Tampere University

MARTIN MAGRIS

Volatility modeling and limit-order book analytics with high-frequency data

Tampere University Dissertations 106

Tampere University Dissertations 106

MARTIN MAGRIS

Volatility modeling and limit-order book analytics with high-frequency data

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Engineering and Natural Sciences
of Tampere University,
for public discussion in the Lecture room K1702
of the Konetalo building, Korkeakoulunkatu 6, Tampere,
on Tuesday, 27 August 2019, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Engineering and Natural Sciences
Finland

Responsible supervisor and Custos	Professor Juho Kanniainen Tampere University Finland	
Pre-examiners	Assistant Professor Matthew Dixon Illinois Institute of Technology USA	Associate Professor Marcelo C. Medeiros Pontifical Catholic University of Rio de Janeiro Brazil
Opponent	Professor Michael McAleer Asia University Taiwan University of Sydney Australia Erasmus University Rotterdam The Netherlands Complutense University of Madrid Spain Yokohama National University Japan	

The originality of this thesis has been checked using the Turnitin Originality Check service.

Cover design: Roihu Inc.

ISBN 978-952-03-1195-7 (print)
ISBN 978-952-03-1196-4 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1196-4

PunaMusta Oy – Yliopistopaino
Tampere 2019

PREFACE

I freshly remember the day I first arrived in Finland. Little I knew about this beautiful country. I've been told the language is impossible and unreadable, the climate severe, and people difficult to hang out with. Not at all, but I still had to find it out. On my first ride from Helsinki to Tampere, I clearly remember exclaiming: "Let's see where this journey will take me", loosely translated. Well, the journey took me all around the globe, and finally here again, writing this Preface and thinking of all the wondrous things happened in the last three years and all the amazing people I met. Going through all the great memories would certainly take something like a hundred pages. Here I want to thank the closest and most important ones who shared this journey with me. If unintentionally I forget someone, please forgive me: I aged a lot by completing this dissertation. For all the others I cannot mention, remember that good memories and great time simply do not vanish and easily come back to mind: believe me, while writing these lines I had a thought for all of you.

I can do no other than thanking Prof. Juho Kanniainen's endeavor in providing the best guidance and supervision one could ever ask for. Words are not enough for sharing my gratitude for all you have done in the past years. Thank you for believing in the first instance in my application, for patiently guiding me towards autonomy in research and writing, for setting deadlines and requiring precision too. Thanks for generously sharing your expertise and ideas and for building every day a fresh, relaxed and informal work environment. Waking up and going to the office has been a pleasure every day. Thank you for being there for whatever issue, for the ubiquitous trust, for the frank and honest yet very caring and respectful, positive and to-do attitude you taught me.

I am particularly thankful to Professor Michael McAleer for serving as an opponent. Thank you for your time and commitment in accepting this role, for going through the

dissertation and traveling up to Finland, despite your undoubtedly busy agenda. I am grateful for the valuable comments from the external examiners Ass. Prof Matthew Dixon and Assoc. Prof. Marcelo C. Medeiros: your thought and insights have been truly appreciated and definitely boosted the quality of the present manuscript, and of the unpublished article as well. In this regard, I would like to thank Ass. Prof. Alexandros Iosifidis for proofreading some technical parts of this dissertation. Furthermore, I thank Prof. Kim Christensen for having supervised and hosted me during the visiting period at Aarhus University.

I wish to thank all my colleagues from the DARE Business Data Research Group, former Financial Engineering Research Group. In particular, the office members of the former FB-108 - Festia Building, room 108 - aka. “beavers”. Thank you "mama" Milla -yes, this is published- for all your infinite patience and guidance towards all the aspects of Finnish life. Without your generous support I would still be figuring out how to book a dentist appointment, driving with foreign plates, and wondering how to check my study record in POP. Thank you Sindhuja for sharing with me all the complaints about the food at the canteen, for your remarkable, but pointless, dedication in teaching me how to run scripts at CSC and, of course, for the amusement we had in India. I am sorry for all my generic excuses for not to swim on a daily basis in cold Finnish lakes, Jimmy. But I will always carry with me all the indecipherable talks we had in the last years and your complaints on how I drink coffee. Thank you Ye for our inspiring “econometrics and traveling talks” and for the several outdoor activities and dinners we had. The research group unofficially extends to Eija: thank you for all the good talks and your guidance. Literally. E.g. when -with some considerable delay- explaining that I missed the turn to Lapland and was driving to Sweden. My most special thanks go to Margarita and Kestutis for all the great office, home, barbecue and hard-to-remember Labor/Herwood/Terrible/Ylä/Taj Hotel & many more, activities and bar-times we shared. The countless number of great moments and memories I had with you in the last years made the most of my stay in Finland. Looking forward for more, and visiting you in Riga soon.

Thanks to all the other academia-related wonderful people I met too. Ji for your reliable help in all the DFA-related issues, your uncommon hospitality and friendliness: I am glad of having met you and Mikko. Thank you Adam for the mutual support for our very first publication and for our many hopeless attempts in running “Hello world” on SPARK. Thanks to Jaakko and Perttu for my very early times at

TUT, for supporting me in moving my first steps towards the earliest data-analyses and publications. Lastly, thanks to the BigData Finance fellows visiting Tampere: Rytis, Sergio and Chiara, for the unforgettable hanging out, which, however, I largely forgot.

And a paragraph is deserved for all the amazing, bright and truly friendly people I lived with in Aarhus. Fede, Cami, and Foteini for the countless number of great days and nights, your everyday smiles, laughs, and support that made the months in Denmark unique. To Luca for our truly work-inspiring talks on daytime and work-complicating practices at nights. My period at BSS would not have been the same without you all. Lastly, to Sanchali, Oscar, Yussef, Marta, Patrizio, Giorgio, Ye, Sigurd and Simon: it's my pleasure to have known you all.

To all the great friends outside the academia, playing baseball together. The informal-but-competitive spirit is our force: "a team having fun, is a winning team". So true, especially thanks to Gabo, Luis, Niko, Tuomo, Tuomas and Mauricio for making each training, game and post-game moment enjoyable, relaxing and, most importantly, full of fun, if not crazy. Finally, a warm hug to all my friends in Italy and around the world, turning all the short moments together into such friendly, sincere and delightful times: Bartolich, Goretex, Marsa, Giulia Reds, Vale, Uba, Jakob, Umbi, Stefania and Lindy.

Finally, to my dear family - Naima, Roberto and Nadja and all the other relatives. I would not be here without you. Thank you for pushing me towards a Ph.D. when most of my applications were rejected, for your support in low and stressful moments, your attention on my overall well-being. For Monday's talks and counseling, and the solid shelter you provide whenever something does not go the right way.

Principio caeli clarum purumque colorem
quaeque in se cohibet, palantia sidera passim,
lunamque et solis praeclara luce nitorem;
omnia quae nunc si primum mortalibus essent
ex improviso si sint obiecta repente,
quid magis his rebus poterat mirabile dici,
aut minus ante quod auderent fore credere gentes?
Nil, ut opinor; ita haec species miranda fuisset.
Quam tibi iam nemo fessus satiate videndi,
suspicere in caeli dignatur lucida templa.
Desine qua propter novitate exterritus ipsa
expuere ex animo rationem, sed magis acri
iudicio perpende, et si tibi vera videntur,
dede manus, aut, si falsum est, accingere contra.
Quaerit enim rationem animus, cum summa loci sit
infinita foris haec extra moenia mundi,
quid sit ibi porro, quo prospicere usque velit mens
atque animi iactus liber quo pervolet ipse.[12]

[1]Lucretius (c. 99 BC — c. 55 BC), *De rerum natura*, Book II, lines 1030-1047.

[2]Look up at the bright and unsullied hue of heaven and the stars which it holds within it, wandering all about, and the moon and the sun's light of dazzling brilliancy: if all these things were now for the first time, if I say they were now suddenly presented to mortals beyond all expectation, what could have been named that would be more marvelous than these things, or that nations beforehand would less venture to believe could be? Nothing, methinks: so wondrous strange had been this sight. Yet how little, you know, wearied as all are to satiety with seeing, any one now cares to look up into heaven's glittering quarters! Cease therefore to be dismayed by the mere novelty and so to reject reason from your mind with loathing: weigh the questions rather with keen judgment and if they seem to you to be true, surrender, or if they are a falsehood, gird yourself to the encounter. For since the sum of space is unlimited outside beyond these walls of the world, the mind seeks to apprehend what there is yonder there, to which the spirit ever yearns to look forward, and to which the mind's emission reaches in free and unembarrassed flight. Translation: Munro, Hugh Andrew Johnstone. *T. Lucreti Cari De rerum natura libri sex: 2*. Vol. 2. Deighton Bell and Company, 1866.

ABSTRACT

The vast amount of information characterizing nowadays's high-frequency financial datasets poses both opportunities and challenges. Among the opportunities, existing methods can be employed to provide new insights and better understanding of market's complexity under different perspectives, while new methods, capable of fully-exploit all the information embedded in high-frequency datasets and addressing new issues, can be devised. Challenges are driven by data complexity: limit-order book datasets constitute of hundreds of thousands of events, interacting with each other, and affecting the event-flow dynamics.

This dissertation aims at improving our understanding over the effective applicability of machine learning methods for mid-price movement prediction, over the nature of long-range autocorrelations in financial time-series, and over the econometric modeling and forecasting of volatility dynamics in high-frequency settings. Our results show that simple machine learning methods can be successfully employed for mid-price forecasting, moreover adopting methods that rely on the natural tensor-representation of financial time series, inter-temporal connections captured by this convenient representation are shown to be of relevance for the prediction of future mid-price movements. Furthermore, by using ultra-high-frequency order book data over a considerably long period, a quantitative characterization of the long-range autocorrelation is achieved by extracting the so-called scaling exponent. By jointly considering duration series of both inter- and cross- events, for different stocks, and separately for the bid and ask side, long-range autocorrelations are found to be ubiquitous and qualitatively homogeneous. With respect to the scaling exponent, evidence of three cross-overs is found, and complex heterogeneous associations with a number of relevant economic variables discussed. Lastly, the use of copulas as the main ingredient for modeling and forecasting realized measures of volatility is explored. The modeling background resembles but generalizes, the well-known

Heterogeneous Autoregressive (HAR) model. In-sample and out-of-sample analyses, based on several performance measures, statistical tests, and robustness checks, show forecasting improvements of copula-based modeling over the HAR benchmark.

CONTENTS

ABBREVIATIONS

ADF	Augmented Dickey-Fuller
CDF	Cumulative Distribution Function
CV-HAR	C-Vine HAR
DFA	Detrended Fluctuation Analysis
ECDF	Empirical Cumulative Distribution Function
HAR	Heterogeneous Autoregressive
IV	Integrated Variance
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
LDA	Linear Discriminant Analysis
LOB	Limit Order Book
MDA	Multilinear Discriminant Analysis
ML	Machine learning
MMS	Market Microstructure
MP	Moore-Penrose
OLS	Ordinary Least Squares
PP	Phillips–Perron
RV	Realized Variance
SLFN	Single Layer Forward-feed Network
TAQ	Trades And Quotes
WMTR	Weighted Multi-channel Time-series Regressor/Regression

ORIGINAL PUBLICATIONS

Publication I Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting* 37.8, 852–866. DOI: 10.1002/for.2543.

Publication II Tran, D. T., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2017). Tensor representation in high-frequency financial data for price change prediction. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/SSCI.2017.8280812.

Publication III Magris, M., Kim, J., Räsänen, E. and Kanniainen, J. (2017). Long-range auto-correlations in limit order book markets: Inter-and cross-event analysis. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/SSCI.2017.8280932.

Publication IV Magris, M. (2019). A C-Vine extension for the HAR model. *Unpublished manuscript*. Submitted to *Journal of Business & Economic Statistics*, May 2019.

Author's contribution

This section describes in detail my personal contribution to each of the above-mentioned publications.

In Publication I, in collaboration with the co-authors, I took part in the discussion about the research objective. Contributing to specifying and defining the current one, based on constraints and feasibility issues implied from the data structure and content. Furthermore, I had the whole responsibility for the data-related part. In particular, I took care of the whole data processing, management, and validation, delivering a well-structured dataset that fits the requirements for the analyses. A major contribution is indeed that of data construction and processing, along with an active role in designing the whole setup for the further methodological applications. Accordingly, I wrote the corresponding data section in Publication I. Finally, I provided comments and suggestions to the main author for the final draft.

Publication II uses the same data as Publication I. Without the careful data-processing already set up for Publication I and made available at the time Publication II was under development, the analyses would not have been possible. As for Publication I, the whole responsibility on the data is solely mine. I took part in reviewing the final draft by suggesting the main author and co-authors improvements in the text, concerning the introduction, description of limit-order books and the methodological part too. I presented the publication at the IEEE SSCI 2017 Conference.

Publication III is a collaboration with researchers at the Department of Physics, who have strong expertise in the methodology used. Although the methods were jointly determined, I defined the exact research objectives by in identifying the actual variables of interest over which the analyses are implemented. Preliminarily, I implemented and tested the method on a limited dataset uncovering those research directions that look promising and worth to investigate. I processed the whole data accordingly and delivered it in a conveniently structured shape for the overall analyses. I took the main responsibility in writing the publication and first drafting for all the sections, including the literature review and methods parts. Later revisions and improvements involved the co-authors as well. Lastly, I presented the publication at the IEEE SSCI 2017 Conference.

Publication IV is entirely an outcome of my own work and ideas. I have full responsibility for the whole project. I defined the main objective, planned the research design to answer the research question and decided over the methodology to tackle it. I processed the data, run the analyses and drawn the conclusion on the results. A first draft of the manuscript and the results therein were presented at the CFE 2018

conference.

Acknowledgments

The research leading to this dissertation received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No. 675044, "BigDataFinance"[3].

[3] https://bigdatafinance.eu

1 INTRODUCTION

This chapter introduces the reader to the topics explored in the publications and their respective research questions. Section 1.1 provides a very general understanding of complexities and opportunities related to nowadays' availability of big datasets. Section 1.2 deepens the discussion addressing the motivation for Publications I-IV and formulates precise research questions. Section 1.3 addresses how the publications are related to each other and what is the web holding them together, as a part of more general research of wider width. Finally, a description of the structure of this dissertation is provided in Section 1.4.

1.1 General background

In the last decade, the world entered the "Big-data era". Every day and at an incredible pace, a huge amount of digital data is created and recorded. As of 2012, about 2.5 exabytes ($2.5 \cdot 10^9$ Gb) of data were created each day (McAfee et al., 2012). The financial sector is part of this trend (Cont, 2011, among the others). By the introduction of more and more sophisticated technologies, trading platforms and algorithms, high-frequency trading exploded. As a consequence, all the high-frequency trading activity taking place on different exchanges all around the world generates a massive amount of digital data. Budish et al., 2015 provides an outlook on the high-frequency trading phenomenon through time and on its quantitative and qualitative impact on the markets, while (Beltran et al., 2005) provides further empirical analyses in the context of cross-market activity. Numbers are impressive, posing both challenges and opportunities.

Opportunities are directly linked to the exceptionally rich information that high-frequency data endows. This can clearly have an impact on different levels, for

instance can uncover new features in how the trading occurs and takes place (Cont, Kukanov et al., 2014; Budish et al., 2015; O'Hara, 2015), can provide behavioural insights on market's participants (see e.g. Pang et al., 2002), can be used to identify early signals e.g. of a forthcoming jump in the prices (e.g. B. Zheng et al., 2012), or used for metrics forecasting (e.g. Kercheval et al., 2015). In other words, the high-frequency data era constitutes a great opportunity for the development of new models and methods capable of providing a more precise understanding on diverse aspects of financial markets, as well as boosting the forecasting ability of concurrent and existing models buy the use of the vast and rich information being nowadays recently available.

