NTNU
Norwegian University of
Science and Technology

Statistical Properties of Opening and Closing Auctions of U.S. Bourses

Nikita Gourianov

Master of Science in Physics and Mathematics
Submission date: July 2017
Supervisor: Jon Andreas Støvneng, IFY
Co-supervisor: Damien Challet, Laboratoire de Mathématiques Appliquées aux Systèmes, centraleSupelec

Norwegian University of Science and Technology
Department of Physics

SAMMENDRAG

Børshandel startes og lukkes typisk med såkalte enkeltprisauksjoner. Disse auksjonene fungerer, for hver notert enkeltaksje, ved å samle inn ordre i en førauksjonsperiode, og rett før kontinuerlig handel starter, utføre alle de oppsamlede ordrene på den endelige auksjonsprisen, som er den prisen som maksimerer det omsatte volumet.
Opp til 10% av total daglig dollar-volum er omsatt under åpnings- og sluttauksjoner i store Amerikanske børser, men likevel har de fått lite oppmerksomhet fra akademia. For å endre på dette, har vi karakterisert de statistiske egenskapene til auksjonene i NYSE, NYSEarca og NASDAQ ved å tilnærme oss dem som om de var systemer fra fysikk.
Vi har fastslått, gjennom bruk av diverse statitiske tester og kurvetilpassningsmetoder, at fordelingen i størrelsen til individuelle handler er preget av fete haler under auksjonene til NYSEarca, og at kvotienten mellom omsatt dollar-volum under åpningsauksjoner og sluttauksjoner er lognormalfordelt for aksjene som er notert på NYSE, NYSEarca og NASDAQ.
Ved hjelp av tidsserieregresjon har vi klart å vise at daglig omsatt dollar-volum er, til en betydelig grad, forutsigbart for aksjer notert på NYSEarca og NASDAQ.
To regulariteter i førauksjonsperioden, altså når auksjonsdeltagerene sender inn ordrene sine, til NYSEarca åpningsauksjoner ble funnet: Den første viser at det totale dollar-volumet til de kjøps- og salgsordene som er matchet vokser lineært med tiden, for aksjene i NYSEarca. Den andre viser at den indikative auksjonsprisen, hvis rolle er å estimere den endelige auksjonsprisen, systematisk overestimerer den endelige auksjonsprisen med rundt 0.05 til 0.1% mellom de 65. og 85. minuttene i førauksjonsperioden, for NYSEarca aksjene.
Til slutt ble prisdynamikken til NYSEarcas åpningsauksjoner sin førauksjonsperioden undersøkt. Dette ble gjort ved å anvende lineære responsfunskjoner med mål å studere hvordan nye ordre, og ordrekanselleringer, påvirker auskjonsprisen. Her ble det oppdaget at ordre som motvirker ubalansen mellom kjøps- og salgsordre har en mye lavere prisrespons enn de ordrene som tilfører ubalanse, at prisresponsen går mot null når førauksjonsperioden nærmer seg slutten, og diverse andre stiliserte faktum.
Dette arbeidet har vært vellykket i å gi en oversikt i de kvantitative egenskapene til åpnings- og sluttauksjoner til Amerikanske børser, og gjennom dette åpnet veien videre for mer detaljerte undersøkelser.

SUMMARY

Major stock exchanges typically open and close by employing so-called single-price auctions. The auctions work, for each noted stock, by gathering orders through a pre-auction period, and just before continuous trading begins, execute the orders at the final auction price, which is the price that maximises the volume traded.

The opening and closing auctions of major U.S. stock exchanges account for up to 10% of total daily trading dollar-volume, yet they have received little attention from the academic community. To amend this, we characterised the statistical properties of the auctions of NYSE, NYSEarca and NASDAQ by approaching them as physical systems.

Using various statistical fitting and testing methods, it was established that the distribution in sizes of individual trades matched at the NYSEarca opening & closing auctions are heavy-tailed, and that the ratio of daily dollar-volume between opening and closing auctions is log-normally distributed for the stocks of NYSE, NYSEarca and NASDAQ.

Next, time-regressive methods were used to show that the auctions' daily dollar-volume is predictable to a significant extent for the stocks of NYSEarca and NASDAQ.

Afterwards, two regularities in the pre-auction period, which is when the auction participants send in their orders, of NYSEarca opening auctions were found: Firstly, the dollar-volume of the buy & sell orders matched with each other exhibits linear growth as a function of time for the stocks of NYSEarca. Secondly, the indicative price of the auction, which functions as an estimate for the final price of the auction, is shown to systematically over-estimate the final price by 0.05-0.1% between the 65th and 85th minutes of the pre-auction period for the stocks of NYSEarca.

Finally, the price dynamics of the pre-auction period of NYSEarca opening auctions were investigated by using linear response functions to study the effect new orders, and order cancellation, have on the auction price. Here it was found that orders which contribute to the current imbalance (between buy and sell orders) cause a much larger price response than orders which counteract the imbalance, that the price response of all order types tend to zero as the pre-auction period nears its end, and other stylised facts.

This work has successfully produced an overview of the quantitative properties of the opening and closing auctions of U.S. stock exchanges, and also outlined the way for more in-depth investigations into the opening and closing auctions.

PREFACE

This Master thesis is the final work of my MS/Siv.ing. degree in Physics and Mathematics at the Norwegian University of Science & Technology (NTNU). Both the research work and the writing was done during the spring 2017 semester. The field of this project is **econophysics** and my thesis advisor was Prof. Damien Challet. I chose this project for two reasons; firstly, I wanted to try something different from physics and because economics and finance are two old interests of mine, econophysics was a natural choice. Secondly, this project would force me to develop skills, in particular a solid understanding of applied statistics, that will be of use for my future career.

The overarching goal of this project was to find empirical regularities in opening and closing auctions. This is a barely explored topic in academia, which made it more challenging, but also more motivating. The results presented here are entirely new and not replications of previous works. In fact, the plan forward is to publish them in a research journal.

This thesis was written with the typical IMRaD structure in mind, though with significant deviations. Due to the nature of this project, a significant part of it constituted of exploratory data analysis instead of well prepared and well-understood experiments; this forced me to split the work done into three mains chapters (4, 5 & 6), and sometimes combine methods with results to make the logical flow behind my actions more clear to the reader.

Finally, I would like to thank Prof. Challet, my supervisor, for taking me in and spending his time and energy on supervising me; I learned a great deal from him. I am also grateful for the warm welcome I received at the Laboratoire MICS of École CentraleSupélec and for the discussions I partook in with the researchers and Ph.D. students there.

NOMENCLATURE

Abbreviations

RMSE Root Mean Square Error

MO Matched Order

IMP Indicative Matching Price

FMP Final Matching Price

SF Survival Function, also known as Complementary Cumulative Distribution Function

ETF Exchange Traded Fund

RV Random Variable

SE Standard Error

Notation

Subscripts are used to denote the coordinates of a variable.

Superscripts are used to label a variable's origin or type.

Ignoring $t_{\texttt{end}}$ When the timestamp of a variable is $t = t_{\texttt{end}}$, meaning the variable lives at the end of an auction, the timestamp is often discarded out of convenience.

Square Brackets [] are used to denote arguments to discrete functions and round brackets for continuous functions. In other words, $p[t]$ is a time-series, while $p(t)$ is a continuous function of t.

Angular Brackets $\langle\rangle$ represent the mean/average of some set of numbers.

Symbols

(s, d, a, t) Stock (which in this report can refer to any security traded on a stock exchange), date, auction type and timestamp, respectively. These are the coordinates of an auction.

$\mathcal{D}$ Denotes the statistical distribution (random or empirical) of a variable.

$m^k_{s,d,a}$ The dollar-volume of the k-th order matched during the auction (s, d, a).

$\{m^k_{s,d,a} : \forall k\}$ The set of dollar-volumes of all the orders matched during the auction (s, d, a).

$\{m^k_{s,d,a} : \forall k \forall d\}$ The set of dollar-volumes of all the orders matched for stock s during auction type a.

$\mathcal{D}_{m_{s,a}}$ The empirical statistical distribution of the MO dollar-volumes of stock s for auction type a that resulted from aggregating across all d, k available in data (notice the missing d, k).

$v_{s,d,a}$ The daily MO dollar-volume of auction (s, d, a).

$\{v_{s,d,a} : \forall d\}$ The set of all daily dollar-volumes for stock s for auction-type a.

$\mathcal{D}_{v_{s,a}}$ The empirical statistical distribution of the daily MO dollar-volumes of stock s during auction type a that resulted from aggregating across all d available in data (notice the missing d).

$\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}}$ The ratio between opening auction and closing auction daily MO dollar-volumes.

$\mathcal{D}_{\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ The empirical statistical distribution of the ratio between opening auction and closing auction daily dollar-volumes aggregated across d for stock s.

$p_{s,d,a}$ The final matching price of auction (s, d, a).

$p_{s,d,a,t}$ The indicative matching price at time t during auction (s, d, a).

$i_{s,d,a,t}$ The total imbalance between all active sell and buy orders at time t during auction (s, d, a).

$v_{s,d,a,t}$ Total MO dollar-volume during auction (s, d, a) at auction-time t.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

Chapter 1

INTRODUCTION

Firstly, the field of my project is motivated, then the system I have chose to work on is presented and finally the goals of the project are defined.

The language used in this thesis is geared towards physicists, thus the finance jargon is kept to a minimum. To this end, the equity securities of both publicly traded corporations as well as exchange-traded funds (ETFs) are referred to simply as "stocks" in this thesis.

1.1 Motivation

My Master thesis is in applied physics, yet during this work I studied a system from finance. But this is not as surprising as might first seem, because the main benefits of an education in physics is that it teaches one to think about and understand phenomena not only from physics, but also in fields that bear similarity to physics. And finance is one such field.

There exist strong parallels between the study of financial and physical systems. In 1900 the French mathematician Louis Bachelier postulated that the price of stocks, i.e. ownership shares of public[1] companies, were perturbed, or "knocked around", by news and events such that their price at a time t was independent of any previous time $t' < t$, and further that these perturbations were normally distributed. This amounts to a random walk in the price dimension. Half a century later, this notion of stocks moving in memoryless random walks turned into a fundamental axiom of modern finance [2] and had profound influence on the Efficient Market Hypothesis (EMH), a widely accepted theory among financial economists which claims that financial products (e.g. stocks) are almost always optimally priced. It is also the foundation under a great deal of financial models. Around the same time, in 1905, Einstein explained Brownian Motion with the same mathematics; Brownian particles were similarly driven into normally-distributed perturbations and a random walk in 3 spatial dimensions, but this time they were "knocked around" by molecules surrounding them instead of stock-market news. This work ended up helping firmly establish the atomic theory of matter.

[1] Companies noted on stock exchanges.

What is interesting here is that both Bachelier and Einstein understood and modelled their systems in quite analogous ways, despite them operating in two completely separate fields. Now, this of course does not mean that these two systems are identical in behaviour or that models from physics can be liberally applied to finance, but this story does illustrate one of the many parallels between physical and financial systems.

Quantitative finance is the subfield of finance that studies financial systems through mathematics. In many ways, quantitative finance is surprisingly analogous to physics, and therefore many physicists have been attracted to this field; indeed, it is the work done by physicists here that is popularly referred to as "econophysics". This research can be everything from physicists reusing approaches of statistical mechanics to exploring data with the goal of understanding how financial systems work.

1.2 Opening and Closing Auctions

This investigation deals with stock exchange behaviour. Because of the dynamic nature of financial markets and the multitude of different, complex agents with various strategies interacting inside it, stock exchanges can be considered to be non-equilibrium systems made up of heterogeneous, imperfect participants; they could be human or machine, long-term or shot-term focused, highly rational or completely emotional, skilled or incompetent, an insider or an amateur; the broadness of the range of participants is difficult to overstate. In addition, every individual stock can by itself be considered as a sub-system of the stock exchange which interacts with many other assets, and stock exchanges themselves can be seen as sub-systems of the *global market*, the grand system that encompasses all economic activity on earth.

Stock exchanges are businesses and so their goal is to create the best environment possible for their clients, the traders/investors. The traders require two central properties from the exchange; sufficient liquidity, which is its ability to absorb trading orders, and a fair price for the trade(s) they want to carry out, namely a so-called equilibrium price/market price. During the open trading sessions, this is achieved through continuous "auctioning"[2], where traders who wish to sell are matched with those who wish to buy in a highly efficient manner, and the execution of the trade is immediate upon a match. But trading is not continuous around the clock; the stock-exchanges are closed during nights, weekends and holidays.

The periods around the opening and closing of exchanges requires special treatment. Right before an exchange is closed for the day, trading activity may become unstable, as many traders wait with carrying out their trades until the final minutes, for a variety of reasons [13]. This can cause buy/sell imbalance, price instability, liquidity shocks and favourable conditions for price manipulation; which both hurts the availability of liquidity towards the end, and creates uncertainty in the values of various external financial products that use the closing prices as reference prices. When the stock exchanges close, the stock prices are freezed but the world moves on; by the time the stock exchange opens, the previous closing price has become largely irrelevant due to the arrival of new information, and so every stock price has to undergo a new price

[2]From now on I will refer to it as simply continuous trading

discovery process (which cannot be continuous trading, as that would lead to price instability, and hence a lack of liquidity until the equilibrium price is found). To solve these issues, stock exchanges employ auctions[3] to open and close their open-market trading [16] in a smooth way.

While the precise rules and regulations for the auctions vary between exchanges, the general principles behind them remain the same. They essentially work by allowing individual traders to place various types of buy/sell orders, e.g. limit orders and market orders (which are defined in section 3.1.1), within the pre-auction period, with the promise that they will get the best price possible on their trade at the conclusion of the auction. This price is found through algorithms that take into account the supply & demand, and then at the conclusion of the auction, all compatible orders are matched and carried carried out at this price, the final market-price of the auction.

The behaviour of stocks during open trading has been extensively explored in academia, but, surprisingly, not their behaviour during auctions; there is in fact very little in the literature on the statistics of auctions. This is a peculiar situation as the auctions make up around 2-10% of the total dollar-volume traded on major U.S. stock exchanges [1]. Therefore we decided to study the statistical regularities of stocks during auctions at three very large exchanges: NYSE, NYSEarca and NASDAQ. Our data originated from Thomson Reuters Tick History.

Auctions exhibit complex dynamics as they take place, but the result at their conclusion is always that a range of buy and sell orders of various sizes[4] are matched and exchanged at a final match price. Thus the distribution in the size of these matched orders (MOs) along with the final price are the two features that largely define the outcome of every auction. Therefore this project revolved around investigating the price-finding process and the properties of the MOs. Before speaking more on this though, let us first describe the auction process in a more mathematical manner.

1.3 Describing Auctions Mathematically

This section will mathematically define the auction process, introduce the most important quantities of the auctions, and establish a notational precedent which will be used throughout this thesis.

1.3.1 Denoting an Auction

The auction-process takes place for every combination of stock s and trading date d, and its type a can be one of two kinds; opening auction or closing auction, thus $a \in \{\texttt{open},\texttt{close}\}$. Every stock is part of an exchange, i.e. $s \in \mathbb{S}_E$ where $\mathbb{S}_E$ represents all the stocks listed on the exchange E. In this project, there was data available on NYSE, NYSEarca and NASDAQ, so $E \in \{\texttt{NYSE}, \texttt{NYSEarca}, \texttt{NASDAQ}\}$.

[3] For the record, they are formally called "single-price auctions", as all the orders matched during the auction are carried out at a single price, the so-called final matching price of the auction.

[4] With the size of a

Thus it is natural to represent an auction of type a for stock s that took place during day d as (s, d, a, t), with t representing the auction's temporal dimension. The so-called pre-auction period takes place during $t \in [t_{\texttt{start}}, t_{\texttt{end}}\rangle$, and at $t = t_{\texttt{end}}$ the auction is concluded. Throughout this thesis, auctions will often be denoted as simply (s, d, a), with t being added only when the auction's temporal properties are relevant.

1.3.2 Auction Variables

At the conclusion of an auction, $n_{s,d,a,t_{\texttt{end}}}$ MOs are executed, with each of these MO being representable by an element in some set $\{m^1_{s,d,a,t_{\texttt{end}}}, m^2_{s,d,a,t_{\texttt{end}}}, .., m^{n_{s,d,a,t_{\texttt{end}}}}_{s,d,a,t_{\texttt{end}}}\}$. The elements $m^k_{s,d,a,t_{\texttt{end}}}$ themselves represent the size of the matched orders (MOs) in terms of *dollar-volume*. This dollar-volume is equal to the volume/turnover multiplied by the final matching price of the auction:

$$m^k_{s,d,a,t_{\texttt{end}}} \equiv q^k_{s,d,a,t_{\texttt{end}}} p_{s,d,a,t_{\texttt{end}}} \tag{1.1}$$

When $p_{s,d,t} \colon t < t_{\texttt{end}}$, the price is called the indicative matching price of the auction. Furthermore, because all the buy orders are matched with sell orders, the elements of every set of MOs $\{m^k_{s,d,a,t_{\texttt{end}}} \colon \forall k\}$ are interrelated with each other through the following manner:

$$\left\{ m^k_{s,d,a,t_{\texttt{end}}} \colon \sum_{k \in \mathbb{O}^{\texttt{Buy}}_{s,d,a,t_{\texttt{end}}}} m^k_{s,d,a,t_{\texttt{end}}} = \sum_{k \in \mathbb{O}^{\texttt{Sell}}_{s,d,a,t_{\texttt{end}}}} m^k_{s,d,a,t_{\texttt{end}}} \right\} \tag{1.2}$$

Where $\mathbb{O}^{\texttt{Buy}}_{s,d,a,t_{\texttt{end}}}$ and $\mathbb{O}^{\texttt{Sell}}_{s,d,a,t_{\texttt{end}}}$ are the lists of buy and sell orders that were matched at the conclusion of the auction $(s, d, a, t_{\texttt{end}})$. In other words, the MOs $m^k_{s,d,a}$ are not independent of each other.

For convenience, t will from now on often be dropped from the variables' subscripts if $t = t_{\texttt{end}}$, i.e. if the variable is at the end of the auction.

Moving on, the aforementioned set $\{m^k_{s,d,a} \colon \forall k\}$ is distributed according to an empirical density $\mathcal{D}_{m_{s,d,a}}$. It turns out that one can think of this empirical density as the realisation of a probability distribution, and as a consequences $q^k_{s,d,a}$, $n_{s,d,a,t_{\texttt{end}}}$ and other empirical quantities are also realisations of random variables (RVs). In other words, the empirical density $\mathcal{D}_{m^k_{s,d,a}}$ can be considered to be a realisation of a probability distribution $\mathcal{D}_{\mathsf{M}^k_{s,d,a}}$, and thus $m^k_{s,d,a}$ is a realisation of some RV $\mathsf{M}^k{}_{s,d,a} \sim \mathcal{D}_{\mathsf{M}^k_{s,d,a}}$. RVs are denoted with a capitalised version of their observables' letter.

Thus from equation (1.1) it follows that $\mathsf{M}^k_{s,d,a}$ can be represented as the product of the RV of the volume of the k-th MO, $\mathsf{Q}^k_{s,d,a}$, and the RV representing the final match price $\mathsf{P}_{s,d,a}$:

$$\mathsf{M}^k_{s,d,a} \equiv \mathsf{Q}^k_{s,d,a} \mathsf{P}_{s,d,a} \tag{1.3}$$

Another quantity of interest is the sum of the individual MOs:

$$v_{s,d,a} \equiv \sum_{k=1}^{n_{s,d,a}} m^k_{s,d,a} \tag{1.4}$$

which will from now on be denoted as the daily MO dollar-volume $v_{s,d,a}$ of the auction (s, d, a). Analogously to the previous definitions, its RV cousin is denoted by $\mathsf{V}_{s,d,a} \sim \mathcal{D}_{\mathsf{V}_{s,d,a}}$.

The variables above characterise the end of the auction, i.e. their $t = t_{\texttt{end}}$, but the notational precedent is extended in a straight-forward way for the pre-auction times $t \in [t_{\texttt{start}}, t_{\texttt{end}}\rangle$. For instance, the daily MO dollar-volume currently matched at time t during the pre-auction period is simply denoted by $v_{s,d,a,t}$, and its RV cousin by $\mathsf{V}_{s,d,a,t}$. And so on so forth for the other variables.

1.3.3 The Auction Process

In formal terms, stock exchanges use so-called single-price auctions/call auctions to find appropriate opening and closing prices. Roughly speaking, the auction-processes can typically be divided into three parts; the pre-auction period, the freeze period and the execution period. During the pre-auction period, traders submit orders to buy or sell given quantities of shares of certain stocks, and at the same time market information is disseminated by the stock exchange; one such piece of disseminated information is the indicative matching price, which is supposed to estimate the final matching price. During the freeze period, which takes places during the final seconds, traders are no longer allowed to modify or cancel their orders, and the only new orders that can be entered are those that counteract any buy/sell imbalance present between the market orders. Finally, during the execution, the final matching price is calculated and all eligible orders are matched at said price.

The precise rules and mechanisms vary between stock exchange, but they are all tâtonnement processes, meaning process that find an equilibrium price that maximises the exchange of stocks. In other words, the goal of the auction-process is to always find a final matching price $p_{s,d,a,t_{\texttt{end}}}$ that maximises the daily MO dollar-volume $v_{s,d,a,t_{\texttt{end}}}$ at the conclusion of every auction (s, d, a). Section 3.1.3 describes the auction process of NYSEarca in detail, and in addition outlines those of NYSE and NASDAQ.

1.4 Goals of Research

When faced with a new system, it is often a good idea to start by characterising the properties of its basic components. In finance, this often means classifying its statistical properties, and so we studied the distribution of MOs at both order-by-order and auction-by-auction scales, and looked for patterns in the pre-auction dynamics. The specific goals of this work are presented here.

1.4.1 Properties of Individual Auctions

Firstly the properties of the auctions for individual days were explored. As previously mentioned, the distribution of the MO dollar-volumes (along with the final matching price) define the outcome of auctions, and so the investigation was centred around $\mathcal{D}_{m_{s,d,a}}$.

At the very beginning, $\mathcal{D}_{m_{s,d,a}}$ appeared to be power-law distributed, or at least

heavy-tailed, which is interesting because many quantities in finance are known to be heavy-tailed (in particular daily volume, price returns, individual order sizes, etc.) [45]. Thus the investigation started by investigating the tail properties of $\mathcal{D}_{m_{s,d,a}}$.

Questions to be Investigated:

1. Is $\mathcal{D}_{m_{s,d,a}}$ heavy-tailed, and if so, power-law distributed?
2. What can be said on the statistical features of $\mathcal{D}_{m_{s,d,a}}$?

The approach will be to systematically fit power-laws and other heavy-tailed distributions to the tails of $\mathcal{D}_{m_{s,d,a}}$, and use the results to uncover the statistical features of $\mathcal{D}_{m_{s,d,a}}$.

1.4.2 Aggregate Properties of the Auctions

The aggregate behaviour, i.e. over long time ranges[5], of the auctions were explored by studying the behaviour of the daily MO dollar-volume $v_{s,d,a}$. Preliminarily, due to the relation defined by equation (1.4), it was checked whether one could expect to infer important properties of $\mathcal{D}_{v_{s,d,a}}$ when knowing the behaviour of $\mathcal{D}_{m_{s,d,a}}$. Further, it turned out that $\mathcal{D}_{\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}}}$ appeared to be log-normally distributed and $v_{s,a}[d]$ was to some extent predictable; both of these qualities were investigated in detail.

Square brackets instead of subscripts will be used when emphasising one of the parameters of a discrete function. For example, $v_{s,d,a}$ will be written as $v_{s,a}[d] \equiv v_{s,d,a}$ below to emphasise that its time-evolution across d is of interest.

Questions to be Investigated:

1. What properties of $\mathcal{D}_{v_{s,d,a}}$ can be inferred from the properties of $\mathcal{D}_{m_{s,d,a}}$?
2. Is $\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}}$ log-normally distributed?
3. Further, what can be said concerning the statistical properties of $\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}}$?
4. Is the time-evolution of $v_{s,a}[d]$ predictable?

The first point will be resolved by a theoretical treatment, while the second and third points will be studied by systematically fitting log-normal distributions. The last question will be resolved by applying various regressive forecasting algorithms on $v_{s,a}[d]$.

1.4.3 Pre-auction Dynamics

Finally, the dynamics of the pre-auction phase were investigated. Firstly, to gain an overview, the available pre-auction data was systematically checked for any regularities across time. Further, one of the premier concerns of large investors is minimising

[5]Aggregate properties here means "properties of the whole system". When this the term is used with regards to with sets, it is meant combining smaller sets into larger ones.

their market impact[6], and so it was of interest to investigate how buy/sell orders shift the stock prices of the auction; this was done by studying the so-called price response function.

Questions to be Investigated:

1. Are there any statistical regularities present in the pre-auction period?
2. What is the price response of individual orders?

The simplest way of investigating market impact [36] is by measuring the price response. To this end, multiple price response functions were defined and applied on the data to measure the the average shift in price due to a new order, conditioned on the type of order. This was done both for the indicative matching price, and much more importantly, the final matching price.

[6]The market impact of a trade on a stock is the change in price it induces; because it makes trading more expensive, it is in the interest of all traders, with the exception of price manipulators, to minimise their market impacts.

Chapter 2

BACKGROUND

This chapter will give an overlook of finance. In particular, different approaches to financial modelling and analysis will be discussed.

2.1 Basics of Finance

Finance is the science of money and investing. Money is fundamentally a tool for facilitating trade, i.e. buying and selling of products, such as financial products. These financial products can be everything from stocks, bonds[1], derivatives [2] to commodities [3] to nice cars and real estate, and so on. Whatever the product, the goal is the same: maximise the the expected profit, and control its risk [4]. This is the core of finance, but there is also the matter of actually carrying out these transactions.

Financial transactions can in principle be carried out in person between two parties who are in agreement and shake hands, but this is not how it is typically done today (though sometimes this is necessary, like when property changes hands or during the acquisition of a company). This because it is comparatively very inefficient for an agent to personally find a buyer or seller, when there exist financial markets with well-programmed computers that can efficiently and fairly do this job instead (for a small price). This is called matchmaking and it is perhaps the most important role of a market.

Financial assets are often traded through exchanges. Each exchange specialises in a given type of asset, with the best known being stock exchanges (which trade in stocks and related equities) and foreign currency exchanges (FOREX).

Due to the enormous sums of money traded on stock exchanges every date, where

[1]This type of security entitles the holder to, at a pre-defined point in time, a debt repayment from the issuer of the bond.

[2]A "derivative", depending on its nature, either obliges or entitles the holder to a certain action. They are essentially either used to bet on the performance of other securities or products, or to "hedge" an investment, i.e. decrease its risk.

[3]An item sold to satisfy a need. E.g. Ammonia sold to a fertiliser-producing company.

[4]In finance, an investment's risk is typically defined as the uncertainty of the return in said investment.

it is not uncommon that single stocks have daily volumes[5] in the order of billions of dollars [4], much effort has been made to study their behaviour. In particular, predictability is a central question. If a trader is able to identify a predictable pattern in the market price fluctuations of a stock, he might be able to carry out trades on this stock that allows him to profit. In fact, finding these patterns is the job of many physicists, mathematicians and engineers today working in (and sometimes owning) hedge funds and banks. In finance jargon these patterns are commonly referred to as *market inefficiencies*, because they are deviations from what otherwise would be an Efficient Market, as per the notorious *Efficient Market Hypothesis* (EMH).

The Efficient Market Hypothesis essentially says that markets are efficient if they reflect all available information, with there being three categories, or forms, of "all available information". These forms are ordered by the strength of the efficiency, and each weaker form is implicit in all stronger forms. In strong-form efficiency, markets price everything perfectly because all information available to humanity is flawlessly employed to price each financial product. In semi-strong-form efficiency, prices completely reflect all publicly available information (meaning all information except insider-information). In weak-form efficiency, prices reflect all previously available *price-information* , which means no autocorrelation in price exists and thus future prices cannot be predicted from past ones; in other words, weak-form efficiency implies that price movement has *no memory* and hence moves in some kind of random walk, which is in-line with Bachelier's original 1900 model.

The strong-form expression of the EMH has a strong consequence. Because all information $\Omega(t)$ at time t is available to the N rational and intelligent traders of security s, the fundamental price[6] $p_s^*(t)$ at time t is in fact in the limit of $N \to \infty$ defined by the expectation value of traders' best guess at $p_s^*(t)$, which can be considered as a RV $\mathsf{P}_s(t)$:

$$p_s^*(t) = \lim_{N\to\infty} \mathtt{E}[\mathsf{P}_s(t)|\Omega(t)] \tag{2.1}$$

This presents a very convenient and simple way to look at things - just put your faith in strong-form market efficiency and you can be safe that whenever you trade anything on the stock market you are getting a fair price. Also known as the market is always right, trust the market, and if you subscribe to neoclassical economics, do as the market commands.

2.2 The Challenge of Financial Markets

These opportunities, market inefficiencies, are far from easy to find however, because every trader is looking for them, and when one is found, it can be mathematically shown that exploiting it removes it. This tendency to hunt for inefficiencies leads to financial markets exhibiting extremely complicated behaviour. Take the BP stock on the London Stock Exchange for instance, which is traded in the order of billions of GBP every day. A large and heterogeneous mix of agents, which includes everyone from small-time investors to robots, perform many different trades on both this stock

[5]In finance, volume either refers to the raw number of shares traded or their value, depending on the context.

[6]The "fundamental price" of a stock defines the real value of the company at the current time.

and the multitude of derivatives stemming from it, with strategies that are based on not only on impressions of how BP itself is doing, but also on what the other agents in the market are doing and on how they might react to other agents' strategies. The trading of this stock alone is thus a hugely complex, chaotic non-equilibrium system with nonlinear feedback loops, many time scales, disordered interactions and heterogeneous and learning agents.

What kind of rich behaviour can we observe from such a system? A financial system's state is said to be in equilibrium when none of the agents can increase his utility[7] by changing his strategy, in other words when the agents exist in a Nash-equilibrium [8]. Nash-Equilibrium is an implicit assumption in the EMH, but real markets are clearly out of equilibrium whenever emotions take over, such as when emotions drive traders to start herding [7], i.e. changing their strategy in response to others changing theirs (also known as peer-pressure). The interaction between traders, and even their own thought processes, are clearly nonlinear due to emotions alone, and this gives rise to many nonlinear feedback loops that can result in all manners of collective phenomena, such as panics and even what has been described by physicists as phase changes.

The inherent instability present in markets has been studied extensively in the literature. For example, in 2003 D. Challet and M. Marsili [25] used minority games models and statistical mechanics to explain a possible origin. They produced a market model which exist in two phases, one where the prices are predictable and one where they are not, and showed that completely unpredictable prices lead to market instability. A later example is Patzelt & Pawelzik's 2013 paper [8], where they show in a generic way that perfectly learning markets, equivalent to perfectly unpredictable markets, must be highly volatile[9].

2.3 Methodologies of Physicists and Economists

The ultimate way of understanding and capturing the behaviour of a system is by creating a model of it, and in finance there are two important paradigms for doing that; financial economics and econophysics. The former evolved from classical economics, while the latter can trace it roots in statistical physics.

Modelling is a central part of physics. When a physicist attempts to create a model, the first step is to observe a phenomena and look for patterns, the second is to identify, or at least guess, the causes and effects, the third is to put this into a mathematical framework, and finally use the tentative model to create predictions and find a way to test them. If the last point turns out good then a phenomena is successfully classified, and if not then it is time to try again. In my opinion, two ingredients in particular makes this approach useful across fields: Firstly, the aforementioned phenomena do not need to be within the realm of physics, or even the natural sciences, they just need to be measurable, which is precisely the case with much of finance. Secondly, physicists model systems from the "ground up"; meaning, if no established theory

[7] In finance, an agent's utility represents his level of satisfaction.

[8] Meaning no agent has an incentive to change their strategy, in this case buy/sell the stock

[9] High volatility means large fluctuations in price.

exists, they create their models on the basis of the empirical behaviour of a system, and only then try to formulate the theory. This latter point is an important one, because economists often do the opposite.

Finance was, and to a certain extent still is, dominated by economists with a way of modelling that is very different from that of physicists. Economists are typically more concerned with deriving mathematically elegant models based on idealised first principles borne out of economical frameworks such as microeconomics and beliefs such as the EMH. Essentially, they derive mathematical models from what they consider truth, analogous to how mathematicians derive results from axioms. Physicists on the other hand tend to focus on creating models that are able to reproduce empirical results to the greatest possible extent, rather than caring whether they fit into any idealised frameworks or are mathematically elegant. In other words, the main goal for a physicist is to understand her systems, with models being a powerful tool for doing this, while an economist focuses on producing models that are already based on a pre-defined set of beliefs of how her system works.

Let us have a closer look both methodologies, starting with the one that was applied to finance first.

2.4 Classical Financial Economics

In essence, financial economics is the extension of microeconomics and the decision theory of rational agents to financial markets.

Modelling in financial economics is based on predefined first principles. Typically, financial models are created so that they exist in so-called "worlds" approximating reality. First postulates concerning the behaviour of the agents in the system are formulated, then the rules of how the system is supposed to work are outlined, and finally, based on these first principles, a model is derived. Some first principles are so often used that they have even become axioms of financial economics, the most prominent of which are rationality of agents and the efficient market hypothesis. The first entails that every agent in the financial market always seeks to maximise his profit over his individual time scale, market efficiency entails that every one of these agents consistently take into account all available information before carrying out any transaction and are in nash-equilibrium with each other. Some first principles have become so powerful and ingrained that they even form the basis of whole economic schools of thoughts. An example of this is the "invisible hand of the market" principle which is the basis of neoclassical economics. It states that when all the agents of the market are in collective pursuit of profit, they end up doing what is best for society as a whole.

Take the case of the BP stock. Classical finance attacks the problem of modelling it by using first principles assumptions to create theoretical models that can capture aspects of the system's behaviour. To illustrate, consider the following two models taken from financial economists; The Capital Asset Pricing Model (CAPM) and the Black-Scholes model. The Capital Asset Pricing Model was introduced in the 1960s and is used to decide on whether buying a stock or not is a rational choice. What is

considered a "fair price" for the European-style put/call options[10] based on the underlying stock can be calculated with the famous Black-Scholes model, introduced in 1974. Both of these models resulted in Nobel Prizes in economics.

