1) from the latter point of view. It clearly indicates that the linear relationship between the volatility of the stock and the ratio $\sqrt{\lambda/D(f)}$ holds.

Moreover, Figure 4.9 confirms the positive correlation induced by the relation (4.1). However, the regression analysis could not statistically confirm this relation, since the residuals are not normally distributed — the Jarque-Bera test rejected hypothesis for all the three cases at a 1% significance level. It is worth to mention that the same regression analysis was realized with equivalent relations such as σ^2 versus $\lambda/D(f)$ and $\log \sigma^2$ versus $\log(\lambda/D(f))$, but none of them presented normally distributed residuals.

Furthermore, by analyzing the components of the factor $\sqrt{\lambda/D(f)}$ as in Figure 4.10, we can see clearly that both components are positively correlated with σ. This result is different from the result in previous section, where $1/\sqrt{D(f)}$ had no correlation with σ.

Moreover, as in the previous case, the Jarque-Bera test also rejected the hypotheses of normality of the residuals. In Figures 4.9 and 4.11, the correlation coefficients show that both the ratio $\sqrt{\lambda/D(f)}$ and the factor $1/\sqrt{D(f)}$ estimated with the second method presented the poorest estimation of σ, while the other methods yielded good results. This result is quite surprising if we consider that the second method did not need to assume the independence between the marginal distributions for f and $\tilde{f}$ and that the third method is clearly detached from the parameter definition — since it was not computed from a distribution conditioned to mid price changes —. It may suggest that the dynamics of the movements when the bid-ask spread is equal to one-tick value is different from the general dynamics.

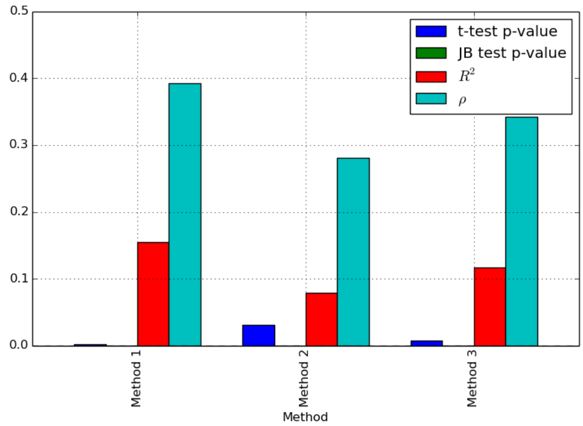

Figure 4.9: Regression analysis p-values, R^2 statistic, and Pearson correlation coefficient related to Figure 4.8.

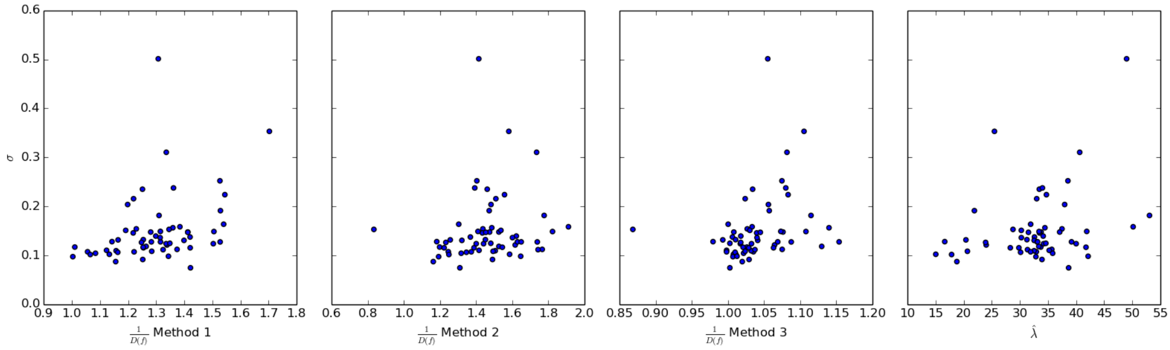

Figure 4.10: Scatter plot of the parameters of the volatility equation versus the 10-minute standard deviation of the mid price log returns.

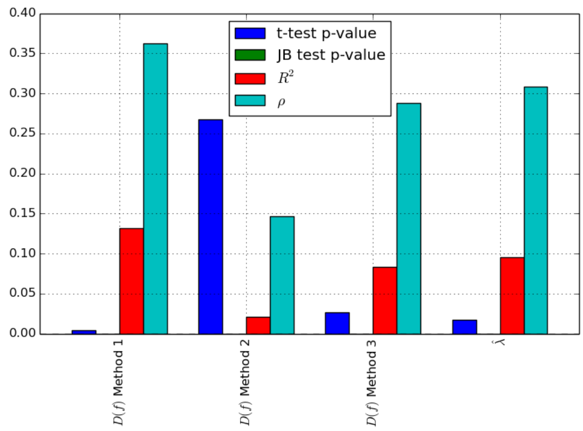

Figure 4.11: Regression analysis p-values, R^2 statistic, and Pearson correlation coefficient related to Figure 4.10.

Dep. Variable:	Var	**R-squared:**	0.151
Model:	OLS	**Adj. R-squared:**	0.120
Method:	Least Squares	**F-statistic:**	4.900
Date:	Mon, 02 Jun 2014	**Prob (F-statistic):**	0.0110
Time:	14:13:18	**Log-Likelihood:**	-73.225
No. Observations:	58	**AIC:**	152.4
Df Residuals:	55	**BIC:**	158.6

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Df	-1.6702	0.538	-3.102	0.003	-2.749 -0.591
Lambda hat	-0.2818	0.243	-1.161	0.251	-0.768 0.205
const	-2.2913	1.644	-1.394	0.169	-5.586 1.004

Omnibus:	18.286	**Durbin-Watson:**	2.209
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	22.876
Skew:	1.273	**Prob(JB):**	1.08e-05
Kurtosis:	4.727	**Cond. No.**	102.