In this large and multidimensional information and data the human eye is clearly lost (the order book data sample in Figure 2.2 is a clear example): high-frequency data is too vast to have an overall outlook and understanding of all the possibly very-complex relationships between all variables. Highly-efficient computer-based methods, processing platforms with a high amount of automatization, requiring little human interaction and high flexibility are important requisites for tomorrow's algorithms. In this regard (Flood, 2012, Sec. 1.2 to 1.4) provides a historical overview on the connection between information technology advances and financial markets' complexity, and their evolving needs. Indeed the classical econometric approach, attempting to precisely identify sets of hypotheses about e.g. variables, their relationships, and possible error terms, is strained. In this context, there is no surprise in the recent attention and popularity that Machine learning (ML) applications gained in finance (indeed, a number of monographs have been written so far on the topic). ML methods are capable of executing complex analyses with little or no human interactions (so-called unsupervised learning), capable of auto-detecting feature and variables of interest to achieve a given (prediction) task, tune their parameters and improve their own efficiency (Michie et al., 1994). As a part of this, also the analysis of the unique properties that characterize high-frequency time-series constitute an opportunity, not only in statistics and econometrics (e.g. Bouchaud, Farmer et al., 2009; Abergel, Anane et al., 2016, as examples of statistical analyses on high-frequency data and their econometric implications and modelling), but also for the possibility of cross-disciplinary applications, e.g. the use of methods historically pertaining to natural sciences and engineering in finance (e.g. Peng, S. V. Buldyrev, Havlin et al., 1994; Ogata, 1998).

With opportunities come challenges. In the first instance, these are related to the enormous amount of data that needs to be handled. Beside hardware issues and software architecture, there is a need for algorithms that are fast and efficient but at the same time accurate and reliable. Second, the data can be very heterogeneous and different. This is a clear problem for science since results would strongly rely on the specific dataset, on the cleaning procedure, and experimental protocol adopted. Third, high-frequency time-series have unique features that require specific and new methodologies to be properly handled (the problem of unbiased volatility estimation under microstructure noise is an example). A more detailed overview over the complexity of nowadays's big data streams and challenges they pose can be found in (Flood et al., 2016). Although the focus in (Flood et al., 2016) is on data-related issues concerning market stability monitoring, most of the discussion broadly generalizes, being easily contextualized and valid over different domains, e.g. limit-order books.

This is the setting where my research moves: high-frequency data, ML methods for specific prediction tasks, time-series properties of the high-frequency data and high-frequency econometric modeling. This dissertation analyzes different topics in high-frequency financial data domain, under different angles. (i) *prediction* , in particular of the mid-price direction different machine-learning approached. (ii) *Descriptive analyses*, in particular on the nature of the long-range autocorrelation of financial duration series extracted from the order book. (iii) *Volatility modeling and forecasting* in an econometric context, based on methods and models of form high-frequency econometrics literature. The common thread to these points is the use of high-frequency data, and disconnections among them, especially with respect to the last points, have to bee seen under the general aim of addressing different topics, namely, prediction, description and modeling. This justifies the title of his dissertation too, recalling volatility modeling, limit-order book analytics (in general, the discovery, interpretation and use of meaningful patterns extracted from the data), and high-frequency data.

1.2 Motivation and research questions for Publications I-IV

Whereas the discussion in Section 1.1 was intentionally general, now I provide a deeper motivation for each of the publications presented in this dissertation outlining

literature gaps, and opportunities for further research. Accordingly, I formulate the relevant research questions Publications I-IV deals with.

Motivation and research objective for Publication I. In the last decades, a number of different ML methods have been applied for tackling prediction problems with high-frequency limit-order book data. In this regard, the outstanding literature is vast and heterogeneous. Overall, there is great variability in the ML methods applied, in the data itself (in quantitative terms such as e.g. data not limited to the best levels only, and qualitative terms, such as stocks involved and periods analyzed), in data processing, in reporting and validation of the results. This implies a general complexity in comparing results and generalizing the findings reported in the literature. Indeed differences in trading platforms, matching rules, transaction costs, possible exchange-specific features, and data aggregation from fragmented markets, pose a general comparability challenge for limit-order-book related research, even for a very same asset (Cont, 2011). Consequently, ML forecasting analyses under different data, methods, and protocols are a real challenge for results' standardization and generalization.

With the purpose of motivating the above discussion about heterogeneity in the data and methods, consider the following example. (Kercheval et al., 2015) addresses the problem of mid-price prediction and spread crossing with support vector machine (SVM) methods, over one-trading-day, 10-levels deep, high-frequency limit-order book data for 5 stocks traded at NASDAQ, performance is evaluated in terms of F1 scores; (J. Liu et al., 2015) uses 42 US securities for a 1-month period, but using only top-level data to estimate a linear model for price prediction over intervals between 30 seconds and one hour. The performance of their model is evaluated in terms of R^2 (indeed the analysis is not framed into a classification problem, as the earlier case). (Pai et al., 2005) uses 50-days data from the year 2002 for 10 stocks to tackle the mid-price prediction with ARIMA-SVM methods, but with a completely different protocol, based on daily-returns, and one-day-ahead forecasts, whose performance is evaluated in terms of mean squared error. This example shows that whether it is quite safe to conclude that indeed we can approach some prediction problem on some data with certain ML methods under specific learning and forecasting schemes, drawing general conclusions, abstracting the results to a general level and e.g. concluding that given ML methods perform better than others is definitely challenging. Furthermore, many different prediction tasks can be considered, e.g. (B. Zheng et al., 2012) uses a Lasso regressor for the CAC40 constituents over a one-month LOB data sample

from the year 2011, for predicting price jumps.

As pointed out so far, the heterogeneity in the available datasets (e.g. period covered, type of events recorded, data-frequency, and order book depth), rather than a research question, stems as a research *issue* specific for the ML literature using LOB data published so far. Among the prediction tasks, a literature gap involving applications of standard ML methods on high-frequency limit-order book data emerges for the prediction of future mid-price movements. Because of this lack, a benchmark of simple ML implementations for a given prediction task (under a uniform) forecasting framework is missing. Implicitly, a clear understanding of which are the standard methods providing a promising direction for their future improving is missing. As a part of it, it is hard to asses the actual need for complex methods, such as that of (Pai et al., 2005), and their improvements in forecasting over much simpler alternatives.

Moreover, the central impact played by different data-standardization schemes on the implementation of the methods, and in driving the forecasting results, has not been addressed. Also, from the current literature, a research gap emerges in the discussion about mid-price movement predictability at different horizons, which is a relevant analysis for understanding to which extent the past LOB flow can be exploited for prediction, and whether the short-term behavior of the mid-price is, on the contrary, largely noisy.

In this context, a publicly-available high-frequency limit-order book dataset that the research community can utilize as a homogeneous basis for future applications is also missing. A preprocessed, well-documented and well-structured dataset over which old and new prediction methods can be implemented, would definitely improve results' comparability and lead to a better understanding of what are the implications of different methodological approaches. In Publication I we provide an example of such heterogeneity in the datasets, motivated also by the fact that the public availability of the rich datasets, is generally constrained by disclosure agreements, and high costs. With this, a set of reference measures related to the forecasting performance of a set of ML methods on a selected prediction task, with uniform experimental setting and data, is also nowadays not available. The research gaps outlined in the above discussion lead to the following research objective (RO), rather than a research question (RQ):

RO 1: Exploring the problem of predicting mid-price movements in the limit-order book with standard machine-learning techniques, under different data-normalization

approaches and for different forecasting horizons. As a part of it, provide a publicly available ultra-high frequency limit-order book dataset for general ML forecasting-related research; comprehensive of a detailed and robust experimental protocol, and inclusive of baseline performance measures for this specific prediction task.

Motivation and research objective for Publication II. Popular ML methods used in financial applications include linear regression (B. Zheng et al., 2012; Panayi et al., 2018), network-based models (Tsantekidis et al., 2017a; Tsantekidis et al., 2017b; Passalis, Tsantekidis et al., 2017), deep learning (e.g. Sirignano, 2019) and many others. Several existing applications utilize models that lean from the data based on vector inputs of features (e.g. Kercheval et al., 2015; Passalis, Tsantekidis et al., 2017, among several others). Vectorization is unable to capture the spatial-temporal information, interactions and inter-links between the input vectors (Tan et al., 2019), thus ML methods based on a tensor-based representation of financial times series constitute an appealing approach, that remains not addresses. In particular, multilinear techniques in ML are not at their early applications and have been widely applied to image and videos classification problems (e.g. He, Cai et al., 2006; Vasilescu et al., 2003). However, their use in high-frequency financial settings is very limited, and not specifically related to the prediction of future mid-price movements (Q. Li, Y. Chen et al., 2016, e.g.).

Besides the growing number of applications involving methods relying on the tensor-representation of a time-series in different fields, a study exploiting this representation for the order book dynamics forecasting (and specifically for the prediction of mid-price movements) is missing. Furthermore, as outlined in Section 1.1 ML models are likely to be very complex and of difficult interpretation[1]. Natural multidimensional (or tensor) generalizations of standard ML models are attractive, but unexplored, alternatives. However in the related ML literature dealing with LOB applications such a discussion is not addressed yet. In this regard, Publication II deals with the following research question:

RQ 1: To which extent can machine-learning techniques based on time-series' tensor-

[1]For instance in terms of the objective function to be minimized and its interpretability. E.g. neural-network -based methods are not of easy estimation, the estimation itself is time-consuming, and the interpretability of the parameters in the (possibly multiple) layers, as well as their impact on the objective function, is not immediate and straightforward.

representation be effectively employed in the prediction of the mid-price movement, boosting forecasts' performance measures over alternative models?

Motivation and research objectives for Publication III. A number of well-established techniques generally used in natural sciences applications have been designed to detect, characterize and quantify the so-called scaling laws and long-range correlation properties of time-series (see e.g. Kantelhardt, 2009). Starting from (Mandelbrot, 1971) this research direction gradually became attractive in finance too, and have been utilized in analyzing e.g. the scaling properties in returns' series (e.g. Mantegna et al., 1995), in currency rates series (e.g. Vandewalle and Ausloos, 1997), and to discuss market efficiency (e.g. Y. Wang et al., 2009). Indeed, the presence of long-range correlation in financial series is not a new concept in econometrics (see e.g. Baillie, 1996, for a review), but its precise characterization is challenging: in this concern, methods from other fields proved to be very valuable and of simple applicability. Indeed the precise understanding of long-memory properties in financial time-series plays a key role in the development of different econometric models, e.g. ARFIMA models, or IGARCH effects are attempts to deal with long-range dependence (Andersen, Bollerslev, Diebold and Labys, 2003), and market hypotheses formulation (e.g. U. A. Müller, Dacorogna, Davé, Pictet et al., 1993). Therefore, analyses on the fractal nature of duration-related time-series in financial data and in the LOB are of high relevance.

The reference methodology in this field is provided by the so-called Detrended Fluctuation Analysis (DFA) (Peng, S. V. Buldyrev, Havlin et al., 1994), aimed at uncovering the presence of long-range autocorrelation features in time-series, and providing a quantitative characterization through the so-called scaling exponent. Although a number of applications involving ultra-high frequency limit-order book data (Ivanov, Yuen, Podobnik et al., 2004; Jiang et al., 2008; G. Cao, Xu et al., 2012; Ivanov, Yuen and Perakakis, 2014), researches exploiting the extensive and rich information that this data contains on different order book events, such as limit orders, market orders, and cancellations, in analyzing long-range properties in duration time-series, are missing. In particular, no joint analyses on the three different message types (events) that affect the order book state have been proposed, nor differences between the side of the book discussed. Also, analyses for cross-series' durations, e.g. between orders' submissions to their respective cancellations, are lacking in the literature.

Furthermore, earlier researches relied on high-frequency datasets that did not span long time periods. Accordingly, earlier studies detected differences in the scaling exponent when switching from intra-day to daily sampling frequencies, but were data-constrained and thus unable to detect differences wider time-scales. This is a sensible gap, considering the theoretical augmentations of (e.g. U. A. Müller, Dacorogna, Davé, Pictet et al., 1993), and several applications (e.g. Corsi, 2009) relying on the assumption of heterogeneous traders competing at different time-scales; namely daily, weekly and monthly horizons. Relying on a broad literature well-documenting on an empirical basis long-range autocorrelations in different financial series, imputable to different factors, including microstructure-related augmentations such as price impact(e.g. Lillo et al., 2004) and the closely related literature reporting strong evidence of long-range correlation in duration time-series extracted from the LOB, we expect our descriptive analysis to detect long-range autocorrelations as well, and successfully address the above-mentioned issues. Based on the literature here introduced, but expanded in 2.4, and on the above-mentioned research gaps, we outline the following research question:

RQ 2: What can be said about the long-range autocorrelation in the inter- and cross- event series of orders, trades, and cancellations in limit-order book data across different stocks, market sides, and sampling frequencies?

Moreover, outlining a set of economic variables that are associated, on a daily, level with the scaling exponent for LOB-extracted duration series, could provide valuable insights for the development of models for the long-range autocorrelation dynamics, and a starting set of variables for future causal analyses. Although some very limited analyses in this direction have been proposed in (Ivanov, Yuen and Perakakis, 2014)[2], this point is largely un-addressed in the current literature. This devises a further research question:

RQ 3: How strong, and consistent across different stocks and side of the book, are the associations between the scaling exponent, characterizing the nature of the long-range autocorrelation in durations' time-series, and general economic variables?

Motivation and research objectives for Publication IV. Volatility modeling and forecast-

[2] And in (Vandewalle and Ausloos, 1997), but for currency exchange rates and macro-economic events.

ing have been one of the most active areas of econometrics research since the seminal paper of (Engle, 1982) and the advent of GARCH-related literature, from (Bollerslev, 1986) onward. The advent of high-frequency data, however, poses new problems to contemporary econometrics. In particular, the effects of market microstructure noise, negligible at low sampling frequencies are a major issue, whose effects lead to bias and inconsistency of common volatility estimators. A number of new techniques for the noise-robust estimation of daily volatility from high-frequency data have then been developed (e.g. L. Zhang et al., 2005; Barndorff-Nielsen, P. R. Hansen et al., 2008; Podolskij et al., 2009). As a part of the very same tale, the introduction of the so-called realized measures (daily-volatility measures relying on high-frequency data) contributed to the development of new techniques for volatility modeling and forecasting. Among them, the Heterogeneous Autoregressive (HAR) model (Corsi, 2009) gained vast popularity in the last years because of its simple structure and effectiveness in predicting (generally) tomorrow's volatility. A motivation for the work of (Corsi, 2009) is to account for the long-range dependence observed in the (realized) volatility time-series (e.g. Andersen, Bollerslev, Diebold and Labys, 2003). Although this phenomenon has been pointed out already in the literature, a *simple* econometric model to account for it was still to be explored. Models such as those of the FIGARCH-class (Baillie et al., 1996) or ARFIMA-class of realized volatility (Andersen, Bollerslev, Diebold and Labys, 2003) are of complex theoretical construction, which affects their attractiveness. The HAR model stands out as a simple linear model, able to capture the long-range dependence in the volatility series, and of easy estimation. Thus, it gained popularity among practitioners as well. However, its construction relies on a set of crucial hypotheses that although supporting the HAR model itself, are generally restrictive. Among these, the linear assumption between the components involved in the HAR mode is particularly critical (and already recognized as such and discussed in the literature, e.g. Hillebrand et al., 2007).

As suggested in (Sokolinskiy et al., 2011) copulas can naturally provide a remedy for a number of methodological and theoretical constraint related to the limitations of the standard HAR model. Historically, the use of copulas has been very extensive in several areas of finance, risk management, and econometrics (see Section 2.6). On the other hand, in the contemporary literature on high-frequency realized-measures modeling and forecasting, the set of copula-based methods is very short. Interestingly, whereas there has been an explosion in the econometric literature about the

robust high-frequency estimation of daily volatility and forecasting in a HAR-based perspective (e.g. Andersen, Bollerslev and Diebold, 2007; A. J. Patton and Sheppard, 2015; Bollerslev et al., 2016), and important advances in (Vine-) copulas literature, cross-applications between these two areas are absent.

By pursuing the research direction of (Sokolinskiy et al., 2011) but setting apart from it methodologically, accommodating a Vine-copula (Aas et al., 2009) expansion for the HAR model of (Corsi, 2009) stems as an unexplored attractive research direction for covering this cross-fields gap and deal with some shortcomings of the HAR specification. This leads to the following research question:

RQ 4: What is the impact in terms of forecasting ability of a copula-based modeling of daily volatility measures over the HAR model?

1.3 Linkages between the publications

The general factor common to all the publications included in this dissertation is the use of high-frequency data. This is the predefined setting of my doctoral studies. As a fellow of the BigDataFinance EU-project project I conducted my research within the "High-frequency econometrics" working package. Therefore of high-frequency is the data I used in all the projects I have been working on. Publications I-IV included in this dissertation are not aimed at analyzing a single aspect and discuss it at different levels, rather they aim at addressing different topics under different angles. Prediction, in Publication I and Publication II, descriptive analyses, in Publication III, and volatility modelling and forecasting in Publication III, also with a cross-disciplinary cut, while sharing the use of high-frequency data.