The CAPM lives in a world mainly described by the following first principles: The stock's price is defined by a normally distributed random walk (as proposed by Bachelier in his original work), all market conditions, except linear growth in returns [11], remain constant, and all agents are risk-averse and in equilibrium. Its purpose is to calculate the appropriate return over some period of time a risk-averse investor should demand from the market given the risk he takes. The equation behind the model is:

$$\mathrm{E}[\mathsf{R}_i] - r_f = (\mathrm{E}[\mathsf{R}_m] - r_f)\left(\frac{\sigma_i}{\sigma_m}\right)^2 \tag{2.2}$$

Here R_i is the RV denoting the (risky) rate of return of the stock (i for investment), r_f is a constant denoting the risk-free rate, R_m is the RV denoting the (risky) rate of return of the market portfolio and σ_i, σ_m are the observed standard deviations of the stock and market portfolio, respectively.

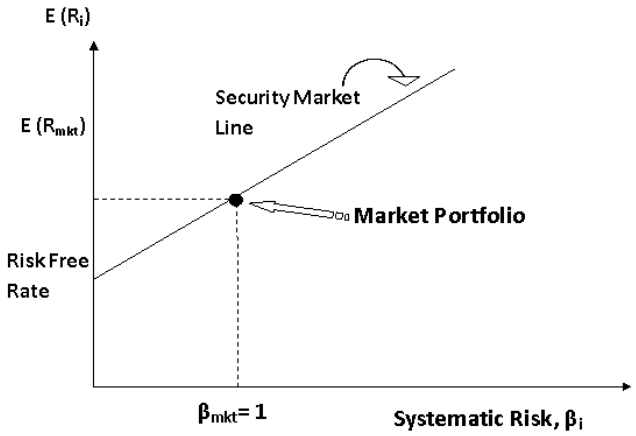

Figure 2.1: An illustration of the CAPM. $\beta_i \equiv \frac{\sigma_i}{\sigma_m}$. *(Augment Systems Pvt. Ltd)*

The most important assumptions in the Black-Scholes world are perfect markets[12], existence of a risk-free interest rate and that the price of the underlying stock goes as a *geometric* random walk (which is a correction of Bachelier's earlier model). Using these assumptions (along with a few other standard ones from economical finance), it can be shown that the fair price V for a European call/put option on an underlying stock without dividends is found from:

$$\frac{\partial V_i}{\partial t} + \frac{1}{2}\sigma_i^2 p_i^2 \frac{\partial^2 V_i}{\partial^2 p_i} = r_f V - r_f p_i \frac{\partial V_i}{\partial p_i} \tag{2.3}$$

[10]Depending on the type of option, a trader who owns an option has the right to sell/buy the underlying security for a certain price at, or until, a certain date.

[11]The return $r_s(t, t_0)$ of stock s between some initial time t_0 and current time t denotes its relative growth (or fall) in price p_s in said time period: $r_s(t, t_0) = \frac{p_s(t)}{p_s(t_0)}$

[12]Perfect markets are completely efficient, have limitless liquidity, zero transaction costs and one can even trade fractions of stocks on them.

Where $p_i = p_i(t)$ is the price of the underlying stock at time t and σ_i is the standard deviation of the stock.

Both of these models can be applied to the BP stock example, and are indeed actively used in real trading. However, the CAPM is so simplified that it can at most be used as a "back of the envelope" calculation to check whether an investment is worth considering. The Black-Scholes model has uses in estimating the prices of options when its serious limitations and known discrepancies (e.g. the option smile) are taken into account and patched up [6].

This first-principles approach dominates financial economics, but it has serious flaws. The worlds the financial economists create are far too simple and rosy, with particularly the central assumptions of rational agents and efficient markets being dubious; the claim that trading agents are rational and unemotional profit-generating machines falls on its own unreasonableness, and the notion of efficient markets, which has even been elevated to worrying heights by many economists [9], forbid the existence of phenomena that are known to happen, such as financial bubbles[13]. Unrealistic assumptions like these lead to models that are strongly limited, yet mathematically elegant and simple and thus convenient to believe in and use.

A main limitation of these models is that they neglect the existence of extreme price movements and, as previously mentioned, financial bubbles, yet this is often ignored by their users. A consequence of this is that when these models are used to price products which are already radically mispriced, the market can become unstable and crash. This was the case with the Black-Scholes model and its effect on the October 1987 crash, and the models used to price the sub-prime mortgage packages that helped trigger the 2007-2008 financial crisis [5].

While this might sound grim, it is not meant to discount economical finance and least of all economics. Economics have met success elsewhere, such as in the study of optimal resource allocation and in creating an intuitive understanding of how most aspects of an economy functions. Many (but not all, as exemplified by the insistence of Fama on the infallibility of the EMH), economists are also empirically-minded enough to accept the serious limitations of the models they use in finance. The point I was trying to make in this section is that economical thinking has it limits and that it can benefit from exchange of thoughts with other fields, most prominently of which is physics.

2.5 Econophysics

Physicists do things differently. The insistence of economists to build models that are mathematically elegant and based on highly idealised worlds is in stark contrast to the empirical, data-driven, approach employed physics, where a model is only as good as the validity of its predictions, regardless of how elegant it is. This lack of empirical attitude among economists, increased competitiveness within theoretical physics and the many exciting properties of economical systems has over the years led to physicists

[13] Trade in assets that are vastly overpriced.

migrating into quantitative finance and creating a new academic field, econophysics. It is worth noting that physicists are not just active in academic finance. Many are paid ample sums by investment banks so that they can use their education to search the market for inefficiencies that can be exploited by the bank's traders, or to control the risk of investments. The research physicists do in industry is however confidential, so I will only talk about the results from academia.

Econophysics applies the thinking of physics to topics of finance. This includes the aforementioned empirical and "ground up" methodology of physics as well as their analytical techniques. One such technique is random matrix theory (RMT), which was first used in nuclear physics to create statistical models of the energy spectrums of large nuclei, and is now similarly used in finance to remove risk from large portfolios[14]. Another prominent example of physics expertise applied to finance is the proliferation of power-laws and heavy-tails in finance. Let us talk about these two achievements.

In the 1950s physicists were scattering neutrons of heavy nuclei, and it was observed that the neutron scattering cross section as a function of neutron energy for each nucleus had distinct peaks that were orders of magnitude above everything around them - see figure 2.2.

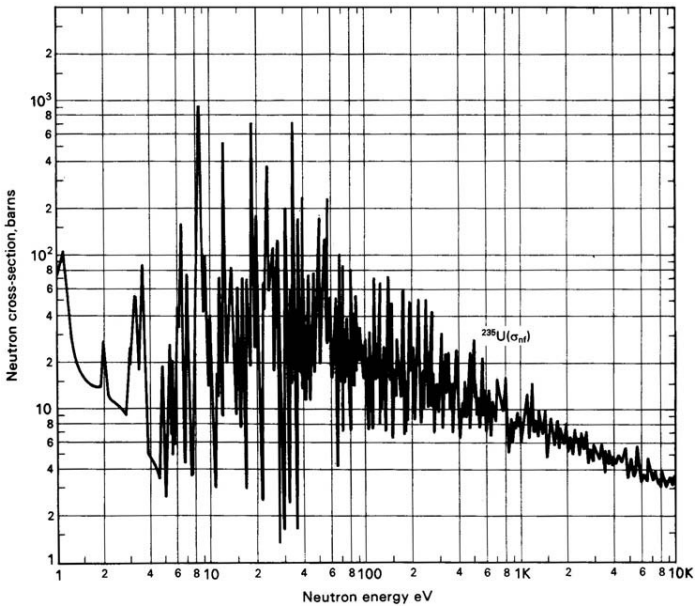

Figure 2.2: The total interaction cross-section for neutrons scattered on U-235 versus neutron energy. Note the resonance peaks. (Kayle& Laby, Ch. 4.7)

Accurately modelling the spectrums of these heavy nuclei from the ground up through quantum mechanics and electrodynamics was (and still is) an insurmountable challenge due to the extreme complexity of their internal structures. The case was not hopeless though; general statistical patterns were observed in the spectra of the heavy nuclei. Eugene Wigner saw this and he reasoned that while it is impossible to analytically calculate the resonance energies, one could treat the extremely complex system

[14]In finance, a portfolio refers to the ownership of a set of weighted securities.

that is the nuclei as a black box which generates energy spectra according to a certain kind of statistics. This black box system would have to be characterised by some kind of Hamiltonian which permitted the existence of all types of nuclei while simultaneously being restricted by the universal laws of quantum mechanics (in particular hermiticity), and Wigner indeed found such a Hamiltonian; in matrix form. The structure of Wigner's matrix reflected the laws of physics, while its entries were RVs that represented possible nuclear structures. Thus, calculating the statistical distribution of its eigenvalues would give the probability distribution $\rho(s)$ of the nearest-neighbour spacing s between the resonance peaks, known as the Wigner Surmise:

$$\rho(s) \sim se^{-\pi/4)s^2} \tag{2.4}$$

This is a successful statistical model that has been empirically verified for isolated peaks, and the mathematics Wigner created to do all of this came to be known as random matrix theory. See figure 2.3 as an illustration; in particular, note the accuracy of the prediction, and more importantly the fact that the empirical spacing distribution is universal, which means it does not depend on the nuclei.

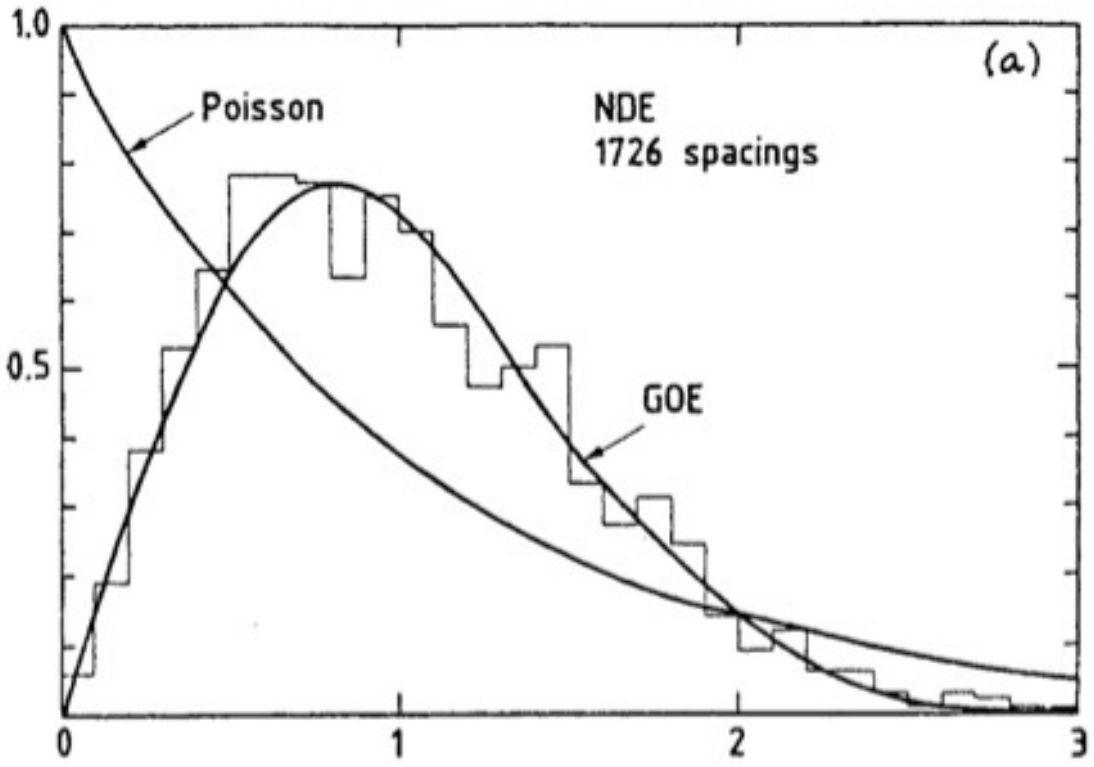

Figure 2.3: The Wigner Surmise superimposed on a histogram of 1407 empirical spacings of isolated resonance peaks taken from a host of different nuclei. (M. Mehta, *Random Matrices 2004, Fig 1.4*)

Stock portfolios, which can be worth $ billions, are generally created with the goal of minimising un-systematic[15] risk. This correlation between all stocks $i, j \in \{1, 2, .., N\}$

[15]Risk that that is not a result of market fluctuations, but due to correlations in the price-movements of the stocks in the portfolio.

making up a portfolio of N stocks can be represented by a correlation matrix $\mathbf{C}$ with elements:

$$C_{i,j} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sigma_i \sigma_j} \tag{2.5}$$

where $x_i[t]$is the time-series of price changes of the stock i and σ_i is its standard deviation. The expected return R_P of this portfolio is defined as weighted sum of the individual expected returns R_i multiplied by the money invested into them, $R_P = \sum_{i=1}^{N} w_i R_i$, and its risk is often expressed as the variance $\sigma_P^2 = \sum_{i=1}^{N} w_i C_{i,j} w_j$. The optimal portfolio then becomes the one that minimises $\sigma_P^2$for any given R_P given the set of N stocks available, and doing this is a standard linear problem.

Finding a reliable empirical determination of $\mathbf{C}$ is very difficult though, and this is where RMT comes into the picture. Using principle values analysis, it can be shown that the eigenvector and eigenvalue pairs of $\mathbf{C}$ represent different classes of risk, with the eigenvectors with the smallest eigenvalues being heavily favoured by a risk-minimising portfolio. However the price x_i is, as previously discussed, a time-series that moves according to an extremely complicated function and is thus very noisy, which carries over to $\mathbf{C}$ and its eigenvalues being noisy too. This leads to the question of how to identify the smallest eigenvalues correctly, in other terms how does one separate the signal from the noise? To answer this, it turns out that it is useful to compare the behaviour of $\mathbf{C}$ with a random-matrix version of it.

If $x_i[t]$ was assumed to be perfectly noisy, all the elements $C_{i,j}$ could be modelled as RVs in a symmetric matrix. This is completely analogous to how Wigner modelled the internal structure of nuclei as randomly distributed entries in a symmetric hamiltonian matrix, and so the whole machinery can be carried over from physics to finance. Doing this, it turns out that information-carrying eigenvalues, and their corresponding eigenvectors, can be isolated and employed in risk optimisation. In an early demonstration of this, a group of physicists [10] managed to reduce the risk of a large portfolio by a factor of 2 compared to the classical Markovitz efficient frontier that economical theory considers to be optimal!

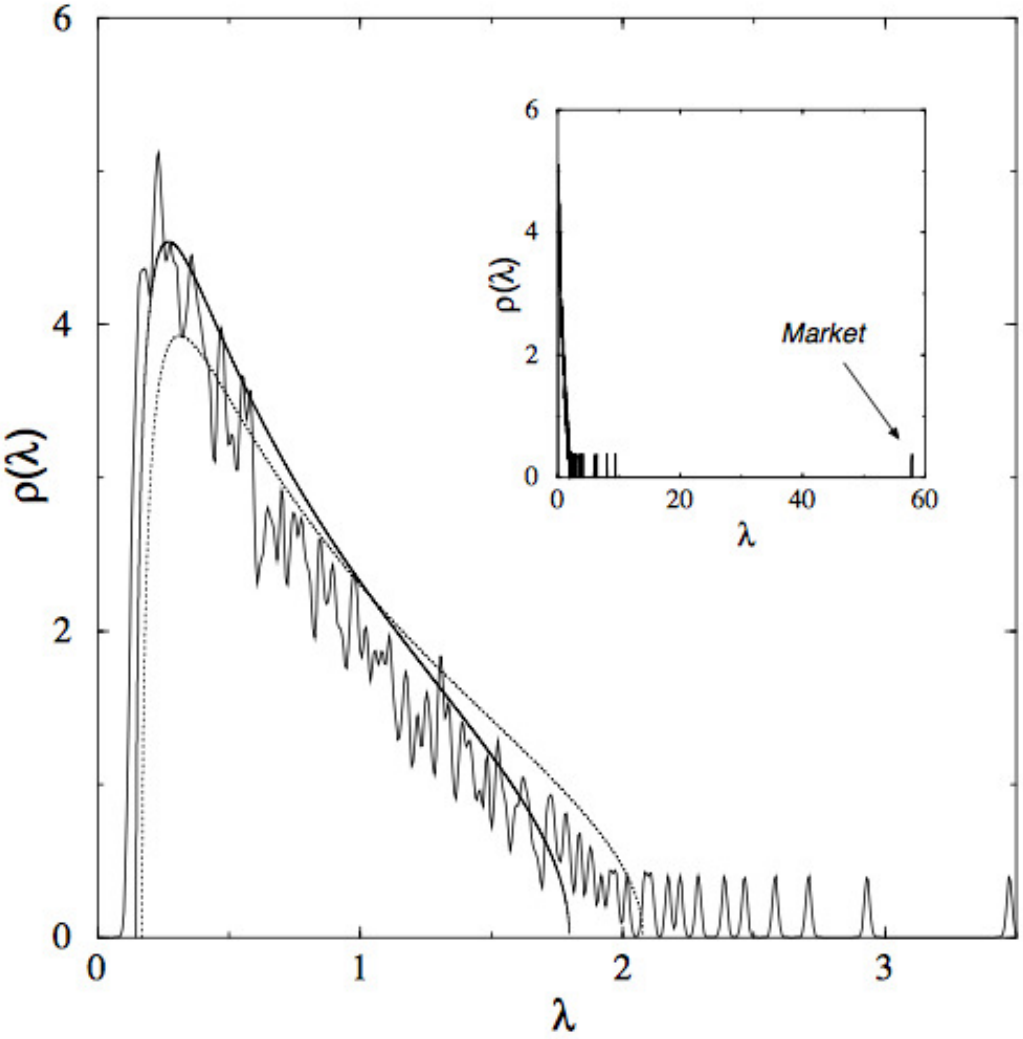

Figure 2.4: Density of the empirical eigenvalues λ calculated from a correlation matrix constructed from a portfolio of $N = 406$ stocks from the $S\&P500$ during the years $1991 - 1996$, along with curves that represent the theoretical probability density distribution of a purely random matrix. The eigenvalues outside of the curves are the signals in the noise. (L. Laloux, P. Cizeau, M. Potters, J. Bouchaud, *Random matrix Theory and Financial Correlations, 2000*)

Another good example of the contribution of econophysicists are power laws. Power laws are functional relationships where one function f varies as a power in some exponent α of another function g such that $f(g) \sim g^{-\alpha}$ (see 3.2.3). The presence of power law tails in the distribution of stock-returns was first suggested by Mandelbrodt in the 1960's [11]; he found evidence for stock price movements being one of his fractals, and therefore that the distribution of stocks' returns were Lévy stable with heavy power-law tails. The last observation was pioneering and led to the introduction of heavy-tails into finance, but Mandelbrodt was wrong in one important aspect; due to a lack of data, he calculated a power-law exponent was far too small. It was so small in fact that it implied an infinite variance in the stock price, and this paradox led to much confusion in the finance community. It was not until the physicists P.

Gopikrishnan, V. Plerou, L. Amaral, M. Meyer and E. Stanley [12] analysed a much larger data-set that this was resolved. Physicists are used to working with huge data sets, and so they managed to firmly establish that there was indeed a power-law in the price returns, but with a larger exponent that did not imply any infinite variances. To be precise, they put the value on this exponent α_r at ≈ 3 for the S&P500, which means that the probability of the log-return of $\mathsf{R}_s(t, t-\Delta t)$ of a stock s rising above some value x being

$$P(|ln(\mathsf{R}_s(t, t-\Delta t))| > x) \sim x^{-\alpha_r} = x^{-3} \tag{2.6}$$

After the exponent was established, the next question physicists started asking was where these power-laws in the distribution of stock returns came from. In other words, what kind of microphenomena are the reasons behind them? Power-laws can have many possible origins (with the origin sometimes revealed by the exponent), and physicists have offered many explanations.

An interesting explanation for them comes from statistical mechanics. In 2001, D. Challet, A. Chessa, M. Marsili & Y. Zhang [15] created an agent-based mean field model of a financial market that turned out to partially reproduce the observed power-laws. To understand why, recall that when a physical system approaches its critical point during a continuous phase transition, its microscopic correlation lengths start diverging and thus the system starts losing its scale. In the case of a sufficiently large system, it becomes completely scale-free, and thus its responses turn into scale-free power-laws. With this line of thinking, these aforementioned physicists created an agent-based model of a financial market that had two phases; one where the market was not efficient and one where it was. This system could be put into criticality by adjusting an independent parameter representing a ratio between the complexity of possible strategies the agents could employ and the number of active agents in the game, and this immediately led to the system exhibiting power-law distributed returns due to scale-invariance with regards to time.

A later explanation, proposed by X. Gabaix, P. Gopikrishnan, V. Plerou & E. Stanley [14], was that optimally-executed, large trades from big market participants was the origin of the power laws.

My hope is that this long text has illustrated the approach employed by physicists doing research in finance as well market some of their successes. Econophysics is far from perfect however [3]: Completely inappropriate physical models have previously been force-fitted to financial markets, physicists have received criticism for wrongly dispensing with statistical tests [3] and it has also been said that many early econophysicists even started studying financial systems without bothering to read the financial economics literature. All in all though, I think the power of statistical physics along with the empirical approach of physicists allows econophysics to provide a valuable perspective on finance.

Chapter 3

THEORY

This chapter explains concepts and methods that were used during, or otherwise relevant to, the research-work.

3.1 Trading

3.1.1 Order Types

Stock exchanges provide a broad array of ways for traders to buy and sell stocks, but the most often used ones are limit orders and market orders.

Market Order

The market order is an order to buy or sell a certain number of assets (e.g. stocks) at the currently best available price.

Limit Order

The limit order is an order to buy or sell a certain number of assets (e.g. stocks) at a specified "limit price", or a price that is better for whoever placed the order.

3.1.2 Order Book

The order book is a list of all the currently active buy and sell orders. Each listing contains an order's type, volume, price and the time at which it was entered.

3.1.3 Exchange-specific Auction Rules

NYSEarca Auctions

NYSEarca carries out three auctions; opening limit order auction, opening market order auction and the closing auction. The opening limit order auction, which is executed at 04:00 ET, is not discussed in this thesis; whenever I write about the NYSEarca opening auction, I am exclusively referring to the opening market order auction.

The order collection & dissemination of market data begins for the opening auction

at 08:00 ET, its freeze period starts at 09:29 ET and finally the auction is executed at 09:30 ET. Orders can be entered for the closing auction at any time after 09:30 ET, but its pre-auction period does not truly begin until NYSEarca starts disseminating market data at 15:00 ET. The closing auction freeze period begins at 16:29 ET, and the auction is executed at 16:30.

The exact mechanisms are described with words by NYSEarca [32], but they could as well be represented through mathematics.

NYSEarca Auction Mechanisms As per section 1.3, at time t into the auction, $v_{s,d,a}(t)$ represents the sum of all currently matched orders, and $p_{s,d,a}(t)$ their matching price. In NYSEarca, they are both defined through a function $f_{s,d,a}(x,t)$, which represents the matchable volume at price x for auction (s,d,a):

$$\begin{aligned} p_{s,d,a}(t) &\equiv \underset{x}{\texttt{argmax}}(f_{s,d,a}(x,t)) \\ v_{s,d,a}(t, p_{s,d,a}) &\equiv f_{s,d,a}(x = p_{s,d,a}, t) \end{aligned} \tag{3.1}$$

$$f_{s,d,a}(x,t) = \texttt{min}(\{v^{\texttt{Buy,L}}_{s,d,a}(x,t) + v^{\texttt{Buy,M}}_{s,d,a}(t), v^{\texttt{Sell,L}}_{s,d,a}(x,t) + v^{\texttt{Sell,M}}_{s,d,a}(t)\})$$

where $v^{\texttt{Buy,L}}_{s,d,a}(x,t)$ and $v^{\texttt{Sell,L}}_{s,d,a}(x,t)$ represent the total tradeable dollar-volumes at price x and time t of the buy and sell limit orders, while $v^{\texttt{Buy,M}}_{s,d,a}(t)$ and $v^{\texttt{Sell,M}}_{s,d,a}(t)$ are the buy and sell market order dollar-volumes available at time t.

In other words, this means that the matching price is simply the price that maximises the daily MO dollar-volume.

NASDAQ Auctions

NASDAQ carries out two auctions; the opening and closing crosses. The opening cross timetable is as follows: it becomes possible to enter orders at 07:00 ET, at 09:25 ET NASDAQ publicises market information, at 09:28 ET it freezes and finally the cross executes at 09:30 ET. The timetable for the closing cross is: orders can be entered at any time after 07:00, the auction freezes at 15:50 and the closing cross is executed at 16:00.

The mechanisms behind the NASDAQ crosses are meant to maximise MO dollar-volume, just as those of NYSEarca. The exact rules are not relevant to this thesis, and so they will not be discussed. However, they are defined through NASDAQ brochures such as [33] [34].

NYSE Auctions

The NYSE carries out an opening and closing auction. It becomes possible to enter orders for the opening auction at 07:30 ET, information on buy/sell imbalance & daily MO dollar-volume starts being disseminated at 08:30 ET, at 09:28 ET the indicative match price information is also disseminated and finally the execution is carried out at 09:30 ET. It is possible to enter orders for the closing auction already from 07:30

ET, but then it gets more complicated as the freeze happens in two parts; after 15:45 only imbalance offsetting orders may be entered, and orders stop being cancellable at 15:58. Then finally the execution takes place at 16:00.

The intricacies of the NYSE auction mechanisms are not relevant to this research work, and so they will not be discussed. The interested reader can, however, read about the precise rules here [35].

3.2 Statistical Distributions

The continuous statistical distributions that are of relevance to this thesis are listed in this section.

3.2.1 Pareto Tails

A distribution is said to be Pareto-tailed if it is asymptotically Lévy stable. Meaning, the distribution $\mathcal{D}_{\mathsf{X}}$ of some RV X is Pareto-tailed if its PDF $\rho(x)$ asymptotically follows the following power-law:

$$\rho(x) \sim C\frac{1}{|x|^{\mu+1}}, \text{ when } x \to \pm\infty \tag{3.2}$$

with $\mu \in \langle 0, 2\rangle$ and C being a constant number.

To understand why Pareto tails are interesting, we need to define Lévy stability; A probability distribution $\mathcal{D}_{\mathsf{X}}$ of some RV X is Lévy-stable if X has the following property: *Let* X_1 *and* X_2 *be two independent copies of* X*. If there exist a positive number* c *and real number* d *such that* $a\mathsf{X}_1 + b\mathsf{X}_2$ *has the same probability distribution as* $c\mathsf{X} + d$ *for all positive numbers* a, b*, then* X *is Lévy-stable.*

This definition implies that an arbitrary number of Pareto tailed RVs can be summed together to a new RV that is also Pareto-tailed; this phenomena is called the Generalised Central Limit Theorem.

3.2.2 Heavy-tailed Distributions

Heavy-tailed, or "fat-tailed", distributions are formally defined to be the set of all functions which are not exponentially bounded [18], or equivalently the set of all functions that decrease more slowly than any exponential. Mathematically, the distribution $\mathcal{D}_{\mathsf{X}}$ on $[0, \infty\rangle$ is formally heavy-tailed i.f.f. its PDF $\rho(x)$ fulfils the following condition $\forall \epsilon > 0$

$$\int_0^\infty \rho(x)e^{\epsilon x}dx = \infty \tag{3.3}$$

Of course, no real statistical distribution goes to infinity, but this definition does define a set of idealised mathematical distributions whose properties are, for practical purposes, also valid for their physical realisations.

3.2.3 Power-law Distributions

The power-law distribution is an extremely important class of heavy-tailed distributions. If the tail of a RV X past some lower bound $x_{\mathtt{min}} > 0$ follows a power-law distribution, then the power-law distribution has the following PDF:

$$\rho(x) = \frac{\alpha - 1}{x_{\mathtt{min}}} \left(\frac{x}{x_{\mathtt{min}}} \right)^{-\alpha} \tag{3.4}$$

where $\alpha > 1$ is the power-law exponent and $x \in \langle x_{\mathtt{min}}, \infty \rangle$.

Further, the survival function[1] (SF) of the power-law is as follows:

$$P(\mathsf{X} \geq x) = \left(\frac{x}{x_{\mathtt{min}}} \right)^{-\alpha+1} \tag{3.5}$$

The power-laws have several interesting features. Firstly, they are Lévy-stable as long as $\alpha \in \langle 1, 3 \rangle$. Second, a power-law (if its domain really goes to infinity) only has a well-defined mean for $\alpha > 2$, a well-defined variance only for $\alpha > 3$ and so on with higher moments. Finally, power-laws have the remarkable and unique feature of being scale-invariant: $\rho(cx) = c^{-\alpha}\rho(x) \sim \rho(x)$; thus the power-law exponent completely defines the scaling of the tails of $\rho(x)$.

Finally, it must be noted that both the SF and PDF of power-laws appear as straight lines in log-log plots. This can be seen by taking the natural log of $P(X \geq x)$: $\ln P(X \geq x) = (-\alpha + 1)(\ln x - \ln x_{\mathtt{min}}) \sim (-\alpha + 1) \ln x$.

Truncated Power-laws

An important subclass of power-laws are the truncated power-laws. In reality, heavy-tails can not go on until infinity; at some point they must stop, i.e. be truncated. This applies for power-laws too, but often one does not have enough data in the tails to notice the truncation and so then the classical power-law expression, i.e. equation (3.5), is used. If a noticeable truncation is present though, the data could be better modelled by an exponentially truncated power-law:

$$\rho(x) = C \left(\frac{x}{x_{\mathtt{min}}} \right)^{-\alpha} e^{-\beta x} \tag{3.6}$$

where C is some normalisation constant and β is the truncation parameter. Another way to truncate power-laws is by means of a "hard" truncation:

$$\rho(x) = \begin{cases} C \left(\frac{x}{x_{\mathtt{min}}} \right)^{-\alpha}, \ x \in \langle x_{\mathtt{min}}, x_{\mathtt{max}} \rangle \\ 0, \ \text{otherwise} \end{cases} \tag{3.7}$$

where the power-law simply stops existing past some upper bound $x_{\mathtt{max}}$.

In theory, the truncation forces the power-laws to have finite means, variance and

[1] The SF is formally known as the complementary cumulative distribution function.

so on regardless of the power-law exponent, but physical power-laws can still *appear* unbounded if the truncation comes into effect at sufficiently large x. And the same goes for the scaling behaviour; over some range, truncated power-laws exhibit scale-free behaviour.

Further, the truncation distorts straight lines in log-log plots expected from power-laws. For example; in the case of hard truncation, the PDF appears as a straight line in log-log plots, but the SF does not, because $P(\mathsf{X} \geq x) = \int_x^{x_{\mathtt{max}}} C \left(\frac{x}{x_{\mathtt{min}}}\right)^{-\alpha} = \frac{C x_{\mathtt{min}}^{\alpha}}{\alpha - 1}(x^{-\alpha+1} - x_{\mathtt{max}}^{-\alpha+1})$ will only produce a linear dependence between $\ln P(X > x)$ and $\ln x$ when $x_{\mathtt{max}} >> x$.

3.2.4 Log-normal Distribution

The log-normal distribution is the distribution of a RV whose logarithm is normally distributed. Thus, if $\ln \mathsf{X} \sim \mathcal{N}(\mu, \sigma^2)$, then it can be shown that $\mathsf{X} \sim \mathtt{Lognormal}(\mu, \sigma^2)$, with the PDF of X being:

$$\rho(x) = \frac{1}{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{3.8}$$

Log-normal distributions arise from multiplicative processes just as normal distributions arise from additive ones; If X is the product of a large number of RVs, then $\ln \mathsf{X}$ must be the sum of a large number of RVs. Then, if the conditions are appropriate, the CLT applies and $\ln \mathsf{X}$ converges towards a normal distribution, and thus X converges towards a log-normal one.

Another interesting feature is that a log-normal distributions is the (Shannon) entropy-maximising distribution of a RV X whose logarithm has a specified mean and variance[37]. In other (and informal) words, if the only consistent behaviour of the RV X is that $\int \ln x \rho(x) dx = u$ and $\int (\ln x - u)^2 \rho(x) dx = s^2$, where the integrals go over the domain of X and $\rho(x)$ is its PDF, then it can be shown that $\mathsf{X} \sim \mathtt{Lognormal}(u, s^2)$.

3.2.5 Exponential Distribution

An exponentially distributed RV X is characterised by an exponentially decaying PDF:

$$\rho(x) = \begin{cases} \lambda e^{-\lambda x}, \ x \geq 0 \\ 0, \ \text{otherwise} \end{cases} \tag{3.9}$$

It is of interest for this thesis because it defines the notion of heavy-tails, as noted in equation (3.3) above.

3.2.6 Empirical Cumulative Distribution Function

The empirical cumulative distribution function is defined as follows. Assume we have a dataset of n points $\{x_k \colon \forall k \in [1, n]\}$; then the associated empirical cumulative distribution function is:

$$F(x) = \frac{1}{n} \sum_{k=1}^{n} I_{[-\infty, x]}(x_k) \tag{3.10}$$

where $I_{[-\infty,x]}$ is the indicator function; it works by returning 1 if $x_k < x$, and 0 otherwise.

3.3 Statistical Methods

The statistical methods (for continuous RVs) that are of relevance to this thesis are presented below.

3.3.1 Maximum Likelihood Estimation

Given a sample of i.i.d. observables $\{x_k : \forall k\}$ along with a prior assumption of their underlying RV X belonging to some family of parametric PDFs $\rho(x; \vec{\theta})$ defined by the parameters $\vec{\theta}$, then one can employ maximum likelihood estimation (MLE) to find an estimator $\hat{\vec{\theta}}$ for these parameters. The estimator is formally defined to be the function that maximises the likelihood L of $\{x_k : \forall k\}$:

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\mathtt{argmax}}(L(x_1, x_2, ..; \vec{\sigma})) = \underset{\vec{\theta}}{\mathtt{argmax}} \prod_{\forall k} \rho(x_k; \vec{\theta}) \tag{3.11}$$

In practice though, one maximises the log-likelihood $\Lambda = \ln L$ because it is more convenient to work with sums (and because maximising it is equivalent to maximising the likelihood, due to the fact that the logarithm is a monotonically increasing function):

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\mathtt{argmax}}(\Lambda(x_1, x_2, ..; \vec{\sigma})) = \underset{\vec{\theta}}{\mathtt{argmax}} \sum_{\forall k} \ln \rho(x_k; \vec{\theta}) \tag{3.12}$$

3.3.2 Kolmogorov-Smirnov Distance

The Kolmogorov-Smirnov (KS) distance $D(P_1, P_2)$ is a metric for the distance between the two cumulative distribution functions (CDFs) $P_1(\mathsf{X} > x)$ and $P_2(\mathsf{X} > x)$:

$$D = \underset{x}{\mathtt{max}} |P_1(\mathsf{X} > x) - P_2(\mathsf{X} > x)| \tag{3.13}$$

in other words, the KS distance is simply the maximum difference between the two CDFs.