Table 4.1: Regression for the model $\log \sigma^2 = \beta_0 + \beta_1 \log \lambda + \beta_2 \log D(f) + \epsilon$.

Additionally, an ANOVA table was produced to analyse the log relationship $\log \sigma^2 \propto \log \lambda - \log D(f)$. The results are shown in Table 4.1. Unfortunately, it rejected of normality for the residuals by the Jarque-Bera test at a 1% significance level.

As a final remark, we should consider that in this analysis we regressed a non observable variable (volatility) against statistics that explains this variable. Traditionally, a regression analysis confronts an observable variable against statistics that hypothetically hold some linear relationship with the variable. Since the latter is not the case, additional noise is produced. Thus, the statistical evidence of dependence between volatility and $\lambda/D(f)$ is probably underestimated in this analysis.

5 Conclusions

This thesis presents an empirical analysis of the assumptions and results of the Markovian model for market microstructure introduced by Cont and de Larrard (2013). The model assumptions were studied in Chapters 2 and 3. Our main findings are the followings:

1. While the mid price is constant, the bid and ask trajectories are not descending in mean, they are mainly convex (Section 2.2);

2. Moreover, the bid and ask quantities present mean-reverting characteristics (Section 2.3);

3. Even for some of the most liquid stocks of the Dow Jones Index, the assumption that the bid-ask spread is exactly one tick is still violated (Section 2.4);

4. When there is a queue depletion, the bid-ask spread rarely preserves the one-tick size — the spread opens or closes when there is a mid price change — (Section 3.3).

5. $\lambda_{\text{bid}} = \lambda_{\text{ask}}$ and $\mu_{\text{bid}} + \theta_{\text{bid}} = \mu_{\text{ask}} + \theta_{\text{ask}}$ (Section 3.2);

6. $\mu + \theta > \lambda$ (Section 3.2);

7. $\mu + \theta \approx \lambda$ (Section 3.2).

8. $f(i, j) = \tilde{f}(j, i)$ for all $(i, j) \in \mathbb{N}^2$ (Section 3.3);

Although our empirical findings partially differs from the model assumptions, overall the model successfully describes many empirical features of the dataset. Theoretical predictions of the model were analyzed in Chapter 4. The conclusions of our analysis is that:

1. The probability of a mid price increase is less informative than its theoretical counterpart (Section 4.1),

2. The negative first autocorrelation is well described by p_{cont} (Section 4.1);

3. Tail indices for the durations distribution clusters around 1 and 2, as predicted by the model (Section 4.2);

4. The independence assumption for the random variables governing f yields best estimates for f (Section 4.3),

5. Finally, we have seen that the simple and elegant formula for the volatility provided in (1.6) has a good fit for different stocks and even for different time periods inside a trading session (Section 4.3).

Therefore, we arrive at the same conclusion as stated in Cont and de Larrard (2013), that "[...] this stylized model as a first step in the analytical study of realistic stochastic models of order book dynamics." Moreover, Cont and de Larrard (2013) stated that "A relevant question is to examine which of the above results are robust to departures from the model assumptions [...]", in which this thesis contributed to answer. However, further research is necessary to better understand the probabilities for mid price increase when the bid and ask quantities are small (Section 4.1) and to study the results of the model which are not covered in this thesis.

Bibliography

Alfonsi, A. and Schied, A. (2010). Optimal trade execution and absence of price manipulations in limit order book models. *SIAM Journal on Financial Mathematics*, 1(1):490–522.

Biais, B., Hillion, P., and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the paris bourse. *the Journal of Finance*, 50(5):1655–1689.

Biais, B. and Weill, P.-O. (2009). Liquidity shocks and order book dynamics. *NBER Working Paper*.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Cont, R. and de Larrard, A. (2013). Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25.

Cont, R., Stoikov, S., and Talreja, R. (2010). A stochastic model for order book dynamics. *Operations research*, 58(3):549–563.

Feller, W. (1971). *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71 – 100.

Gourieroux, C., Jasiak, J., and Le Fol, G. (1999). Intra-day market activity. *Journal of Financial Markets*, 2(3):193–226.

Lawler, G. F. and Limic, V. (2010). *Random walk: a modern introduction*, volume 123. Cambridge University Press.

Madhavan, A., Richardson, M., and Roomans, M. (1997). Why do security prices change? a transaction-level analysis of nyse stocks. *Review of Financial Studies*, 10(4):1035–1064.

Obizhaeva, A. A. and Wang, J. (2012). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*.

Parlour, C. A. (1998). Price dynamics in limit order markets. *Review of Financial Studies*, 11(4):789–816.

Robert, C. Y. and Rosenbaum, M. (2011). A new approach for the dynamics of ultra-high-frequency data: The model with uncertainty zones. *Journal of Financial Econometrics*, 9(2):344–366.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.

Stoikov, S., Avellaneda, M., and Reed, J. (2010). Forecasting prices from level-i quotes in the presence of hidden liquidity. Technical report, working paper (http://ssrn.com/abstract= 1691401).

Tsay, R. S. (2005). *Analysis of financial time series*, volume 543. Wiley. com.

Tsoukalas, G., Wang, J., and Giesecke, K. (2013). Dynamic portfolio execution. *Available at SSRN 2089837*.