My research has been focusing on gaining new insights on different problems by use of the massive intra-day information characterizing high-frequency data. In particular, I explored two main directions. First, my research deals with the prediction of limit-order book related events in the LOB markets by the use of machine learning (ML) methods. This is precisely the setting of Publication I and Publication II where the join use of different ML methods together with the rich information of high-frequency financial data is exploited to address the problem of mid-price prediction. This aims at shedding light over the general possibility of exploiting high-frequency

data and ML in effectively predicting future events. This setting is entirely data-driven in a way that it differs with respect to the standard methodology adopted in econometrics. Therefore in Publication IV I investigated the prediction problem also under a properly-called econometric approach. Motivated by the recent literature growth about volatility estimation with high-frequency data and the traversal relevance of volatility forecasting in finance, in Publication IV I suggest a possible improvement of a well-established econometric method for daily-volatility forecasting. Publication I, Publication II, Publication IV thus share the same motivation, which is that of forecasting different aspects of financial markets with high-frequency data, although the methodological approach is different. (i) Publication I and Publication II rely on ML methods, whereas Publication IV relies on econometric methods[3] and attains the econometric literature; moreover (ii) Publication I and Publication II investigate the prediction of the mid-price movement, whereas Publication IV the modeling and forecasting of daily volatility.

At this stage, it needs to be pointed out that the above-mentioned forecasting problems, necessarily rely on some data, and in particular in the time-series of one or more variables of interest. The dynamics of the underlying variables are of central importance for developing descriptive and eventually forecasting models: therefore new insights on the properties of high-frequency financial time-series are of high relevance in this context. And of particular importance is having an understanding of the complexity beneath the observed (long-range) correlations and persistence in the time-series (in terms of the relationship between by the current state or variable with its lagged values). Indeed the reference model of Publication IV has been developed also with the purpose of dealing with long-memory effects observed in volatility, however, the assumption about other aspects of the volatility times series are questioned, leading to a revised model. Publication III precisely focus on the specific dynamics observable in some selected high-frequency financial time-series extracted from the LOB data. Moreover, a considerable part of the research presented in this dissertation is cross-disciplinary. Indeed whereas Publication I and Publication II merges ML methods with finance, Publication III represents and application of a method developed in natural sciences.

[3]In a broad sense, not only in the models themselves but for instance also in the backtesting and validation. Publication I and Publication II report measures such as accuracy and precision, typical in the ML filed, while Publication IV uses in-sample an out-of-sample analyses and tests typical to econometric literature.

1.4 Dissertation structure and outline of the original publications

This dissertation continues with five more chapters. Copies of the manuscripts are appended at the very end. In the following Chapter 2, I introduce the reader to the main concepts addressed in the research publications here presented, and expand the literature cited so far. In particular, the discussion in Chapter 2 provides Introduction and general discussions around the concepts of (i) Limit order book markets, (ii) High-frequency financial data and econometrics, (iii) Machine learning application in finance, (iv) Detrended fluctuation analysis, (iv) Volatility modeling and forecasting, and (v) Copula applications in finance. Chapter 3 introduces the two datasets used in Publications I-IV, while Chapter 4 presents the mathematical details of the main methods used throughout Publications I-IV: (i) ML algorithms, (ii) DFA, (iii) HAR model, and (iv) Vine-copulas. Chapter 5 summarizes the findings for all the research publications and answers the research questions postulated in the Introduction. Finally, Chapter 6 concludes by addressing the contribution of the current research, its validity, its reliability, and importantly, its limitations, suggesting directions for future work.

2 KEY-CONCEPTS AND RELATED RESEARCH

The main concepts behind Publications I-IV are here presented to the reader. Sections include the relevant literature review, defining the background for Publications I-IV. The discussion is intentionally kept non-technical, although not generic. However, general mathematical ingredients are provided as well for the purpose of a clear and exhaustive exposition, but of simple understanding. This chapter has to be intended as a complement to Publications I-IV, enriching the discussion therein addressed.

2.1 Limit order book

The vast majority of modern stock exchanges are nowadays order-based electronic markets (Bloomfield et al., 2005; C. Cao et al., 2009; Cont, 2011; Gould, Porter et al., 2013). In limit order markets all the investors can participate in the market: by submitting limit and market orders they can provide or absorb liquidity. On the other hand, in quote-driven markets, the so-called dealers (market makers and liquidity providers) are affecting the liquidity provision (L. Harris, 2003). Whether in quote-driven markets only the ask and bid are on display to market makers or other designates specialists, with clear issues of regarding market transparency, in order driven-markets all the prices at which participants are willing to trade are available, although there is no guarantee of order execution.

The key-ingredient in limit order markets are commitments of investors of buying or selling a predefined quantity of shares at a predefined price. These are respectively called buy and sell limit-orders. Until an investor cancels a submitted limit order or the limit order is completely filled (executed in this whole quantity), the order is

valid and stays in the book. Thus, we refer to limit order book as to the collection, book-keeping of all the outstanding limit orders, waiting to be executed or perhaps canceled. Execution takes place whenever an investor submits a market order. This is a command to either buy or sell a given quantity of shares at the best available price. Importantly, some limit order markets are characterized by a price-time priority rule: an incoming market order is first executed against an outstanding limit order book the best (ask or bid) price[1], while among a number of limit orders with the same best price, priority is given to those first placed (submitted and thus annotated earlier in the book). Market-specific rules determine, for instance, the finest grid over which prices can be placed, generally increments of 1-5 cents (P. Hansen et al., 2017), and the size of the minimum quantity chunk associated with an order. As a consequence of this mechanism, only limit orders at the current best (ask or bid) prices can be matched with a market order[2], while the others stay idle. This means that although a market order gives the certainty of execution, the price at which the execution occurs is not guaranteed (market moves and best prices too); on the other hand, a limit order is not guaranteed to be executed, but if so, the price is settled. The example in Figure 2.1 illustrates the mechanism driving a limit order book[3]. Further information about limit order book markets can be found for instance in (Biais et al., 1995; Cont, 2011; L. Harris, 2003; Gould, Porter et al., 2013) or in (Abergel and Jedidi, 2013; Abergel, Anane et al., 2016; Cont, Stoikov et al., 2010) for a more mathematical description.

2.2 High-frequency financial data and econometrics

The floor-based quote-driven markets have been in the last 20 years replaced by order-driven platforms. Indeed, historically established exchanges such as NYSE, Tokyo Stock Exchange, London Stock Exchanges and more have gradually introduced order-driven platforms. At the same time, and directly related to the mechanism governing

[1] We refer to bid (ask) price as the highest (lowest) price of all the buy (sell) limit orders.

[2] Unless the market order quantity is such that it cannot be entirely filled by the outstanding limit orders at the best price, and thus orders at the second, third and deeper levels are executed too.

[3] In the example and the publications I-III, only time-instances within the continuous-time trading hours are considered. Indeed before the market opening time, an auction period with a structurally different mechanism, where e.g. no market orders can be submitted, determines the price discovery and thus the initial state of the book, after which regular continuous-time trading happens, with the mechanism described above.

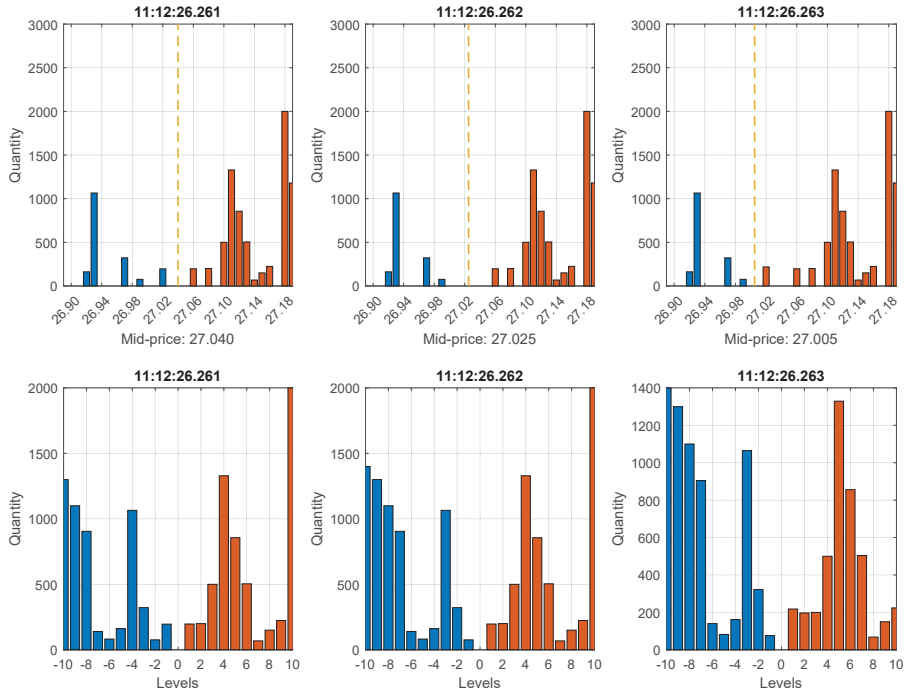


Figure 2.1 Order book state example. Data refers to ticker KESKOB (ISIN code FI0009000202) on June 14, 2010. The first row provides representation in terms of prices and quantities, second rows in terms of levels (limited to the first best ten). Headers are time-stamps, with millisecond accuracy, referring to the epoch t of event arrival. Plots depict the state of the order book at t^+: the updated book state immediately after the event in t. The initial state at $t^+ =$ 11:12:26.261 am is depicted in the first column. In the next millisecond a buy market-order is submitted to the book, the outstanding quantity at level -1 (negative signs stand for the bid side) is therefore traded and the configuration of the book changes (second column), leading to a new mid-price, a wider spread, and new bid and ask prices (respectively 26.99 and 27.06). The spread gap, however, is immediately filled by the submission of a limit-order on the first level on the ask side (third column), which updates the ask price to (27.02), affecting the mid-price.

order-driven markets the frequency of the data increased (Cont, 2011). In other words, the number of events per unit of time dramatically increased, up to thousands per seconds for today's most liquid securities. Nowadays data generated on trading exchanges represent the largest volume outside any other source (Sewell et al., 2008). The trading frequency has increased at the point that nowadays the idea concept of frequency itself is becoming obsolete: indeed data is recorded at its maximum resolution, that is event-by-event, and even when the time scale is extremely small

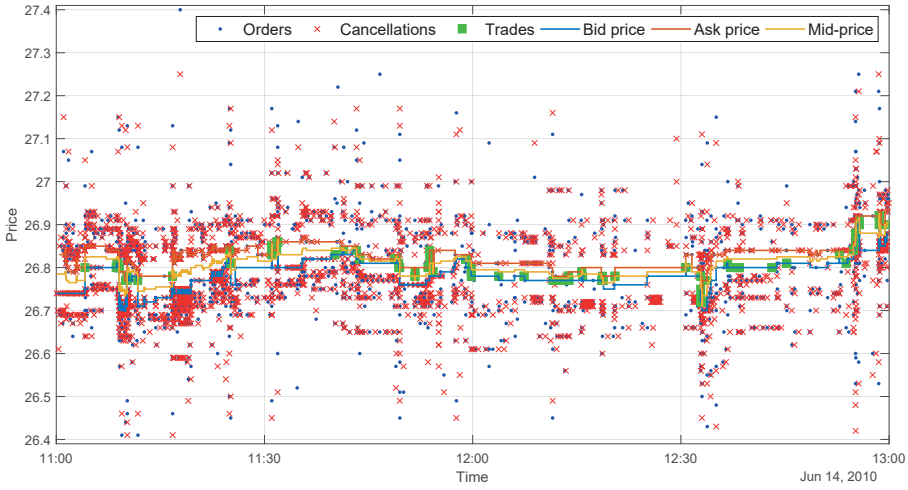


Figure 2.2 Order book flow example for KESKOB June 14, 2010, between 11:00 am and 1:00 pm, illustrating the complexity and multi-dimensionality of limit-order book datasets. Bid and ask levels have been extracted following the procedure explained in Chapter 3. Their average constitutes the mid-price, here included for reference. Note the massive amount of limit order submissions and cancellations, characterizing high-frequency datasets.

(milliseconds scale) and shrinking (Budish et al., 2015). By virtue of the time-priority rule, the precise order of event arrival is known and no information is missing. This is an extraordinary source of data and a complex challenge as well. All this tick-by-tick data being nowadays to researches and practitioner even on real times basis constituting what we call high-frequency data or databases. Besides the specific data-generating process, e.g. order-driven markets, futures markets or currency exchange markets, there are common statistical, econometric, data-handling and processing challenges that the next future needs to address. Statistical analyses on the high-frequency financial data can provide insights on the properties of order inflows, e.g. in microstructure characterization (e.g. P. R. Hansen and Lunde, 2006), uncover the interaction between different trading frequencies and investment horizons (e.g U. A. Müller, Dacorogna, Davé, Pictet et al., 1993; Lux et al., 1999; Alfarano et al., 2007), improve volatility estimation (e.g. Barndorff-Nielsen and Shephard, 2002) and modelling (e.g. P. R. Hansen, Z. Huang et al., 2012). There has been a "race for speed" driven by the sniping opportunities that the mechanism ruling the order-driven markets naturally embed. Despite the social utility and impact of higher speed (Budish et al., 2015), it seems that big-data trend over time-series of higher and higher frequencies with even more dense data is what we should expect from

the future. This trend can however be beneficial for price discovery and market efficiency (Brogaard et al., 2014). Therefore challenges that high-frequency data pose are extremely relevant, contemporaneous and stem for more research to be done.

To further motivate the complexity of the order book data it has to be noticed that an order book high-frequency dataset is indeed made of two subsets. The trades data and the quotes data. The first keeps track of prices and quantities that were actually traded, while the second one keeps track of the outstanding limit orders, i.e. collects all the expressions of markets participants of buying or selling a given quantity at a given price. Furthermore, quote data is affected by cancellations: those can be full or partial depending on whether the quantity associated with a limit order is entirely depleted or not. Figure 2.2 aggregates this complexity: trades correspond to trades data, while all the other marks come from quotes data. Within the two datasets, there are several differences, and with respect to their corresponding time-series sampled at lower frequencies.

Among the most relevant peculiarities of the trade data, is the well-known negative autocorrelation in price changes (well explained in Roll, 1984), which is a direct consequence of price discreetness over of a grid of admissible prices on which orders can be placed. Another aspect is the well-known U-shape intra-day seasonality pattern observed in a number of time-series (e.g. Abergel, Anane et al., 2016, section 2.7 as an example), widely confirmed in the literature. As a consequence, one must either correctly address seasonality (e.g. by obtaining its profile by averaging several days and removing it) or adopt modeling solutions for which stationarity is not critical. Lastly, in recent years many papers analyzed the problem of modeling the duration between the inflow of events. Models such as the ACD (Engle and Russell, 1998) and its variants, or Hawkes-like models e.g. Ogata, 1998 gained popularity since the duration series between transactions is not uniform, nor exponentially distributed but its likely endogenous and depending on complex mechanisms both (i) market-related, to the current state, e.g. spread level and Limit Order Book (LOB) volume (Muni Toke et al., 2017) and the past states (in a sense of filtration generated by the time-series up to a certain time) and (ii) participant-related mechanisms, e.g. how participants react to other participants' trading activity. Furthermore, duration series are not independent but show considerable autocorrelation, e.g. (P. R. Hansen and Lunde, 2006) from the econometrics literature, and (Ivanov, Yuen and Perakakis, 2014) from the econophysics side.

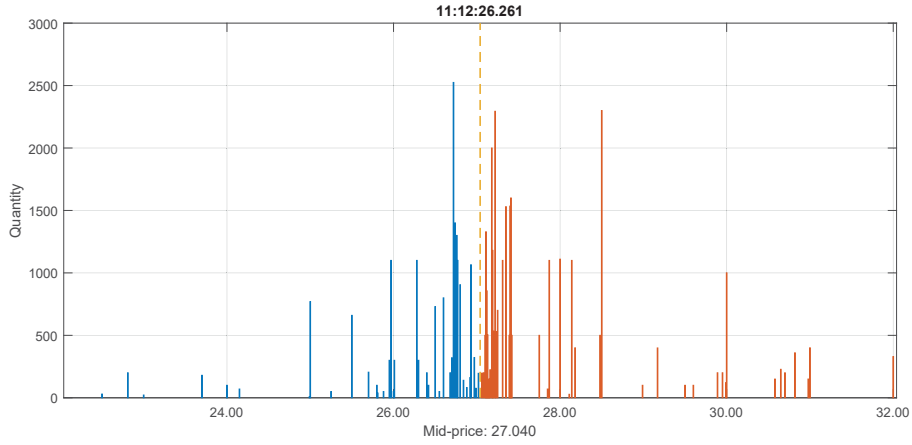


Figure 2.3 An example of limit order-book state for the full-depth of a Level-II LOB dataset (KESKOB, June 14, 2010). There are 40 and 46 price levels on the bid and ask side respectively.