3.3.3 Hypothesis Testing

Three hypothesis tests that are of relevance to the research work are described in this section.

Log-likelihood Ratio Test

The log-likelihood ratio test (LLRT) is used to compare fitted distributions against each other, and the variant used by the `python.powerlaw` package [39] is described here. Assume we have a dataset of n samples $\{x_k : \forall k \in [1, n]\}$ and that two distributions were fitted to this dataset, with PDFs $p_1(x)$ and $p_2(x)$. Then the respective likelihoods functions are $L_1 = \prod_{\forall k} p_1(x_k)$ and $L_2 = \prod_{\forall k} p_2(x_k)$, and thus their log-likelihood ratio Λ is:

$$\Lambda = \ln \frac{L_1}{L_2} = \sum_{\forall k} \ln p_1(x_k) - \ln p_2(x_k) = \sum_{\forall k} l_k^1 - l_k^2 \tag{3.14}$$

where l_k^1 and k_k^2 are the log-likelihoods for the k-th observable. Now, if the samples are sufficiently well-behaved (i.e. close to i.i.d.), then by the CLT Λ approaches the normal distribution $\mathcal{N}(\mu, \sigma^2)$ as the sample size tends to infinity. Assuming from now on that the sample is sufficiently well-behaved and numerous for Λ to follow $\mathcal{N}(\mu, \sigma^2)$, we have that $\mu = \langle l_k^1 - l_k^2 \rangle_{\forall k}$ and $\sigma^2 = n\langle((l_k^1 - l_k^2) - \mu)^2\rangle_{\forall k}$.

Now comes the core of the LLRT: Let the null-hypothesis be that $\mu = 0$, so that neither fit is considered better than the other. Then, under this null-hypothesis, the probability p to see a fluctuation more extreme than Λ is simply given by the standard normal distribution due to the assumed normality of Λ:

$$p = P\left(\frac{\Lambda - 0}{\sigma} \geq |z|\right) = 1 - \texttt{erf}\left(\frac{\Lambda - 0}{\sigma}\right) \tag{3.15}$$

where $P(\mathsf{Z} \geq z) = 1 - \texttt{erf}(z)$, with $\texttt{erf}(z)$ being the error function of the standard normal distribution, whose values can of course be looked up in any statistical table.

If the null-hypothesis is that $\mu = 0$ and thus none of the fits are better than each other, then the alternative hypothesis must be that one of the fits is better than the other, and which one is better can be decided from the sign of Λ.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) tests the similarity between two distributions. Its strength lies in it making no assumptions about the nature of the distributions it is testing, i.e. it is "distribution free" [40]. The test uses the Kolmogorov distribution $\mathcal{K}$ to decide whether the two tested distributions are similar. The two (very similar) versions of the KS test are presented below.

The one-sample version tests whether the n observables in the sample $\{x_k : \forall k \in [1, n]\}$, with empirical CDF $F_n(x)$, are realisations of some RV X with CDF $P(\mathsf{X} > x)$. To this end, the test-statistic $\sqrt{n}D_n(F_n, P)$ is used, where $D_n(F_n, P)$ is the KS distance from equation (3.13). Under the null-hypothesis, F_n converges in distribution to P when $n \to \infty$, meaning the observables are indeed realisations of X, and thus $\sqrt{n}D_n(F_n, P) \sim \mathcal{K}$. The alternative hypothesis is then that the samples are not observables of X.

The two-sample version is practically the same, except now it tests the similarity of the samples $\{x_k \colon \forall k \in [1, n]\}$ and $\{x'_k \colon \forall k \in [1, m]\}$. The test-statistic this time is $\sqrt{\frac{nm}{n+m}}D_{n,m}(F_n, F'_m)$, and it operates under the null-hypothesis of the two samples being drawn from the same distribution. The alternative hypothesis must then be that the two samples are not drawn from the same distribution.

Student's t-Test

Student's t-test is an umbrella term for all hypothesis tests that employ Student's t-distribution. In this thesis, only the one-sample version was used and it is described below:

The one-sample t-test tests the null-hypothesis that the mean μ of some sample of n i.i.d. observables $\{x_k \colon \forall k \in [1, n]\}$ is equal to some μ_0 with the test-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{3.16}$$

with $\bar{x}$ being the sample mean and s being the sample standard deviation. The test is carried out by calculating the p value for t by using Student's t-distribution with $n - 1$ degrees of freedom. The alternative hypothesis is that the mean of the sample is in fact not equal to μ_0.

3.3.4 Bootstrapping

Bootstrapping is a Monte-Carlo method used to empirically estimate the properties, such as variance, of complex statistical functions. More precisely, given a set of i.i.d. observables $\{x_k \colon \forall k \in [1, n]\}$ used to calculate some statistic $\hat{y}(x_1, x_2, ..)$, one can estimate various properties of this statistic by drawing (with replacement) m random sub-samples $\{x_k \colon \forall k \in [1, l]\}$, and using these sub-samples to construct a "bootstrapped" distribution $\mathcal{D}_{\hat{y}}$ from the bootstrapped set of statistics $\{\hat{y}_j(x_1, x_2, ..) \colon \forall j \in [1, m]\}$. One must typically find appropriate m, l through trial and error, but the rule-of-thumb guideline for choosing appropriate values is to simply use sufficiently large m, l so that $\mathcal{D}_{\hat{y}}$ converges in distribution to a peaked distribution [38].

3.3.5 Distribution Fitting

Distribution fitting is the process of fitting probability distributions to a sample of data. The procedure can essentially be broken down into three steps:

Procedure

1. Based on the properties of the sample, one must find an appropriate distribution, or family of distributions to fit.
2. Then the parameters of the fitting distribution must be found by some means.
3. And finally the quality of the fit must be evaluated.

Four different heavy-tailed distributions and the normal were fitted to data in this thesis; the methods used to do that are descried below.

Fitting Heavy-Tails

The `python.powerlaw` package [50] is capable of fitting, and testing against each other, 4 kinds of distribution to the tails of some data-sample $\{x_k \colon \forall k\}$; power-laws, exponentially truncated power-laws, log-normal tails and exponential distributions (see section 3.2.2 for an overview of their statistical properties).

Fitting Procedure Its fitting algorithm employs MLE (which is described in section 3.3.1) and the KS distance (which is described in section 3.3.2), and it works as follows [39]:

1. Find an $x_{\texttt{min}}$ by minimising the KS distance $D = \underset{x \geq x_{\texttt{min}}}{\texttt{max}} |F(x) - P(\mathsf{X} \geq x)|$, where $F(x)$ is the empirical CDF of the data and $P(\mathsf{X} \geq x)$ is a power-law fitted through MLE.

2. The $x_{\texttt{min}}$ now represents the starting point of the tail, and $\hat{\alpha}$ is the fitted power-law exponent. Next, fit the tails $\forall x_k \colon x_k \geq x_{\texttt{min}}$ with exponentially truncated power-laws, log-normal tails and exponential distributions by again employing MLE

3. Finally, return $x_{\texttt{min}}$ along with the parameters of the fitted distributions.

Goodness of Fit The variance of the fitting parameters can be estimated through bootstrapping (which is described in section 3.3.4), but when one is dealing with heavy-tails it is typically more useful to simply compare tail distributions with each-other; for this, the `python.powerlaw` package employed Log-likelihood Ratio Test (which is described in section 3.3.3).

Fitting Normal Distributions

The `R::MASS` package [49] is capable of fitting a variety of distributions to datasets, and among them is the normal distribution. The fitting procedure for the normal distribution is rather simple; `R::MASS` simply employs MLE (as described in section 3.3.1) to find the normal-distribution parameters μ and σ^2 that fit the data best.

3.4 Time-series

A time-series $y[t]$ is simply a series of data-points sorted by an index representing their timestamp (here t).

3.4.1 Stationarity

Time-series are often divided into two types; stationary, and non-stationary. A time-series $y[t]$ is stationary if it is generated by some stochastic process $\{Y_t\} \colon F_Y(y[1], y[2], .., y[t])$ whose cumulative joint probability distribution F_Y is independent from time; in other words, $F_Y \colon F_Y(y[1], y[2], .., y[t]) = F_Y(y[1+\tau], y[2+\tau], .., y[t+\tau]), \forall \tau$. If this is not the case, then the time-series is not stationary.

The expectation value, variance and autocorrelation pattern of a stationary time-series are all constant across time.

3.4.2 Autocorrelation Function

The autocorrelation function (ACF) was often used to check for stationarity, predictability. and independence of time-series across time.

The theoretical ACF of a time-series $x[t]$ is defined as:

$$R(\tau) = \frac{\mathtt{E}[(x[t] - \mu_t)(x[t+\tau] - \mu_{t+\tau})]}{\sigma_t \sigma_{t+\tau}} \tag{3.17}$$

where τ is the lag, μ_t and $\mu_{t+\tau}$ are the time-series' mean at time t and $t+\tau$ respectively, and similarly σ_t and $\sigma_{t+\tau}$ are the standard deviations.

In practice however, one does not know what the theoretical means and variances are as functions of time; thus the ACF, as used in the `R::stats` package, is defined [41] to be:

$$R(\tau) = \frac{1}{n} \sum_{t=\mathtt{max}(1,-\tau)}^{\mathtt{min}(n-t,n)} \frac{(x[t+\tau] - u)(x[t] - u)}{s} \tag{3.18}$$

where n is the length of $x[t]$, s is its standard deviation and u is its mean.

If the absolute value of the ACF is statistically significant at lag τ (meaning larger than what one would expect under the null-hypothesis of white noise), it implies that knowing the time-series' amplitude time t gives you prior information on its amplitude at $t+\tau$. This is equivalent to saying its movement exhibits a degree of predictability, and thus one can use time-regressive forecasting algorithms to capture this predictability (more on them below). For example, periodic movements in the ACF across τ imply the presence of seasonality.

3.5 Time-series Forecasting

In chapter 5 a significant amount of time-series forecasting is performed, and this section was written as a reference for the forecasting methods and statistical concepts used there.

3.5.1 The Concept of Forecasting

Forecasting is the process of analysing past responses of a time-series with the aim of predicting its future responses. This is a regressive process that is carried out by fitting a model $\hat{y}[t'], t' \in \mathbb{T}$ to the time-series' past responses $y[t'], t' \in \mathbb{T}$, where $\mathbb{T}$ is a training interval made up of past responses, and then use this model to make predictions on future responses $y[t], t \in \mathbb{P}$, where $\mathbb{P}$ is a prediction interval for future times.

In this thesis, three different forecasting methods were employed; ARIMA, Facebook Prophet and Random Forests.

3.5.2 ARIMA with Fixed Effects

ARIMA is a generative (meaning it defines a stochastic process) linear model that is capable of modelling autoregressive and moving average processes, as well as account for certain non-stationarities by differencing. Mathematically, it is denoted as ARIMA(p, d, q) and defined as:

$$y[t] = c + \sum_{i=1}^{p} a_i \Delta^d L^i y[t] + \sum_{j=1}^{q} b_j L^j \epsilon[t] + \epsilon[t] \tag{3.19}$$

where c, a_i, b_j are parameters of the model, Δ is the difference operator, L is the lag operator, d is the degree of differencing and p, q define the orders of the model. (p, d, q) together define a stochastic process.

In this research work, the `R::forecast` [46] package was used to find the appropriate ARIMA(p, d, q) model, which is built on the Hyndman-Khandakar (2008) algorithm [31]. But it was not used in isolation; a fixed-effects model was applied in conjunction to stationarize[2] the time-series.

This was carried out in the following manner: Before ARIMA was applied, the training interval $y[t']$ was transformed to

$$\tilde{y}[t'] = y[t'] - \hat{y}^{\texttt{FE}}[t']$$

and after ARIMA was applied on $\tilde{y}[t']$, the resulting predictive model was transformed back from $\hat{\tilde{y}}^{\texttt{ARIMA}}[t]$ to

$$\hat{y}^{\texttt{ARIMA}} = \hat{\tilde{y}}^{\texttt{ARIMA}}[t] + \hat{y}^{\texttt{FE}}[t]$$

Fixed Effects Model

The fixed effects model accounted for certain seasonal non-stationarities that ARIMA was unable to model. In essence, it works as just a rolling conditional average, which mathematically means the fixed effects model $\hat{y}^{\texttt{FE}}[t]$ is defined as:

$$\hat{y}^{\texttt{FE}}[t] = \begin{cases} \frac{1}{|\mathbb{T}^*|} \sum_{\forall t' \in \mathbb{T}^*} y[t'], & \text{if } t \text{ is a seasonality} \\ 0, & \text{otherwise} \end{cases} \tag{3.20}$$

where $|\mathbb{T}^*|$ is a lists of all the seasonalities in the training set that are of the same type as the seasonality at t.

3.5.3 Facebook Prophet

Facebook Prophet [22] is a non-linear, additive-decomposable model used for time-series analysis. With additive-decomposable, it is meant that the model is simply the sum of a trending function $g[t]$, a function representing seasonalities $s[t]$, a "holidays" function $h[t]$ and a gaussian error-term $\epsilon[t]$:

$$\hat{y}^{\texttt{Prophet}}[t] = g[t] + s[t] + h[t] + \epsilon[t] \tag{3.21}$$

$g[t]$ can represent a range of growth models/trends, such as piecewise linear growth or logistic growth. $s[t]$ is a fourier sum capable of modelling various seasonalities. $h[t]$ is essentially a fixed-effects model almost identical to the one described above; the only difference is that rolling conditional average in $h[t]$ is smoothed by a gaussian kernel to avoid overfitting. The Prophet model is implemented in the `R::Prophet` package [47].

[2] That is, transform time-series data such that it becomes more stationary

Finally, it is worth mentioning that unlike ARIMA, Prophet is not a generative method, but a purely regressive one. This means it essentially is a curve-fitting algorithm and does not define any specific stochastic process and is thus of limited use for inferencing, i.e. understanding the physical reasons for why the data behaves as it does.

3.5.4 Random Forests

Random Forests is a machine-learning algorithm that is fundamentally different from the two previous methods. While ARIMA & Prophet are parametric models whose behaviour is precisely defined through parametric functions, Random Forests is an ensemble learning method which is defined algorithmically.

The regressive version of Random Forests inputs a training set $\mathbb{D} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_k, y_k)\}$, with $\vec{x}_i \in \mathbb{R}^d$ being a vector of predictors of d features (features are variables related to the response[3]) and $y_i \in \mathbb{R}$ being the corresponding response, and then Random Forests creates a model which it uses to forecast the response $\hat{y}_{\texttt{new}}$ for any vector of predictors $\vec{x}_{\texttt{new}}$.

At its core, the Random Forests algorithm works by segmenting the data $\{(\vec{x}_i, \hat{y}_i) \colon \forall i\}$ according to the data's dependence with its features, it does this multiple times and every time in a random fashion, and then it uses decision trees to navigate the resulting segmentations to produce forecasts. More precisely, the regressive version of the algorithm works as described below:

Algorithm 1 Regressive Random Forests

Input: $\mathbb{D}$, $n_{\texttt{tree}}$, $\vec{x}_{\texttt{new}}$, $m_{\texttt{try}}$
Output: $\hat{y}_{\texttt{new}}$

1: **for** $(\forall b \in \{1, 2, ..., n_{\texttt{tree}}\})$ **do**
2: Bootstrap a sample $\mathbb{D}_b$ from $\mathbb{D}$ of sufficient size
3: From the d features, randomly draw $m_{\texttt{try}} < d$ features
4: Grow a binary decision tree T_b by segmenting the data according to the drawn .....features by using a greedy algorithm which minimises a least squares error .....metric
5: Travel down all the decision trees using $\vec{x}_{\texttt{new}}$, and return the average of their individual forecasts: $\hat{y}_{\texttt{new}} = \frac{1}{n_{\texttt{tree}}} \sum_{b=1}^{n_{\texttt{tree}}} T_b(\vec{x})$

This is the exact algorithm used in the `R::randomForest` package [48] (which is the package used in this project).

[3] Examples of features include time, first order difference of the response, some past response, day of the week/month/year etc.

Chapter 4

PROPERTIES OF INDIVIDUAL AUCTIONS

This chapter concerns itself with the properties of the orders at the end of the auctions, i.e. at $t = t_{\texttt{end}}$. The investigation revolved around studying the properties of the tails of $\mathcal{D}_{m_{s,d,a}}$ for NYSEarca.

4.1 Available Data

The available data consisted of $13,221,820$ unique datapoints for 1408 stocks traded on the NYSEarca exchange, across the years of 2010 up to and including 2014.

Each datapoint represents a concluded auction (s, d, a) with an associated dataset of MO dollar-volumes $\{m^k_{s,d,a} \colon \forall k\}$ along with the price at which they orders were carried out, i.e. the final matching price of the auction $p_{s,d,a,t_{\texttt{end}}}$. Thus each datapoint can be classified as $(s, d, a, k, t_{\texttt{end}}, p_{s,d,a,t_{\texttt{end}}})$.

4.2 Behaviour of $m_{s,d,a}$

4.2.1 Power-laws Spotted

When first exploring the data, power-law like behaviour was observed in $\mathcal{D}_{m_{s,d,a}}$, the empirical density of the dollar-volumes of MOs. The linear nature of the data in the log-log plot of figure 4.1 exemplifies this.

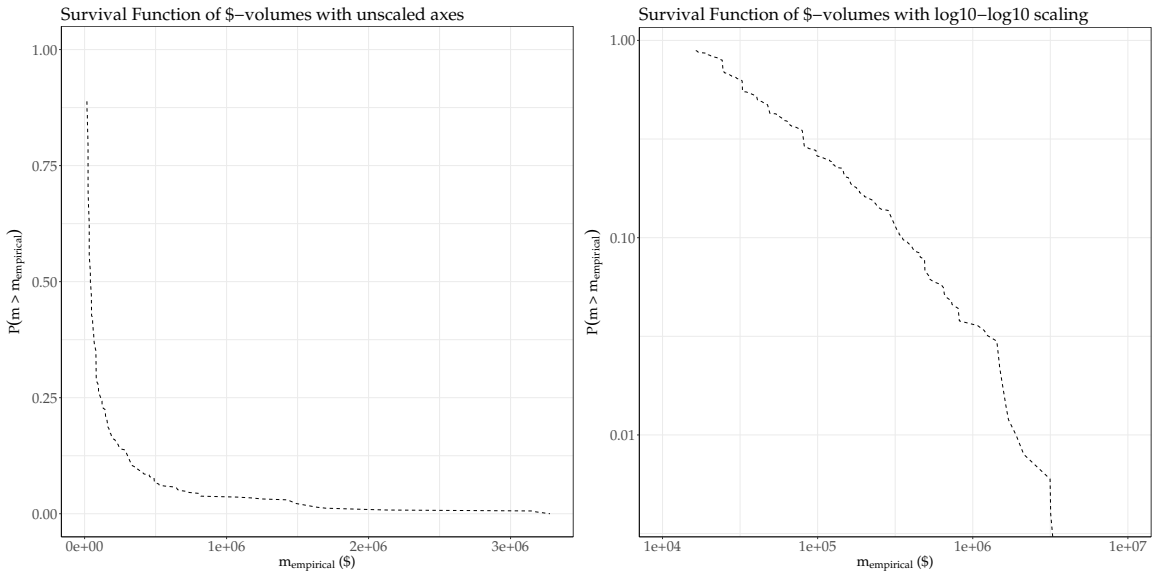

Figure 4.1: Empirical survival function of the MO dollar-volumes for the opening auction of SPY during 2013-06-03.

Due to the interesting properties of power-laws, as describes in section 3.2.3, let us first focus on them.

4.2.2 Fitting a Power-law

Finding Power-law Fit Parameters

A way to fit power-law distributed tails to data was needed before there was any point in seriously investigating whether power-law tails were present. For this purpose, the `python.powerlaw` package was employed to find the power-law parameters $\alpha_{s,d,a}$, which is the exponent characterising the power-law observed for auction (s, d, a), and $m^{\texttt{min}}_{s,d,a}$, the dollar-volume value at which the power-law/fat-tail starts. The package employs maximum likelihood estimation (MLE) and minimisation of the Kolmogorov-Smirnov (KS) distance to find these parameters. The theory behind power-laws is described in section 3.2.3, and the fitting procedure in 3.3.5.

To illustrate, the dataset from figure 4.1 is fitted with a power-law in figure 4.2 below.

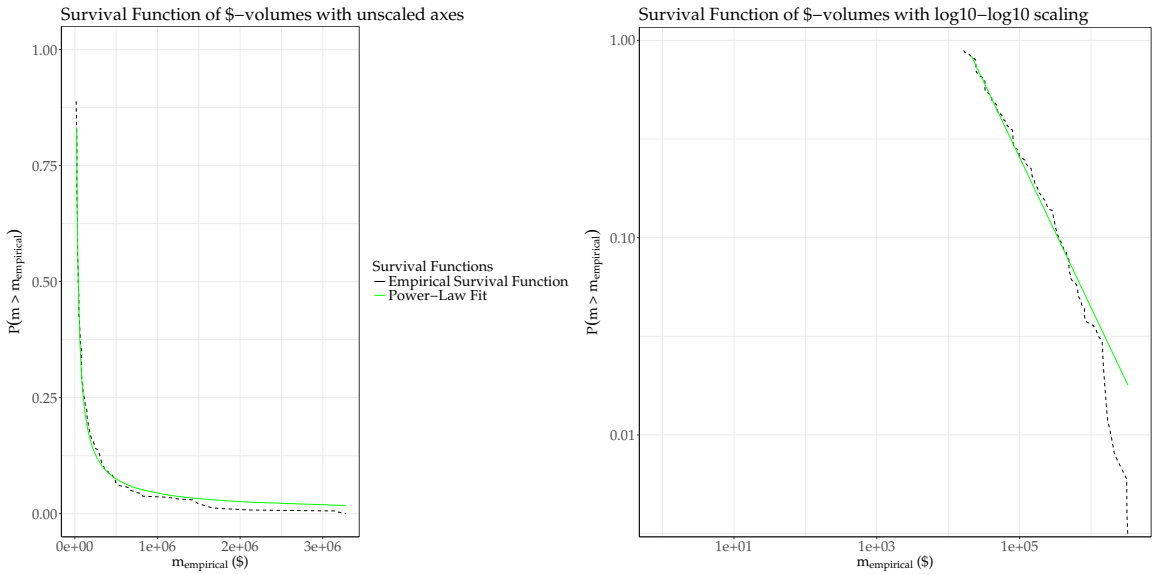

Figure 4.2: Power-law fit to empirical survival function of the MO dollar-volumes for the opening auction of SPY during 2013-06-03.

Numerical Stability of Fitting and the Question of I.I.D.

A cursory investigation into the numerical stability of the fitted $(\alpha_{s,d,a}, m^{\texttt{min}}_{s,d,a})$ pairs was also undertaken. The mathematics employed by `python.powerlaw` to find $(\alpha_{s,d,a}, m^{\texttt{min}}_{s,d,a})$ (and test its goodness of fit) in principle require i.i.d. data to be theoretically guaranteed to converge. Alas, equation (1.2) already established that the data was not independent, and there were reasons to suspect it was not identically distributed either (as will be explained in section 5.2). Therefore a Monte-Carlo procedure called bootstrapping (see section 3.3.4) was employed to ensure that MLE, which `python.powerlaw` uses for fitting, can be expected to converge.

Bootstrapping is a computationally intensive method, and so the numerical stability was investigated by looking at the stability of a handful fitted $(\alpha_{s,d,a}, m^{\texttt{min}}_{s,d,a})$ pairs for randomly selected auctions (s, d, a). A demonstrative example of this investigation is presented in figure 4.3 below, where the estimated numerical uncertainty corresponding to the fit-parameters in figure 4.1 was estimated. 500 bootstrap samples were used.

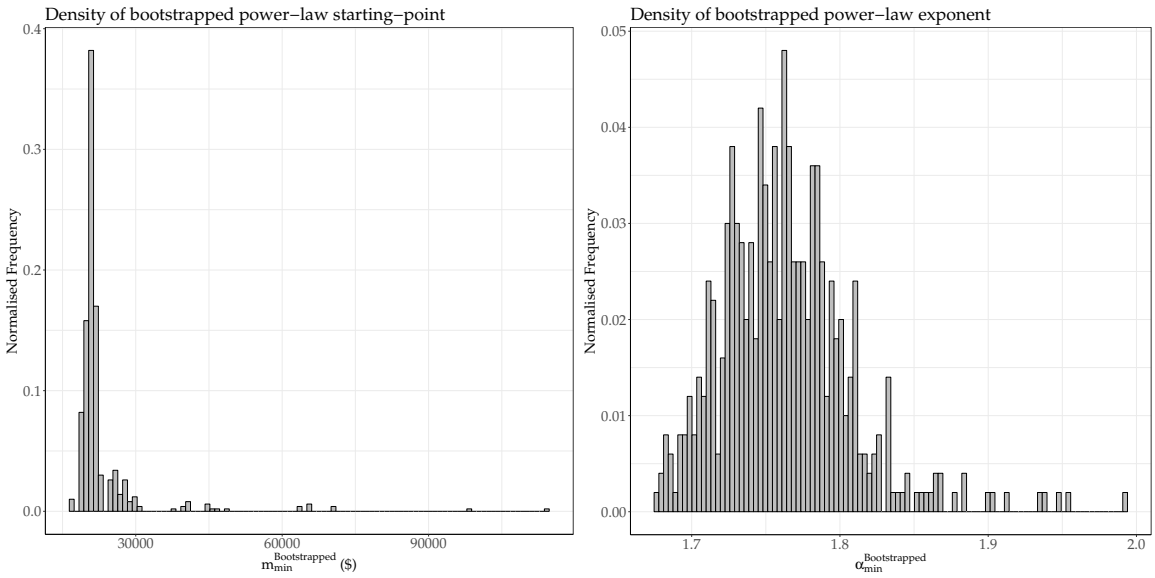

Figure 4.3: Result of bootstrap-estimation of the numerical uncertainty in the power-law fit parameters fitted to the opening auction of SPY at 2013-06-03.

The success of a bootstrapping run is judged on whether the parameters of interest cluster around a typical value. In our case, 500 samples was sufficient for them to do that, as both the distributions in figure 4.3 above demonstrate. Further, the bootstrapping runs consistently produced *peaked distributions*, which means MLE can indeed be expected to converge. Therefore, the investigation continued by retaining MLE, and assuming the lack of i.i.d. was sufficiently mild to be discounted as noise, as far as this section is concerned[1].

4.2.3 Confirming Power-law Presence

Further power-law fitting consistently revealed the presence of straight lines in the log-log plots, hinting towards the power-law relationships being a general phenomena $\forall s \in \mathbb{S}_{\texttt{NYSEarca}}$. Therefore, the investigation continued, with the next step being fitting power-laws to every available dataset $\{m_{s,d,a}^{k} : \forall k\}$.

However, the resulting plots, including the one above, also firmly established that the power-laws were truncated in their tails. In principle this is not a cause for worry because, firstly, physical power-laws must get truncated at some point, and secondly, truncation is always exaggerated in log-log SF plots (see section 3.2.3 for why). Still, the truncations were heavy enough to make us doubtful on whether the power-law fit was superiour at explaining the data compared to other heavy-tailed distributions. Proclaiming tenuous power-laws relationships has been a common problem in science [17], and so to avoid past mistakes it was decided to make sure our suspected power-law behaviour was a better fit than the other prominent candidates.

[1]It should be noted that while a great deal of statistical theory makes the assumption of i.i.d., real data is almost never completely i.i.d.; thus in cases such as this, one must simply assume the lack of i.i.d. is sufficiently mild and proceed with caution, as the alternative is to either use very complex statistical theory, or give up.

4.2.4 Other Candidates

The power-law fit could be compared against many heavy-tailed distributions, but the most important alternative candidates are the exponential and log-normal distributions. As defined in 3.2.2, heavy-tails are absolutely required to not be exponentially bounded. Therefore, as a minimum condition, a power-law must be a better fit than an exponentially distributed tail; otherwise it cannot even be claimed that heavy-tailed behaviour is present. Log-normal distributed tails are also important to test against due to their close relation to power-law distributions [19]. In addition to these two, it was also of interest to check whether a power-law fit was better than the fit of an exponentially truncated power-law, which is just a power-law relation multiplied with an exponentially decaying function (see section 3.2.3 for more). Thus came the final part of the investigation: systematically fitting power-laws and testing them against the alternative distributions.

4.2.5 Procedure for Fitting & Testing $\forall(s, d, a)$

The systematic fitting was carried out by employing the `python.powerlaw` package to fit the tails of all the datasets with at least 100 datapoints, i.e.
all$\{m^k_{s,d,a} \colon \forall k\} \colon |\{m^k_{s,d,a} \colon \forall k\}| > 100$, with exponential, log-normal, power-law and exponentially truncated power-law distributions. Then the Log-likelihood Ratio Test (which is explained in section 3.3.3) was used to separately test each of the alternative distributions against the power-law distribution on the tails of $\{m^k_{s,d,a} \colon \forall k\}$, with the null-hypothesis being that both the tested distributions fit the tails of $\{m^k_{s,d,a} \colon \forall k\}$ equally well.

4.2.6 Results of the Fitting & Testing

The result was a set of fitted power-law exponents $\{\alpha_{s,d,a} \colon \forall(s, d, a)\}$, and for each of these power-law exponents there were 3 p-values, one for each of the alternative distributions it was tested against. The resulting datasets covered 263 stocks, and had cardinalities of $|\{\alpha_{s,d,a} : \forall(s, d, a)\}| = 24962$ datapoints. The starting points of the power-laws, $\{m^{\mathtt{min}}_{s,d,a} \colon \forall(s, d, a)\}$, were of course also found, but their distributions provided no interesting information; therefore the following discussion will solely revolve around the fitted power-law exponents and the results from Log-likelihood Ratio Test against alternative distributions.

Distribution of Power-law Parameters

The distributions of the fitted power-law exponents are illustrated in the following figures across stocks, auctions and years.

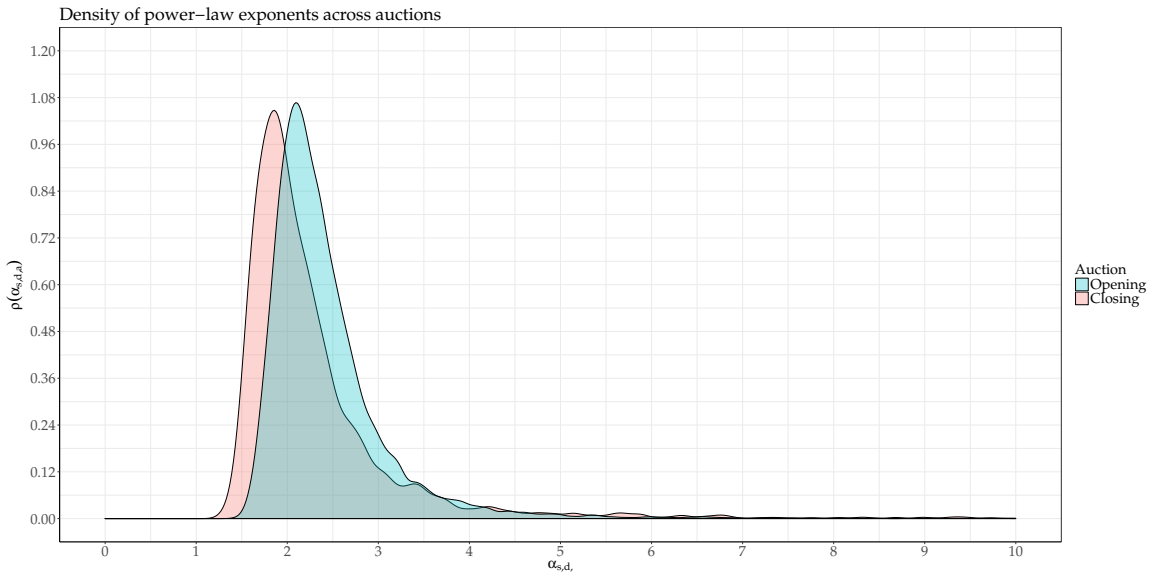

Figure 4.4: Density of the open-auction and closing-auction fitted power-law exponents aggregated across all stocks and dates.

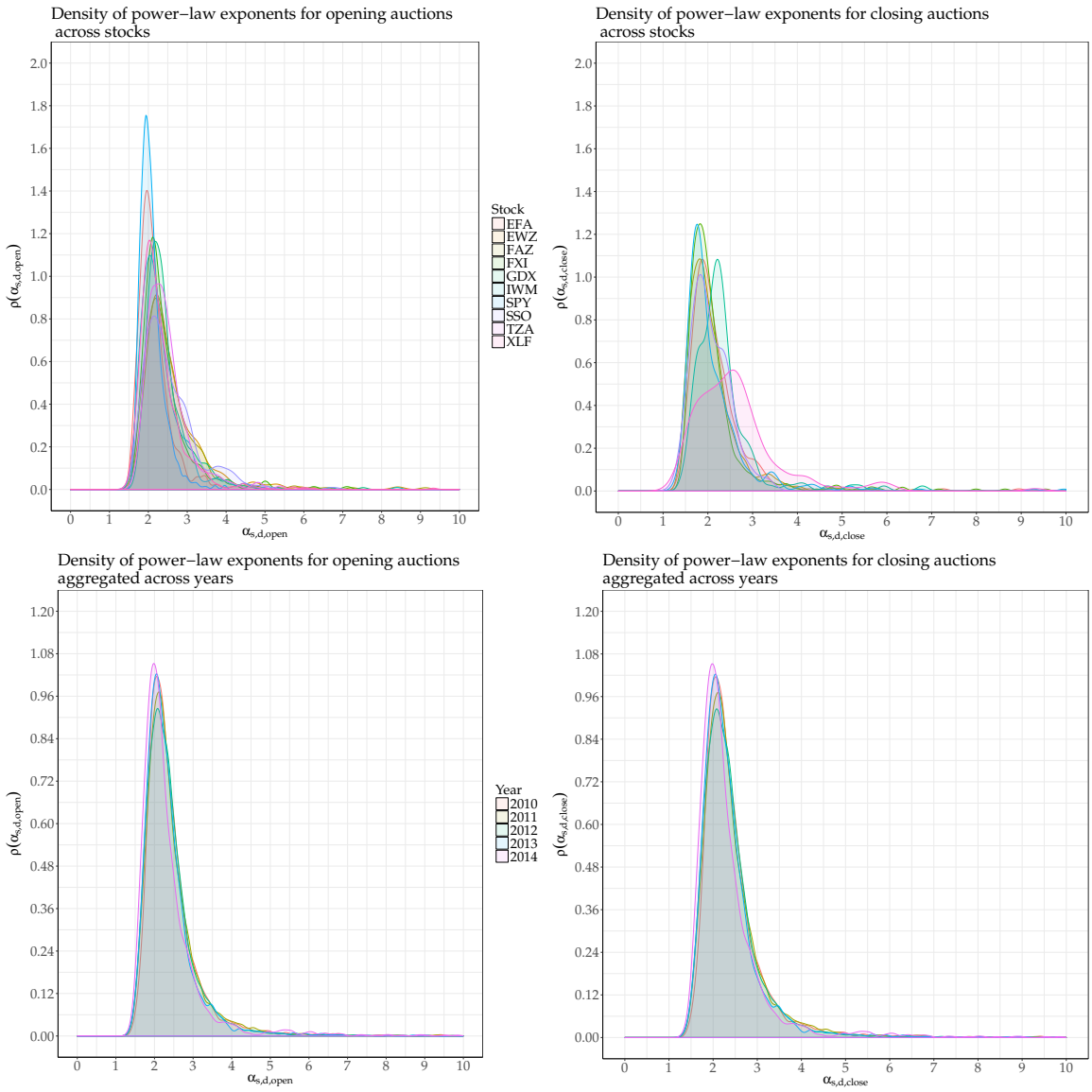

Figure 4.5: Top: Density of open-auction (left) and close-auction (right) fitted power-law exponents aggregated across dates for the 10 stocks with most datapoints available. Bottom: The same distributions, except aggregated across stocks and dates for each year.