Quote data is of higher complexity. The simplest quote data is the so-called Level-I data consisting of five variables, namely timestamp, bid price, ask price, bid quantity ask quantity. Level-I quote data, therefore, is a complete track of the dynamics of the bid and ask levels (best levels or level I) only. Nevertheless, the information Level-I data conveys is limited and partial, indeed updates in the order book happen ad any distance from the current best quotes. A deeper limit order book data, comprehensive of the all information for the ten best levels, is referred to as Level-II data, see Figure 2.3. It is clear that either Level-I and Level-II data is of great complexity and constitutes a challenges since the data therein contained corresponds of a sorted list of limit order submission and their updates (deletion because of a matching market order or cancellation), which does not provide any *direct* information about the current best levels and the deeper (non-empty) levels in the book. Information such as the bid price or quantity at the second bid level needs to be inferred and reconstructed from the raw feed. For very liquid assets this corresponds to a complex and time-consuming operation[4]. Clearly, levels can be sparsely filled and the reconstruction needs to be separately done for the ask and bid side. A visual representation of the full-depth LOB data in its flow domain is provided in Figure 2.2.

[4]To reconstruct the book one needs to separate ask and bid orders. Sort them by price and time, the number of unique prices correspond to the number of levels of the book at a particular time, the lowest one to the bid price (for bid orders). Within the hundred or thousands of orders sharing the same price, the sum of their respective quantities gives the total quantity at a given level. This is how e.g. Figures 2.1and 2.3 are obtained

As Figure 2.2 remarks, modeling the dynamics and forecasting limit order books is, clearly a complex task. Two major approaches have been investigated: machine learning methods and econometric models (see Section 2.5). Cont, 2011 provides a comprehensive introductory overview of the major modeling approaches for the order book dynamics. This generally focuses on queue models and point-processes models (Cont, 2011, and references therein). Among the first ones, as an example of econometric modeling, (Cont, Stoikov et al., 2010) provides a mathematically tractable and analytic method for the prediction of various events (although relying on simplistic hypothesis, as Poissonian arrival rates and stationarity for the processes therein involved).

There is a further filed in which the availability of high-frequency data has been shown to be critical. This is the field of econometrics generally known as "market microstructure". Market microstructure examines how markets work in practice and what are the mechanism driving, determining and affecting, prices, volumes, costs, and the overall aggregated trading behavior (Hasbrouck, 2007). Here we identify the econometric literature going back to the works of Roll, 1984; O'hara, 1995; R. D. Huang et al., 1997 about the price formation, price impact and spread components. Barndorff-Nielsen and Shephard, 2002 were among the first suggesting the use of high-frequency data for volatility estimation. However, with this purpose the impact of microstructure effects is of central importance and in general have disruptive effects on common volatility estimators proper of non-high-frequency literature where there is no distinction between the so-called efficient price and the observed price (Hasbrouck, 2007, among the others). Indeed the problem of robust volatility estimation under market microstructure frictions (noise) and the characterization of the microstructure effects at high frequencies is one of the most prolific fields in the contemporary econometrics. For a review on volatility estimation see for instance (McAleer et al., 2008b; Bucci et al., 2017), for a discussion on market microstructure and microstructure noise, see in this regard e.g. (P. R. Hansen and Lunde, 2006; Hasbrouck, 2007). Importantly, advances in high-frequency volatility estimation, market microstructure noise characterization, contributed also to other areas, such as that of jump detection and estimation (e.g. Barndorff-Nielsen and Shephard, 2006; Lee et al., 2007, among the major contributions).

2.3 Machine learning for price prediction with limit-order book data

The use of Machine learning (ML) in finance is becoming more apparent every day. Among the broad field of data science, we can identify a fast-growing sub-filed that by use of statistical methods aims at the automated and powerful detection of patterns in the data, converting expertise (input data) into knowledge. The use of machine learning in finance is attractive both in replacing routine tasks so far performed by humans and in performing tasks that go beyond the human capability, such as the analysis of complex datasets and challenging prediction problems (Shalev-Shwartz et al., 2014, Chapter 1). Refer to Table 1 in (Bose et al., 2001) for a comprehensive list of applications, not only limited to the financial sector.

2.3.1 Machine learning for mid-price prediction

Concerning Publication I and Publication II, different approaches have been developed with respect to the mid-price prediction task. Several applications have discussed network-based methodologies. Recurrent Neural networks have shown to be effective for predicting future price movements in (Tsantekidis et al., 2017b), while (Tran, Iosifidis et al., 2018) introduces in the network an attention mechanism that allows the layer to detect and retain only the crucial information in the temporal mode, of high efficiency and suitable for practical real-time applications. The impact of data-normalization is discussed in (Passalis, Tefas et al., 2019) where a neural layer able of adaptively normalize the input data is proposed. Automation of feature extraction is also addressed under a Convolutional Neural Network (CNN) setting in (Z. Zhang et al., 2019), interestingly pointing out the possibility of a universal set of features: indeed their model generalizes well to instruments that were not included in the training data. (Nousi et al., 2018) however argues that a combination of hand-crafted features and automatically extracted ones improves the mid-price prediction the most. (Sirignano, 2019) provides an extensive analysis of over 500 stocks supporting a new neural network methodology taking advantage of the spatial structure of the limit order book ad a model for price movements. (Tsantekidis et al., 2017a) bases the

prediction of mid-price movement on CNNs with an improved data normalization technique (updated on daily basis and differentiated between prices and volumes), showing its superiority in forecasting with respect to methods like support vector machine (SVM) and Multilayer neural networks. Classification of mid-price movements has been also addressed by the use of Recurrent neural networks in (e.g. Dixon, 2018a; Dixon, 2018b) and deep neural networks (e.g. Dixon et al., 2017). In (Dixon et al., 2017) the applicability of complex and multi-layered architectures is showed to be scalable yet feasible for high-dimensional analyses involving up to 9895 features extracted for 43 symbols and FX mid-prices (sampled every 5-min) aimed at capturing their historical co-movements observed over almost 20 years of data. Although the network involves more than 1.2×10^7 weights, it's shown to be of overnight-trainability with proper C++ implementation and a performant processor, and easily applicable to buy-hold-sell intraday trading strategies. SVM applications include (Kercheval et al., 2015), designing input-vectors of time-sensitive and time-insensitive features to address different prediction tasks, including mid-price movements, by use of data for a single trading day for five stocks traded at NASDAQ. A hybrid classifier combining SVM and ARIMA that models nonlinearities in the class structure is provided by (Pai et al., 2005). The relationship between limit-order book variables and mid-price movements have been analyzed in (Cenesizoglu et al., 2014), different order book variables extracted from a one-year-long, 20-levels-deep high-frequency LOB show that it is possible to obtain economic gains from the relation between limit order book variables and mid-price return. Furthermore, they provide a Granger-causality between mid-price movements and lagged order book variables, supporting the use of lagged LOB dynamics for forecasting purposes. Short-term forecasting of the mid-price change is also discussed in (Palguna et al., 2016), where non-parametric predictors implementing features conditioned on the state of the order book show significant gains when incorporated into existing order execution strategies. Besides the above-mentioned methodologies, Passalis, Tsantekidis et al., 2017 develop an extension of the BoF (bag of features) model with a neural layer, which receives the features extracted from a time-series and gradually builds its representation, as a radial basis function (RBF) layer and an accumulation layer. This methodology is demonstrated to have a great positive impact on classification accuracy w.r.t. BoF for the mid-price prediction problem, also on a second data set, being highly scalable mixable with different classifiers. A different direction in improving the BoF methodology

is discussed in (Passalis, Tefas et al., 2018) where different RBF and accumulation layers enable to capture both short-term and long-term dynamics of the time-series. Discriminant analysis techniques for the same prediction task have been discussed in (Tran, Gabbouj et al., 2017).

2.3.2 The role of deep-in-the-book data and market microstructure in mid-price predictability

The above subsection pointed out that the effective prediction of mid-price movement in LOB markets is a task that can be successfully accomplished under different ML approaches. But to which extent wide datasets including deep-in-the-book information are relevant? I.e. how informative are the best levels themselves? Furthermore, are ML methods indeed discovering informative hidden and complex patterns in the data or, rather, there are some inborn microstructure features (e.g. price impact and toxicity) that implying some predictability congenital to high-frequency domains? With respect to the order flow in capability in conveying useful information for predictive purposes, i.e. to which extent data and features referring to deeper levels of the book are informative. C. Cao et al., 2009 by using one-month data with a one-hundredth of a second resolution from the 100 most actively traded stocks at the Australian Stock Exchange, applying the information share method of (Hasbrouck, 1995), find that 78% of the contribution to price discovery is from the best bid and offer prices on the book and the last transaction price. Indeed, they find that especially the section of the book that is near the top, helps investors to predict future short-term returns, thus being directly relevant for mid-price prediction, while deeper levels are found being moderately informative, but however significantly related to future returns, as well as imbalances between demand and supply, the latter also observed in (Cont, Kukanov et al., 2014). Also (Glosten, 1994; Seppi, 1997) similarly conclude about the mild informativeness of the order flow, and e.g. (Gould and Bonart, 2016) obtains very satisfactory results on the prediction of one-tick-ahead mid-price movements with simple logistic regression by using the information at the best-levels queues, following study of (Yang et al., 2016). On the other hand, several studies suggest that predictive power is conveyed in the order book beyond the best levels (including e.g. Parlour, 1998; Bloomfield et al., 2005; L. E. Harris et al., 2005; Kaniel et al.,

2006; Cont, Stoikov et al., 2010; B. Zheng et al., 2013; Cont, Kukanov et al., 2014; Cenesizoglu et al., 2014; Kercheval et al., 2015; Dixon, 2018b; Sirignano, 2019). It has to be pointed out that however, on a general level, these studies do not point out the contribution of the first levels to the mid-price prediction ability as a separate analysis on a reduced set of features or variables, so do not not exactly discern the contribution of the first levels to the overall forecasting results. E.g. Kercheval et al., 2015 discuss and present their results for the overall set of features they consider, including those related to the best levels; (Cont, Stoikov et al., 2010) develop an analytically tractable model that is shown to closely reproduce empirical probabilities observed in the order book, but the corresponding results over a reduced Level-I dataset involving best levels only are not reported. To provide a satisfactory yet limited outlook on the informativeness of the order flow, the so-called order flow toxicity cannot be left out. An order flow where the quick filling of passive orders when they should fill slowly is regarded as toxic (Easley, López de Prado et al., 2012). The above-mentioned researches, generally adopting high-frequency data, confirm the insight from the microstructure literature that the order arrival process is informative for subsequent short-term price moves and order flow toxicity (Easley, López de Prado et al., 2012). Metrics related to the order flow toxicity have been showed to be useful indicators for short term market variables. E.g (Easley, López de Prado et al., 2012) show that their VPIN toxicity measure has significant forecasting power over toxicity-induced volatility, and since toxicity may cause the withdrawal of (uninformed) market participant from trading causing liquidity issues, has a close connection with financial crashes and risk management (Easley, De Prado et al., 2011). (Van Ness et al., 2017) find VPIN measure[5] being a good predictor of high frequency traders' liquidity supply and demand changes (directly related to mid-price movement, given the above discussion on the role that supply and demand play wrt. short-term mid-price movements). In tick-by-tick limit order book data from the Australian stock exchange, Wei et al., 2013 recently found evidence of Granger causality from VPIN quote imbalance, price volatility ad inter-trade duration, extending the findings of (Easley, López de Prado et al., 2012)[6].

Finally, price impact has been an active area of econometric research, closely related

[5]Generally accepted as the standard measure for toxicity, although discussed to be inadequate in (Andersen and Bondarenko, 2014).

[6]Which involves in its computation the order flow imbalance, so is not surprising that both VPIN and imbalances in the supply and demand have been shown to convey predictive information.

to market microstructure and the information trades convey about future market movements (e.g. seminal works are those of Kyle, 1985; Hasbrouck, 1991; Dufour et al., 2000; Hasbrouck, 2007). Price impact refers to the correlation between an incoming order (to buy or to sell) and the subsequent price change: that a buy (sell) trade should push the price up (down) seems at first sight obvious and is easily demonstrated empirically (Bouchaud, 2010). Price impact problem is complex and manifold, a complete review of the price impact problem and modeling is here out of scope, a comprehensive review is provided in e.g. Bouchaud, Farmer et al., 2009. (Hasbrouck, 2007, section 5.6) remarks the relevance of market impact for predictive applications, "orders do not impact prices. It is more accurate to say the orders forecast prices.". Indeed, in the light of the above definition and intuition, the use of market impact as a valuable ingredient conveying predictive informativeness is clear, e.g. a clear connection between price impact and price adjustment is provided by (Kyle, 1985, e.g.), or in (Hardiman et al., 2013) under the assumption that order arrivals follow a Hakes process. Market design, trading rules, participants' behavior, trading strategies and different levels of information, all contribute to price impact, intuitively, but also empirically, associated with the incoming flow and Price dynamics: implicitly market microstructure defines a relationship between the current trading and market (yet very complex and difficult to model), which in some extent informative of future movements. This can be exploited, and in some parts motivates mid-price prediction applications, although price impact mainly concerns best quotes dynamics, last order's price and volume, rather than deep levels of the book, and related to factors such as information asymmetric in the participants which are difficult to capture.

Price impact is however an aspect that traders would like to avoid. Think of selling n shares over a certain time horizon h. Placing a market order immediately, if n is large this will give progressively worse prices for subsequent shares as the buy side of the book is eroded. Alternatively, one might split the n is different chunks and submit a series of market orders, however the price might move and for each order transaction costs are paid. Furthermore, patterns in order submission when splitting n in smaller chunks may be detected by other investors, which become informed and may want to exploit this information. A balance between market impact (trading too quickly and moving the price), market risk from exogenous price movements (if trading too slowly), urgency of trading within h and transaction costs has to be accounted

for. Note that liquidity takers, dividing their volume in smaller quantities lead to empirically well-documented long-range correlations in trade signs (e.g. Bouchaud, Gefen et al., 2004).

In this light, optimization of the trade execution becomes important: several authors discussed this problem (e.g. Almgren, 2003; Hewlett, 2006; Nevmyvaka et al., 2006; Engle and Ferstenberg, 2007; Hasbrouck, 2007; Bouchaud, Farmer et al., 2009, among several others). The optimization strategy between competing goals with a clear objective (execution at best possible prices, possibly rapidly, and minimization of the costs) is that of finding a suitable action model that would maximize the total reward of the agent. This flawlessly fits reinforcement learning's (RL) realm (see e.g. Sutton et al., 2018, as classical reference). In a trial-an-error setting this ML *algorithm* finds the *policy* that for an agent in a certain *environments* maximize the *reward* of taking an action in a particular *state*. Italicized elements are the key-ingredients in RL modelling, all having an immediate correspective in the financial problem above discussed. Applications of RL in limit order book are those of (e.g. Nevmyvaka et al., 2006; S.-H. Chen et al., 2011; He and Lin, 2019).

2.3.3 Mid-price prediction in high-frequency setting: an opportunity for latency-based arbitrages

Latency defines windows of time for faster traders to use their information advantage to trade with orders at stale prices, earning a low-risk profit at the expense of slower investors. Exchanges may digest incoming data-flows slowly than the fastest traders, which can execute at stale prices. Similarly, market states as at the exchange may appear at investors' display with a considerable delay, based on communication lines' transmission performance and vicinity to the exchange (e.g. Hasbrouck and Saar, 2013; Budish et al., 2015). The information advantage triggering latency-motivated trades is based on knowledge of how the price of a stock is about to move, and on the sniping of lower-speed traders. Reasonably, latency jointly with accurate forecasts on price movements are low-risk and perhaps almost instantly profitable. The latency phenomenon and arbitrage opportunities high-frequency trading can exploit, are well analyzed in (Budish et al., 2015), along with historical analysis on empirical data, showing clear trends towards increasing trading speed. Therefore the effective use

of market direction forecasts obtained from any algorithm and method are clearly bounded to this framework. In particular on how efficiently and readily they and process new data from the information stream and return outputs and on the speed these outputs are transformed into action and submitted to the exchange, in relation to other investors' speed. Whereas computing infrastructures matter, as well as the physical vicinity to the exchange and incoming traffic, the main player in this game is the prediction algorithm itself. Fast trainability and the efficient update of the prediction given the new information inflow are elements of concern. Indeed, when a full-model calibration can be performed overnight (e.g. Dixon et al., 2017), it rather important that given the estimated parameter the processing of the new information is efficient (assuming stationarity for the overall relationship between features and mid-price dynamics). It has to be pointed out that however efficient implementations (e.g. in `C++`, perhaps the fastest language) and a well-engineered and optimized parallel-computing infrastructure can exploit short arbitrage windows arising from latency. Remarkable are the implementations of recurrent neural network for mid-price movement classification applications of (Dixon, 2018a; Dixon, 2018b) (both relying on a 32-features set), rapidly estimated in around 10^{-4} seconds, allowing for actually profitable trading strategies. Response time for fastest low-latency traders were observed to be around 2-3 milliseconds already a decade ago (Hasbrouck and Saar, 2013, using data from 2007), while nowadays typical latency windows between 0.1 and 1 milliseconds (Dixon, 2018b, e.g.). Further estimates of latencies, as of a decade ago, on different markets are reported in (e.g. Ende et al., 2011). The SVM application of (Kercheval et al., 2015), is also promising this direction: its (un-optimized) classification of unseen data point takes less than 10^{-3} seconds.

Note that a broad-mindset inspired research aimed at addressing both satisfactory and effective prediction of market movements while looking at the actual employability of the method for practically exploiting arbitrages, i.e. committed to efficient implementation and attention on the implementation and selection of the computing facilities to implement trading strategies within typical latency windows (as in e.g. Dixon, 2018b), is of particular relevance for practitioners. Aspects being easily neglected from experts in either machine learning (overlooking the importance of application) or finance experts (overlooking the importance of implementation). Clearly, this is just one side of latency-based arbitraging. A suitable strategy needs to be adopted and the effects of type I and II errors must be established to evaluate an overall profitability.

Furthermore, a realistic settings accounting for transaction costs, the possibility of insufficient liquidity and sufficient cash in the brokerage account, independent on the realization of the profit, must be taken into consideration.

2.4 Long-range autocorrelation and fractality

This section first introduces the reader the concepts of long-range autocorrelation, fractality, self-affinity and scaling exponent (common references are Feder, 1988; Kantelhardt, 2009), then presents the relevant financial literature regarding the long-range dependence detection and analysis.

2.4.1 Long-range autocorrelation and scaling exponent

The dynamics of the times series depicting a complex system is often characterized by scaling laws, over a continuous range of time scales and frequencies (Kantelhardt, 2009). A system characterized by self-similar structures (*self affinity*) over different time-scales is called a *fractal*, i.e. a magnification of a small part is statistically equivalent to the whole (Kantelhardt, 2009). Financial time-series are an example of such complexity: their highly stochastic nature of complex dynamics and their susceptibility to exogenous factors often emerges as time-dependent properties or, more generally, non-stationarity. Self affinity and persistence are closely related: persistence holds for all time scales, where self-affinity holds. A key ingredient in determining the degree of persistence is provided by the autocorrelation function. For stationary data (therefore for data where mean and variance do not change over time), the autocorrelation function for the increments Δx_i, of some process $\{x_t\}_{t=0,\dots,N}$:

$$C(s) = \frac{1}{N-s} \sum_{i=0}^{N-s} \Delta x_i \Delta x_{i+s} \tag{2.1}$$

characterizes two types of correlations: short-range correlation and long-range correlation. An exponentially declining autocorrelation function of the type $C(s) = e^{-s/T}$ catheterizes short-range autocorrelation (examples are found for instance in AR

processes), whereas a power-law decay characterizes long-range autocorrelation:

$$C(s) \propto s^{-\gamma} \tag{2.2}$$

with $0 < \gamma < 1$. The exponent γ is referred to as the *scaling exponent*, and more generally a scaling law with a scaling exponent α, describes the behaviour of a certain quantity F in terms of the parameter α: $F(s) = s^{\alpha}$. A system characterized by a non-integer scaling exponent is called *fractal*. Whether persistence, in general holds at least for a range of time-scales, in long-range correlation the decay is sufficiently slow such that a characteristic correlation time scale (or range) where the persistence holds cannot be defined, i.e. $C(s) \propto s^{-\gamma}$, at least asymptotically.
Finally the notion of self-affinity. The data is self-affine, whenever for a given factor a the following scaling relation holds(Kantelhardt, 2009):

$$x(t) \mapsto a^{H} x(at)\,. \tag{2.3}$$

Re-scaling time t by a factor a, requires re-scaling $x(t)$ by a factor a^{H}, to obtain a statistically equivalent magnification (Kantelhardt, 2009). The type of self affinity is driven by the Hurst exponent H. For long-range correlation, $H = 1 - \frac{\gamma}{2}$. Therefore self affinity with $\frac{1}{2} < H < 1$ corresponds to long-range autocorrelation. In strict sense, self-affinity encompasses fractality, i.e. the terms are not equivalent. In practice, literature refers to "fractal" whenever H can be defined. Similarly long-range dependence, long-range correlation/autocorrelation and power-law decay are used instinctively, all referring to eq. (2.2).

Several series are not characterized by a unique exponent H, thus they are not self-affine: we talk of crossovers in the scaling exponent, identified by different H applicable at different time-scales. Crossovers can be be caused by non-stationarity in the data as well. Violations of either weak stationarity or strong stationarity can impact on the estimate of H (and γ, or in general α), therefore methods robust to e.g. trends are needed to correctly estimate the scaling exponent. Traditional approaches for estimating the scaling exponent under stationary time-series are (i) autocorrelation function analysis,not advisable since affected by the generally superimposed noise on the underlying process, and because at large time scales is of high variability (Kantelhardt, 2009), (ii) Hurst re-scaled range analysis (classical references are Hurst, 1951; Mandelbrot and Wallis, 1969; Feder, 1988), (iii) Spectral analysis (Taqqu et al., 1995;

Rangarajan et al., 2000; Hunt, 1951) and (iv) fluctuation analysis (FA) (Peng, S. V. Buldyrev, Goldberger et al., 1992). Complex financial series can potentially preset non-stationary features (e.g. due to seasonal patterns or structural breaks). Therefore estimation methods robust to non-stationarity are preferable options, among these, there are methods relying on wavelet analysis (Koscielny-Bunde et al., 1998; Kantelhardt, Roman et al., 1995) and the DFA (Peng, S. V. Buldyrev, Havlin et al., 1994). The DFA (a detrended, non-stationarity robust version of FA) constitutes the methodological method of Publication III and constitutes the most used methods for the estimation of the scaling coefficient. Chapter 4 discusses DFA in detail. Comparisons between the methods are provided in a number of studies (among them Heneghan et al., 2000; Delignieres et al., 2006; Mielniczuk et al., 2007; Bashan et al., 2008; Shao et al., 2012).

2.4.2 Applications in finance

The earliest discussion about long-range dependence properties in financial prices is motivated by the empirical evidence that the very early theories were failing to capture phenomena such as non-normality of returns, their fat tails, correlation in prices, cyclical patterns, and general evidence that prices do not behave as random walks. (Mandelbrot, 1997) is a collection and discussion over the early contributions to the development of a following wide literature in this direction. Following analyses are for instance those of (Mandelbrot, 1971), where investigating the problem of market inefficiency the slow-decaying correlation of in the price process is discussed. However, is much later when a financial literature based on these early findings exploded, i.e. when the early contributions to economic literature in the field of time dependence, cyclical pattern, and speed of the autocorrelation decay, met the math-physics literature on fractals (which goes back to e.g. Hurst, 1951; Mandelbrot and Wallis, 1969), also thanks to new methodological developments, such as (Peng, S. V. Buldyrev, Havlin et al., 1994; Taqqu et al., 1995). The following reviews some of the applications in finance, while the latter part focuses on DFA and analyses related to duration time-series, since relevant with respect to Publication III.

Mantegna et al., 1995 analyze six-years S&P 500 index data, finding that the scaling exponent of the power-law is constant over the same period and, that the distribution

of the differences in the index is not Gaussian. Returns have been analyzed in (Grau-Carles, 2000), where a number of methods are utilized to conclude that little temporal correlation is observed in return's series. Similar conclusion on the returns' scaling properties are drawn in e.g. (Grau-Carles, 2001; G. Oh et al., 2006; W.-X. Zhou, 2009), confirming the intuition that financial returns are of difficult predictability, neither weakly dependent nor related to most of the economic variables (Cont, 2005), while (Carbone et al., 2004) expands the complexity in the discussion by devising variability in the scaling exponent on high-frequency returns from the German market. On the other hand, (Grau-Carles, 2000) finds strong long-range dependence arising in the volatility series, aligned with a number of later analyses (e.g. Cizeau et al., 1997; Y. Liu et al., 1999; Yamasaki et al., 2005), Y. Wang et al., 2009 point out that under long-range dependence, standard econometric models such as GARCH and EGARCH are inadequate. Early application on exchange rates are those of Vandewalle, Ausloos and Boveroux, 1997; Vandewalle and Ausloos, 1997. More recently (G.-J. Wang et al., 2013) expanded the analysis to cross-correlation between a basket of currencies as well, while G. Cao, Xu et al., 2012 study the cross-correlation between Chinese stock and exchange markets. Applications in studying market efficiency via DFA include, e.g. Stošić et al., 2015; Tiwari et al., 2017; Y. Wang et al., 2009. Interestingly, (Lillo et al., 2004) shows that although the long-range correlation found in order placement in London Stock Exchange would stem for market inefficiency, there are long-range anti-correlations in trade size and liquidity whose overall netting effect drives the market closer to efficiency. Related analyses on coexisting factors implying long-range correlations and anti-correlations within the order book have also been proposed in (Bouchaud, Gefen et al., 2004). Furthermore, (e.g. Lillo et al., 2004; Bouchaud, Gefen et al., 2004; Bouchaud, Farmer et al., 2009; Tóth et al., 2011), discussed long-range autocorrelations, in particular those found in trade signs, relating it to price impact and on its persistence. Long-range correlation analyses has also concern business cycles, market periods (Czarnecki et al., 2008; Qian et al., 2004) and can identify forthcoming crashes (e.g. Grech et al., 2004). Connections with the common econometrics time-series literature are discusses for instance in (Torre et al., 2007; Podobnik et al., 2008). Further references on long-range autocorrelation analyses can be found in (e.g. Lillo et al., 2004, Section II), while to (e.g. Baillie, 1996) for a general review on the econometric approach for long-range correlation modelling.

2.4.3 Applications in duration analysis

While the above is a general overview of the different application of long-range analyses in finance, the following focuses on the most relevant literature related to inter-event long-range correlation analyses, concerning Publication III.

Inter-trade times for a three-year sample of 30 stocks extracted from the TAQ database, were analyzed in (Ivanov, Yuen, Podobnik et al., 2004), which analyzes the scaling properties of the density function of inter-trades duration, finds that the inter-trade times exhibit power-law correlated behavior within a trading day, devising the possibility of universal scaling patterns common to different industry sectors. Importantly, their results provide the first crossover evidence in the scaling exponent extracted via DFA. Inter-trade durations were also considered in (Ivanov, Yuen and Perakakis, 2014) for stocks traded at NYSE and NASDAQ, power-law correlations in inter-trade times are influenced by the market structure and coupled with the power-law correlations of absolute returns and volatility, motivating the association analysis in Publication III. For inter-trade durations, also (Jiang et al., 2009) finds strong evidence of crossovers between two different power-scaling regimes from 23 Chinese. Fractal properties of inter-cancellations durations have been analyzed for 18 stocks in Shenzhen exchange in (Gu et al., 2014), using different variants of standard DFA, and in (Ni et al., 2010), both confirming long-range autocorrelation in cancellation series.

2.5 Volatility estimation and modelling in high-frequency settings

2.5.1 Volatility estimation

High-frequency data provide a valuable source of intra-day information. Intuitively all the intra-day information nowadays available can be exploited to refine and improve estimates on volatility over a given period, generally set to a trading day. This section provides a review on the volatility estimation problem at high-frequencies, i.e. under market microstructure noise.

Rather than on the volatility diffusion term (usually denoted with σ_t) involved in a generic price model, which is difficult to capture, the recent econometric literature has focused on a closely-related quantity, known as Integrated Variance (IV). The IV naturally constitutes a volatility measure, synthesizing in a single quantity (the IV itself) the complex dynamics of the underlying random and unobserved process σ_t over a time window.

To fix the ideas, consider a standard price model of the form $dS_t = \mu S dt + \sigma_t dW_t$ (geometric Brownian motion), where μ is a (negligible) drift term, σ_t the volatility diffusion and W a Brownian motion. The integrated variance over an interval $[0, t]$ (generally a day) is thus defined as:

$$IV_t = \int_0^t \sigma_s^2 ds \ . \tag{2.4}$$

With σ_t finite and bounded. The stochastic theory of quadratic variation naturally identifies a simple and feasible estimator for the IV[7], the so-called Realized Variance (RV):

$$RV_n = \sum_{i=1}^{n} r_i^2 \ , \tag{2.5}$$

where the price over $[0, t]$ is sampled in n intervals and r_i is the return over the i-th interval $r_i = p_i - p_{i-1}$. As n grows, therefore as prices are sampled more and more frequently, RV_n converges in probability to IV_t. The idea of defining a volaility estimator through a sum of squared returns goes back to (Merton, 1980), however the term "Realized" was first introduced by(Andersen, Bollerslev, Diebold and Labys, 2000a; Andersen, Bollerslev, Diebold and Ebens, 2001) The realized variance therefore is an immediate estimator for the IV, our volatility measure, and is importance of high-frequency settings is clear: the higher the sampling frequency the better is the estimate RV_n provides of IV, so the smaller the error (Barndorff-Nielsen and Shephard, 2002):

$$\operatorname*{plim}_{n \to +\infty} RV_n = IV_t \ . \tag{2.6}$$

The results in eq. (2.6) generally holds for for an underlying log-price processes p_t defined as a semi-martingale. At high frequencies however, the log-price p follows a different dynamics. A disturbance term affects the efficient semi-martingale price.

[7] Most of textbooks in stochastic analysis address this point.

In the high-frequency econometrics literature the observed price p_t is therefore decomposed in two parts, the efficient price (p_t^*) (which follows a semi-martingale, such as the geometric Brownian motion) and an error term ε_t (e.g. Hasbrouck, 2007; P. R. Hansen and Lunde, 2006; L. Zhang et al., 2005; Aït-Sahalia and Jacod, 2014, for instance):

$$p_t = p_t^* + \varepsilon \,. \tag{2.7}$$

The error term, negligible at low sampling frequencies, becomes relevant at high-frequencies and is broadly imputed to market microstructure noise (Hasbrouck, 2007) and includes (i) frictions, e.g. bid-ask bounce, price-discreteness and truncation, issues related to trading on different networks (ii) informational effects, e.g. different price-responses to block-trades or inventory control costs, and (iii) errors, e.g. entries with zero-price, misplaced decimal points (Aït-Sahalia, Mykland et al., 2011). As a consequence, the estimation of the IV via RV becomes biased and inconsistent (e.g. L. Zhang et al., 2005), because of the noise term, although it holds for the efficient price p_t^* which is however unobservable. The volatility signature plot (Andersen, Bollerslev, Diebold and Labys, 1999) in Figure 2.4 clearly highlights a bias in RV when approaching higher sampling frequencies.

The impact of the error at low frequencies is negligible in the IV estimation through RV (e.g. Andersen, Bollerslev, Diebold and Labys, 2000b) in such a way that its random effect is canceled out (ε is usually taken to be of mean zero) as its impact on low-frequencies returns is considerably small. However at high frequencies ε biases the RV measure. Effects such as the bid-ask bouncing will inevitably generate sequences of high-frequency returns as wide as the spread: these price movements are not imputable to price volatility, rather to microstructure effects (a clear statement in the context of Roll, 1984).

Whereas a zero-mean hypothesis on the error term is reasonable, an iid. framework (including by Bandi et al., 2008; L. Zhang et al., 2005) is simplistic. Empirical investigations have shown that, for instance, the variance of the noise is time-varying, that the ε_t process is negatively auto-correlated, and that is correlated with the efficient price itself (P. R. Hansen and Lunde, 2006; Aït-Sahalia, Mykland et al., 2011). Besides non-stationarity, there are also diurnal effects (Jacod, Y. Li and X. Zheng, 2017; Kalnina and Linton, 2008) and heterogeneities among stocks (P. R. Hansen and Lunde, 2006; Aït-Sahalia and Yu, 2008) to complicate the discussion. The empirical

properties of the market microstructure and consequent theoretical assumptions drawn on it play a central role in designing estimators for the integrated variance that are noise-robust with desirable properties (e.g. P. R. Hansen and Lunde, 2006; Barndorff-Nielsen, P. R. Hansen et al., 2008).