Result of Log-likelihood Ratio Test

The results of the log-likelihood ratio test of the power-law fits versus the fits of exponential distributions, log-normal distributions and exponentially truncated power-laws are shown in figure 4.6 below. There, the p-values represent the probability for both the power-law and alternative distributions being equally good fits. Each of them is calculated from its corresponding log-likelihood ratio Λ, which are also shown in the two bottom plots of the figure. The log-likelihood Λs are defined as:

$$\Lambda \equiv \ln \frac{L_{\texttt{Power-law}}}{L_{\texttt{Alternative}}} \tag{4.1}$$

where $L_{\texttt{Power-law}}$ & $L_{\texttt{Alternative}}$ represent, respectively, the likelihood functions of the power-law fit and the fit of the alternative distribution to some dataset $\{m^k_{s,d,a} : \forall k\}$. See section 3.3.3 for the theory behind the Log-likelihood ratio test.

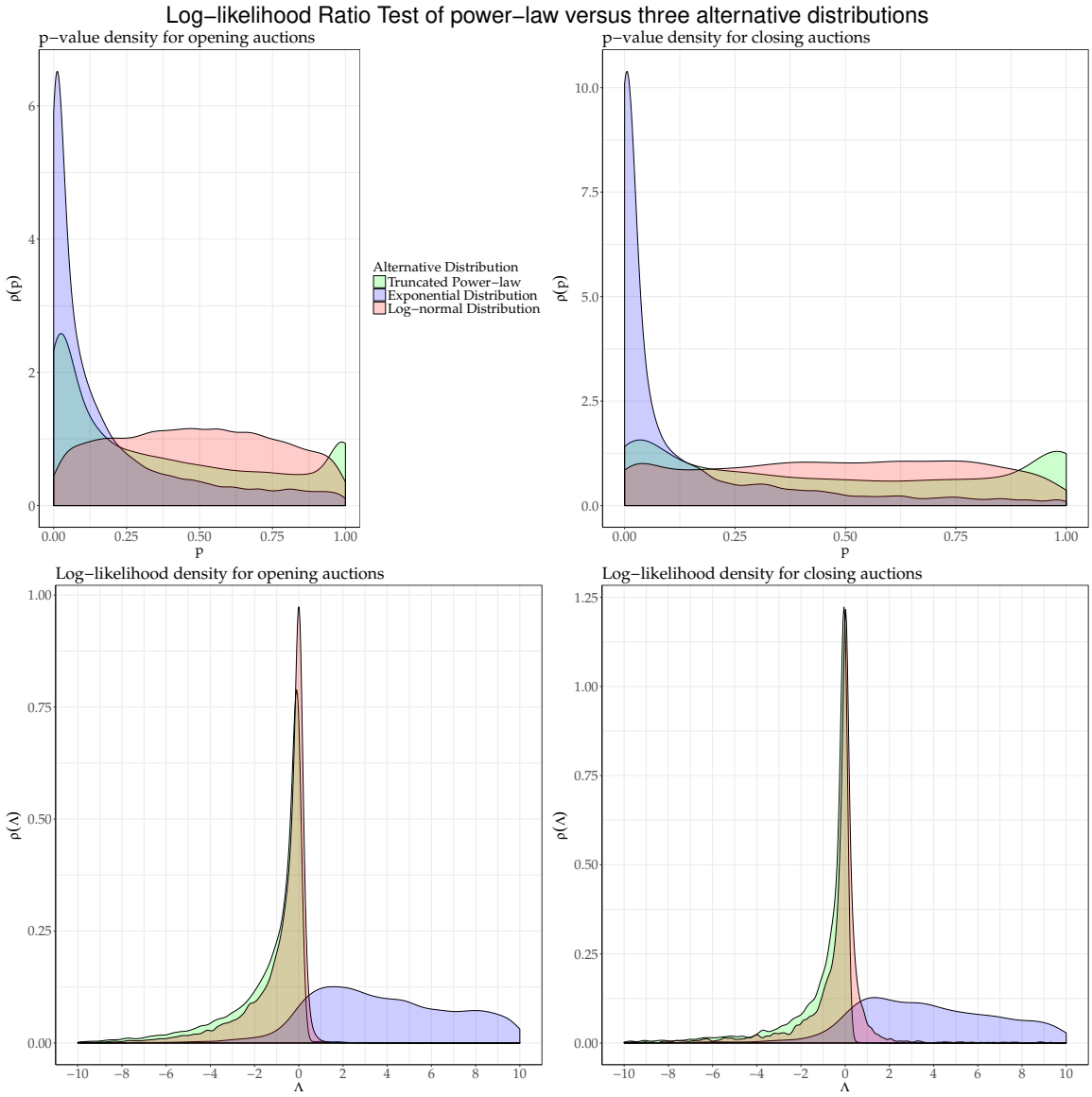

Figure 4.6: The log-likelihood ratios (bottom) and their associated p-values (top) found through Log-likelihood Ratio Test. The ratios and p-values are aggregated across days and stocks for opening (right) and closing (left) auctions. The test compared power-law fits to the MO dollar-volume distributions against various alternative fits (see legends).

4.2.7 Discussion of Results

Log-likelihood Ratio Test

The Log-likelihood Ratio Test was used to test the power-law fit to the tails of $\mathcal{D}_{m_{s,d,a}}$ against that of three other alternative distributions; exponential, log-normal and exponentially truncated power-law. Under the null-hypothesis (which was defined in

section 4.2.5), its p-values represent the probability that both the power-law and alternative distributions are equally good fits. So, judging from their distributions in figure 4.6, it is clear that none of the alternative distributions described the data much better. However, the exponential distribution is clearly a far worse fit than power-laws. This firmly establishes that $\mathcal{D}_{m_{s,d,a}}$ is not exponentially bounded, and thus is concluded that $\mathcal{D}_{m_{s,d,a}}$ is indisputably heavy-tailed. A brief discussion on the other two fits follows below:

The log-likelihood values of the log-normal fits are skewed towards the left, which implies the log-normal does tend to fit the data better. However, the corresponding p-values are roughly evenly distributed, which is what one would expect under the null-hypothesis. In addition, the log-normal distribution has more fitting parameters than the power-law distribution and so it more vulnerable to overfitting. Therefore, when taking these arguments into consideration, it is firmly concluded that the tails of $\mathcal{D}_{m_{s,d,a}}$ are not significantly better described by log-normal distributions than power-laws, and vice versa.

Concerning the exponentially truncated power-laws; judging from the distributions of the p-values and log-likelihood ratios, they fit the data somewhat better than pure power-laws for the opening auctions, but not closing auctions. This implies that the opening auctions are more exponentially truncated than the closing ones. The fits are hardly significantly better though, and as with log-normals, the truncated power-laws have more parameters and are thus easier to overfit. Therefore, it is firmly concluded that exponentially truncated power-laws do not explain the data significantly better than pure power-laws, and vice versa.

Analysis of Power-law Fits

$\mathcal{D}_{m_{s,d,a}}$ is not a simple distribution, as the above discussion shows, and therefore no clean, general conclusions can be reached on its behaviour, except it consistently being heavy-tailed. That said, the results did reveal rule of thumb behaviour that are often valid for the datasets $\{m^k_{s,d,a} \colon \forall k\}$, and so they will be discussed in the following paragraphs for the sake of completeness.

The power-law exponentials found are typically in the asymptotically Lévy-stable regime. As explained in section 3.2.1, a power-law tail is stable if its $\alpha \in \langle 1, 3\rangle$; and the figures 4.4 and 4.5 reveal that the great majority of the fitted $\alpha_{s,d,a}$ are indeed within that regime. While by itself this would imply that also the tails of $\mathcal{D}_{m_{s,d,a}}$ are Lévy-stable, it was established above that two non Lévy-stable distributions, exponentially truncated power-laws and log-normal distributions, fit equally well as pure power-laws. Therefore it cannot be concluded that the distributions $\mathcal{D}_{m_{s,d,a}}$ are Lévy-stable in their entireties. At most, there are some regions of $\mathcal{D}_{m_{s,d,a}}$ where there exists approximate Lévy-stability.

The two figures also illustrate the dependence of $\mathcal{D}_{\alpha_{s,d,a}}$ on s, d & a. In figure 4.4, the fitted power-law exponents are typically smaller for closing auctions than the opening ones, and figure 4.5 illustrates that $\mathcal{D}_{\alpha_{s,d,a}}$ is largely independent of d for yearly time-horizons. The influence of s is difficult to discern, but it appears significant; which is

not surprising, as the stocks traded on NYSEarca are largely ETFs based on wildly different financial assets; after all, one can not expect an ETF such as SPY, which is based on a basket of very large, very stable companies, to exhibit the same behaviour as an ETF like e.g. VXX, which is based on highly speculative futures. Therefore in conclusion, as a rule of thumb, the scaling properties/fatness of the tails of $\mathcal{D}_{m_{s,d,a}}$ vary by auction type, but are constant through time.

4.2.8 Data Limitation

The investigation was done on about 300 large (in terms of market capitalisation) NYSEarca stocks, and it covered half a decade of time. Therefore, there is no reason to believe that these results are not representative of the general behaviour of NYSEarca auctions (for stocks that are liquid under said auctions) under market conditions similar to those seen between 2010 and 2014.

Chapter 5

AGGREGATE PROPERTIES OF THE AUCTIONS

The aggregate properties of the MOs at the end of the auctions were explored next. First a theoretical treatment of $\mathsf{V}_{s,d,a}$ was undertaken to see how its distribution was linked with that of $\mathcal{D}_{\mathsf{M}_{s,d,a}}$. Then, the ratio $\log_{10} \frac{v_{s,\texttt{open},d}}{v_{s,\texttt{close},d}}$ was examined in detail. Finally, the predictability of $v_{s,a}[d]$ across d was examined thoroughly with forecasting algorithms.

5.1 Available Data

The available data consist of $840,379$ unique (s, d) datapoints for 2871 stocks traded on the NYSE, NYSEarca and NASDAQ stock exchanges. The data covers the time-period of 2009 to 2016.

Each datapoint contains the day's closing price $p_{s,d,\texttt{close}}$, the daily MO dollar-volumes for the open-market trading $o_{s,d}$ and the daily MO dollar-volume for both the opening auction $v_{s,d,\texttt{open}}$ and for the closing one $v_{s,d,\texttt{close}}$.

5.2 Theoretical Relation between $\mathcal{D}_{\mathsf{V}_{s,d,a}}$ and $\mathcal{D}_{\mathsf{M}_{s,d,a}}$

As defined in equation (1.4), the empirical daily MO dollar-volume $v_{s,d,a}$ is simply the sum of the $n_{s,d,a}$ elements in $\{m^k_{s,d,a} \colon \forall k\}$, and so their RV analogues are related through the same sum:

$$\mathsf{V}_{s,d,a} = \sum_{k=1}^{\mathsf{N}_{s,d,a}} \mathsf{M}^{\mathsf{k}}{}_{s,d,a} \tag{5.1}$$

Therefore, can one link the investigations done on $\{m^k_{s,d,a} \colon \forall k\}$ with the investigation that will now be done on $v_{s,d,\texttt{close}}$? The answer to this question largely depends on whether the sum of $\{m^k_{s,d,a} \colon \forall k\}$ can be expected to converge towards a distribution that is not overly messy.

To this end, it would be tempting to assume that $\mathsf{M}^k_{s,d,a}$ was i.i.d. $\forall k$, as then the sum would imply that $\mathcal{D}_{\mathsf{V}_{s,d,a}}$ was completely defined by just two other distributions; $\mathcal{D}_{\mathsf{M}_{s,d,a}}$ and $\mathcal{D}_{\mathsf{N}_{s,d,a}}$. Further, if $\mathsf{N}_{s,d,a}$ were sufficiently peaked and the variance of $\mathsf{M}^k_{s,d,a}$

finite, one would expect $\mathcal{D}_{\mathsf{V}_{s,d,a}}$ to converge towards the normal distribution by the CLT, and if the variance is not finite then at least the tails of $\mathcal{D}_{\mathsf{V}_{s,d,a}}$ would converge towards a single power-law by the generalised CLT (see section 3.2.1).

Alas, the mechanics of the auction defined a mathematical dependency between the empirical sets $\{m^k_{s,d,a} : \forall k\}$ with equation (1.2), which means the $\mathsf{M}^k_{s,d,a}$ must be correlated with each other through:

$$\left\{ \mathsf{M}^k_{s,d,a} : \sum_{k \in \mathbb{O}^{\texttt{Buy}}_{s,d,a}} \mathsf{M}^k_{s,d,a} = \sum_{k \in \mathbb{O}^{\texttt{Sell}}_{s,d,a}} \mathsf{M}^k_{s,d,a} \right\} \tag{5.2}$$

with the cardinalities of the list of matched buy orders $\mathbb{O}^{\texttt{Buy}}_{s,d,a}$ and matched sell orders $\mathbb{O}^{\texttt{Sell}}_{s,d,a}$ being yet another two RVs with unknown distributions. In addition to this, the traders are heterogeneous and interact with each other, which implies the possible presence of yet another significant interdependence between $\mathsf{M}^k_{s,d,a}$ across k. Further, one cannot assume $\mathsf{M}^k_{s,d,a}$ is identically distributed either, as traders are heterogeneous and this might well mean that $\mathsf{M}^k_{s,d,a}$ has significantly different distributions across k. Thus, in conclusion, one can by no means assume i.i.d. In addition to this, it was clearly established in section 4.2.7 that the realisations $m^k_{s,d,a}$ are heavy-tailed over several orders of magnitude, and are thus of extreme variance; this makes any convergence towards the normal distribution, for all practical purposes, impossible. In addition, section 4.2.7 also established that the heavy-tails can not be considered Lévy stable either; thus convergence in the tails cannot be expected either. Thus one must conclude that neither the CLT nor the generalised CLT can be applied on the sums of $\mathsf{M}^k_{s,d,a}$.

This means that the statistical properties of $\mathcal{D}_{m_{s,d,a}}$ and $\mathcal{D}_{v_{s,d,a}}$ are not linked in any simple, clear way, with the exception of the observed heavy-tailed and complex behaviour of $\mathcal{D}_{m_{s,d,a}}$ carrying over to $\mathcal{D}_{m_{s,d,a}}$; this means $\mathcal{D}_{m_{s,d,a}}$ can neither be well-behaved nor convergent towards any simple distribution. Therefore, the investigation of $v_{s,d,a}$ was carried out in a completely separate fashion from the investigation of $\{m^k_{s,d,a} : \forall k\}$ in the previous chapter.

5.3 Properties of $\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}}$

5.3.1 Statistical Regularities of $\mathcal{D}_{\log_{10} \frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$

Preliminary Investigation

When carrying out the preliminary investigations into $v_{s,d,a}$, it was observed that the datasets $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ were approximately normally distributed for many s (the $\log_{10}$ operator had to be employed due to the heavy-tailed nature of $v_{s,d,a}$). This normality is illustrated through figure 5.1 for the Apple stock dataset $\{\log_{10} \frac{v_{\texttt{AAPL},d,\texttt{open}}}{v_{\texttt{AAPL},d,\texttt{close}}} : \forall d\}$.

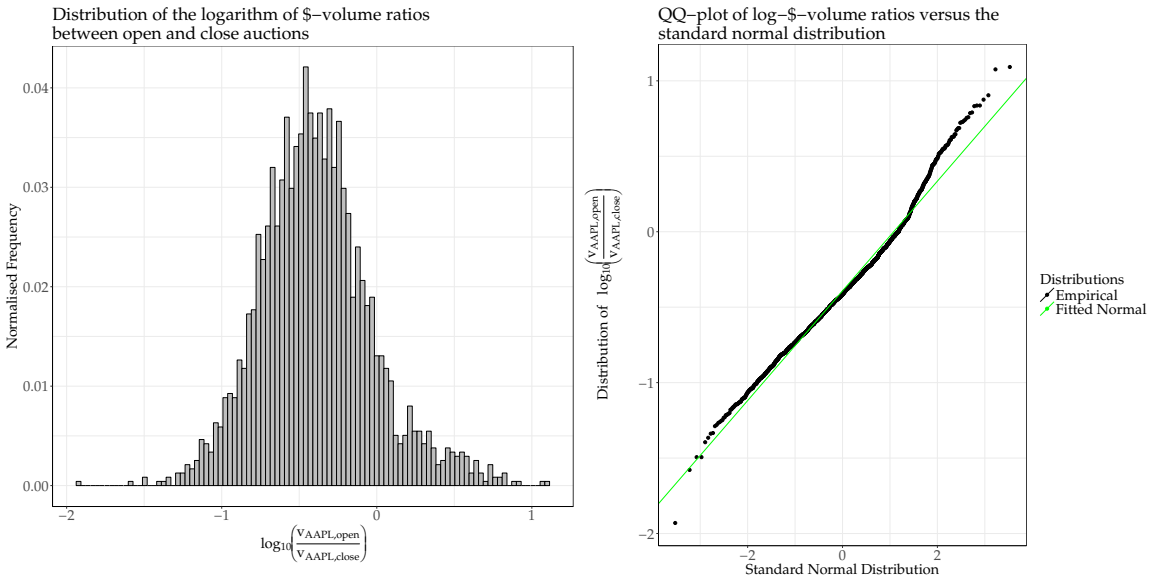

Figure 5.1: A histogram and QQ plot of the opening vs closing auction dollar-volume ratios aggregated across all dates for the AAPL stock.

If this was a general phenomena, it would mean that $\frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}$ would be approximately log-normally distributed. This would be an interesting statistical regularity, and so this possible log-normality was investigated next.

Fitting Log-normal Distributions

The indications of log-normality were investigated by using the `R::MASS` package to systematically fit normal distributions across $s \in \mathbb{S}_E$ and d to every dataset $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ with at least 100 datapoints. The fitting parameters μ_s and σ_s were found through MLE, as described in section 3.3.5.

Resulting Fits

This resulted in successful fits for 2779 stocks across, which is illustrated in the two figures below, which show the distributions of the fit parameters μ_s and σ_s.

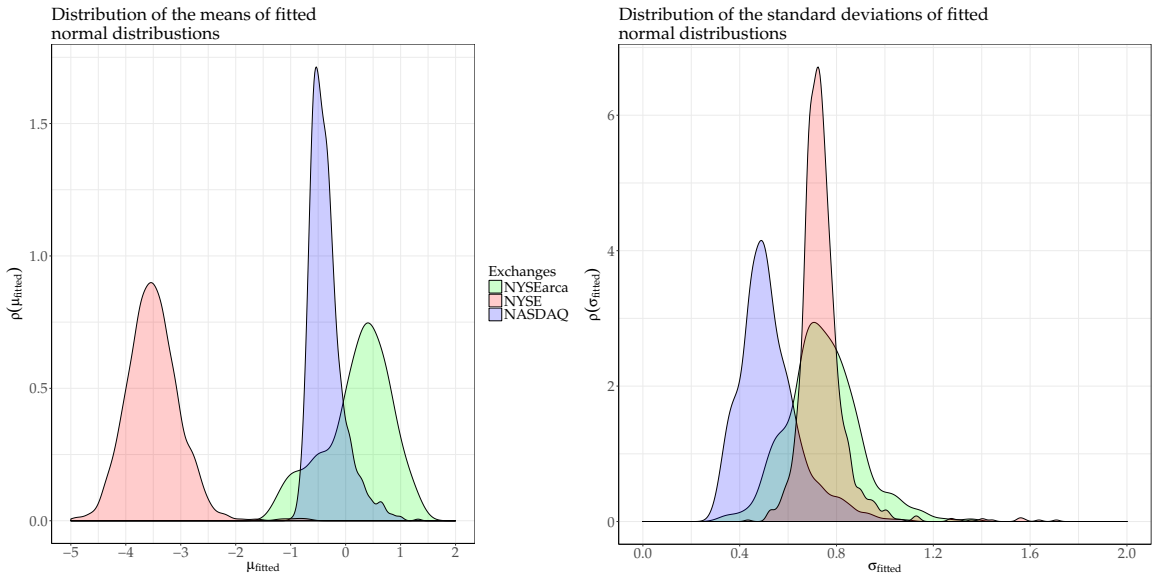

Figure 5.2: Density of the parameters of normal distribution fitted to log MO daily dollar-volume ratios aggregated across stocks, for each exchange. The distribution of fitted means is on the left, and the distribution of fitted standard deviations is on the right.

Figure 5.2 shows that the parameters of the fitted log-normals are strongly dependent on the exchanges. Further, it is clear that, in almost all cases, σ_s is significantly larger than 0, which implies the log-ratios are noticeably skewed and kurtotic in shape; this is unsurprising in light of the heavy-tailed nature of $m^k_{s,d,a}$.

Aggregation by Exchange

Moving on, aggregation was also done over s and d by exchange. This resulted in three datasets $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d, \forall s \in \mathbb{S}_E\}, E \in \{\texttt{NYSEarca}, \texttt{NYSE}, \texttt{NASDAQ}\}$, and they are distributed according to the empirical densities $\mathcal{D}_{\log_{10} \frac{v_{E,\texttt{open}}}{v_{E,\texttt{close}}}}$ plotted in figure 5.3 below.

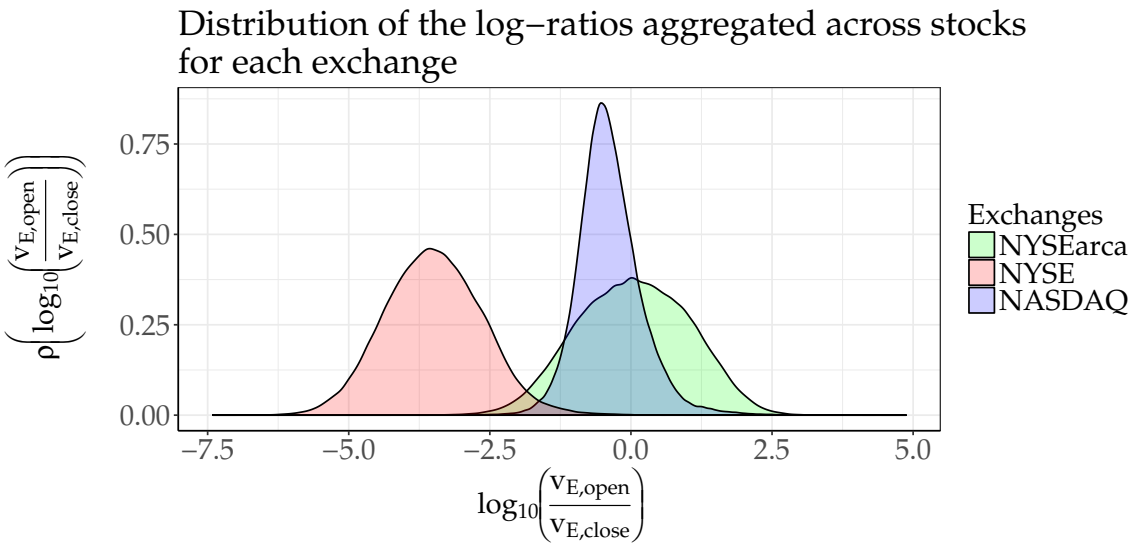

Figure 5.3: Density of the log MO daily dollar-volume ratios aggregated across dates and stocks for each exchange.

Figure 5.3 underlines the fact that the log-ratios vary significantly across exchanges.

5.3.2 Testing Fits and Comparing Distributions

Testing Procedure

The goodness of fit of the fitted normal distributions had to be investigated, and it was also of interest to see whether the log-ratios had identical distributions across all stocks for each exchange. For this the KS test, which is described in section 3.3.3, was used.

The goodness of fit was evaluated by using the one-sample KS test to test whether the empirical densities $\mathcal{D}_{\log_{10} \frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ and their respective fitted normal distributions $\mathcal{N}(\mu_s, \sigma_s^2)$ were significantly different. The null-hypothesis here was that the samples $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ were indeed drawn from $\mathcal{N}(\mu_s, \sigma_s^2)$.

Afterwards, the two-sample version of the KS test was used to check whether the log-ratios of the stocks were all distributed according to the aggregated distribution of their exchange, i.e. according to $\mathcal{D}_{\log_{10} \frac{v_{E,\texttt{open}}}{v_{E,\texttt{close}}}}$. The null-hypothesis here was that $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ and $\{\log_{10} \frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d, \forall s \in \mathbb{S}_E\}, E \in \{\texttt{NYSEarca}, \texttt{NYSE}, \texttt{NASDAQ}\}$ were indeed drawn from the same distribution.

Testing Results

The distribution of the resulting KS p-values are shown in figure 5.4 below. In the right plot of the figure, the p-value distributions were scaled instead of normalised for aesthetic reasons.

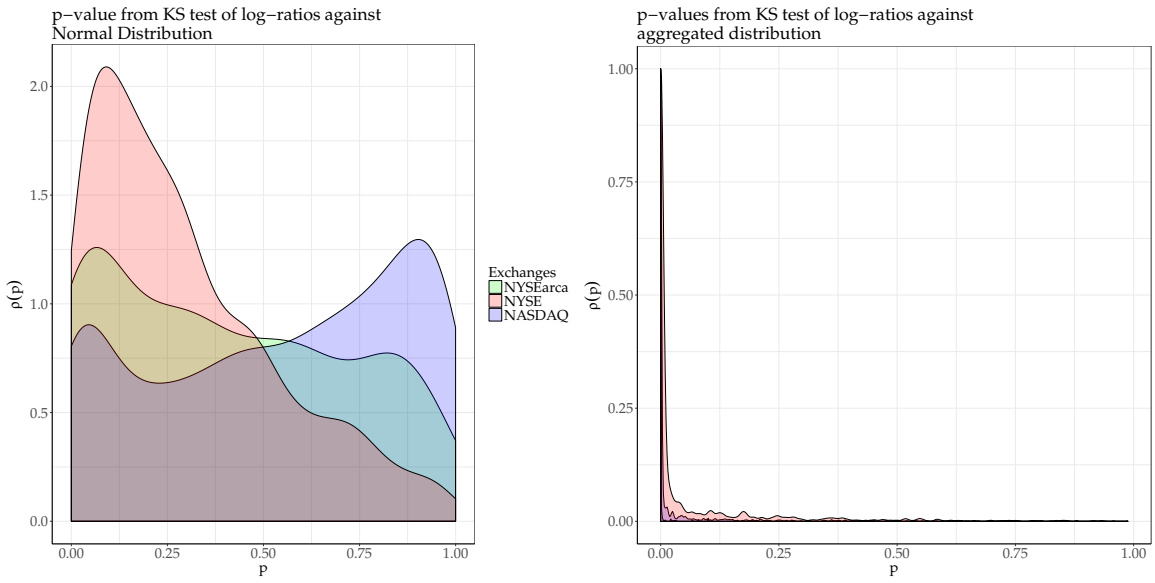

Figure 5.4: p-values of Kolmogorov-Smirnov tests performed on the distribution of log MO daily dollar-volume ratios aggregated across days for each stock. These stocks' distributions were tested against both fitted normal distributions (left) and the aggregated distribution of log-ratios of the exchange the stock belonged to.

5.3.3 Discussion of Observations

Stylised Facts

Synthesising the information given in figures 5.2 and 5.3 tells us several stylised facts on the auctions of the exchanges; firstly, the traded dollar-volume during the closing auctions of NYSE are, typically, several orders of magnitude more than that of its opening auctions. Further, the ratios of both NYSEarca and NYSE are enormously spread out, while that of NASDAQ has much less variability. Finally, the trading volume of the great majority of NASDAQ stocks and all NYSE stocks is larger during closing than opening auctions.

Dependence of $\mathcal{D}_{\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ on Exchange and Stock

The figures 5.2 and 5.3 clearly show that the stocks' exchanges affect the ratios, but they are evidently far from the only effect; because figure 5.4 shows that the KS test firmly rejected the null-hypothesis of $\{\log_{10}\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ being distributed according to $\mathcal{D}_{\log_{10}\frac{v_{E,\texttt{open}}}{v_{E,\texttt{close}}}}$. It is therefore concluded that the distributions of the datasets $\{\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ are not only dependent on the exchange, but also on the stock itself.

$\mathcal{D}_{\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ and Log-normality

The distributions of p-values in the left plot of figure 5.4 show that the KS test failed to reject the null-hypothesis of $\mathcal{D}_{\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ being a log-normal distribution. In fact, for NASDAQ and NYSEarca, the distribution of p-values are roughly what would be

expected if the null-hypothesis was correct for these exchanges. Therefore, it is concluded that the datasets $\{\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} : \forall d\}$ are, approximately, log-normally distributed.

Explaining the Log-normality & Further Work Assuming that the reached conclusion on log-normality is correct, one can form hypothesises on why the log-normality is present for future investigation.

Log-normal distributions arise from the multiplication of many i.i.d. RVs, and so the first thought that might enter one's mind is that $\mathcal{D}_{\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}}$ is the affected by one such multiplicative process. However, decomposing $\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}$ gives:

$$\frac{v_{s,d,\texttt{open}}}{v_{s,d,\texttt{close}}} = \frac{\sum_{k=1}^{n_{s,d,\texttt{open}}} m^k_{s,d,\texttt{open}}}{\sum_{k=1}^{n_{s,d,\texttt{close}}} m^k_{s,d,\texttt{close}}} = \frac{p_{s,d,\texttt{open}}}{p_{s,d,\texttt{close}}} \cdot \frac{\sum_{k=1}^{n_{s,d,\texttt{open}}} q^k_{s,d,\texttt{open}}}{\sum_{k=1}^{n_{s,d,\texttt{close}}} q^k_{s,d,\texttt{close}}} \tag{5.3}$$

which is essentially a sum of heavy-tailed variables divided by another sum of heavy-tailed variables, with both sums being heavily correlated with each-other. There are only two obvious multiplicative processes present here, namely the division of the volumes by each other and multiplication with the price ratio, which rarely deviates from unity on a daily time-horizon; and these can obviously not explain the log-normality by themselves. Thus it is very difficult to see how any multiplication of many i.i.d. RVs can be involved, and why a hypothesis based on that would be reasonable.

Another possibility is that maximisation of (Shannon) entropy[1] is involved. The log-normal distribution is one of the entropy-maximising distributions, which means if the expectation value and variance of a RV $\log_{10} \frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}$ were constrained for each s such that:

$$\begin{aligned} \texttt{E}[\log_{10} \frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}] &= \mu_s \\ \texttt{Var}[\log_{10} \frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}] &= \sigma_s^2 \end{aligned} \tag{5.4}$$

and some process (e.g. trading activity) forced $\frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}$ to tend towards maximum uncertainty, then it would mathematically follow [37] that $\frac{\mathsf{V}_{s,\texttt{open}}}{\mathsf{V}_{s,\texttt{close}}}$ must be log-normal. In physical terms, this process would have to be causing maximum disorder in the MO dollar-volumes under the constraints of equation (5.4). The success of this section's fitting of log-normal distributions to $\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}$ suggests that the mentioned constraints are indeed present, though it is remains to be understood why. Further, the alleged process that maximises the uncertainty of $\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}$ remains unknown (though due to equation (5.3), it would have to be related to whatever process causes the heavy-tails of the individual $m^k_{s,d,a}$ in the first place). In summary then, disorder maximisation given the constraints remains a viable route of attack, but further work must be done to understand the system before a rigorous hypothesis can be formulated.

Thus, in conclusion, no rigorous explanation for the observed log-normality can be offered with the results available. All that can be said, for now, is that it is clear $\frac{v_{s,\texttt{open}}}{v_{s,\texttt{close}}}$ is approximately log-normal, but it is a truly hard case to understand why that is so.

[1] Which can be interpreted as being equivalent to thermodynamic entropy.

5.3.4 Data Quality

The data covered all the major (in terms of market capitalisation) stocks of NYSEarca & NASDAQ and it ran across 8 years, 2009-2016. It should be noted that the data was unreliable for NYSE, which could explain why the log-normality was less pronounced for it than the other exchanges. Thus, all in all, the results should be considered valid in general for NYSEarca and NASDAQ, and dubious for NYSE, for market conditions of the given time-frame.

5.4 Predictability of $v_{s,a}[d]$

5.4.1 Autocorrelation of $v_{s,a}[d]$

It was discovered that the daily volume $v_{s,d,a}$ was strongly autocorrelated across days for multiple stocks, and that its autocorrelation function varied periodically with time. See figure 5.5 as an example. The presence of such autocorrelation implied that $\log_{10} v_{s,a}[d]$ was highly non-independent, and therefore, to some extent, also predictable through analysis of its previous realisations $\log_{10} v_{s,a}[\delta], \delta < d$. The daily MO dollar-volume are very important as they are closely related to the auction's liquidity, and so this was investigated next.