A number of estimators have been developed to disentangle the contribution of the microstructure noise in the observed high-frequency returns and thus to provide a consistent and unbiased estimator for the IV, our target volatility measure. Several approaches have been developed. Besides the choice of the best sampling method, for instance, transaction-time sampling (e.g. R. C. Oomen, 2005; R. C. A. Oomen, 2006; Pooter et al., 2008), the presence of noise practically implies a trade-off between sampling frequency, which we would like to be as high as possible in virtue of eq. (2.6) and the effect of the microstructure error. A remedy is to sample at moderate frequencies, where the noise impact is small: sparse sampling at e.g. 5 minutes (or higher frequencies) is a commonly adopted scheme (L. Zhang et al., 2005; Ait-Sahalia et al., 2005). Several works, therefore, have investigated the effect of sampling and the possibility of determining an optimal sub-sampling scheme (e.g. Bandi et al., 2005; Bandi et al., 2008; L. Zhang et al., 2005; Kalnina, 2011). Among these we find the two-scales estimator (L. Zhang et al., 2005), and its multi-scale extension (L. Zhang et al., 2006; Aït-Sahalia, Mykland et al., 2011). Pre-averaging estimators (Podolskij et al., 2009; Jacod, Y. Li, Mykland et al., 2009; Jacod, Podolskij et al., 2010) on the other hand, pre-average a number of returns to reduce the microstructure impact, and use them in estimating the IV with a (properly re-scaled) RV-like approach. A further approach for the robust estimation of the IV is the kernel approach (B. Zhou, 1996; P. R. Hansen and Lunde, 2004), leading to the so-called realized kernel of (Barndorff-Nielsen, P. R. Hansen et al., 2008). Nevertheless, several other approaches have been developed, e.g. the Fourier-based approach (Malliavin et al., 2009; Barucci et al., 2002), and pre-filtering methods for the intraday-returns (e.g. Bollen et al., 2002; Andersen, Bollerslev, Diebold and Ebens, 2001). Figure 2.4 shows how some selected estimators behave at increasing sampling frequencies, underlying the impact of the MMS noise as a source of bias. Note that whereas the observed price process is in the truth of unknown exact nature, realized measures have been developed under precise noise assumptions: considerable deviations of high-frequency measures w.r.t. to low-sampled and noise-free RV are possible.

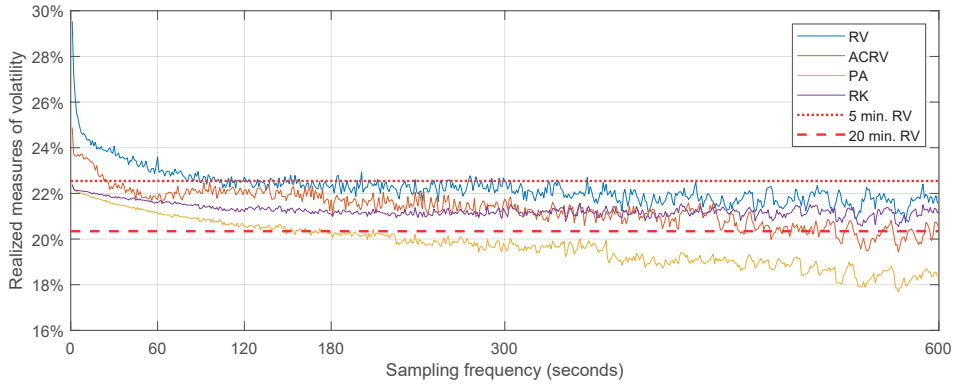


Figure 2.4 Volatility signature plot, depicting the behaviour of different realized *volaility* measures w.r.t. the sampling frequency. Realized measures include (B. Zhou, 1996) - ACRV, (first-order) autocorrelated realized variance, the (Jacod, Y. Li, Mykland et al., 2009) - PA, pre-averaging estimator, and (Barndorff-Nielsen et al., 2009) - RK, realized kernel. Realizes Volatility measures, i.e. square-roots of the corresponding realized variance measures, are annualized for convenience. 5 and 20 min. RV are provided for reference (justified e.g. in L. Zhang et al., 2005; Barndorff-Nielsen et al., 2009, respectively). Average series over twenty days for stock AAPL, corresponding to May 1, 2012; Trades And Quotes data.

2.5.2 Volatility modelling

Different approaches for high-frequency volatility modeling have been proposed in the literature as well. The long-memory properties in the time-series of (realized) log-volatility have first been investigated in (Andersen, Bollerslev, Diebold and Ebens, 2001; R. Oomen, 2001; Andersen, Bollerslev, Diebold and Labys, 2003). These studies pose the basis for the development of the so-called fractionally integrated models (ARFIMA). Note that the idea of a fractionally integrated process is closely related to the concept of scaling exponent (discussed in Publication III), (Torre et al., 2007; Podobnik et al., 2008; Leite, Rocha, Silva et al., 2007; Grau-Carles, 2000) and not only relevant for financial literature (e.g. Leite, Rocha and Silva, 2009). ARFIMA-class models are however complex from a theoretical perspective and not of immediate implementation. The long-term memory feature observed in volatility is also captured through the much simpler Heterogeneous Autoregressive (HAR) model (Corsi, 2009). Motivated by the above-mentioned research and earlier findings (e.g. U. A. Müller, Dacorogna, Davé, Pictet et al., 1993) the HAR model decomposes the realized variance into three (lagged) factors, the daily, weekly and monthly volatility

components:

$$RV_{t+1}^{(d)} = c^{(d)} + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d} \,, \qquad (2.8)$$

where d, w, m respectively represent daily, weekly and monthly volatility components, $RV^{(w)}$ and $RV^{(w)}$ are the (re-scaled) averages of the daily volatility terms involved in the specific component, e.g. $RV_t^{(w)} = \frac{1}{5}\sum_{i=0}^{4} RV_{t-i}^{(d)}$ and e.g. $RV_t^{(m)} = \frac{1}{20}\sum_{i=0}^{19} RV_{t-i}^{(d)}$, and ε_t an error term. The HAR model takes a simple linear form can be estimated with OLS regression. For its simple linear structure, immediate estimation and for its remarkable performance in out/in -sample analyses, the HAR model nowadays constitutes a well-established benchmark for volatility forecasting with realized (high-frequency) measures. Several extensions to the HAR model have been proposed, for instance (Andersen, Bollerslev and Diebold, 2007) includes a jump component in the regressors, (A. J. Patton and Sheppard, 2015) includes asymmetries based on the return sign. Non-linear models of log-RV have been proposed as well (Hillebrand et al., 2007; McAleer et al., 2011; Gallo et al., 2015), and models allowing for structural breaks (Martens et al., 2004; McAleer et al., 2008a; Wen et al., 2016; Gong et al., 2018), as well as extensions allowing for time-varying parameters (e.g. Bollerslev et al., 2016). Among them, based on (Blasques et al., 2014), (Buccheri et al., 2017) acknowledge that their SHAR model of can be employed as an alternative specification for a general non-linear HAR model through time-varying coefficients. Multivariate extensions of (Corsi, 2009) are for instance discussed in (Chiriac et al., 2011; Lyócsa et al., 2016).

Setting apart from the HAR model baseline, there are different approaches, not related to the discussion in Publication IV. These include the realized-GARCH (P. R. Hansen, Z. Huang et al., 2012) where, with respect to the classical GARCH, realized measures enter in the conditional variance equations, HEAVY models (Shephard et al., 2010) where an equation explicitly models the dynamics of the conditional expectation of the realized measure, and MEM models (Engle and Gallo, 2006; Gallo et al., 2015). Also, the so-called MIDAS class of models, involving regressors sampled over different frequencies (Ghysels et al., 2007; Andreou et al., 2010; Bai et al., 2013), have been effectively applied in modeling and forecasting volatility, also in combination with a GARCH-modelled short-term component around a long-term trend (Engle, Ghysels et al., 2013). ARFIMA or stochastic volatility (SV) models extension that includes

realized measures are for instance discussed in (Andersen, Bollerslev, Diebold and Labys, 2003; Koopman et al., 2005).

2.6 Copulas in finance

The analysis of the dependence structure between two or more variables is of importance in a wide number of problems. Whereas dependence can be captured by the so-called dependence measures (Rényi, 1959; Schweizer et al., 1981) (e.g. Pearson's correlation coefficient, Kendall's tau or Spearman's rho), an alternative approach is given by copulas (M. Sklar, 1959; A. Sklar, 1973).

Consider a random vector $\mathbf{X} = (X_1, ..., X_d)$,with marginals CDFs of X_i, $i = 1, ..., d$ being $F_i(X_i) = Pr[X_i \leq x_i]$. The random vector $\mathbf{U} = (U_1, ..., U_d)$ obtained by applying the probability integral transform to each component[8] is distributed in $[0,1]^d$ with uniform marginals. Its joint cumulative distribution C is the copula of $\mathbf{X}$:

$$C(u_1, ..., u_d) = Pr[U_1 \leq u_1, ..., U_d \leq u_d],$$

that is, $\mathbf{U} \sim C$, the copula of $\mathbf{X}$.
The Sklar theorem (M. Sklar, 1959) provides the theoretical foundation for copula applications. Be H a multivariate CDF of a random vector $\mathbf{X}$, $H(x_1, ..., x_d) = Pr[X_1 \leq x_1, ..., X_d \leq x_d]$ with margins $F_i(x_i) = [X_i \leq x_i]$. For every $\mathbf{x} \in \mathbb{R}^d$, H can be expressed in terms of its margins and a copula C:

$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

Whereas H entails all the univariate and multivariate information on $\mathbf{X}$, C contains all the information about the dependence between the components $X_1, ..., X_d$. Conversely, for a copula $C : [0,1]^d \to [0,1]$ and d margins F_i, $C(F_1(x_1), ..., F_d(x_d))$ is a d—dimensional CDF with margins F_i,$i = 1, ..., d$. Importantly, if F_i are continuous, on the Cartesian products of the ranges $\text{Ran}(F_1) \times ... \times \text{Ran}(F_d)$, C is uniquely defined. A general reference for a rigorous introduction, comprehensive of all the mathematical background and preliminaries to the Sklar theorem is given by (Nelsen,

[8] I.e. $(U_1, ..., U_d) = (F_1(X_1), ..., F_d(X_d))$.

2007). Analogous results hold for densities (e.g. Nelsen, 2007), and for conditional distributions (e.g. A. Patton, 2013).

Since the analysis of dependence is the main purpose of copulas, their applicability involves a wide range of fields and analyses. From the influential paper of (Embrechts, McNeil et al., 2002), there have been extensive applications in finance, econometrics, risk management and actuarial sciences (see e.g. Bouyé et al., 2000; Cherubini, Luciano and Vecchiato, 2004; A. Patton, 2013). The main motivation for using copulas in finance comes from evidence of non-normality in the dependence (e.g. Malevergne et al., 2003, amongy many others) between asset returns (A. Patton, 2013). This is critical in several directions, especially in risk management. Application of copula methods in value-at-risk evaluations are among the earliest (e.g. Cherubini and Luciano, 2001; Embrechts, Höing et al., 2003). The use of copulas in the study of financial contagion became relevant in recent years: the complex connection and dependence between markets are strongly simplified by a normal-dependency assumption, and importantly the strength of dependencies increase during periods of crisis (Erb et al., 1994; A. J. Patton, 2006). Early studies in this direction are the switching copula model of (Rodriguez, 2007), focusing on the Asian and Mexican crises. This research direction closely relates to the recent explosions of credit derivatives products and multiple underlying products such as basket default swap and collateralized debt obligations. These pools of assets and liabilities are of high complexity whose dependence structure has been studied in a copula perspective in several papers (see e.g. D. X. Li, 1999; Jouanin et al., 2001; Laurent et al., 2005). More about derivative pricing using copulas can be found, in e.g. Cherubini, Luciano and Vecchiato, 2004. On the other hand, seminal applications in optimal portfolio decision are those of (e.g. A. J. Patton, 2004; Kole et al., 2007). Early contributions from (A. J. Patton, 2001b; A. J. Patton, 2001a; A. J. Patton, 2006) explore the use of copula in time-series modelling. Conditional on a set of past information, copula parameters are allowed for time-variation in an autoregressive way. The time-varying dynamics in parameters naturally represents a suitable feature for financial time-series data. In a similar perspective, Jondeau et al., 2006 provide a copula-based extension of the framework in (B. E. Hansen, 1994), first advocating a Copula-GARCH model class.

The the discussion (Sokolinskiy et al., 2011), hints the development of Publication IV. One-day-ahead forecasts extracted from the HAR model (Corsi, 2009) are compared with those of copula-based realized-volatility forecasts. This is achieved by decom-

posing the joint distribution of the integrated volatility estimator (RR) and its first lag into marginal distributions F (estimated non-parametrically) and a copula density c (estimated via maximum likelihood). The conditioning the density f of RR_t on RR_{t-1}: $f\left(RR_t|RR_{t-1}\right) = c\left(F(RR_t), F\left(RR_{t-1}\right)\right)f(RR_t)$, forecasts are obtained via Monte Carlo (MC) simulation. (i) Draws of today's volatility (captured with the realized range estimator) are simulated from the fitted copula given yesterday's volatility (its lagged value) (ii) the uniform draws are transformed via inverse empirical distribution function (iii) their mean is taken as a conditional forecast of today's volatility given yesterday's. Under a Gumbel copula, their specification is shown to beat the HAR model in out-of-sample analyses. The authors also implement a conditional-copula version of their models (relying on A. J. Patton, 2006) with time-varying parameters which however seems not to improve the accuracy of volatility forecasts. A generalization of their model to account for a four-dimensional copula that resembles the modeling spirit of the HAR model (dealing with four volatility variables), and that possibly allows for an evaluation of the conditional expectation without relying on MC methods is the task developed in Publication IV. However, the multidimensional extension of bivariate copula models is not immediate. For instance, long-memory properties observed in volatility would suggest that the dependence strength and type (as empirically observed) are different among pairs of variables involving volatility measures at different scales: a multivariate copula that is flexible in this concern would represent an attractive solution.

Extending of copula-based models in high dimension is most obvious but difficult problem (A. Patton, 2013). Flexible yet parsimonious and feasible directions are Factor-copulas (D. H. Oh et al., 2017) and Vine-copulas (see Section 4.3.2). For a comprehensive overview on copula construction methods see (e.g. Joe, 2014, chapter 3). Vine copulas applications are widespread, examples are risk-management and value-at-risk applications (e.g. Weiß et al., 2013; Reboredo et al., 2015), volatility modelling (e.g. Vaz de Melo Mendes et al., 2014; So et al., 2014; E. C. Brechmann, Heiden et al., 2018) - but under a HAR-like approach, and in the analysis of financial returns (e.g. Chollete et al., 2009; Nikoloulopoulos et al., 2012; Dissmann et al., 2013; Joe, 2014).

3 DATA

The two datasets utilized in the publications are here introduced to the reader. Refer to the Publications I-IV for further details on variables and time-series that each of them deals with.

3.1 Order book data

Publication I relies on Level-II order book data (see Section 2.2, extracted for 5 securities traded at the Helsinki stock exchange. The task of moving from the unstructured raw-data to a workable format suitable for feature extraction and ML applications is however complex: the procedure is here described. The original raw ITCH flow data provided by NASDAQ OMX operating at Helsinki exchange is distributed in day-specific and market-wide files[1]. The data in each file consists in formatted strings[2] where texts and numbers at specific positions from the beginning of each row have a specific meaning. Rows correspond to different events that affect the state of the order book, either for a specific security, either for the whole market. An identifier available for each row defines the event the row is referring to, for instance, limit or market order submissions, partial and total cancellations or market events such as the beginning of a trading halt. This flow of events is often referred to as the message book. Although the message book is comprehensive of *every* event that occurred on the exchange, the data in this format is of difficult use. First of all, in the raw ITCH flow events for different securities are mixed together within the same file. Second, events can occur at any distance (level) from the current bid or ask prices and can

[1] NASDAQ's Historical TotalView-ITCH files.