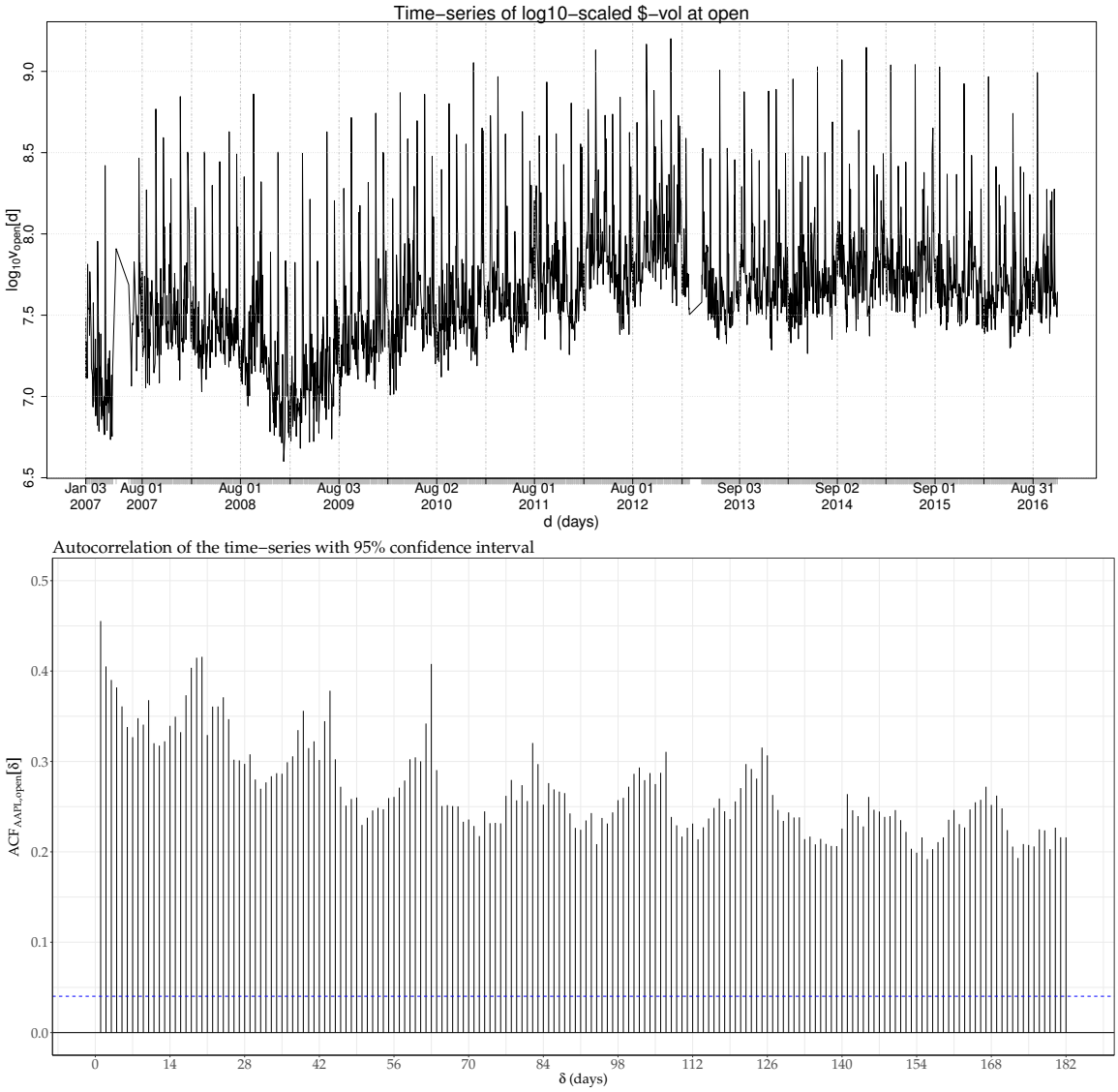

Figure 5.5: Top: Time-plot of the log-scaled total opening auction MO dollar-volume for the AAPL stock. Bottom: Autocorrelation function of the same time-series with a maximum lag of 182 days.

5.4.2 Pre-forecasting Considerations

Forecasting is the art of using historical information to make predictions into the future (see section 3.5.1 for a more detailed explanation). In our case, this meant using the available data to construct a model for the dynamics of the time-series $v_{s,a}[d]$ that could be used for forecasting. The previous paragraph revealed a clear interdependence between the historical amplitudes of $v_{s,a}[d]$, and so we started by analysing the $v_{s,a}[d]$ time-series in isolation. Before a statistical model could be found though, the unique features of the data had to be considered.

Large Variability

$v_{s,a}[d]$ had an enormous variability due to its heavy-tails, and this would make fitting models directly to it difficult. Specifically, the residuals of the fits would end up heavy-tailed instead of Gaussian-shaped, which would imply bias and hence invalidate the results. Therefore, it was decided to log-transform the dollar-volume from $v_{s,a}[d]$ to $\log_{10} v_{s,a}[d]$ beforehand to significantly bring down the variability. This is equivalent to making what is known as a Box-Cox transform in statistics [21]. To illustrate, let us look at this from the perspective of the random variables.

Let f be the PDF of $V_{s,d,a}$, and g be the PDF of its log-transform $\log_{10} V_{s,d,a}$. Then, dropping the subscripts and using conservation of probability we get that

$$f(V)dV = g(\log_{10} V)d(\log_{10} V) \Rightarrow g(\log_{10} V) = \frac{1}{\log_{10} e} V f(V) \tag{5.5}$$

and therefore the new mean $\mu_{\log_{10} V_{s,d,a}}$ and SD $\sigma_{\log_{10} V_{s,d,a}}$, when dropping the subscripts, become:

$$\begin{aligned} \mu &= \int_0^\infty \log_{10} V g(\log_{10} V)d(\log_{10} V) = \frac{1}{\log_{10} e}\int_0^\infty \log_{10} V f(V)dV \\ \sigma &= \int_0^\infty (\log_{10} V - \mu)^2 g(\log_{10} V)d(\log_{10} V) = \frac{1}{\log_{10} e}\int_0^\infty (\log_{10} V - \mu)^2 f(V)dV \end{aligned} \tag{5.6}$$

This mean that the log-transform of $v_{s,a}[d]$ similarly log-transforms its SD, curbing its extreme variability and make fitting models much easier. The negative consequence of this though, is that a predictive model found under some criteria to be optimal for $\log_{10} v_{s,a}[d]$ can not be expected to also be optimal for $v_{s,a}[d]$. In the end though, the most important goal was to confirm predictability of the MO volumes and not necessarily find a perfect predictive model, and so let us continue while using the log-transformation.

Lack of Stationarity

Second, the time-series $\log_{10} v_{s,a}[d]$ was highly non-stationary, making the modelling more difficult. The non-stationarity can be both observed empirically from the irregularities in figure 5.5, and understood from theoretical considerations; the volume is a function of many exogenous variables that are dynamic over time, such as the performance of the underlying company, market conditions and the dynamical nature of the traders and their strategies.

3rd Friday of the Month

There was one very important cyclostationary, i.e. seasonal, effect; equity futures (e.g. options) on our stocks (almost) always expire on the 3rd Friday of every month (with the exception of when it falls on holidays). This lead to very large jumps in volume of the underlying stocks, as can be observed by the peaks in figure 5.5. Extreme events such as this have to be identified and handled manually, thus fixed effects were employed to handle the 3rd Friday of month jumps (see 3.5.2 for a description of fixed effects).

5.4.3 Forecasting $\log_{10} v_{s,a}[d]$

After taking the above into consideration, automatic statistical learning methods in conjunction with analytical reasoning was employed to find appropriate models. Three different methods were used; ARIMA, Facebook Prophet and Random Forest. For the two former methods, the analytical reasoning revolved around accounting for effects that would be difficult for the methods to handle, such as the 3rd Friday of the month phenomena. In the case of Random Forest, the analytical reasoning was used to identify features (& combinations of features) important for predicting the response.

The methods were applied in practice in `R`.

Methods Used

ARIMA The family of ARIMA (see section 3.5.2 for theory) models are able to represent a broad set of linear processes, and the model-finding algorithm used in the `R::forecast` package, i.e. the Hyndman-Khandakar algorithm, is computationally fast and requires little tuning. Therefore, it was decided to start by applying the ARIMA method from the `R::forecast` package in combination with 6 fixed effects; one for each weekday, and one for the 3rd Friday of the month. The latter fixed effect was absolutely necessary for reasons already explained, and the former 5 fixed effects were added to remove any non-stationarities present due to the specific weekday, such as the effect of stock exchange opening after a weekend on a Mondays and closing before a weekend on Fridays.

It was hoped that some ARIMA model would be able to explain the fluctuations not explained by the fixed effects, however that turned out not to be the case. After the fixed effects were accounted for in $\log_{10} v_{s,a}[d]$, ARIMA ended up not being able to describe the residuals as anything other than white noise. This meant the dynamics involved were too subtle for ARIMA, and thus more powerful, non-linear models were required; hence Prophet was tried next.

Facebook Prophet Facebook Prophet (see section 3.5.3 for theory) was advertised [22] as a powerful forecasting method for time-series analysis, and it was also available in `R` through the `R::prophet` package. Unlike ARIMA, Prophet is able to uncover a range of non-linear processes. Particular advantages Prophet had over ARIMA was its claimed robustness against gaps in data and ability to identify weekly and yearly seasonalities through Fourier transforms.

Therefore, it was decided to employ the Prophet method while accounting for 3rd Friday phenomena with fixed effects (also known as "holidays" in `R::prophet`). It turned out that Prophet did provide more accurate forecasts than ARIMA, albeit at the cost of being much more computationally intensive and also exhibiting instability (i.e. it was prone to crashing).

Random Forests It was next decided to use a powerful, fast and versatile machine learning method; Random Forests (see section 3.5.4 for theory). Random Forests is in particular very good at finding accurate models in high-dimensional feature-spaces [23], which gave us two opportunities. Firstly, it allowed us to incorporate a broad

range of information during modelling, such as historical open-market liquidity or price movements in the underlying stock. Second, it allowed us to test many different combinations of features and evaluate their importance for the response.

Random Forests was implemented through the `R::randomForest` package. Various combinations of features were used to train models and then use them for forecasting. While the primary goal was to optimise the forecast, a secondary goal was to understand which features were the most strongly related to the response. The selection of features was done through trial-and-error, i.e. by adding a feature and see whether it improves the accuracy, as well as using the Breiman-Cutler importance measure [24] as a supporting guide.

Implementation in `R`

Application across (s, a) The algorithms behind the methods were in the `R` packages, but there was still the question of calibrating them with the fixed effects/features, and applying them systematically through the datasets. For this, an algorithm was written which, for every auction (s, a), moved through its response function $\log_{10} v_{s,a}[d]$ along d in two intervals simultaneously; an in-sample training interval where a model was trained, and an out-of-sample prediction interval where the prediction(s) was made. This is illustrated in more detail, albeit still highly simplified compared to the actual implementation, in the pseudocode below.

It works in the following manner: For each iteration, a method $\mathcal{F}$ is first calibrated by applying fixed effects or defining features, secondly it is used to find a model $\mathcal{M}$ which describes the in-sample interval, and finally $\mathcal{M}$ is used to generate a forecast for the out-of-sample interval. The forecasts are then all combined into a dataset $\{\hat{v}^{\mathcal{F}}_{s,a}[d] : \forall d \in \hat{\mathbb{D}}^{\mathcal{F}}_{s,a}\}$ and stored.

The user input the algorithm required was the length of the in-sample interval $n_{\texttt{train}}$, the length of the prediction-interval $n_{\texttt{pred}}$ and the machine-learning method $\mathcal{F}$.

Algorithm 2 Fit & Forecast Algorithm

Input: $n_{\texttt{train}}, n_{\texttt{pred}}, \mathcal{F}$, `Available Data`
Output: $\{\hat{v}^{\mathcal{F}}_{s,a}[d] \colon \forall d \in \hat{\mathbb{D}}^{\mathcal{F}}_{s,a}\}$

for $(\forall (a, E) \in \{\texttt{open},\texttt{close}\} \times \{\texttt{NYSEarca},\texttt{NYSE},\texttt{NASDSAQ}\})$ **do**
  for $(\forall s \in \mathbb{S}_E)$ **do**
    $d^{\texttt{begin}}_{\texttt{train}} \equiv$ First element in $\mathbb{D}_{s,a}$
    $d^{\texttt{end}}_{\texttt{train}} \equiv n_{\texttt{train}}$-th element after $d^{\texttt{begin}}_{\texttt{train}}$ in $\mathbb{D}_{s,a}$
    $d^{\texttt{begin}}_{\texttt{pred}} \equiv$ First element after $d^{\texttt{end}}_{\texttt{train}}$ in $\mathbb{D}_{s,a}$
    $d^{\texttt{end}}_{\texttt{pred}} \equiv n_{\texttt{pred}}$ -th element after $d^{\texttt{end}}_{\texttt{train}}$ in $\mathbb{D}_{s,a}$
    $\{\hat{v}^{\mathcal{F}}_{s,a}[d] : \forall d \in \hat{\mathbb{D}}^{\mathcal{F}}_{s,a}\} \equiv \emptyset$
    while ($d^{\texttt{end}}_{\texttt{pred}}$ is well defined) **do**
      Calibrate $\mathcal{F}$
      $\mathcal{M} = \mathcal{F}(\{v_{s,a}[d] : \forall d \in (d^{\texttt{begin}}_{\texttt{train}} \leq \mathbb{D}_{s,a} \leq d^{\texttt{end}}_{\texttt{train}})\})$
      Use $\mathcal{M}$ to create the forecast $\{\tilde{v}_{s,a}[d] : \forall d \in (d^{\texttt{begin}}_{\texttt{pred}} \leq \mathbb{D}_{s,a} \leq d^{\texttt{end}}_{\texttt{pred}})\}$
      Insert $\{\tilde{v}_{s,a}[d] : \forall d \in (d^{\texttt{begin}}_{\texttt{pred}} \leq \mathbb{D}_{s,a} \leq d^{\texttt{end}}_{\texttt{pred}})\}$ into $\{\hat{v}^{\mathcal{F}}_{s,a}[d] : \forall d \in \hat{\mathbb{D}}^{\mathcal{F}}_{s,a}\}$
      $d^{\texttt{begin}}_{\texttt{train}} = n_{\texttt{pred}}$-th element after $d^{\texttt{begin}}_{\texttt{train}}$ in $\mathbb{D}_{s,a}$
      $d^{\texttt{end}}_{\texttt{train}} = n_{\texttt{pred}}$-th element after $d^{\texttt{end}}_{\texttt{train}}$ in $\mathbb{D}_{s,a}$
      $d^{\texttt{begin}}_{\texttt{pred}} = n_{\texttt{pred}}$-th element after $d^{\texttt{begin}}_{\texttt{pred}}$ in $\mathbb{D}_{s,a}$
      $d^{\texttt{end}}_{\texttt{pred}} = n_{\texttt{pred}}$-th element after $d^{\texttt{end}}_{\texttt{pred}}$ in $\mathbb{D}_{s,a}$

5.4.4 Evaluating the Forecast

After the algorithm had run its course for all three methods, the resulting forecasts were plotted and their accuracy was measured.

Measuring Accuracy

The forecasting accuracy was measured with the root mean square of the residuals between the forecast and empirical response, i.e. with root mean square error (RMSE). The residuals are defined as:

$$r^{\mathcal{F}}_{s,d,a} = \log_{10} \hat{v}^{\mathcal{F}}_{s,a}[d] - \log_{10} v_{s,a}[d] = \log_{10} \frac{\hat{v}^{\mathcal{F}}_{s,a}[d]}{v_{s,a}[d]} \tag{5.7}$$

Theoretically, the expectation value of the residuals must be 0 if the forecast is unbiased, and not autocorrelated across d if all the systematic movements have been captured. Moving on, the RMSE of method $\mathcal{F}$ for auction a was defined by:

$$\texttt{RMSE}^2 = \langle (r^{\mathcal{F}}_{s,d,a})^2 \rangle_{\forall (s,d)} = \frac{1}{\sum_{\forall s \in \hat{\mathbb{S}}^{\mathcal{F}}} |\hat{\mathbb{D}}^{\mathcal{F}}_{s,a}|} \sum_{\forall s \in \hat{\mathbb{S}}^{\mathcal{F}}} \sum_{\forall d \in \hat{\mathbb{D}}^{\mathcal{F}}} (r^{\mathcal{F}}_{s,d,a})^2 \tag{5.8}$$

Cross-validation

The mortal danger of forecasting is overfitting. Overfitting, as the name suggests, occurs when a model follows the dataset it is trained with too closely, leading to the model becoming optimal at describing the noise of the dataset instead of its underlying dynamics. Such a model would perform poorly at describing new data and hence also be bad at forecasting. There exist multiple techniques for avoiding overfitting, but

the simplest and most common way is evaluating the model's performance on a new dataset with dynamics identical to those of the old one. This is called cross-validation.

Cross-validation was employed in two ways. Firstly, when evaluating the quality of the fits, only the accuracy of the forecasts were considered - the goodness of fit of the models to the respective datasets they were trained on was ignored. This can be considered cross-validation across d. Second, when comparing the fits of different methods with each other, they were cross-validated across s.

5.4.5 Results of Forecasting

In this section, the results of the forecasts are presented through plots and tables.

Technical Notes

- All the methods were employed on identical data.
- $n_{\texttt{train}} = 480$ training points were chosen, which is nearly 2 years worth of trading days (252 trading days per year). The training window was this large because Facebook prophet required well over a year of data to refrain from regular crashing.
- The predictions were done only one day in advance to maximise the forecasting accuracy, so $n_{\texttt{pred}} = 1$.
- A significant quantity of stocks had to be dropped because they had less than the required 480 datapoints available for training. This included all the stocks in NYSE.
- The set of forecasted datapoints $\hat{\mathbb{D}}^{\mathcal{F}}_{s,a}$ were identical for the Random Forests and ARIMA methods, and of length 198886 for both opening and closing auctions. The equivalent for Prophet was 153875 datapoints. *Therefore, all the results shown in the tables below are calculated only for the* 153875 *datapoints which all three methods managed to fit.*
- The cross-validation across s, as mentioned in section 5.4.4, was done by splitting the forecasts of each method into two segments, and calculating the RMSE for each of them separately. The splitting was done by alphabetically sorting the set of stocks forecasted and then cutting them in half down the middle.
- Finally, the naive method, marked as "0. Naive;" is present in the tables 5.1 and 5.2 as a benchmark. The naive method "forecasts" the daily-dollar volume by simply setting it equal to the previous trading day's daily dollar-volume.

Comparison between Methods

Across Auctions The two tables below represent the results of the forecasts across auctions. In the first column, Δ is the differencing operator, L is the lag operator and $o_s[d]$ is the open-market dollar-volume traded on date d for stock s. So, for example, including $L^1 \log_{10} o_s[d]$ in the rows of the first column means the log of the previous day's open-market dollar-volume was one of the variables used to predict the next

day's response. The $RMSE'$ and $RMSE''$ columns represent the RMSE of the two cross-validation segments defined in the above list. The $RMSE_{tot}$ is the RMSE of the whole dataset of forecasts. μ_r is the mean of the residuals of the forecast.

Opening Auctions

Method & fixed effects/Features	RMSE$'$	RMSE$''$	$RMSE_{\texttt{tot}}$	μ_r
0. Naive; New \$-volume same as last	0.417	0.400	0.409	0.000
1. ARIMA; d, Weekdays, 3rd Friday	0.330	0.341	0.336	-0.030
2. Prophet; d, 3rd Friday	0.284	0.292	0.288	-0.011
3. RF; d, Weekdays, 3rd Friday	0.283	0.290	0.287	-0.000
4. RF; d, 3rd Friday	0.278	0.285	0.282	0.006
5. RF; Weekdays, 3rd Friday	0.330	0.341	0.336	-0.031
6. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$	0.274	0.282	0.278	0.011
7. RF; Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$	0.297	0.305	0.300	-0.017
8. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} o_s[d]$	0.269	0.277	0.273	0.010
9. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$	0.274	0.281	0.278	0.012

Table 5.1: Result of forecasting the log MO daily dollar-volumes for the opening auction.

Closing Auctions

Method & fixed effects/Features	RMSE$'$	RMSE$''$	$RMSE_{\texttt{tot}}$	μ_r
0. Naive; New \$-volume same as last	0.534	0.556	0.543	0.002
1. ARIMA; d, Weekdays, 3rd Friday	0.459	0.489	0.473	-0.103
2. Prophet; d, 3rd Friday	0.413	0.435	0.423	-0.035
3. RF; d, Weekdays, 3rd Friday	0.410	0.430	0.419	-0.030
4. RF; d, 3rd Friday	0.407	0.426	0.415	-0.023
5. RF; Weekdays, 3rd Friday	0.458	0.488	0.472	-0.103
6. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{close}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$	0.409	0.428	0.417	0.003
7. RF; Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{close}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$	0.429	0.452	0.439	0.072
8. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{close}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$, $\log_{10} o_s[d]$	0.398	0.417	0.406	-0.001
9. RF; d, Weekdays, 3rd Friday, $\log_{10} o_s[d]$	0.398	0.418	0.407	-0.026

Table 5.2: Result of forecasting the log MO daily dollar-volumes for the closing auction.

Across Exchanges The methods are also illustrated for each exchange, in the tables below.

Opening Auctions for each Exchange

	NYSEarca		NASDAQ	
Method & fixed effects/Features	RMSE	μ_r	RMSE	μ_r
1. ARIMA; d, Weekdays, 3rd Friday	0.376	-0.085	0.294	-0.051
2. Prophet; d, 3rd Friday	0.335	-0.001	0.237	-0.021
3. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} o_s[d]$	0.323	0.009	0.217	0.011
4. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{open}}[d]$, $L^1 \log_{10} v_{s,\texttt{open}}[d]$	0.327	0.011	0.224	0.011

Table 5.3: Result of forecasting log MO daily dollar-volumes for the opening auction across exchanges.

Closing Auctions for each Exchange

Method & fixed effects/Features	NYSEarca RMSE	μ_r	NASDAQ RMSE	μ_r
1. ARIMA; d, Weekdays, 3rd Friday	0.589	-0.092	0.369	-0.115
2. Prophet; d, 3rd Friday	0.532	-0.027	0.288	-0.042
3. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{close}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$, $\log_{10} o_s[d]$	0.517	-0.008	0.266	0.006
4. RF; d, Weekdays, 3rd Friday, $\Delta^1 L^1 \log_{10} v_{s,\texttt{close}}[d]$, $L^1 \log_{10} v_{s,\texttt{close}}[d]$	0.527	-0.005	0.281	0.010

Table 5.4: Result of forecasting log MO daily dollar-volumes for the closing auction across exchanges.

Illustrations of Forecasts

In this section, the methods are graphically illustrated through scatterplots of the forecasted versus empirical response, histograms of the residuals and an autocorrelation chart of the residuals. The autocorrelation charts were produced by calculating the autocorrelation of the Apple stock from NASDAQ as a demonstrative example, with a maximum lag $\delta_{\texttt{max}} = 182$ days.

ARIMA As previously mentioned, ARIMA was used in conjunction with the 6 fixed effects; 5 to denote weekdays, and 1 to denote 3rd Friday of the month.

One important result is not showed in these plots; It turned out that ARIMA (with the 6 fixed effects) without exception only managed to fit ARIMA(0,0,0) models.

Opening Auctions:
Illustration of 1st row in table 5.1

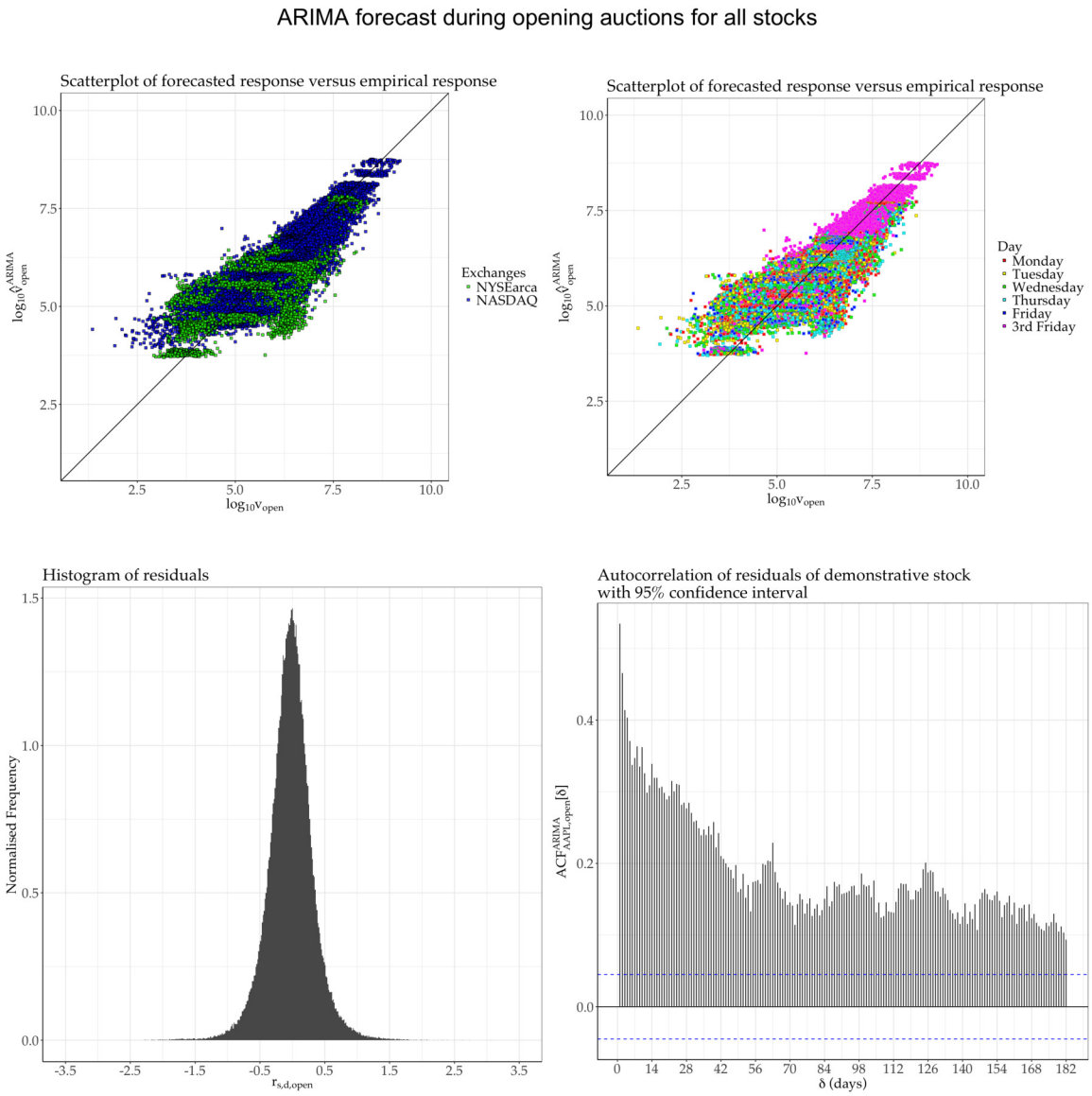

Figure 5.6: Result of ARIMA prediction of the log MO daily dollar-volumes for the opening auctions. d, Weekdays and 3rd Friday were used as features.

Closing Auctions:
Illustration of 1st row in table 5.2

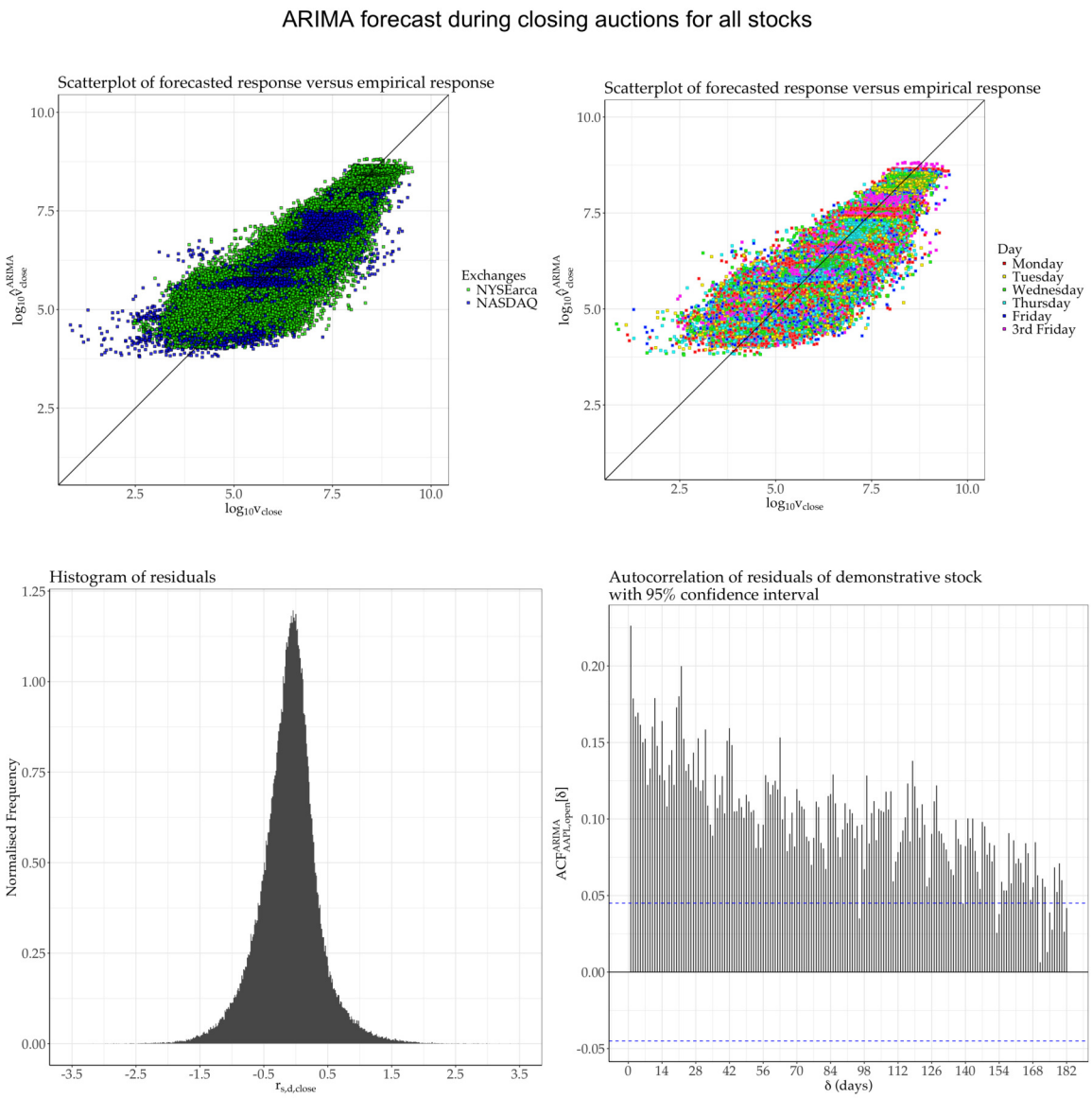

Figure 5.7: Result of ARIMA prediction of the log MO daily dollar-volumes for the closing auctions. d, Weekdays and 3rd Friday were used as features.

Facebook Prophet Prophet was used in conjunction with a fixed effect to denote the 3rd Friday of the month.

Opening Auctions:
Illustration of 2nd row in table 5.1

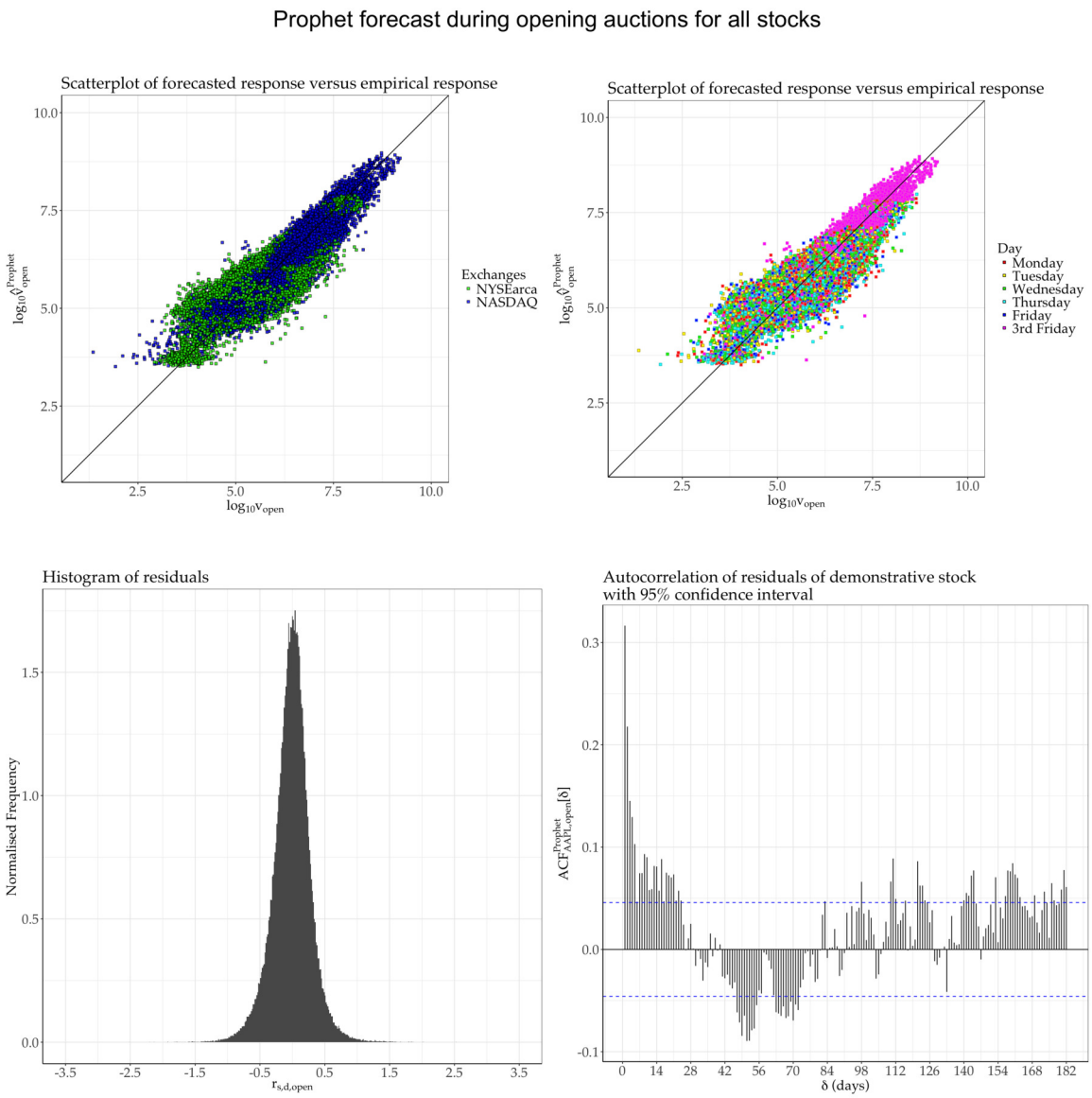

Figure 5.8: Result of Facebook Prophet prediction of the log MO daily dollar-volumes for the opening auctions. d and 3rd Friday were used as features.