[2] For more information consult the official documentation available at: `http://www.nasdaqtrader.com/Trader.aspx?id=DPSpecs`

be of little impact (and thus interest) in the short-term mid-price dynamics[3], third, simple information about e.g. the number of outstanding limit orders at any time-instance is not directly accessible: the whole flow needs to be processed, submitted orders matched with cancellations and transactions and only the outstanding limit orders counted. Similarly, no immediate information about the current best levels is available: up to a certain time instance all the messages from the beginning of the day needs to be processed, only those corresponding to active orders (orders that have not been fully canceled -by either a single or a sequence of partial cancellations of the originally submitted quantity- or traded) sorted in descending or ascending order depending on the book side. In general, the message book does not provide an immediate outlook at the order book states, i.e. at the best bid and ask and deeper levels of the book in terms of prices and quantities. Although complete the message book is unhandy for analyses, so the first step in data analysis is a whole message book preprocessing so that order book states can be easily reconstructed. Figure 2.2 provides a graphical illustration of the message-book dynamics, Figures 2.1 and 2.3 clearly refer to sample order book states at given epochs.

A `C++` converted has been implemented to achieve this task. In particular, messages for a specific stock are extracted from NASDAQ's raw market-wide and day-specific files. For each row information such as the message type, unique ID of the entry that the current event is referring to (e.g. in case of partial cancellation, the ID of the earlier event (order) that is updated, or a new ID in case of a new-order submission), the timestamp, order/cancellation side, price and volume. Timestamps are provided with millisecond precision, below this resolution the ordering of the message book reflects the order at which the events have been submitted to the platform. A `C++` converter thus generates day-by-day and stock-specific tables grouped by message type (order, trades, cancels) where the information for each event is stored in separate columns with numeric formats. Because of the data size and the convenience of tabular structures, the convenient .h5 format is chosen for saving the processed data. Importantly cross-references between the tables are implemented. For each limit-order submission available in the "orders" table the trade or cancel events that removed it from the order book are uniquely referenced and its time-stamp available: for each order the edges t_{in}, t_{out} the period in which it has been active in the book are thus readily available. In this way, retrieving the order book state at a time t is

[3]Think for instance about a limit-order submission at a wrong price and its immediate cancellation.

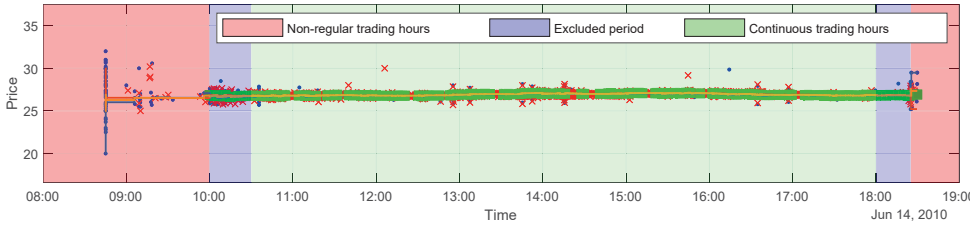


Figure 3.1 Illustration of a trading day (June 14, 2010, for KESKOB). Red areas define non-regular trading hours (pre-trading and post-trading periods). The whole blue and green area denotes continuous-time trading hours. However, analyses exclude their the first 30 and last 25 minutes, focusing on the green area only. Other marks are of analogous interpretation of Figure 2.2, here intentionally shrank to display the anomalous flow during non-continuous trading hours.

trivial: take the orders such that $t_{out} > t$ (i) the maximum (minimum) price of all the orders is the bid (ask) price (ii) the number of unique prices at each side corresponds to the number of levels (iii) for a given price level, the sum of the quantities of all the order gives the total quantity at that level. The exact reconstruction of the order book state at any t is in this way immediate. A simple Matlab function performs this evaluation, taking as an input the .h5 files and returning tables of prices and quantities at an ordinary number of levels at the ask and bid side.

In the auction period aimed at price discovery before the beginning of the so-called continuous trading period and the post-opening period, message submission follows a different mechanism that is structurally different and with different regulations w.r.t. to the normal continuous trading. Moreover "edge-effects" in the proximity of market opening and closing times are likely to alter the flow dynamics observed in the central part of the trading day (Siikanen, Kanniainen and Luoma, 2017; Siikanen, Kanniainen and Valli, 2017), e.g. around the closing time, liquidity providers would adjust their inventories, altering the regular flow. Thus this decision aims at avoiding clear confounding effects (e.g. C. Cao et al., 2009), visible in 3.1. Accordingly, the analysis in Publications I-III ignore the trading activity outside 10:30 am and 6:00 pm (UTC/GMT +2), whereas the effective continuous-time trading period goes from 10 am to 6:25 pm[4]. The pre-opening session runs from 9:00 am to 10:00 am and the post-trading from 6:25 pm. to 6:30 pm. See 3.1 for an illustration.

The analysis in Publication I focuses on the five Finnish stocks from different sectors reported in Table 3.1. The order book state data for the analysis is limited to the first

[4]However the flow in the first 30 minutes is used for computing any order book state.

ten best bid and ask levels. This means 40 order book state entries at each time-stamp, 10 prices 10 quantities for the two sides of the book. These entries, along with time-sensitive and time-insensitive features described in Table 4 of Publication I constitute the set of variables upon which the ML methods are implemented. As pointed out in Publication I full-depth data is not common, and data-oriented order-book analyses have often used Level-II data (limited to the first 5 levels), thus the data made available in Publication I doubles this threshold. The period covered by the data corresponds to 10 trading days, between June 1, 2010 and June 14, 2010, and collects approximately $4 \cdot 10^6$ events. This very same dataset has been used for the analyses in Publication II as well.

Ticker	ISIN Code	Company	Sector	Industry
KESKOB	FI0009000202	Kesko Oyj	Consumer Defensive	Grocery Stores
OUT1V	FI0009002422	Outokumpu Oyj	Basic Materials	Steel
SAMPO	FI0009003305	Sampo Oyj	Financial Services	Insurance
RTRKS[5]	FI0009003552	Rautaruukki Oyj	Basic Materials	Steel
WRT1V	FI0009003727	Wärtsilä Oyj	Industrials	Diversified Industrials

Table 3.1 Stocks used in in Publication I.

On the other hand, order book data for Publication III covers a longer period of 752 trading days (from June 1, 2010 to May 31,2013). The analysis here is limited to the best bid and ask levels only. I.e. the variables involved in the analyses such as order-to-order durations are calculated as durations between consecutive orders submitted at the best level on the same book side. Here the securities involved are not limited to the Helsinki exchange only, but include stocks traded in Stockholm and Copenhagen, see Table 3.2. For a complete list of variables extracted from this data, see Table II in Publication III.

Ticker	ISIN Code	Exchange	Company	Sector	Industry
NRE1V	FI0009005318	Helsinki	Nokian Renkaat Oyj	Consumer Cyclical	Rubber & Plastics
METSO	FI0009007835	Helsinki	Metso Oyj	Industrials	Conglomerates & Steel
ATCOA[6]	SE0000101032	Stockholm	Atlas Copco AB	Industrials	Diversified Industrials
VOLVB	SE0000115446	Stockholm	Volvo B	Industrials	Truck Manufacturing
VWS	DK0010268606	Copenhagen	Vestas Wind Systems	Industrials	Diversified Industrials

Table 3.2 Stocks used in Publication II.

[5]Historical ISIN and ticker: ISIN delisted on 20 November, 2014.
[6]Historical ISIN and ticker: ISIN changed on April 28, 2015, and further changed on May 9, 2018.

Stationarity and unit-root testing are among the most important analyses in econometrics. Not taking these into consideration nor discussing them at any level would represent an egregious omission for a dissertation in this field. This concluding part report results for the unit-root testing applied to a sample from the duration series Publication III uses. The following analyses are completed by the critical discussion in Chapter 6 to which I refer for and understanding of how unit-root testing and stationarity in general related to DFA. For a trading day of a sample stock, Table 3.3 reports rejections for most common stationarity tests, namely the Augmented Dickey-Fuller (ADF) test (Dickey et al., 1979), the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests (Kwiatkowski et al., 1992), and the Phillips–Perron (PP) test (Phillips et al., 1988). The last row in 3.3 reports the statistics for the whole daily time series. In virtue of the DFA algorithm described in Section 4.2, windows of different lengths are considered and their detrended fluctuation averaged: is important to address whereas stationarity holds for the disjoint intervals drawn by these windows. The number of intervals in which the day is divided is given in the second column, corresponding to the number of events, while the third column reports the average time-length for the specific window. From moderately low frequencies to half-day windows the joint analysis of the three tests indicates trend stationarity, thus over this domain, the application of DFA can be considered robust and reliable. At moderately low frequencies (10 seconds and less) tests lead to discrepant conclusions[7]. This is a well-known caveat that does not allow to conclude whether stationarity is to be rejected or not. In this case, heteroscedasticity may play a role (indeed well captured by the Bartlett test, besides the questionable normality of the samples, and by use of boxplots), a variance-ratio test could provide further insights. To which extent these analyses are relevant and to which extent they determine drive and limit the DFA applicability is however not addressed in the literature. Given the DFA procedure outlined in Section 4.2 stationarity appears to be a sensible hypothesis playing a relevant role. However, interestingly, the literature is vague in this regard and econometric analyses aimed at clarifying both the exact DFA hypotheses, defining its applicability, and

[7]Note that the PP test is based on asymptotic theory, i.e., in short time-series its validity is stretched e.g. ≤ 30 (Kwiatkowski et al., 1992). Furthermore, note that the three tests are based on testing the significance of an autoregressive coefficient given specific alternative hypotheses about the time-series dynamics, built over specifications involving an AR-like term. Disturbances are generally allowed to be either i.i.d. or of finite-lag correlation. A detailed discussion here is out of scope, however, it would be relevant to clearly frame the role long-range autocorrelation might play in offsetting testing powers and testing distributions.

how unit-root testing relates to it, are missing. This is a critical aspect of the DFA methodology. A broader and complementing discussion is addressed in Chapter 6.

Intervals			Rejection of H_0 (blank: no, 1: yes)								
			Order-to-order			Trade-to-trade			Cancel-to-cancel		
Average length (sec.)	No. of intervals	No. of events	ADF	KPSS	PP	ADF	KPSS	PP	ADF	KPSS	PP
4.71	90	33									
5.25	80	37									
6.05	70	43									
7.05	60	50									
8.48	50	60									
10.52	40	75	1		1	1		1	1		1
14.02	30	100	1		1	1		1	1		1
20.92	20	151	1		1	1		1	1		1
41.75	10	302	1		1	1		1	1		1
83.82	5	605	1		1	1		1	1		1
140.01	3	1009	1		1	1		1	1		1
216.93	2	1514	1		1	1		1	1		1
419.85	1	3028	1	1	1	1	1	1	1	1	1

Table 3.3 Stationarity testing for different duration time-series (June 06, 2010, VWS). Hypotheses read as follow. ADF & PP. H_0: the process has a unit-root, H_1: the process has no unit root, it's either stationary or trend stationary. KPSS. H_0: the process is trend-stationary, H_1: the process has a unit root. The multiple-testing nature is taken into account by adjusting the significance thresholds (Bonferroni correction).

3.2 TAQ Data

NYSE Trades and Quote data is a reference database for market research Publication IV containing intraday transactions data (trades and quotes) for all securities listed on the New York Stock Exchange (NYSE) and American Stock Exchange (AMEX), as well as NASDAQ National Market System (NMS) and SmallCap issues. The analyses in Publication IV involve data for ten stocks constituting the DOW30 index, during the period from 1 January, 2012 to 30 June, 2018. The long span of 1634 days is motivated by the necessity of a sample suitable for both in-sample and out-of-sample analyses. The Trades And Quotes (TAQ) database contains raw data as submitted to the platform and is known for having errors and mis-recordings. A detailed procedure for handling and preprocessing the data is provided in (Barndorff-Nielsen et al., 2009), accordingly, the applicable steps in the cleaning procedure there outlined have been

implemented.

Although the availability of the quote data, the analysis in Publication IV involves trades only. Indeed, the analysis deals with volatility estimation through appropriate realized measures computed from transaction prices robust to microstructure effects that may arise for instance from the bid-ask bouncing. The data has not been re-sampled: the full tick-by-tick information has been retained to implement the estimators. However, analogous restrictions as those depicted in Figure 3.1 have been applied.

Ticker	Exchange	Company	Industry
AAPL	NASDAQ	Apple	Information technologies
AXP	NYSE	American Express	Financial services
BA	NYSE	Boeing	Aerospace and defense
CAT	NYSE	Caterpillar	Construction and mining equipment
CSCO	NASDAQ	Cisco Systems	Information technologies
CVX	NYSE	Chevron	Oil & gas
DIS	NYSE	Disney	Broadcasting and entertainment
GS	NYSE	Goldman Sachs	Financial services
HD	NYSE	The Home Depot	Retail
IBM	NYSE	IBM	Information technologies

Table 3.4 Stocks used in Publication IV.

4 METHODS

The variety of methods used throughout Publications I-IV are here described in details, and in a general setting. Section 4.1 describes the machine learning methods utilizes for Publication I and Publication IV. Section 4.2 describes the methodological framework for Publication III. Finally, Sections 4.3.1 and 4.3.2 are committed to precisely address the main ingredients for Publication IV.

4.1 Machine learning methods

The next three subsections provide a description of the three machine learning methods used in papers Publication I (ridge regression and Single Layer Forward-feed Network) and Publication II (discriminant analysis).

4.1.1 Ridge regression

Consider a $(n \times p)$-dimensional data-matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$ and a response variable $\mathbf{Y} = (Y_1, ..., Y_n)^\top$. By use of a functional relationship like $Y_i = f(\mathbf{x}_i)$ the aim of a regression is to explain $\mathbf{Y}$ in terms of $\mathbf{X}$. Since there is in general no knowledge about the exact functional form of f, whenever the relationship is assumed linear we have a linear regression model:

$$\begin{aligned} Y_i &= \mathbf{x}_i \beta + \varepsilon_i \\ &= \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i \,, \end{aligned}$$

with $\beta = \left(\beta_1, ..., \beta_p\right)$ being the regression parameter and ε_i the error, representing the part of Y_i not explained by $\mathbf{x}_i\beta$ and $i = 1, ..., n$. Error terms are assumed normally distributed, with constant variance σ^2 and independent, i.e. $\varepsilon_i \sim N\left(0, \sigma^2\right)$ and $\text{Cov}\left(\varepsilon_i, \varepsilon_j\right) = \sigma^2$ for $i = j$ and zero otherwise. This leads to $\mathbb{E}(Y_i) = \mathbf{x}_i\beta$ and $\text{Var}(Y_i) = \sigma^2$, thus $Y_i \sim N\left(\mathbf{x}_i\beta, \sigma^2\right)$.
In compact matrix notation, the linear regression reads:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \,,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$ and $\boldsymbol{\varepsilon} \sim N\left(0_p, \sigma^2\mathbf{I}_{nn}\right)$, so $\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}_{nn}\right)$. The Ordinary Least Squares (OLS) estimator that minimizes the loss function

$$\mathscr{L}(\beta) = ||\mathbf{Y} - \mathbf{X}\beta||^2$$

is given by:

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} \,. \tag{4.1}$$

Collinearity in the data-matrix $\mathbf{X}$ implies that the rank of the $(p \times p)$-dimensional matrix $\mathbf{X}^\top\mathbf{X}$ is less than p, consequently, its determinant is zero and the matrix is called singular. A singular matrix is not invertible, thus $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$ is not a feasible estimator for β. In large dimensions, although perfect collinearity in the columns of $\mathbf{X}$ might not arise, situations where $\mathbf{X}$ is close to rank deficiency may often occur. This corresponds to a large variance in the estimates of the regression parameters for the collinear variables.

A remedy is provided by the rigde loss function (Hoerl et al., 1970). The standard loss function for the linear regression here includes a penalty tuned by a penalty parameter λ:

$$\mathscr{L}(\beta, \lambda) = ||\mathbf{Y} - \mathbf{X}\beta||^2 + \lambda||\beta||^2 = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \,. \tag{4.2}$$

With $\lambda > 0$ the penalty term affects the loss function and its minimum, with $\lambda = 0$ the OLS estimator is retrieved, given that $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ is well-defined. The penalty terms acts by controlling the norm of β in such a way that the larger β the larger the contribution of the penalty function to $\mathscr{L}(\beta, \lambda)$ and to shrink the regression

coefficients towards zero.