Closing Auctions:
Illustration of 2nd row in table 5.2

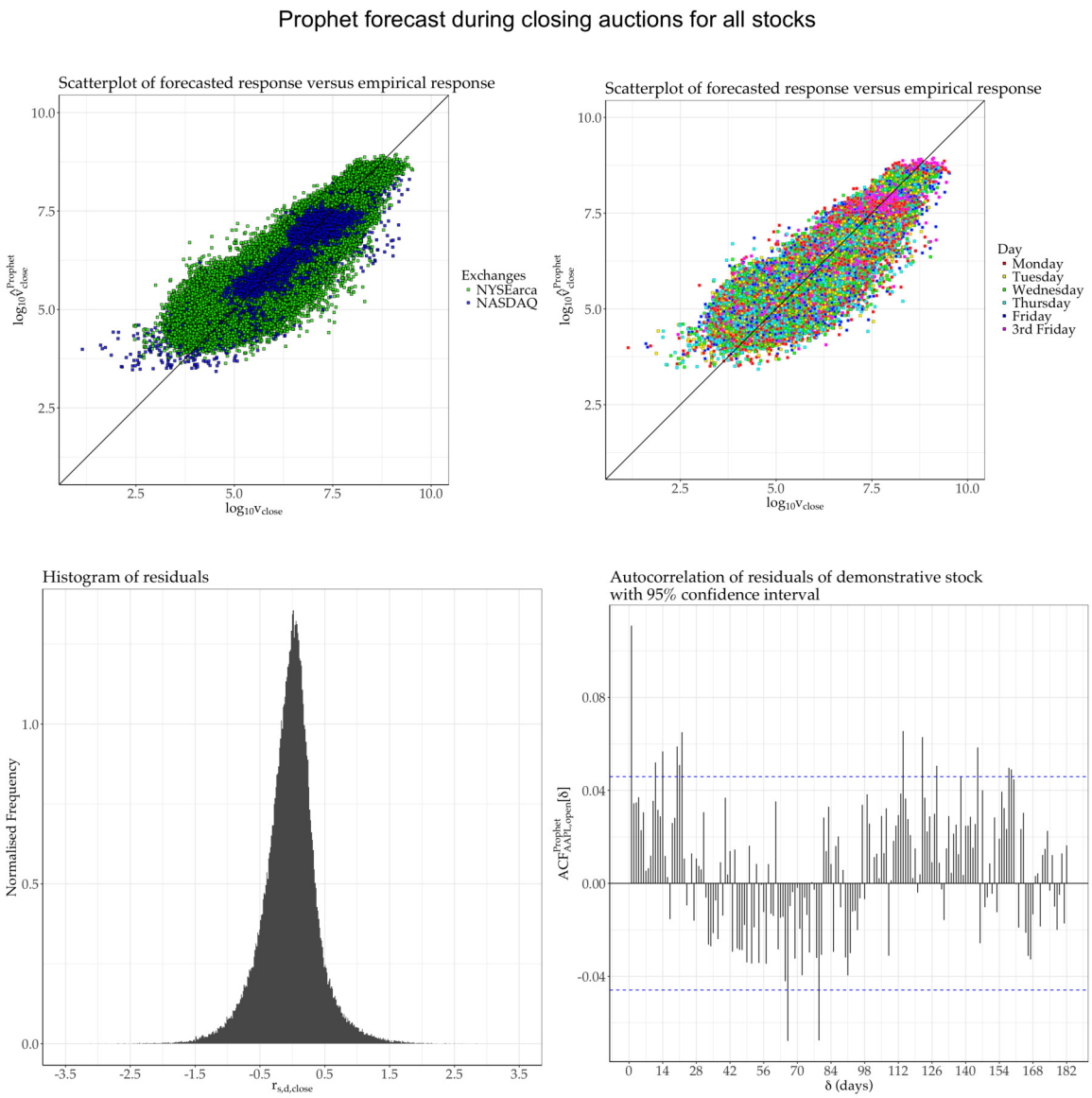

Figure 5.9: Result of Facebook Prophet prediction of the log MO daily dollar-volumes for the closing auctions. d and 3rd Friday were used as features.

Random Forests Various combinations of features were used during the RF forecasting. To illustrate the effect on the forecasting accuracy of adding/removing features, some of the methods are illustrated here.

First, the two methods that achieved the lowest RMSE for each auction are plotted below.

Optimal Features, Opening Auctions:
Illustration of 8th row in table 5.1

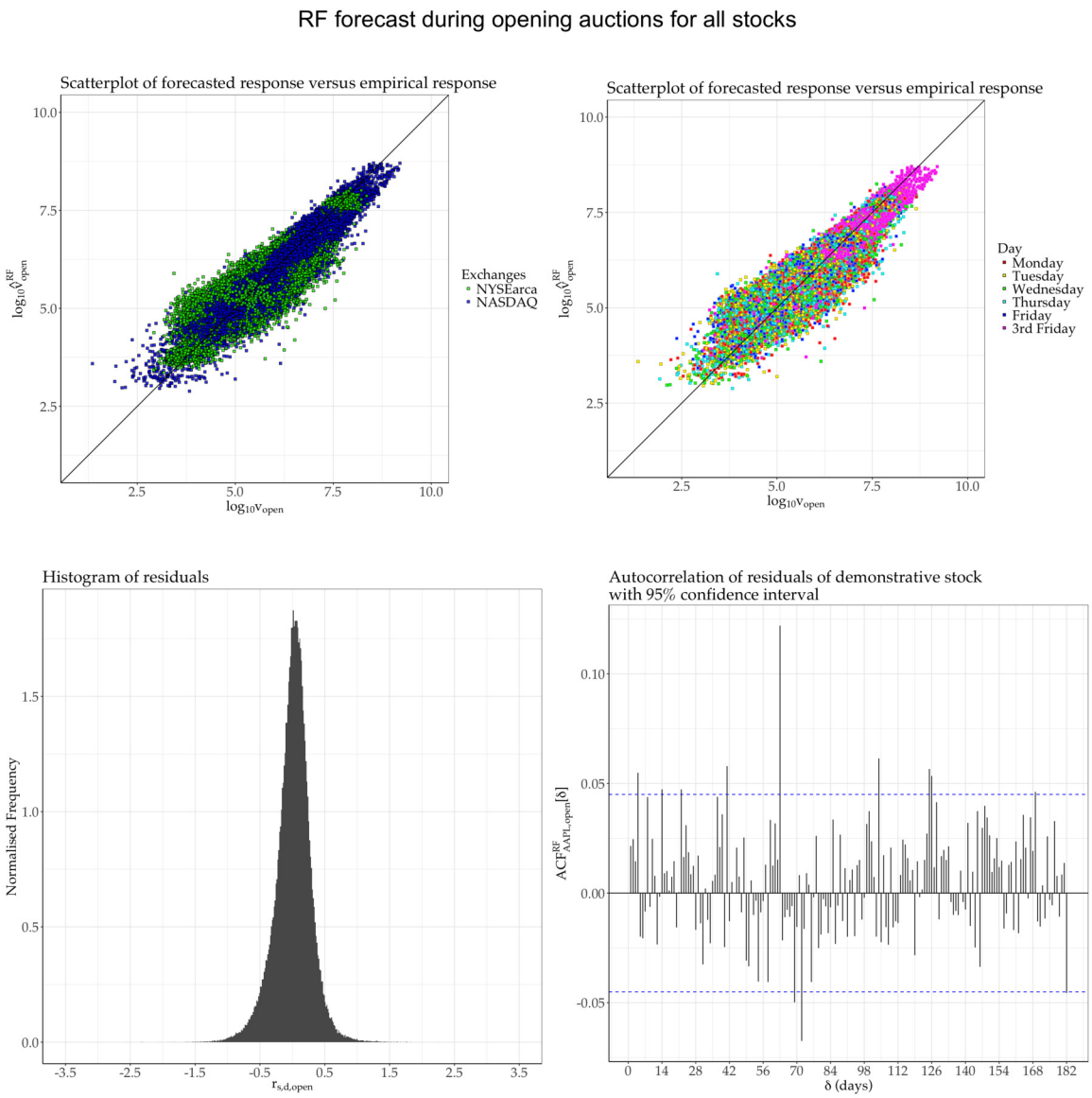

Figure 5.10: Result of Random Forests prediction of the log MO daily dollar-volumes for the opening auctions. d, Weekdays, 3rd Friday, the previous day's response, the previous day's first difference of response and the previous day's open market volume were used as features.

Optimal Features, Closing Auctions:
Illustration of 8th row in table 5.2

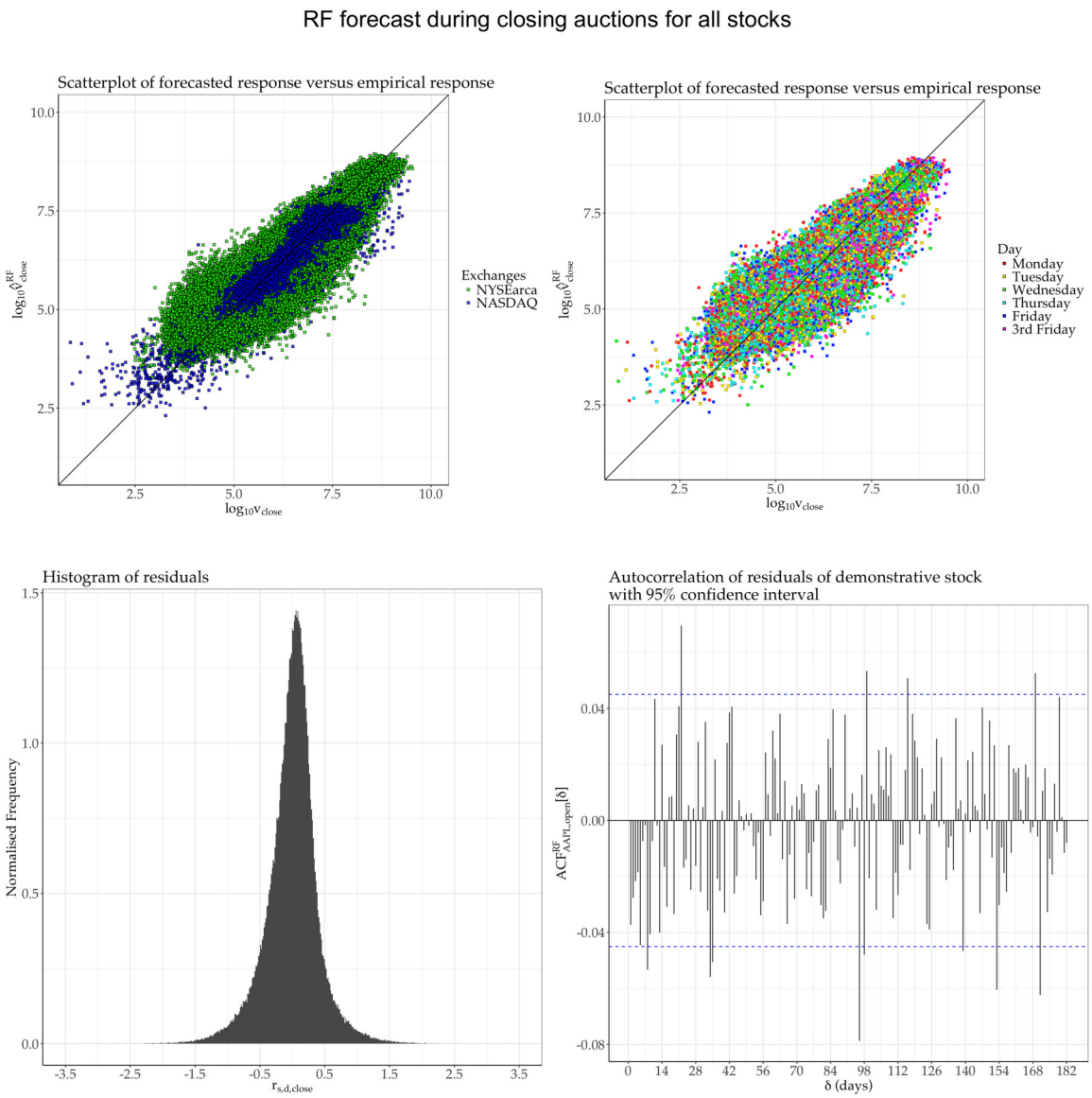

Figure 5.11: Result of Random Forests prediction of the log MO daily dollar-volumes for the closing auctions. d, Weekdays, 3rd Friday, the previous day's response, the previous day's first difference of response and the current day's open market volume were used as features.

It turned out that the autocorrelation between the residuals could be step-wise reduced, and the forecasting accuracy increased, by adding d and then the previous day's response combined with the 1st difference of the previous day's response as features. This is illustrated by plotting the effects on the forecasting accuracy by adding each of these features for the opening auctions.

Only Weekdays & 3rd Friday as Features:
Illustration of 5th row in table 5.1

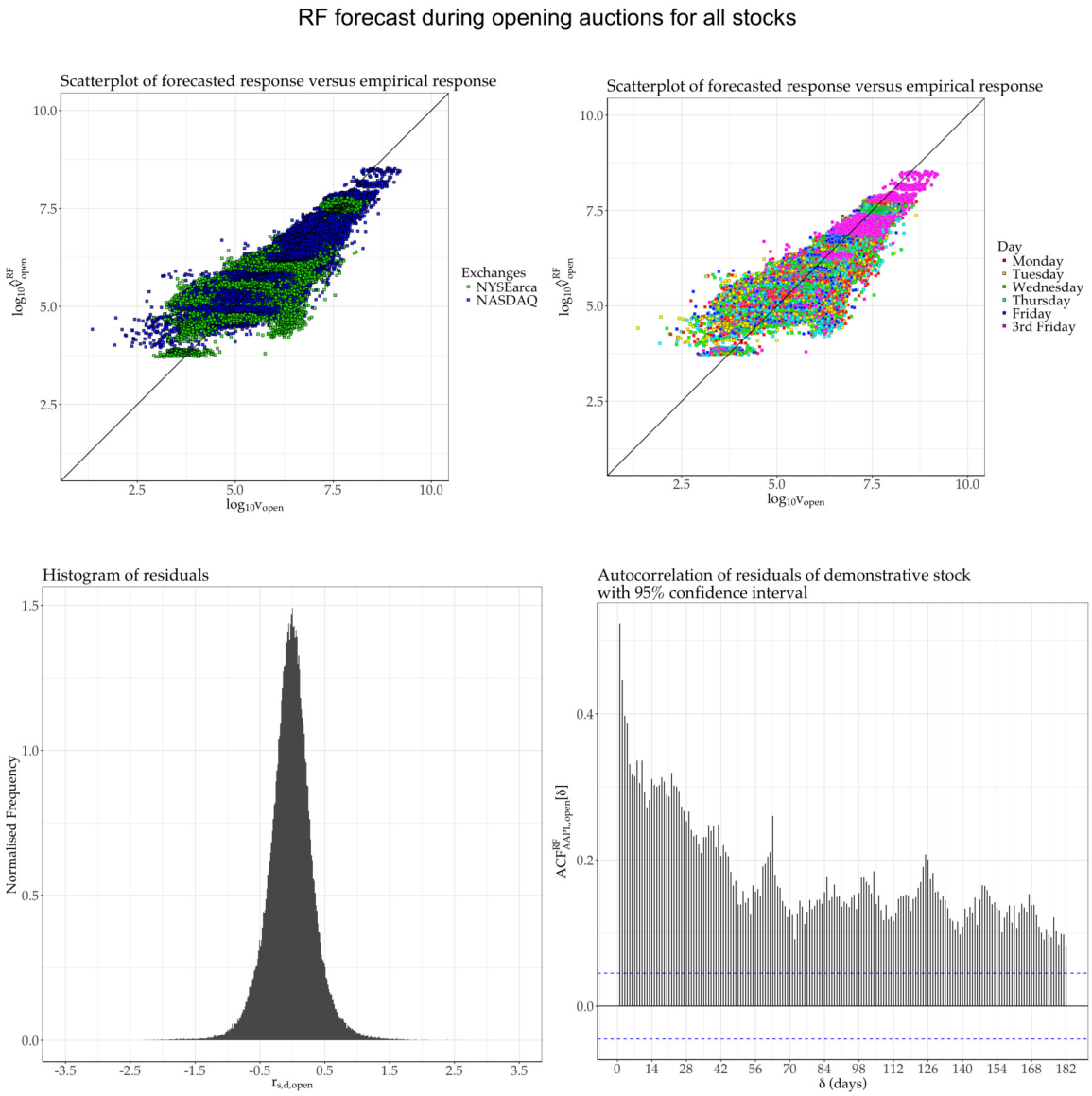

Figure 5.12: Result of Random Forests prediction of the log MO daily dollar-volumes for the opening auctions. Weekdays & 3rd Friday were used as features.

Effect of adding d to the Features:
Illustration of 3rd row in table 5.1

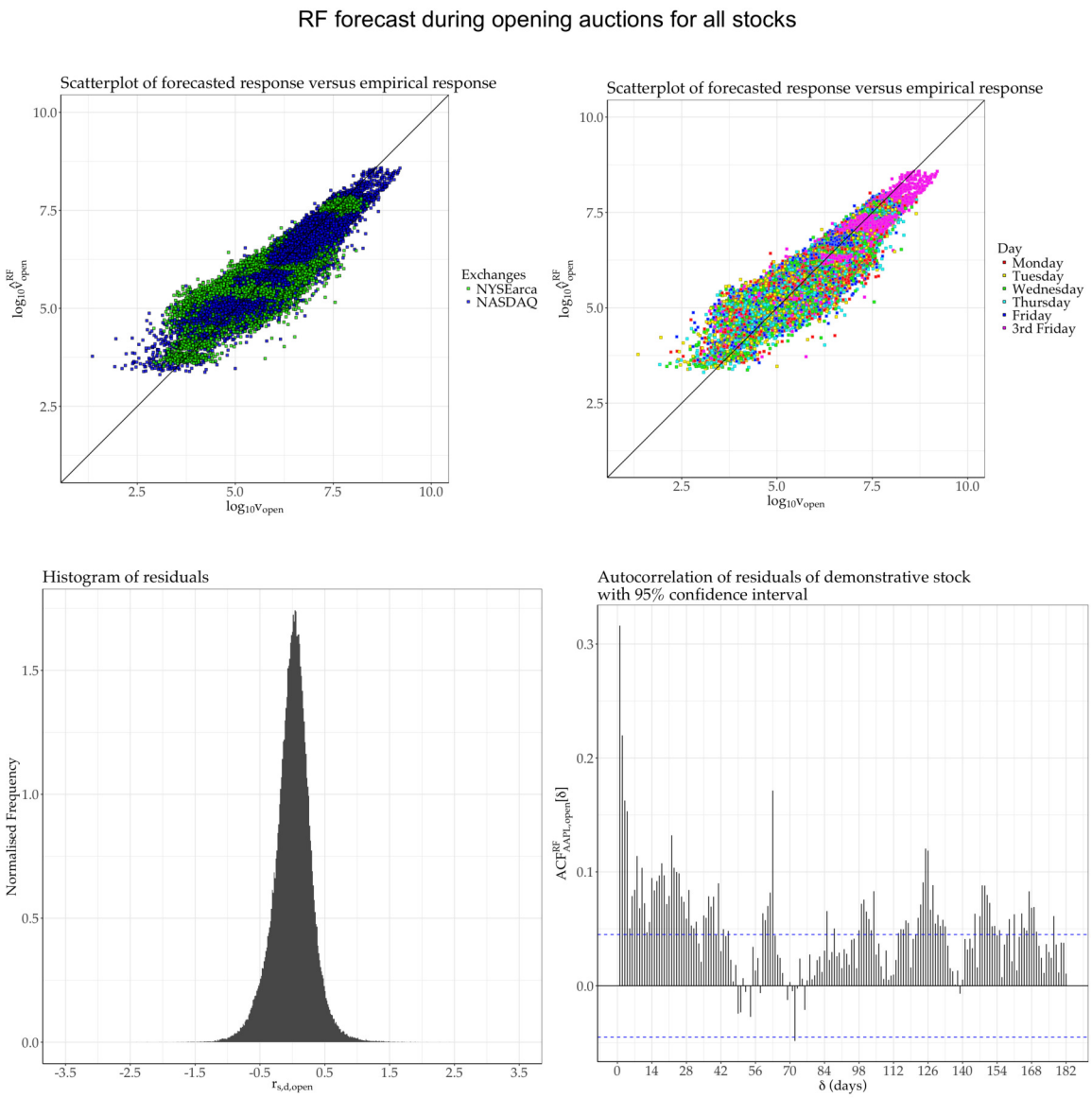

Figure 5.13: Result of Random Forests prediction of the log MO daily dollar-volumes for the opening auctions. d, Weekdays, 3rd Friday were used as features.

Effect of adding d, previous day's response & 1st difference of the previous day's response to the Features:
Illustration of 7th row in table 5.1

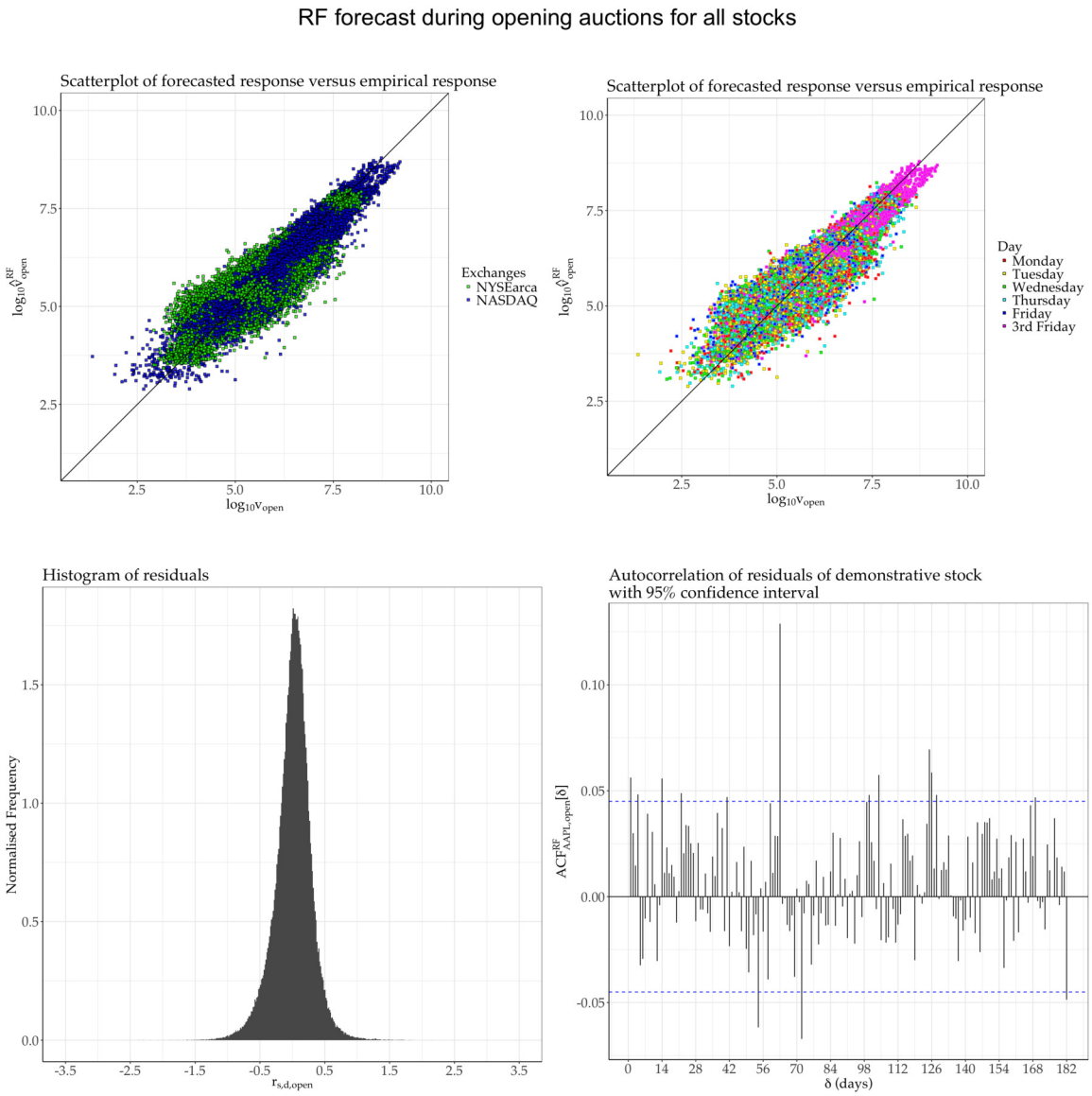

Figure 5.14: Result of Random Forests prediction of the log MO daily dollar-volumes for the closing auctions. d, Weekdays, 3rd Friday, the previous day's response & the previous day's first difference of response were used as features.

5.4.6 Discussion of Forecasting Accuracy

ARIMA

It turned out that after adjusting the data for the weekdays and 3rd Friday fixed effects, ARIMA was unable to model the time-series as anything other than ARIMA(0,0,0), which is a white noise process. Therefore, it must be concluded that the dynamics of $\log_{10} v_{s,a}[d]$ are not generated by any of the family of stochastic processes ARIMA is able to identify.

In practice, ARIMA(0,0,0) along with the 6 fixed effects works as a rolling conditional average in each fixed factor (as described in theory section 3.5.2). In other words, the prediction for e.g. Monday is the rolling average of the previous Mon-

day responses from the in-sample interval, for Tuesday it is the rolling average of the previous Tuesday responses from the in-sample interval, and so on. This is a naive method that only captures the variation in the mean of the time-series, and thus it had, unsurprisingly, a relatively high RMSE and significant autocorrelation of residuals, as can be seen from all the tables and figures 5.6 and 5.7.

Facebook Prophet

While the RMSE of Prophet was lower than ARIMA, it only managed to find models for $\approx \frac{3}{4}$ of the datapoints that ARIMA and Random Forests did. In addition, it turned out to be relatively slow compared to the other methods. Prophet did however capture significant amounts of the variations in the datapoints it managed to forecast, judging by its relatively low RMSE and the small autocorrelation of its residuals.

Random Forests

As can be seen from the results, Random Forests turned out to by far be the superiour method of the ones used. The optimal Random Forests feature combination was found by incorporating an exogenous feature, the prior open-market dollar-volume, though Random Forests also beat the other methods on "fair terms", i.e. without needing to use exogenous information; this can be seen from the surprisingly good forecasts in line 6 and even 4 of tables 5.1 & 5.2.

The Importance of Using d Figures 5.12, 5.13 and 5.14 show that adding d and the previous day's response combined with its 1st difference, strongly reduces the autocorrelation of the residuals. Tables 5.1 and 5.2 also show that the forecasting accuracy is particularly improved by adding d. This was at first surprising to us, as d is a strictly increasing variable and it was at believed that the algorithm was unable to handle it, because during the prediction phase (line 5 in algorithm 1) all the decision trees would split every time in the "larger than" direction when they checked d against the previous days, hence discarding half the tree every time.

However, while this is true, d is in fact very important for two reasons. Firstly, it is extremely important during the segmentation phase of the data, i.e. when the trees are being created (line 4 in algorithm 1), as segmenting along the d feature-dimension allows the model to capture the the fact that the variance and means of the responses vary by time; so in short, it improves the segmentation of the data and thus quality of the trees. The second reason is that consistently splitting the trees in the "larger than" direction is in fact a good thing, as it ensures that each tree uses the newest, and hence most relevant, data segments it has for forecasting; in other words, splitting forward means splitting in the correct direction.

Comparing the Methods against Each Other

Prophet versus the Rest As previously mentioned, Prophet only managed to forecast $\approx \frac{3}{4}$ of the datapoints that ARIMA and Random Forests did, and in addition it was beat by most of the Random Forests methods. Thus it is concluded that Prophet performed comparably poorly at this task.

ARIMA and Random Forests Almost all Random Forests produced significantly better forecasts than ARIMA, with the exception of one; the Random Forests method of row 5 in tables 5.1 & 5.2 produced forecasts that had practically identical accuracy to that of ARIMA (whose results can be seen in the 1st line of the same tables). The reason for this is that the Random Forests method of interest only used weekdays and 3rd Friday as features, leading to it segmenting the data solely by weekday and 3rd Friday, and hence to the averaging of its trees simply being a conditional average by weekday and 3rd Friday; which is equivalent to the rolling conditional average model ARIMA was able to find. This thus serves as a good reality check that everything is working as it is supposed to.

5.4.7 Discussion of Results

Predictability of $v_{s,d,a}$

The results show that $v_{s,d,a}$ is predictable[2] up to factors of $10^{0.217} \approx 1.6$ and $10^{0.266} \approx 1.8$ for the opening and closing auctions of NASDAQ, and $10^{0.323} \approx 2.1$ and $10^{0.517} \approx 3.3$ for the opening and closing auctions of NYSEarca. This means opening auctions are more predictable than closing auctions, that the difference in predictability between open and closing auctions is larger for NYSEarca than for NASDAQ, and that NASDAQ is in general more predictable than NYSEarca. That said, these errors are far, far smaller than what the naive method gives, and so one can safely conclude that the daily MO dollar-volume can be readily forecasted. Further, it must be noted that some stocks are more predictable than others, so the errors given here can be considered as conservative estimates.

The fact that NYSEarca's closing auctions are more unpredictable is unsurprising in the light of the discussion in section 4.2.7. There it was established that the tails of the MO dollar-volume distributions for NYSEarca are systematically heavier for closing auctions than opening auctions; and heavier tails means more extreme behaviour which again means more unpredictability and this is precisely what is observed.

Correlations between the Volumes

Comparing the RMSE of the different Random Forests methods also brings perspective on the interrelations between the open and close MO dollar-volumes, as well their relation to the open-market dollar-volume. Adding the current day's open-market volumes to the features significantly improves the prediction for the closing volume, as can be seen in rows 8 and 9 of table 5.2; this is unsurprising, as the open-market trading volume carries fresh information. Further, based on rows 8 and 9 in table 5.1, information on the previous day's open-market volume seem to be more valuable than that of the previous day's closing volume. These two results imply that the open/close auction volumes are more closely related to the open-market volumes than they are to each other.

[2]To a physicist, such error rates might seem extreme. However, it must be kept in mind that due to the heavy-tailed fluctuations present in financial systems, accurate forecasting of financial time-series is *very* difficult.

5.4.8 Further Work

The work done so far is extensive, but it can be taken further.

Improving Forecasting Accuracy

Figures 5.10 & 5.11 represent the optimal forecasts achieved. These figures, along with 5.14, reveal that the residuals no longer exhibit statistically significant autocorrelation; this means the patterns in the time-series have largely been exploited, and so the only viable way to improve the forecasts is by incorporating exogenous information in Random Forests. Here are two ways to do that:

- It has been known for many years [28][29] that trading volume and market price are dependent, so a good place to start is by trying to use the stock price return, or sign of return, and so on, to improve the prediction of the MO dollar-volume. On this note, one can also go the other way and see whether the stocks' price movements are correlated with the movement of their dolllar-volumes.
- Minor technical improvements can always be done. For example, the first order difference of the open-market volume, i.e. $\Delta^1 \log_{10} o_{s,d,a}$, could be incorporated as a feature for Random Forests. Another improvement could be fine-tuning the effect of the futures expiration date. The futures expire on the 3rd Friday only if the 3rd Friday does not fall on a holiday; if it does, the expiration date is pushed to the earliest date before the holiday.
- New forecasting methods could also be employed; such as for example Deep Learning, which has seen extensive publicity as of late [30].

Miscellaneous

It was mentioned in section 5.4.3 that the Breiman-Cutler importance measure was used to aid in the selection of features for Random Forests. One could also use it as a way to estimate the dependence of the daily MO-dollar volumes with the features used to predict them (e.g. weekdays, 3rd Friday of the month, open-market volume, etc.).

As noted, NASDAQ was remarkably more predictable than NYSEarca; this could be due to differences in the types of stocks they trade, as the bulk of NYSEarca stocks are ETFs while the bulk of NASDAQ stocks are traditional corporate stocks, and/or it could also be due to different auction mechanisms. There is great potential for further investigation here.

Finally, one can always investigate the predictability for individual stocks, or classes of stocks. For example, one could look at the predictability of stocks with very large market capitalisations compared to the smaller ones, or try to understand why some stocks are more predictable than others.

5.4.9 Data Quality

It must be kept in mind that these results are only valid for the stocks traded on NYSEarca & NASDAQ, as none of the stocks in $\mathbb{S}_{\texttt{NYSE}}$ contained enough data for

forecasting. Other than that, the data covered all the largest (in terms of market capitalisation) stocks on NASDAQ and NYSEarca over half a decade of time, and without excessive amounts of gaps in the data; therefore the results are representative for these two exchanges.

Chapter 6

PRE-AUCTION DYNAMICS

The final part of this thesis is an exploration of the pre-auction dynamics of NYSEarca opening auctions. First, statistical regularities in the price-finding and order matching processes were observed and investigated. Afterwards, the dynamics of order-placing was investigated with the aim of understanding the price response of sell and buy orders.

The notation used in this chapter is somewhat different from the rest of the thesis.

- Because the investigation solely concerned opening auctions, the a subscript will from now on be dropped.
- As explained in section 3.1.3, the opening auctions of NYSEarca start at $t_{\texttt{start}} =$ 08:00:00 and end at $t_{\texttt{end}} =$ 09:30:00, ET (Eastern Time). However, out of convenience t will from now on be measured from 08:00, such that $t_{\texttt{start}} = 0$ and either $t_{\texttt{end}} = 90\text{min}$, or $t_{\texttt{end}} = 90 \cdot 60\text{s} = 5400\text{s}$, depending on whether the minute or second is the unit used.
- The matching price $p_{s,d,t}$ will be referred to as the indicative matching price (IMP) if $t < 90\text{min}$ and final matching price (FMP) when $t = 90\text{min}$, as per conventions.
- Analogously to the IMP vs FMP, the daily MO dollar-volume is $v_{s,d,t_{\texttt{end}}}$, and $v_{s,d,t}$ will referred to as the total MO dollar-volume when $t < t_{\texttt{end}}$.
- Out of convenience, the total dollar-volume imbalance (which is defined in the section below) will often be shortened to just imbalance.

6.1 Available Data

We had access to data disseminated by NYSEarca during its pre-auction period of the opening auctions. The data covered 37 stocks over a time-period of 2016-09-29 to 2017-04-07. It consisted of $10,331,607$ (s,d,t) datapoints, and each datapoint contained data on three variables; the total MO dollar-volume (at time t) $v_{s,d}(t)$, matching price $p_{s,d}(t)$ and total dollar-volume imbalance between sell and buy orders $i_{s,d}(t)$. The former two variables are defined in section 3.1.3, and the definition of the total imbalance is [42]:

$$i_{s,d}(t) \equiv (v_{s,d}^{\texttt{buy,L}}(0,t) + v_{s,d}^{\texttt{buy,M}}(t)) - (v_{s,d}^{\texttt{sell,L}}(\infty,t) + v_{s,d}^{\texttt{sell,M}}(t)) \tag{6.1}$$

where $v_{s,d}^{\texttt{buy},\texttt{L}}(0,t)$ is the cumulative dollar-volume of all buy limit orders (meaning every buy order starting at \$0 and upwards is counted) and $v_{s,d}^{\texttt{sell},\texttt{L}}(\infty,t)$ is analogously the cumulative dollar-volume of all sell limit orders.

6.2 Regularities

6.2.1 Preliminary Considerations

As mentioned in section 1.2, the outcome of an auction (s,d,a) is defined by the FMP and the distribution of the MO dollar-volumes; therefore the investigation revolved around the dynamics of $v_{s,d}(t)$ and $p_{s,d}(t)$. $i_{s,d}(t)$ was also investigated, as one expects the total dollar-volume imbalance to be relevant to the dynamics of price formation. These variables can be considered to be the state-variables of the dynamical, time-evolving systems that the auctions (s,d,t) are.

The aim here was to find regularities that are independent of stock and date. After all, chapter 4 shows that the behaviour of the NYSEarca opening auctions at their conclusion is reasonably stable across stocks and days, and thus it is not far-fetched to imagine that the same is true for the pre-auction dynamics. The search for these general regularities was carried out by normalising the variables, averaging them across (s,d) and then studying the time-evolution of the results. The normalisation is necessary to make the stocks comparable with each other.

The time evolution was studied through 1-minute slices. This technically means the variables' t-domains were discretised with the flooring function $\texttt{floor}(t) = \lfloor t \rfloor : \mathbb{R} \geq 0 \to \mathbb{Z} \geq 0$, and afterwards for each minute $t \in [0, 89\text{min}]$ the latest available datapoint was used to set the values of the variables. For example, if for some auction (s,d) there were two datapoints available at the 3rd minute, at eg times 08:03:29 and 08:03:56, then the "last available one", the one at 08:03:56, would be used to set the values of the variables for $t = 3\text{min}$. Only using the last available datapoint was necessary to ensure the results were not biased towards stocks that had more datapoints/were of higher data-quality

6.2.2 Time-evolution of $p_{s,d}[t]$

To start off, the IMP was discretised $p_{s,d}(t) \to p_{s,d}[t]$, as described above, and then normalised with the FMP $p_{s,d}[t_{\texttt{end}}]$:

$$\tilde{p}_{s,d}[t] \equiv \frac{p_{s,d}[t]}{p_{s,d}[t_{\texttt{end}}]} \tag{6.2}$$

Afterwards, the average price across all (s,d) was calculated for each minute:

$$\bar{p}[t] = \langle \tilde{p}_{s,d}[t] \rangle_{\forall (s,d)} \tag{6.3}$$

and plotted. The result revealed that $\tilde{p}_{s,d}[t]$ was not fluctuating randomly around 1, but apparently growing, as figure 6.1 shows.

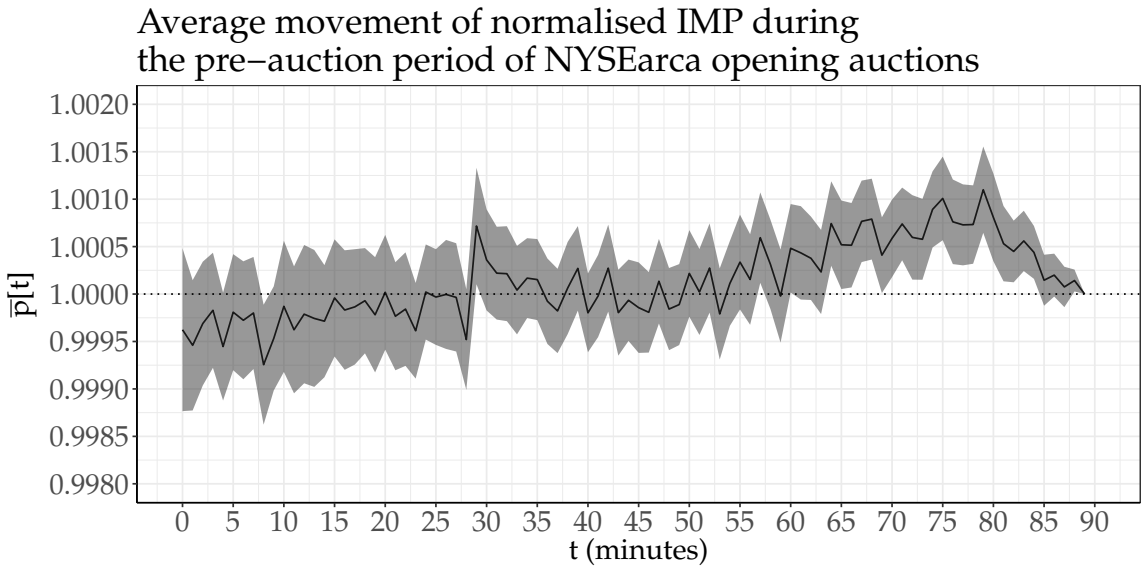

Figure 6.1: The time-evolution of the averaged indicative matching price normalised by the final matching price during the pre-auction period of NYSEarca opening auctions. The grey shade is a two standard error confidence interval for the measured average normalised indicative matching price.

To check whether this growth was statistically significant, a t-test (see section 3.3.3 for theory) was ran on the underlying datasets $\{\tilde{p}_{s,d}[t] \colon \forall(s,d)\}$ used to calculate $\bar{p}][t]$, with the null-hypothesis being that their mean followed Student's t-distribution and was equal to 1. The result is plotted in figure 6.2 below.

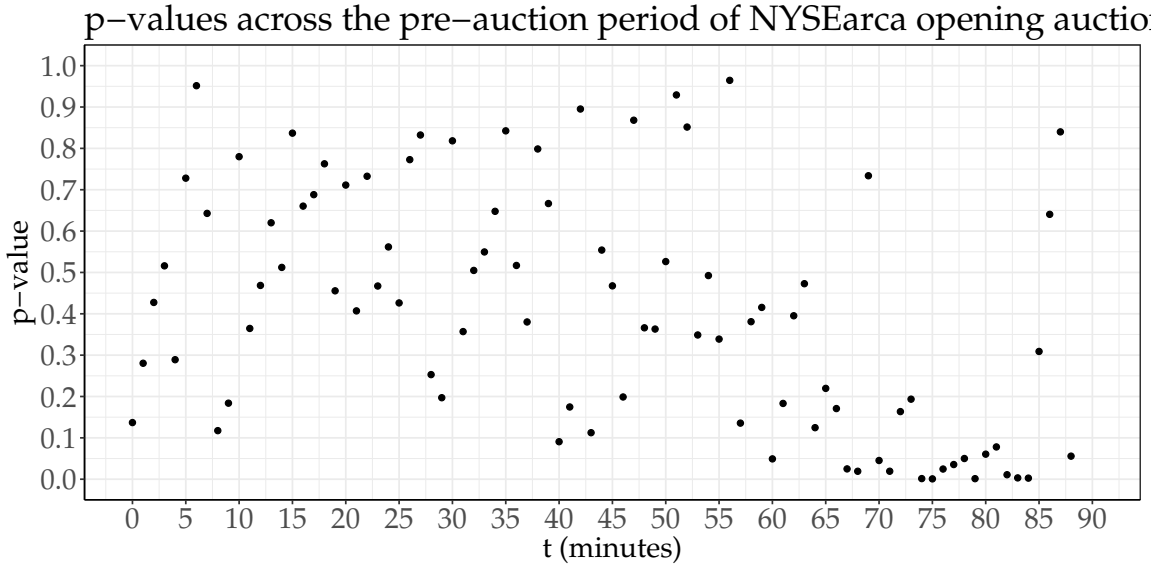

Figure 6.2: P-values from the t-test that checked whether the sample mean of the normalised indicative matching price as a function of time significantly deviates from 1.

The t-test implies that between 65th and the 85th minutes the IMP is roughly 0.05-0.1% higher than the FMP, but it does not support the notion that the IMP is growing

throughout the auction.

The t-test is dependent on the sample mean of its observables converging to the normal distribution, which will happen if the observables are i.i.d. In this specific case, there is no reason to believe that $\tilde{p}_{s,d}[t]$ significantly deviates from i.i.d.; while the traders are heterogeneous between stocks and days, the auction rules are common for all stocks and therefore one can expect $\tilde{p}_{s,d}[t]$ to be similarly distributed for each t. Further, every $\tilde{p}_{s,d}[t]$ came from a different auction (s, d) and thus their interdependence under regular market conditions is certainly limited, if it at all exists.

6.2.3 Time-evolution of $v_{s,d}[t]$

The total MO dollar-volume was studied in an analogous way. It was first normalised:

$$\tilde{v}_{s,d}[t] \equiv \frac{v_{s,d}[t]}{v_{s,d}[t_{\texttt{end}}]} \tag{6.4}$$

and then averaged across all (s, d) for each minute:

$$\bar{v}[t] = \langle \tilde{v}_{s,d}[t] \rangle_{\forall (s,d)} \tag{6.5}$$

Plotting $\bar{v}[t]$ by t showed, as can be seen in figure 6.3, that it was increasing almost linearly. "Almost", because it appears to exhibit slight acceleration during the final ≈ 15 minutes of the auctions.

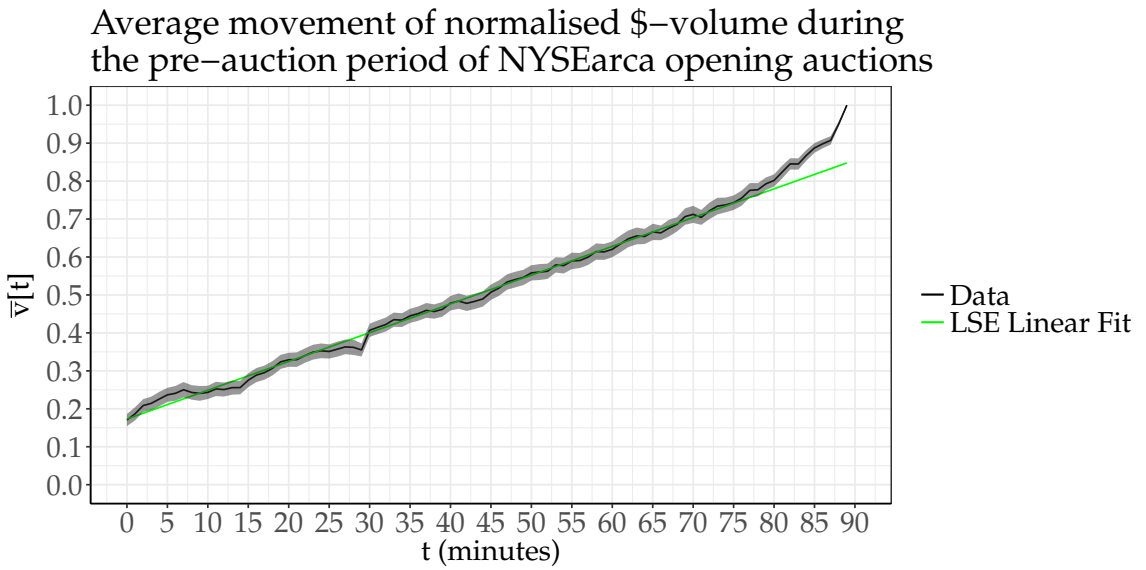

Figure 6.3: The time-evolution of the total MO dollar-volume at time t normalised by the daily MO dollar-volume, along with a least square errors linear fit to the first 80 minutes of the data. The grey shade is two standard errors of confidence interval for the measured average normalised MO dollar-volume.

The line was generated by fitting the first 80 minutes of the data with a least square errors linear model, and it is intended to illustrate the upwards curvature of $\bar{v}[t]$ during the last minutes of the auction.

6.2.4 Time-evolution of $i_{s,d}[t]$

No clear pattern in the time-evolution of the total imbalance through the 1-minute slices was found, and no correlation between the imbalance and the other two state variables in the 1-minute slices was identified either.

However, it should be noted that when the temporal resolution is increased from 1-minute slices to the maximum that the available data allows, one can make educated guesses as to whether a new *in the money*[1] buy or sell order was entered, or cancelled, between two datapoints. Using equations (6.1) along with the pair of equations (3.1), one must conclude the following:

$$\begin{aligned}
&\Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] > 0 \Rightarrow \text{Likely Sell Order at } t \\
&\Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] > 0 \Rightarrow \text{Likely Buy Order at } t \\
&\Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] < 0 \Rightarrow \text{Likely Cancelled Buy Order at } t \\
&\Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] < 0 \Rightarrow \text{Likely Cancelled Sell Order at } t \\
&\Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] = |\Delta i_{s,d}[t]| \Rightarrow \text{Likely Sell Market Order at } t \\
&\Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] = |\Delta i_{s,d}[t]| \Rightarrow \text{Likely Buy Market Order at } t
\end{aligned} \tag{6.6}$$

and so on, with Δ being the difference operator acting in time. These results will be put to use in section 6.3.

6.2.5 Discussion of Observations

Observed Regularities

The results imply that the time-dynamics of $v_{s,d}(t)$ and $p_{s,d}(t)$ contain regularities in their behaviour during the pre-auction periods of the opening auctions of NYSEarca. $\bar{v}_{s,d}[t]$ exhibited linear growth through time until the final few minutes, and the IMP, according to the t-test from figure 6.2, showed largely statistically insignificant price-fluctuations around the FMP (with the exception of when $t \in [65, 85]$min). These regularities imply that the time-evolution of the auctions exhibit a degree of stability and predictability.

The total imbalance, on the other hand, did not exhibit any such clear regularities. This is, however, not unexpected, because while the matching price and total MO dollar-volumes are regulated by auction mechanisms, which are represented through the pair of equations (3.1), the imbalance is just a reflection of whatever orders the traders have put in. From another perspective; it is much more difficult for a trader to move the the matching price or MO dollar-volume than it is for her to move the imbalance[2], and so one should expect that the two former variables to exhibit more stability than the imbalance. Further, it was mentioned that no strong correlation was found between the imbalance and the IMP and total MO dollar-volume; this is not surprising for the same reasons, as the total imbalance is a global variable that can contain LOs far outside the IMP.

[1] In the money means the orders were entered at a price that allowed the orders to be matched.

[2] She just needs to send in a large limit order at an outrageous price, which would have a large effect on the imbalance but no effect on the price or the MO dollar-volume.

Final Minutes of the Auctions

Around the final minutes, $\bar{p}[t]$ started exhibiting interesting behaviour. As seen in figure 6.2, $\bar{p}[t]$ showed statistically significant convergence towards the FMP at exactly the 85th minute. However, while statistically significant, it should still be noted that the average price deviation was quite small, only 0.05-0.1%. The underlying cause of this behaviour could be the fact that NASDAQ starts disseminating price information at exactly $t = 85$, i.e. 9:25am [26]. Because ETFs are the main asset class listed on NYSEarca and many of these funds derive their market-values from the prices of stocks traded at exchanges such as NASDAQ, then naturally the new information on the prices of the stocks underlying the ETFs will cause movement in the price and updates in the trading positions of these ETFs.

Further, $\bar{v}[t]$ was observed to accelerate during the roughly 15 last minutes of the auctions; the explanation might be human behaviour in the face of deadlines. This phenomena is reminiscent of how conference participants behave as registration and fee payment deadlines approach, as observed by V. Alfi, A. Gabrielli & L. Pietronero [27]. In their paper, it was hypothesised that scientists delay committing to a conference to avoid having to cancel their visit and lose the money that paid for the registration fee. One could make a similar hypothesis for our system; a trader has a clear disincentive to commit to a trade too early because that would reveal his intentions to competing traders, and so it is reasonable for him to delay it for as long as possible; this latter point would explain the observed acceleration in the MO dollar-volume.

6.2.6 Further Work

In this section general regularities were studied by normalising and then averaging. The issue with this approach is that while it finds general behaviour that applies across stocks and days, significant amounts of information and subtleties are lost when averaging. Thus, this work can be continued by seeing whether the regularities in $\bar{v}_{s,d}[t]$ and $\bar{p}_{s,d}[t]$ express themselves differently for individual stocks and during different time periods.

Further, it must be mentioned that the regularity of $\bar{v}_{s,d}[t]$ can be used to improve the forecasts of the daily MO dollar-volumes done in the previous chapter. Namely, one can add the total MO dollar-volume matched at some specified t into the auction as a feature to Random Forests for predicting the final daily MO dollar-volume of the auction.

6.2.7 Data Limitations

It must be noted that the regularities were only present in the available sample. As described in section 6.1, the sample consisted of only 37 stocks across a time-period lasting little than more than half a year, 2016-09-29 to 2017-04-07, which is far too little given the variability of market conditions through time. Therefore, while the results of this section are interesting, they cannot be considered declarative until reproduced in a broader range of data.

6.3 Price Response

The simplest way [36] to study market impact (which is the price shift a new order causes) is by measuring a stock's price change in response to a new order. Thus as a first step towards understanding the dynamics of order placing, the price response due to a new order was studied. This was done by defining the concept of price response, and then using different price response functions to measure the effect different types of orders have on the price. First the IMP was measured, and then the much more important FMP.

6.3.1 Measuring Price Response

A price response function measures the expected fluctuations in stock price conditioned on the sign of a trade (i.e. whether it was a buy or sell order) carried out as a function of time τ after the trade was entered. The response functions used in this investigation can be segmented into two types: unsigned response functions $R[\tau]$, and signed response functions $R^{\eta}[\tau]$. The concept behind both of them is the same, and in form they are almost identical; the only difference between them is that the signed variant adds the addition condition of whether the order contributes to the imbalance or counteracts it.

Unsigned Response Function

In the case of our auctions, the unsigned response can be mathematically formulated as:

$$R[\tau] \equiv \left\langle \epsilon(s,d,t) \frac{p_{s,d}[t+\tau] - p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall (s,d,t):\, \epsilon(s,d,t) \neq 0} \tag{6.7}$$

here t is the timestamp of the order and $\tau \geq 0$ is the time after an order was entered. Further, $\epsilon[s,d,t]$ is a function that discriminates between new buy and sell orders by returning $+1$,-1 or 0 depending on whether it guesses[3] $[s,d,t]$ contains a new buy order, new sell order or neither. It is based on equation (6.6), and its precise definition follows below:

$$\epsilon(s,d,t) \equiv \begin{cases} +1, \text{ if } \Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] > 0 \\ -1, \text{ if } \Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] > 0 \\ 0, \text{ otherwise} \end{cases} \tag{6.8}$$

where Δ is the difference operator applied on the t parameter.

The response function measures the systematic correlation between order sign and price change. If it is positive, then on average the price largely moves "in the expected direction"; i.e. buy orders push the price up, and/or sell orders push it down. If the response function is largely negative however, it means the price moves in the opposite way one would expect it to; i.e. buy orders in fact push the price down, while sell orders push it up.[4]

[3] The term "guessed" was used because the datapoints were not consistently updated whenever a new order came in, and so it was impossible to classify every datapoint with complete certainty.

[4] As a side-note, it should be mentioned that a price-response function is simply a linear response

Signed Response Function

The signed response $R^{\eta}[\tau]$, with the *label*[5] $\eta \in [-1, +1]$ being its sign, is defined as follows:

$$R^{\eta}[\tau] \equiv \left\langle \zeta(s,d,t,\eta) \frac{p_{s,d}[t+\tau] - p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall (s,d,t):\, \zeta(s,d,t,\eta) \neq 0} \tag{6.9}$$

where the new function $\zeta(s,d,t,\eta)$ is similar to $\epsilon(s,d,t)$ except it also discriminates between the orders that increase the imbalance (i.e. $i_{s,d}[t]\Delta i_{s,d}[t] > 0$) and those that decrease it ($i_{s,d}[t]\Delta i_{s,d}[t] < 0$) according to η:

$$\zeta(s,d,t,\eta) \equiv \begin{cases} +1, \text{ if } \Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] > 0 \wedge \texttt{sgn}(i_{s,d}[t]\Delta i_{s,d}[t]) = \eta \\ -1, \text{ if } \Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] > 0 \wedge \texttt{sgn}(i_{s,d}[t]\Delta i_{s,d}[t]) = \eta \\ 0, \text{ otherwise} \end{cases} \tag{6.10}$$

with $\texttt{sgn}\colon \mathbb{R} \to [-1, +1]$ being the signum function.

The signed response function measures the systematic correlation between the price change of an order, and the order type and direction; with $\eta = +1$ representing orders that push the imbalance upwards, and $\eta = -1$ those that push it downwards.

6.3.2 Indicative Matching Price Response

The response functions are used here to measure the IMP response. Before continuing, it is worth mentioning that market impact of any new order on the IMP can in principle be predicted with complete certainty from the pair of equations (3.1), but only when the order book is known (which we did not have, hence why we were forced to use the response function).

The effect of an order can rise and fall within seconds (or even less due to high frequency trading), thus 1-second slices were used to calculate the response functions. Further, only the latest available datapoints for each second were used to set the values of $v_{s,d,t}$,$i_{s,d,t}$ and $p_{s,d,t}$ (similarly to what was done in section 6.2) to avoid bias towards the stocks with most data.

Moving on, as written in section 3.1.3, the NYSEarca opening auctions freeze at 09:29, i.e. at the $t = 89 \cdot 60 = 5340$th second. Therefore, it was decided to impose the criterion $t + \tau < 5340$s on the response functions of the IMP, and so the IMP response functions are defined as:

$$\begin{aligned} R^{\texttt{IMP}}[\tau] &\equiv \left\langle \epsilon(s,d,t) \frac{p_{s,d}[t+\tau] - p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall (s,d,t):\, (\epsilon(s,d,t) \neq 0 \wedge t+\tau < 5340)} \\ R^{\eta,\texttt{IMP}}[\tau] &\equiv \left\langle \zeta(s,d,t,\eta) \frac{p_{s,d}[t+\tau] - p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall (s,d,t):\, (\zeta(s,d,t,\eta) \neq 0 \wedge t+\tau < 5340)} \end{aligned} \tag{6.11}$$

function that measures the average change in relative price (often referred to as "return" in finance) resulting from new orders of a certain type.

[5]Please note that η is a label, not an exponent.

with the averaging being done over all available data.

Measured Price Response for all Stocks

Figure 6.4 below illustrate the above response functions in two cases; in the top row, the detailed behaviour of the response functions are plotted for the first 60 seconds in τ. In the bottom row, the response function is plotted over the next 10 minutes; plotting over a longer period was not possible due to the nature of the data. The by-minute plot were created by averaging the response function, whose temporal resolution was seconds, for each minute and thus make the plots more clear.

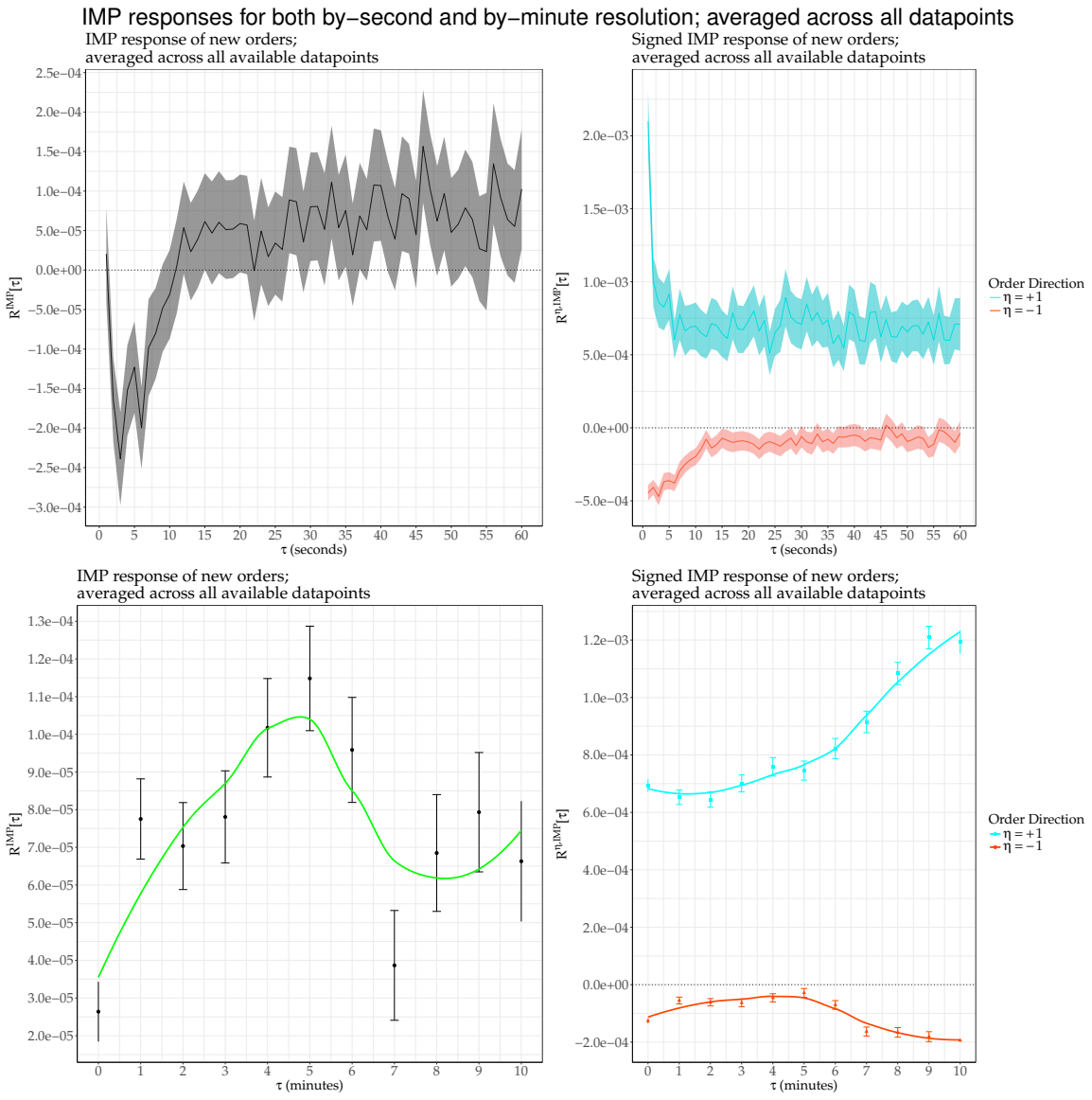

Figure 6.4: The time-evolution of the measured IMP responses averaged across all data. The shades and error bars are two standard errors of confidence interval for the measured responses. The two left figures are of the unsigned responses, while the two right ones are plots of the signed responses. The smooth curves are polynomial fitted to the data for illustrative purposes.

A few observations can be made. Firstly, and most importantly, it is clear that the signed response function segments the data into two groups with significantly different properties; the imbalance-increasing responses are globally positive and of relatively large amplitude, while the imbalance-decreasing ones are globally weak and have a negative price response. Secondly, the upper charts show that the orders immediately incite a strong, transient peak in the IMP response. For imbalance-increasing orders, it lasts for roughly 5s, and 10s for the imbalance-decreasing ones. Further, after the peak has subsided, the amplitudes of $R^{\eta,\texttt{IMP}}$ remain stable for the next 5 or so minutes for both imbalance-increasing and -decreasing orders.

6.3.3 Final Matching Price Response

While the IMP responses are interesting, the FMP response is much more important; after all, it is at the FMP, not IMP, that the MOs are exchanged at.

The FMP response was investigated differently from the IMP:

- No clear patterns were found by averaging across all stocks, so instead the FMP response was measured on a stock-by-stock basis.
- Further, it is also easy to measure the FMP response due to the cancellation of an order; thus the cancellation-responses $C^{\texttt{FMP}}[\tau]$ and $C^{\eta,\texttt{FMP}}[\tau]$ were introduced and studied for the sake of completeness.
- The measurements were done (for aesthetic reasons) with 2-minute slices instead of 1-second slices. Further, because bias towards stocks with the most data was no longer an issue due to the investigation being performed on a stock-by-stock basis, no data was discarded; all available datapoints $v_{s,d,t}$, $i_{s,d,t}$ and $p_{s,d,t}$ were aggregated for each 2nd minute and used in the calculation of the responses.

The FMP can be measured by imposing the criterion $t + \tau = t_{\texttt{end}}$. Doing this, the definitions for the cancellation-responses and the new order responses become:

$$\begin{aligned}
R_s^{\texttt{FMP}}[\tau] &\equiv \left\langle \epsilon(s,d,t)\frac{p_{s,d}[t+\tau]-p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall(d,t):\,(\epsilon(s,d,t)\neq 0 \wedge t+\tau=t_{\texttt{end}})} \\
R_s^{\eta,\texttt{FMP}}[\tau] &\equiv \left\langle \zeta(s,d,t,\eta)\frac{p_{s,d}[t+\tau]-p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall(d,t):\,(\zeta(s,d,t,\eta)\neq 0 \wedge t+\tau=t_{\texttt{end}})} \\
C_s^{\texttt{FMP}}[\tau] &\equiv \left\langle \epsilon^{\texttt{C}}(s,d,t)\frac{p_{s,d}[t+\tau]-p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall(d,t):\,(\epsilon^{\texttt{C}}(s,d,t)\neq 0 \wedge t+\tau=t_{\texttt{end}})} \\
C_s^{\eta,\texttt{FMP}}[\tau] &\equiv \left\langle \zeta^{\texttt{C}}(s,d,t,\eta)\frac{p_{s,d}[t+\tau]-p_{s,d}[t]}{p_{s,d}[t]} \right\rangle_{\forall(d,t):\,(\zeta^{\texttt{C}}(s,d,t,\eta)\neq 0 \wedge t+\tau=t_{\texttt{end}})}
\end{aligned} \tag{6.12}$$

where the averaging is done over all the available data for stock s at time $t_{\texttt{end}}$, i.e. at the end of the auction, and $\epsilon^{\texttt{C}}(s,d,t)$, $\zeta^{\texttt{C}}(s,d,t,\eta)$, analogously to $\epsilon(s,d,t)$ and

$\zeta(s,d,t)$, guess on the cancellation of orders:

$$\epsilon^{\mathtt{C}}(s,d,t) \equiv \begin{cases} +1, \text{ if } \Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] < 0 \\ -1, \text{ if } \Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] < 0 \\ 0, \text{ otherwise} \end{cases}$$
$$\zeta^{\mathtt{C}}(s,d,t,\eta) \equiv \begin{cases} +1, \text{ if } \Delta i_{s,d}[t] > 0 \wedge \Delta v_{s,d}[t] < 0 \wedge \mathtt{sgn}(i_{s,d}[t]\Delta i_{s,d}[t]) = \eta \\ -1, \text{ if } \Delta i_{s,d}[t] < 0 \wedge \Delta v_{s,d}[t] < 0 \wedge \mathtt{sgn}(i_{s,d}[t]\Delta i_{s,d}[t]) = \eta \\ 0, \text{ otherwise} \end{cases} \tag{6.13}$$

Both $\epsilon^{\mathtt{C}}(s,d,t)$ and $\zeta^{\mathtt{C}}(s,d,t,\eta)$ return $+1$ or -1 if they guess, respectively, a sell order or buy order being cancelled, and 0 if they guesses on no cancelled order.

The two cancellation response functions are almost identical to the old response functions, except they measure the effect of cancelling an order instead of the effect of inserting a new one. More precisely, the unsigned cancellation-response function measures the expected price fluctuations conditional on the type of cancelled order, while the signed variant measures the expected price fluctuations conditional on both the type and direction of the order. Just like with the price response, when the cancellation-response is positive, then the price moves in the expected direction (e.g. a cancelled sell order leads to a price increase), and when it is negative the price moves in the opposite direction of what one would expect.

Measured Price Responses for Select Stocks

The results for the 5 stocks with the most datapoints available, i.e. SPY, IWM, GLD, EWZ and VXX, are presented below. Note that the y-axes have different ranges for each figure, and the x-axes is inverted; it was natural for τ to be decreasing along the x-axis, as τ is equivalent to the time until auction execution and so a lower τ implies the order was entered at a later point in time.

The SPY stock represents the SDPR S&P500 ETF.[6]

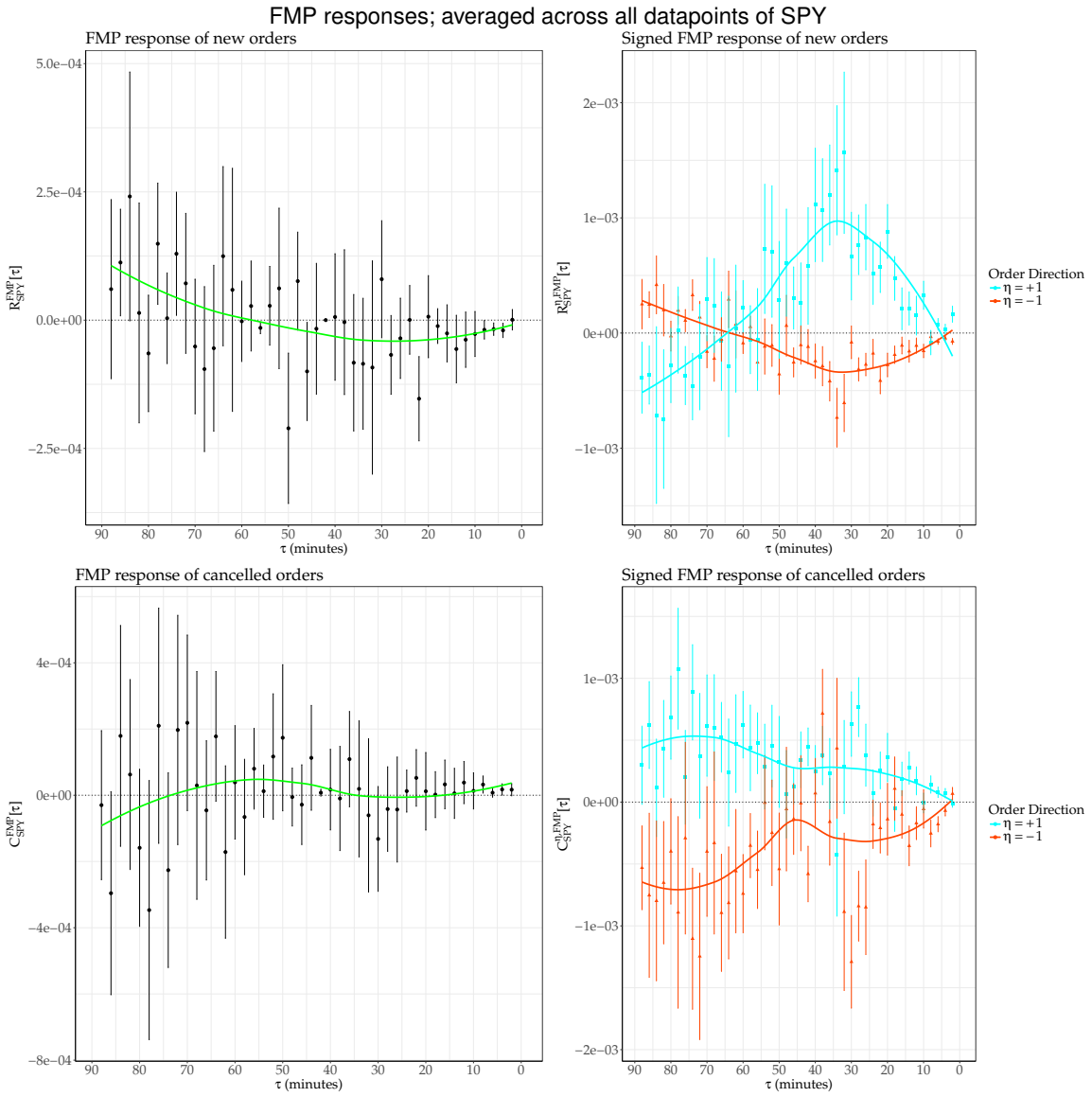

Figure 6.5: The time-evolution of the measured FMP responses averaged across all data available for SPY. The error bars are two standard errors of confidence interval for the measured responses. The top left chart shows the unsigned response, top right shows the signed response, bottom left shows the unsigned cancellation-response and bottom right shows the signed cancellation-response. The smooth curves are polynomials fitted to the data for illustrative purposes.

[6]This ETF tracks the top 500 U.S. registered public companies by market capitalisation.

The IWM stock represents the iShares Russel 2000 ETF.[7]

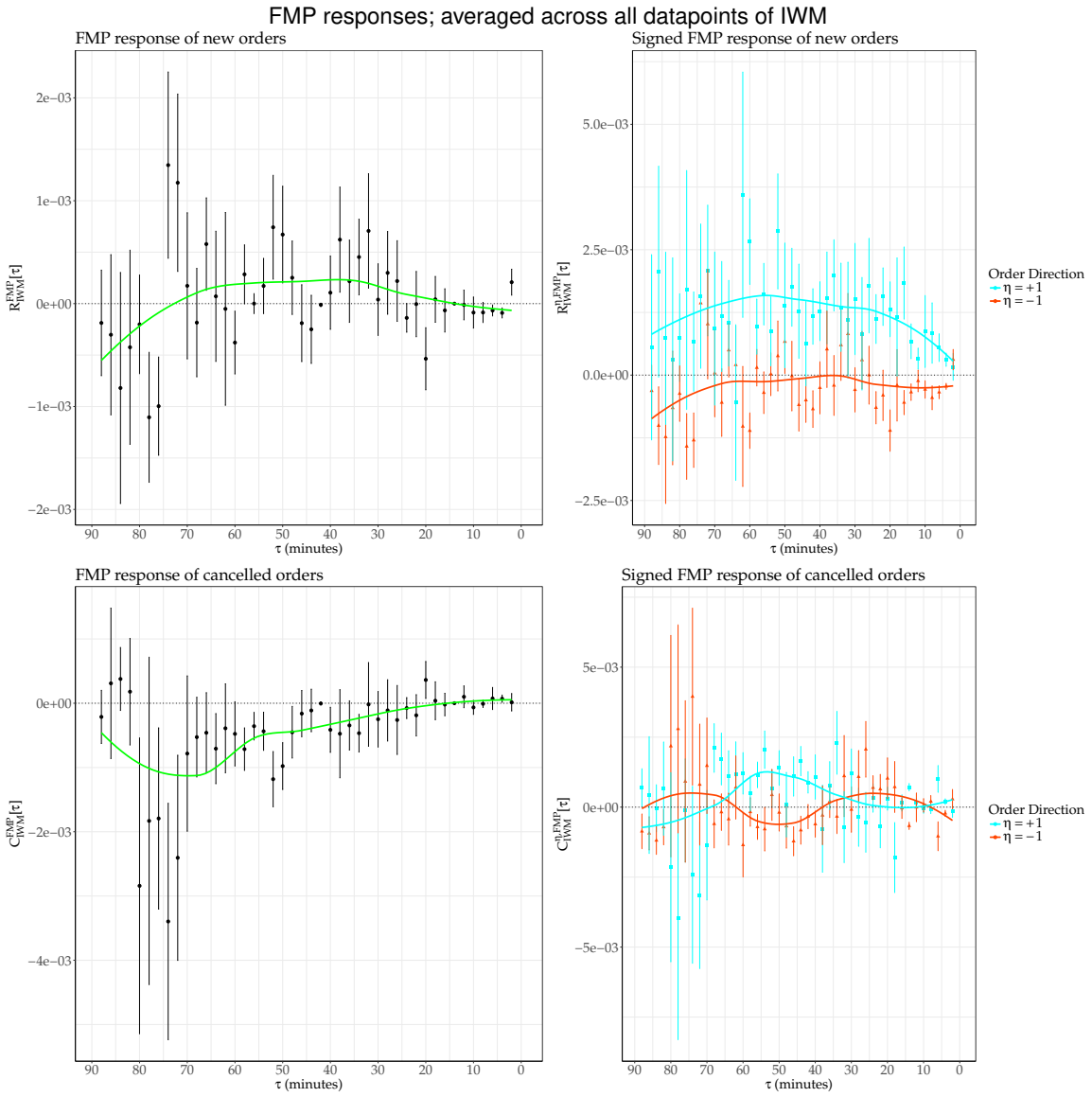

Figure 6.6: The time-evolution of the measured FMP responses averaged across all data available for IWM. The error bars are two standard errors of confidence interval for the measured responses. The top left chart shows the unsigned response, top right shows the signed response, bottom left shows the unsigned cancellation-response and bottom right shows the signed cancellation-response. The smooth curves are polynomials fitted to the data for illustrative purposes.

[7]This ETF tracks various public companies with small market capitalisations.

The GLD stock represents the SPDR Gold Shares ETF.[8]

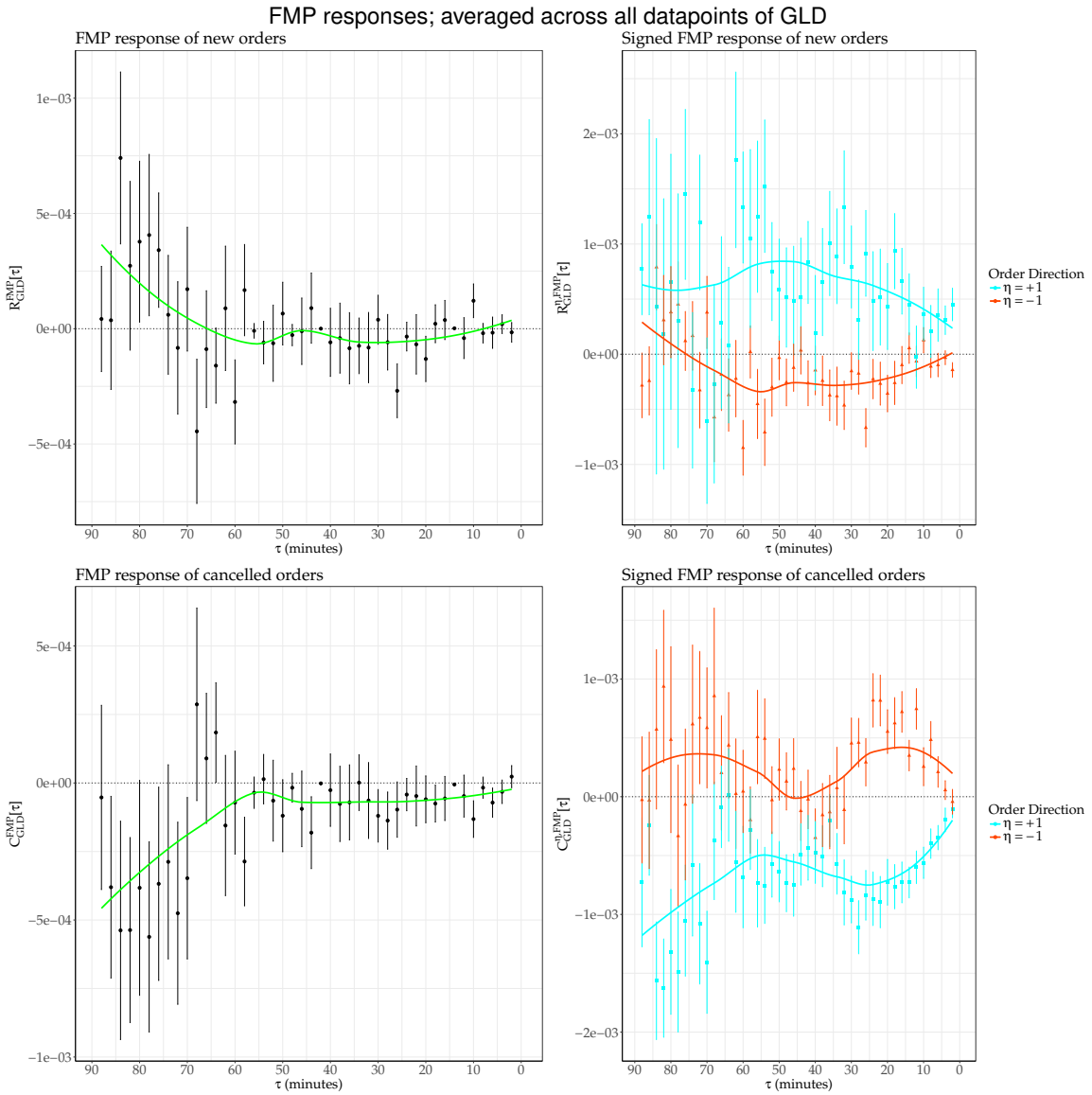

Figure 6.7: The time-evolution of the measured FMP responses averaged across all data available for GLD. The error bars are two standard errors of confidence interval for the measured responses. The top left chart shows the unsigned response, top right shows the signed response, bottom left shows the unsigned cancellation-response and bottom right shows the signed cancellation-response. The smooth curves are polynomials fitted to the data for illustrative purposes.

[8]This ETF denotes a share of gold bullion; meaning its underlying asset is gold.

The EWZ stock represents the IShares MSCI Brazil ETF.[9]

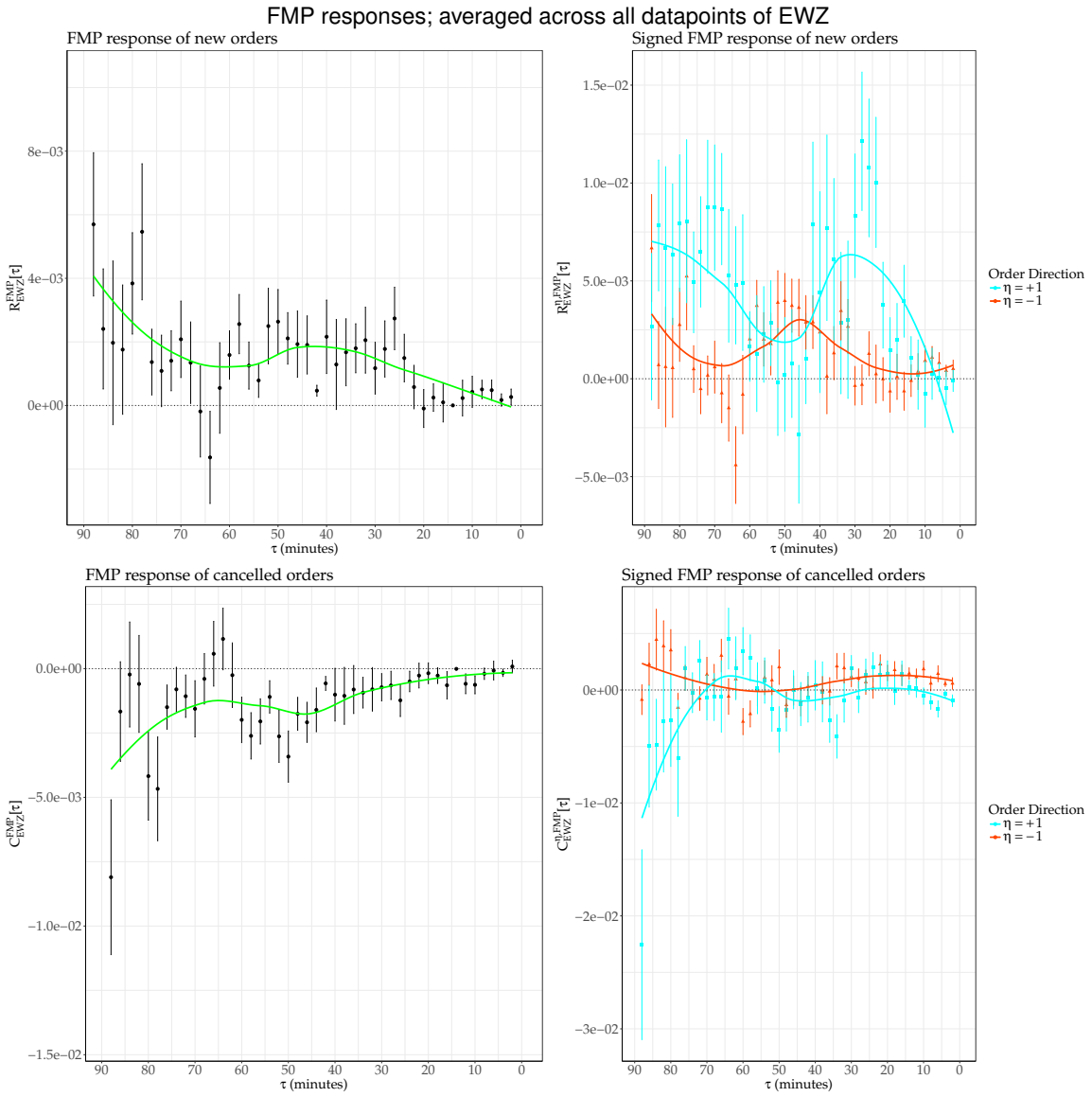

Figure 6.8: The time-evolution of the measured FMP responses averaged across all data available for EWZ. The error bars are two standard errors of confidence interval for the measured responses. The top left chart shows the unsigned response, top right shows the signed response, bottom left shows the unsigned cancellation-response and bottom right shows the signed cancellation-response. The smooth curves are polynomials fitted to the data for illustrative purposes.

[9]As the name suggests, this ETF tracks various Brazilian public companies.

The VXX stock represents the iPath S&P500 VIX ST Futures exchange traded note.[10]

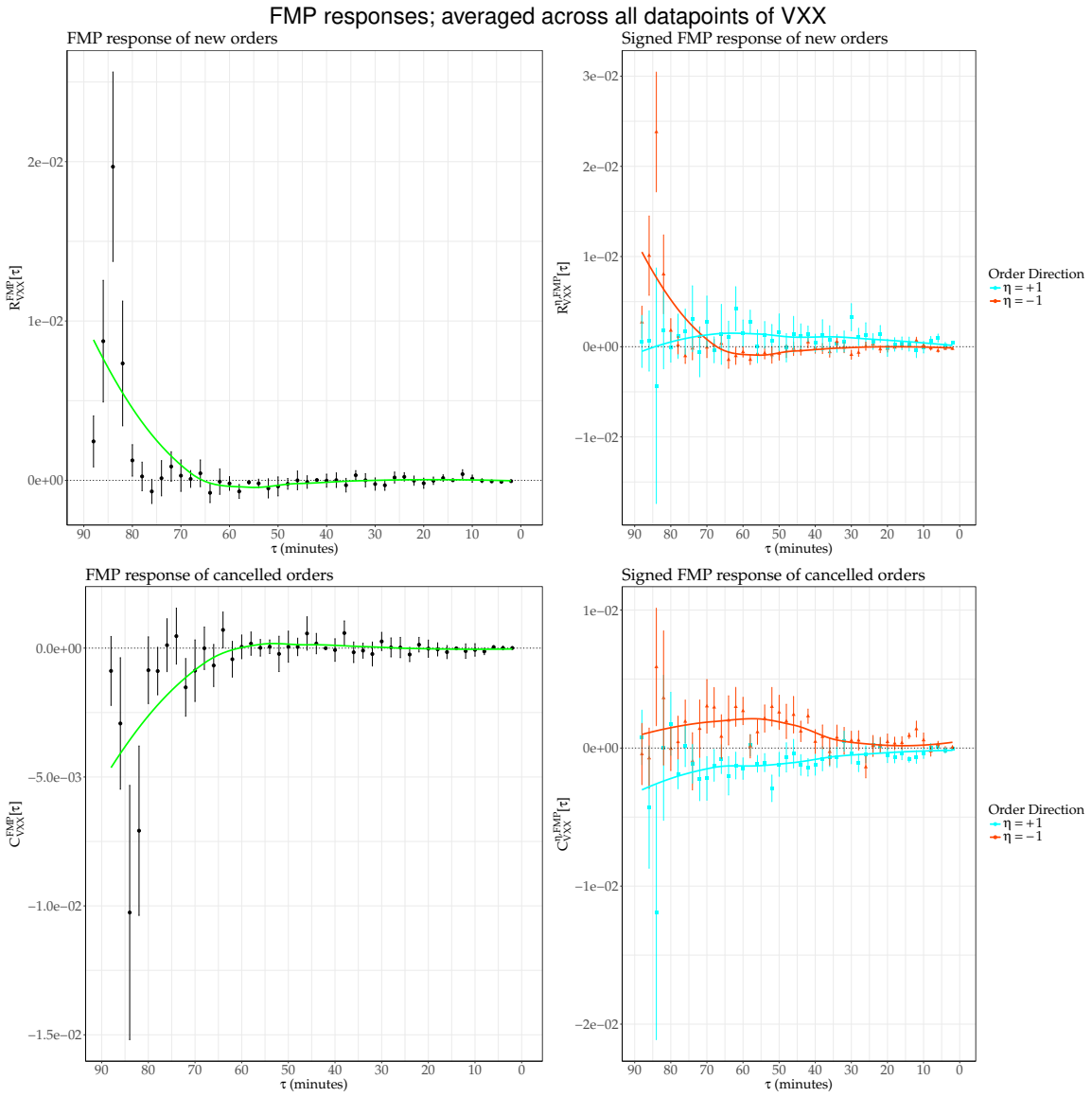

Figure 6.9: The time-evolution of the measured FMP responses averaged across all data available for VXX. The error bars are two standard errors of confidence interval for the measured responses. The top left chart shows the unsigned response, top right shows the signed response, bottom left shows the unsigned cancellation-response and bottom right shows the signed cancellation-response. The smooth curves are polynomials fitted to the data for illustrative purposes.

[10]VXX holds a portfolio of various futures.

The above results are difficult to interpret due to the large amount of noise, but a handful of clear observations can be made. Firstly, the level of noise progressively decreases for all the response measures as the pre-auction period progresses. Secondly, the cancellation response $C_s^{\texttt{FMP}}[\tau]$ typically behaves opposite to what one would expect; judging from $C^{\texttt{FMP}}[\tau]$, removing buy orders lead to higher prices while removing sell orders lead to lower ones. Thirdly, as with the IMP response, $R_s^{+,\texttt{FMP}}[\tau]$ is on average positive and larger in amplitude than $R_s^{-,\texttt{FMP}}[\tau]$. Further, the $\eta = -1$ responses appear to be negatively correlated with the $\eta = +1$ ones. And finally, all measured responses decrease in amplitude as the auction progresses.

6.3.4 Discussion of Results

IMP Response

The investigation into the IMP response revealed a collection of stylised facts.

Firstly, the signed response function clearly segments the data into two groups with significantly different price responses; so orders that add to the present imbalance cause a larger IMP response than the ones that contribute to decreasing the imbalance. This implies that, e.g., in a buy-side imbalance ($i[t] > 0$) sellers can enter sell-orders at high prices and so not cause drops in the IMP; which implies a small price response for their orders. Indeed, if the buy-side imbalance is very heavy, it is possible for them to push the price upwards by matching with buyers who are willing to pay a larger price than the IMP. The same reasoning applies to buyers during a sell-side imbalance.

Secondly, the measured IMP response exhibited a strong, transient response during the first seconds. The fact that the peak arrives within the first second tells us that the reaction speed of the system is $\leq$ 1s. Further, dissipation of the peak tells us that it takes on average roughly 5s to find a new equilibrium price for imbalance-increasing orders and 10s for imbalance-decreasing ones.

FMP Response

As already mentioned, it is the FMP that is important, as this is the price the stocks are actually exchanged at when the auction executes. With regards to this a handful of stylised facts were found.

Firstly, as with the IMP, orders that increase the imbalance cause larger FMP responses than those that decrease it. This firmly establishes that counteracting an imbalance leads to a lower, and sometimes even negative, price response; this implies that liquidity providers, who act to stabilise the imbalance, are systematically rewarded by the auction mechanism.

Secondly, the cancellation-response appeared to act oppositely to expectations; e.g. cancelling sell orders leads to lower-prices (and vice versa for buy orders). This might appear surprising at first glance, but it can be explained by simply traders adjusting their positions; e.g. if a trader who wishes to sell notices that the FMP will be higher than he expected, he could simply cancel his current order and insert a new sell order

at a higher price, which would not decrease the FMP.

And finally, the amplitudes of all responses decrease with the time of the auction; this reflects that the IMP becomes a better and better predictor for the FMP as the auction time approaches $t_{\texttt{end}}$. This is not surprising in light of the fact that the average MO dollar-volume is known to increase as the auction progresses, as previously established in section 6.2.3, which makes the IMP more difficult to perturb.

6.3.5 Further Work

The work done in this section has largely established stylised facts on the order-placing dynamics, and naturally one can go further. The results were achieved by employing data that was quite limited both in scope and completeness; the data on closing auctions was not available, and the many gaps in the data that was available caused much unnecessary noise in the results. Thus the first step to better understand the order-placing dynamics is to gain access to the whole order book for both opening and closing auctions, and then repeat the work done here. Once this is sorted, one could continue the investigation by defining a price *impact* function by conditioning the current price response function with the volume of the order, and study the measured impacts as functions of volume and time. Afterwards, as has already been achieved by econophysicists in the case of continuous trading [45], it should become possible to formulate a dynamical model that is able to predict the shift in some stock's FMP due to a new order, given a volume and the current state of the auction.

6.3.6 Data Limitations

The data was quite incomplete. In particular, many event updates were missing, and the updates that we had were often made up of multiple events; e.g. some of the differences Δi & Δv that were assumed to be single orders, were in reality a combination of several consecutive orders of unknown type. The timeperiod studied and the number of stocks was also insufficient, as mentioned in section 6.2.7.

Chapter 7

CONCLUSION

In this project the opening and closing auctions of stock exchanges were analysed. The work was split into three parts, and each part is summarised below.

7.1 Properties of Individual Auctions

In the first part, individual opening and closing auctions at NYSEarca were studied by thoroughly investigating the distributions of the MO dollar-volumes. The investigation was done with two goals in mind; (1) check whether the MO dollar-volumes are distributed as power-laws, or at least heavy-tails, and (2) investigate the tail properties of the distribution of MO dollar-volumes.

It turned out that the distributions of the MO dollar-volumes are not power-laws, but they are heavy-tailed. Secondly, it was found that the scaling properties of the heavy-tails are independent of date, but they are dependent on stock and auction type; in particular, it was found that the tails are typically heavier for closing auctions than for the opening ones.

In essence, this means the values of the trades done during both opening and closing auctions of NYSEarca are characterised by heavy-tailed fluctuations, with the closing auctions exhibiting heavier tails compared to opening auctions.

7.2 Aggregate Properties of Auctions

The second part concerned itself with the properties of the daily MO dollar-volumes. Here, the opening and closing auctions of NYSE, NYSEarca and NASDAQ were investigated. There were four goals; (1) ascertain what the properties of the individual MO dollar-volumes imply for the properties of the daily MO dollar-volumes. Then (2) check whether the ratio between the daily open-auction MO dollar-volume and that of the closing-auction is log-normal and (3) understand its statistical properties. Finally, (4) check whether the time-evolution of the daily MO dollar-volume is predictable.

It was firstly found that due to the complex behaviour of the individual MO dollar-volumes, no regularities except heavy-tails can be expected in the distribution of the

daily MO dollar-volumes. However, interesting statistical regularities in the distribution of the ratio of the daily MO dollar-volumes were found and investigated; firstly, the ratios are approximately log-normally distributed, and secondly the statistical properties of the ratios varies strongly by exchange. Finally, the time-evolution of the daily MO dollar-volume was studied through forecasting. Three regressive time-series analysis methods were used; ARIMA, Facebook Prophet and Random Forests, and, out of these, it turned out that Random Forests was by far the superiour method. The forecasting was done using the $\log_{10}$ daily MO dollar-volume, and judging from the resulting RMSE, NASDAQ opening and closing auctions are typically predictable up to factors of $10^{0.217} \approx 1.6$ and $10^{0.266} \approx 1.8$, respectively, while the equivalent cases for NYSEarca is $10^{0.323} \approx 2.1$ and $10^{0.517} \approx 3.3$.

The main discoveries here are that the ratio of the daily MO dollar-volumes between the opening and closing auctions is approximately log-normal, and that the time-evolution of the daily MO dollar-volume is, to an extent, predictable. The log-normality is interesting because it was unexpected, and because understanding its origins would further our theoretical understanding of market behaviour. The predictability of the daily MO dollar-volumes is, on the other hand, of *practical* importance; one can easily imagine [54] that being able to predict the daily MO dollar-volume is useful for tasks such as scheduling the executions of large orders and detecting (and hopefully profiting from) abnormal market conditions, as well as for being useful input to various high frequency trading algorithms [52] [53].

7.3 Pre-auction Dynamics

The final part concerned the pre-auction dynamics of NYSEarca opening auctions. There were two goals: (1) identify any statistical regularities present, and (2) investigate the price response of the auction.

Two regularities were found; firstly, the MO dollar-volume grows linearly as a function of time throughout the first roughly 75 minutes, and secondly the IMP is, on average, elevated by 0.05-0.1%, relative to the FMP, between the 65th and 85th minutes of the pre-auction period. Next, by introducing various price response/cancellation-response functions, the price fluctuations that resulted from a new order, conditional on the type of order (i.e. new buy, new sell, cancel buy or cancel sell) were measured. This revealed that imbalance-increasing orders typically have a significantly larger FMP response than imbalance-counteracting orders, that the FMP cancellation-response is often negative, that the amplitudes of the FMP response/cancellation-responses tend towards 0 as the pre-auction period approaches its end, and finally, judging from the IMP response, that it takes the auction system on average roughly 5s to locate a new equilibrium price after an imbalance-increasing order and 10 seconds for an imbalance-decreasing one.

Here the dynamical properties of the pre-auction period of NYSEarca opening auctions were characterised. Once this is also done for the closing auctions, the path to finding a good dynamical model of the pre-auction period for NYSEarca will be open. Finding such a model will certainly be a challenge due to the complexity of

the system, but as can be seen by what has been achieved concerning the dynamics of continuous trading [45], it is not an insurmountable one for sufficiently motivated physicists.

7.4 Final Words

The success of this work underlines the viability of a physicist's approach in finance. The auctions were approached as unknown physical systems, which means that rather than using axioms of financial economics to derive how the auctions are supposed to behave, we instead found patterns in the auction-generated data, and then used these patterns to draw conclusions on how the auctions actually behave in reality. This has resulted in a broad overview on the quantitative properties of opening and closing auctions, which is both of practical relevance to the financial industry, and sets the stage for further academic work towards properly understanding the observations made in this project, and creating dynamical models that reproduce them.

Bibliography

[1] The Thomson Reuters Tick History data we had available for our research.

[2] C. Read, The Efficient Market Hypothesis, *ISBN 9781349324354*

[3] D. Challet, Regrets, Learning and Wisdom, *article in European Physical Journal may 2016 edition*

[4] Official data from the New York Stock Exchange

`http://www.nyxdata.com/nysedata/asp/factbook`

[5] J-P Bouchaud, Economics Needs a Scientific Revolution, *NATURE, Vol 455, 30 October 2008.*

[6] J-P Bouchaud & M. Potters, Welcome to a non Black-Scholes world, *Quantitative Finance, 1:5, 482-483.*

[7] M. Torrecillas, R. Yalamova & B. McKelvey (2016) Identifying the Transition from Efficient-Market to Herding Behavior: Using a Method from Econophysics, *Journal of Behavioral Finance, 17:2, 157-182.*

[8] F. Patzelt & K. Pawelzik, An Inherent Instability of Efficient Markets, *NATURE, Scientific Reports 3, Article number: 2784, 2013.*

[9] Paul Krugman,

`http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html`

[10] L. Laloux, P. Cizeau, M. Potters, J-P Bouchaud, *International Journal of Theoretical and Applied Finance, Volume 3, issue 3, July 2000.*

[11] Mandelbrot, *B. B. J. Bus. 36, 394–419 (1963).*

[12] P. Gopikrishnan, V. Plerou, L. Amaral, M. Meyer, E. Stanley, *Phys. Rev. E 60, 6519–6529 (1999).*

[13] B. Hagstromer, L. Norden, *Closing Call Auctions at the Index Futures Market, European Financial Management Association*

[14] X. Gabaix, P. Gopikrishnan, V. Plerou & E. Stanley, *A theory of power-law distributions in financial market fluctuations, Letters to Nature, vol. 423, May 2003.*

[15] D. Challet, A. Chessa, M. Marsili & Y. Zhang, From Minority Games to real markets, *Quantitative Finance 2001, 1:1, 168-176.*

[16] G. Ibikunle, Opening and closing price efficiency: Do financial markets need the call auction?, *Int. Fin. Markets, Inst. and Money 34 (2015) 208–227*

[17] M. Stumpf & M. Porter, Critical Truths About Power Laws, *Science, Vol. 335, Issue 6069, 10 Feb 2012.*

[18] S. Asmussen, Applied probability and queues, *2003, pp. 412, Berlin: Springer.*

[19] M. Mitzenmacher, A Brief History of Generative Models for Power-law and Log-normal Distributions, *Internet Mathematics Vol. 1, No. 2: 226-251.*

[20] A. Clauset, C. Shalizi, M. Newman, POWER-LAW DISTRIBUTIONS IN EMPIRICAL DATA, *SIAM Rev., 51(4), 661–703.*

[21] G. Box, D. Cox, An analysis of transformations, *Journal of the Royal Statistical Society, Series B. 26 (2): 211–252, 1964.*

[22] S. Taylor, B. Letham,

`https://facebookincubator.github.io/prophet/`

[23] R, Caruana, N. Karampatziakis, A. Yessenalina, An Empirical Evaluation of Supervised Learning in High Dimensions, *ICML 2008, 96-103*

[24] L. Breiman, A. Cutler,

`https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`

[25] D. Challet, M. Marsili, Criticality and market efficiency in a simple realistic model of the stock market, *Physica Review E 68, 2003*

[26] U.S. Securities and exchange commission,

`https://www.sec.gov/rules/other/nasdaqllcf1a4_5/e_sysdesc.pdf`

[27] V. Alfi, A. Gabrielli & L. Pietronero, How people react to a deadline: time distribution of conference registrations and fee payments, *Central European Journal of Physics, 7:483*

[28] J. Karpoff, *The Relation between Price Changes and Trading Volume: A Survey, Journal of Financial and Quantitative Analysis, 1987*

[29] C. Hiemstra, J. Jones, Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation, *1994*

[30] J. Heaton, N. Polson, J. Witte, Deep Learning in Finance,

`https://arxiv.org/pdf/1602.06561.pdf`

[31] R. Hyndman, Y. Khandakar, Automatic Time Series Forecasting: The forecast Package for R, *Journal of Statistical Software, 2008*

[32] NYSEarca Auction Rules Brochure

`https://www.nyse.com/publicdocs/nyse/markets/nyse-arca/NYSE_Arca_Auctions_Brochure.pdf`

[33] NASDAQ Opening & Closing Crosses Quick Reference Guide

http://www.nasdaqtrader.com/content/technicalsupport/specifications/TradingProducts/openclosequickguide.pdf

[34] NASDAQ Open/Close Crosses FAQs

https://www.nasdaqtrader.com/content/ProductsServices/Trading/Crosses/openclose_faqs.pdf

[35] NYSE Official Rulebook

http://wallstreet.cch.com/NYSE/Rules/

[36] J. Bouchaud, J. Kockelkoren & M. Potters, *Random walks, liquidity molasses and critical response in financial markets, Quantitative Finance, 6:02, 115-123*

[37] S. Park & A. Bera *Maximum entropy autoregressive conditional heteroskedasticity model, Journal of Econometrics 150 (2009) 219–230*

[38] M. Chernick *Bootstrap Methods: A Guide for Practioners and Researchers*

[39] J. Alstott, E. Bullmore & D. Plenz, *powerlaw: a Python package for analysis of heavy-tailed distributions*

[40] MIT lecture notes

http://www.mit.edu/~6.s085/notes/lecture5.pdf

[41] W. Venables & W. Ripley, *Modern Applied Statistics with S, 2002*

[42] NYSEARCA XDP IMBALANCES FEED

https://www.nyse.com/publicdocs/nyse/data/NYSE_Arca_XDP_Imbalances_Feed_Client_Specification_V1.0b.pdf

[43] J. Bouchaud, Y. Gefen, M. Potters & M. Wyart *Fluctuations and response in financial markets: the subtle nature of 'random' price changes, Quantitative Finance, 4:2, 176-190, 2004*

[44] F. Lillo, S. Mike & J. Farmer, *Theory for long memory in supply and demand, Physical Review E 71, 2005*

[45] J. Bouchaud, J. Farmer & F. Lillo *How markets slowly digest changes in supply and demand, 2008*

[46] R's forecast package:

https://cran.r-project.org/web/packages/forecast/forecast.pdf

[47] R's (facebook) prophet package:

https://cran.r-project.org/web/packages/prophet/prophet.pdf

[48] R's randomForest package:

https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[49] R's MASS package:

https://cran.r-project.org/web/packages/MASS/MASS.pdf

[50] Python's powerlaw package:

http://pythonhosted.org/powerlaw/

[51] Torben G. Andersen, *Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility, THE JOURNAL OF FINANCE, MARCH 1996*

[52] V. Satish, A. Saxena & M. Palmer *Predicting Intraday Trading Volume and Volume Percentages*

[53] D. Alparslan, M. Borkovec, D. Ho & K. Tyurin*Predictions of Intraday Volume, Volatility and Spread Profiles and Their Applications*

[54] M. Chlistalla of Deutsche Bank Research, *High-frequency trading: Better than its reputation? 2011*