The ridge regression estimator is thus formulated as the solution of the minimization problem:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} ||\mathbf{Y} - \mathbf{X}\beta||^2 + \lambda ||\beta||^2 \ . \tag{4.3}$$

By taking the derivative of eq.(4.2) with respect to β one obtains the normal equations for the rigde regression and by solving $\frac{\partial}{\partial \beta}\mathscr{L}(\beta, \lambda) = 0$, the ridge regression estimator is obtained:

$$\hat{\beta}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \ . \tag{4.4}$$

Note that the unconstrained minimization problem in eq. (4.3) is equivalent to the constrained optimization $\arg\min_{||\beta||^2 \leq c} ||\mathbf{Y} - \mathbf{X}\beta||^2$ where $||\beta||^2 \leq c$, where c is a constant. Indeed through Lagrange multipliers eq. (4.4) can be equivalently obtained. In this perspective, is very clear that the penalty λ constrains the norm of β, and in $\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}\right)$ guarantees that the eigenvalues are all non-zero (a matrix is singular if and only if the determinant -which is the products of eigenvalues of the matrix- is zero), so that the original ill-posed problem $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ under rank deficiency, in ridge regression is well-defined. The inversion problem eq. (4.4) involves, can be optimized with the Moore-Penrose (MP) pseudo-inverse method, numerically very stable especially when (λ) shrinks the coefficient close to zero.

There are two further notes to add to this exposition. First the estimator eq. (4.4) depends on λ, so the problem of choosing an appropriate λ. Noticing that $\lambda = 0$ corresponds to standard liner regression, while with $\lambda \to +\infty$, in eq. (4.4) $\mathbf{X}$ has no impact (indeed the variance of the estimator would be 0_{pp}). A strategy to fix a value for the penalty parameter in this spectrum, is to choose λ such that it minimizes the mean squared error on a (fixed) training set. Second, with respect to the estimator in eq. (4.1) which is unbiased, is simple to show that the expectation of the ridge-regression estimator is $\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}\right)^{-1}\left(\mathbf{X}^\top \mathbf{X}\right)\beta \neq \beta$, but interestingly its variance is less than that of eq. (4.1) for $\lambda > 0$.

In the setting of Publication I, $\mathbf{X}_i$ corresponds to a 144-dimensional vector of features extracted from the limit-order book data, Y_i is a 3-dimensional vector specifying the classification class $\mathbf{x}_i$ belongs to (mid price increases, decreases or stays still). Once the ridge regression model is estimated on the training set, i.e. $\hat{\beta}(\lambda)$ is solved, observations are classified according to the maximum component of the projections

$\mathbf{x}_i \hat{\beta}(\lambda)$.

4.1.2 Single layer forward feed network

The idea of Single Layer Forward-feed Network (SLFN) goes back to (Rosenblatt, 1958) and represents the oldest neural network in literature. SLFN are machine learning methods that using examples (i.e. supervised learning), assign n-dimensional input vectors to different classes by use of a stochastic gradient-descent algorithm that minimizes the classification error in attempting to linearly separate the set of training data (see e.g. Minsky et al., 2017).

Neurons constitute a network, elementary building blocks that can be thought of as processing units. Each neuron processes the input information computing the weighted sum of the input signals received from the neurons that are connected to it (or external outputs) and generates an output (to other neurons or a final output). Whenever the information flow between neurons has no feedback (the output of a network is not fed back to itself), in the sense that information flows from the input to the neurons producing output the network, the network is referred to as *feed-forward*. Neurons are arranged in layers (if there is a single layer, we talk of *single layer* network), layers are generally referred to as *hidden layers* since they stand between the input and the output (the "tangible" information, input samples and their classification). Figure 4.2 provides a graphical illustration of a network.

Two important aspects are still missing. Take $\mathbf{x}_d$ as the input column vector of the d-th sample and $\mathbf{w}^1_{\cdot i}$ as the row vector of weights associated with the i-th node in the first layer. For any weighted sum $\mathbf{w}^1_{\cdot i}\mathbf{x}_d$, the network applies the so-called activation function g. General choices are sigmoid functions $g(x) = 1/(1+e^{-x})$ or radial basis functions (see e.g. Park et al., 1991). This shrinks and rescales the input weighted sum (e.g. to $[0,1]$ with a sigmoid), specifying and limiting the output of a node to $g\left(\mathbf{w}^1_{\cdot i}\mathbf{x}_d\right)$. Here, the importance of a proper normalization of the input data, depending on the specific form of the activation function g utilized. However not to limit the output domain (e.g. to $[0,1]$, whether the desired target output/label is for instance 3) usually a bias b is added to each weighted sum. Therefore each node's output corresponds to a value computed as $g\left(\mathbf{w}^1_{\cdot i}\mathbf{x}_d + b\right)$. Both weights and biases at each node need to be estimated (properly tuned) to minimize the classification

error. Generalizing this introduction, it has to be noticed that biases and activation functions apply to any node in any layer of the network, as addressed below[1].

Importantly, the most relevant feature of a network is its capacity of *learning*, this corresponds to the ability to improve its outputs (performance in classification) by self-tuning its nodes connections, i.e. by learning how to classify. Learning algorithms of neural networks use a learning problem, described by a set of training data and iteratively update the parameters (weights and biases) of a network such that some error measure is decreased or some performance measure is increased Suzuki, 2011, page 256.

The following discusses how the estimation of a SLFN works, with the so-called back-propagation algorithm (P. J. Werbos et al., 1990)[2]. Let θ denotes the full set of parameters (weights w and biases b) and, in the in the following, the superscript k indicates the k-th layer, which is referred to as the current layer. w_{ij}^k denotes the weight between the node i in the previous layer l_{k-1} and the node j in the current layer l_k. b_i^k is the bias at node i, a_i^k the weighted sum plus bias b_i^k for node i, o_i^k the output of node i, parsed through the activation function common to all the nodes in the hidden layer g:

$$o_i^k = g\left(a_i^k\right) = g\left(b_i^k + \sum_{l=1}^{r^{k-1}} w_{li}^k o_i^{k-1}\right), \tag{4.5}$$

with r^k being the number of nodes in the layer k. All the above definition refer, as pointed out, to the node k, double subscripts for w_{ij}^k refer respectively to nodes in layer $k-1$ and k. Let assume that the number of layers is m, with m thus corresponding to tie output layer and m to the first layer, directly feed by $\mathbf{x}_i$. While this exposition is general, note that for a SLFN there is only a single layer, and that within the nodes there are no connections. Furthermore, call g_0 the activation function for the output layer nodes. Figure 4.1 provides a sketch of how a node works.

The data consists of vectors $\mathbf{z_i} = (\mathbf{x}_i, \mathbf{y}_i)$, with $\mathbf{x}_i$ representing the input and $\mathbf{y}_i$ the output (target values) for $i = 1, ..., n$ samples. Be $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1,...,n}$ their collection.We

[1]indeed, as a reference for the exposition below, this corresponds to the term $g\left(a_i^k\right)$ for the first node, thinking of a_i^0 as the weighted sum involving the sample observations, ranter than node output.

[2]Historically the method goes back to (Linnainmaa, 1970; P. Werbos, 1974).

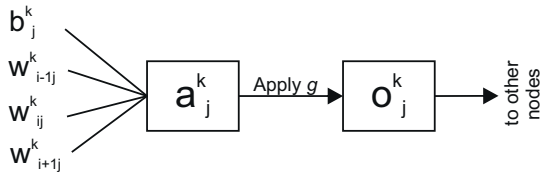


Figure 4.1 Representation of the operations within a node. The current node is j and layer is k, connections between the node j in layer k and nodes in layer $k-1$ are represented by the lines $w^k_{\cdot j}$. These inputs are processed in the first box, computing the weighted sum a^k_j. Function g is applied to a^k_j, which determines the output o^k_j, sent to nodes at layer $k+1$.

define the error function $E(X,\theta)$ at a particular parameter θ as the mean squared error between the target values and the output $\hat{\mathbf{y}}_i$ of the network for the sample $\mathbf{x}_i$:

$$E(\mathbf{Z},\theta) = \frac{1}{2n}\sum_{i=1}^{n} ||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 \ . \tag{4.6}$$

Note that this is an average across all the n samples while the norm captures the error associated with each output vector $\hat{\mathbf{y}}_i$.
The objective is that of minimizing $E(\mathbf{Z},\theta)$ by appropriately tuning θ. A local minimum for $E(\mathbf{Z},\theta)$ can be found by gradient descend method: we expect to find a local minimum where $\nabla E(\mathbf{Z},\theta) = \mathbf{0}$, therefore at each iteration the gradient of the objective function is determined and the parameter θ (or better, its elements) is updated according to:

$$\theta \longmapsto \theta - \eta \frac{\partial E(\mathbf{Z},\theta)}{\partial \theta} \ .$$

The little abuse in notation means that this update is performed element-wise for each component of θ at each iteration. η is called *learning rate* and rescales the magnitude of the step ($\Delta\theta$) towards the updated parameter, driven by the magnitude of the partial derivatives. The focus is therefore on the gradient of $E(\mathbf{Z},\theta)$ and on the partial derivatives $\frac{\partial E(\mathbf{Z},\theta)}{\partial w^k_{ij}}$ [3].

The minimization of the objective (4.5) can be achieved by considering the partial derivatives individually for each sample $\mathbf{x}_d$, indeed summation and deviation exchangeable in their order, so that first the partial derivative of the error for each $\mathbf{z}_d$ is

[3] To simplify the notation, we set $w^k_{0i} = b^k_i$, so we treat the bias as a weight corresponding to a node zero in layer $k-1$ with fixed output set to one, i.e. $o^{k-1}_0 = 1$. This is equivalent to eq. (4.5), since a^k_i rewrites as $\sum_{j=0}^{r^{k-1}} w^k_{ij} o^{k-1}_j$, by practically having $r^{k-1}+1$ node in layer $k-1$.

computed and then averaged:

$$\frac{\partial E(\mathbf{Z},\theta)}{\partial w_{ij}^k} = \frac{1}{n}\sum_{d=1}^{n}\frac{\partial}{\partial w_{ij}^k}\frac{1}{2}||\mathrm{y}_i - \hat{\mathrm{y}}_i||^2 = \frac{1}{n}\sum_{d=1}^{n}\frac{\partial E_d}{\partial w_{ij}^k} \tag{4.7}$$

with $E_d = \frac{1}{2}||\mathrm{y}_d - \hat{\mathrm{y}}_d||^2$ being the error associated with the d-th sample pair $\mathbf{z}_d$, the parameter is omitted to keep the notation compact. The chain rule is applied to the partial derivatives $\frac{\partial E_d}{\partial w_{ij}^k}$: the change in E_d caused by weight w_{ij}^k is the change in E_d caused by the activation a_j^k times the change in the activation a_j^k due to w_{ij}^k:

$$\frac{\partial E_d}{\partial w_{ij}^k} = \frac{\partial E_d}{\partial a_j^k}\frac{\partial a_j^k}{\partial w_{ij}^k} \tag{4.8}$$

To obtain a short-hand notation, we solve the second term on the right side of the above equation:

$$\frac{\partial a_j^k}{\partial w_{ij}^k} = \frac{\partial}{\partial w_{ij}^k}\sum_{l=0}^{r^{k-1}} w_{lj}^k o_l^{k-1} = o_i^{k-1}$$

and by defining $\delta_j^k = \frac{\partial E_d}{\partial a_j^k}$, we rewrite eq. (4.8) as:

$$\frac{\partial E_d}{\partial w_{ij}^k} = \delta_j^k o_i^{k-1}\,. \tag{4.9}$$

The partial derivative $\frac{\partial E_d}{\partial w_{ij}^k}$ is therefore the product of an error term δ_j^k at node j and layer k and the output of the node i in the layer $k-1$. Indeed, w_{ij}^k connects the node i in layer $k-1$ to node j in the layer k. It is important to notice that δ_j^k is up to now not solved, however it depends on the error term in the next level $k+1$, but independent on the error term at layer $k-1$, so that is possible to impute the error term at each layer backwards from the output layer down to the input layer.

To simplify the notation we shall assume that there is only a single output layer, so that the norm in eq. (4.6) simplifies into a squared value, also this calls for j being equal to one (a single node) in the last layer, here denoted by m. The error is therefore:

$$E_d = \frac{1}{2}(y-\hat{y})^2 = \frac{1}{2}(y - g_0(a_1^m))^2$$

and so δ_1^m turns to be:

$$\delta_1^m = -(g_0(a_1^m) - y) g_0'(a_1^m) = (\hat{y} - y) g_0'(a_1^m) \,.$$

Finally, the partial derivative of the error w.r.t. w_{i1}^m turns to be:

$$\frac{\partial E_d}{\partial w_{i1}^m} = \delta_1^m o_i^{m-1} = (\hat{y} - y) g_0'(a_1^m) o_i^{m-1} \,. \tag{4.10}$$

In the output layer the derivation was quite straightforward, in the hidden layers the notation gets more complex but the idea is the very same: recall the chain rule applied involving the weighted sum at layer $k+1$. For $1 \leq k < m$,

$$\delta_j^k = \frac{\partial E_d}{\partial a_j^k} = \sum_{l=1}^{r^{k+1}} \frac{\partial E_d}{\partial a_l^{k+1}} \frac{\partial a_l^{k+1}}{\partial a_j^k} \,. \tag{4.11}$$

Two observations, (i) the sum involves all the r^{k+1} nodes at the layer $k+1$, but (ii) not the node-zero since the output o_0^k leading to w_{0j}^{k+1} is independent on all the previous layers, so the summation starts from $l = 1$. The rightmost partial derivative term in the above equation is immediately solved. Since $a_l^{k+1} = \sum_{j=0}^{r^k} w_{jl}^{k+1} g'\left(a_j^k\right)$,

$$\frac{\partial a_l^{k+1}}{\partial a_j^k} = \sum_{j=0}^{r^k} w_{jl}^{k+1} g\left(a_j^k\right) \,,$$

and recalling that $\frac{\partial E_d}{\partial a_l^{k+1}} = \delta_l^{k+1}$ and eq. (4.11), the back-propagation formula is obtained:

$$\delta_j^k = g'\left(a_j^k\right) \sum_{l=1}^{r^{k+1}} \delta_l^{k+1} w_{jl}^{k+1} \,. \tag{4.12}$$

Eq. (4.12) allows to compute all the partial derivatives for any element in θ:

$$\frac{\partial E(\mathbf{z}_d, \theta)}{\partial w_{ij}^k} = g'\left(a_j^k\right) o_i^{k-1} \sum_{l=1}^{r^{k+1}} \delta_l^{k+1} w_{jl}^{k+1} \,.$$

In this way all the partial derivatives can be obtained and the gradient descend method

implemented:

$$\Delta\theta = \theta - \eta \frac{\partial E(\mathbf{Z},\theta)}{\partial \theta} .$$

Since any term $\frac{\partial E(\mathbf{Z},\theta)}{\partial w_{ij}^k}$ in ∇E is evaluated by averaging the contributions to the overall gradient component across all the n samples as in eq. (4.7):

$$\frac{\partial E(\mathbf{Z},\theta)}{\partial w_{ij}^k} = \frac{1}{n}\sum_{d=1}^{n}\left(\delta_j^k o_i^{k-1}\right)_d = \frac{1}{n}\sum_{d=1}^{n}\left(g'\left(a_j^k\right) o_i^{k-1} \sum_{l=1}^{r^{k+1}} \delta_l^{k+1} w_{jl}^{k+1} \right)_d ,$$

where for $k = m$ eq. (4.10) applies, and $(\cdot)_d$ are to be interpreted as quantities computed under the sample $\mathbf{z}_d$. In case of multiple outputs K, the norm of the vectors is considered n the error E_d. E_d is therefore decomposed in the sum of the K squared component-wise differences between $\hat{\mathbf{y}}$ and $\mathbf{y}$, representing $k = 1,...,K$, partial errors $E_d^{(k)}$. So E_d writes $E_d = \sum_{k=1}^{K} E_d^{(k)}$. Partial derivatives of $E_d^{(k)}$ solve with the chain rule, analogously as above for E_d. Equation (4.9) is thus simply updated to include a summation over all the K possible outputs.

Publication I uses this SLFN to classify mid-price movement direction. The output layer is made of three nodes, while the input layer accommodates 144-dimensional vectors of features. The initialization of the weights and biases is made according to a k-mean clustering (leading to K clusters) of the initial input data to determine initial weighting vectors for each node in the hidden layer. In this context, a radial basis function is used as an activation function. The overall objective is specified (as in the example above) as a norm minimization problem (where a penalization is introduced to control for large weights). According to the above illustration, classification is performed according to the maximal network value.

4.1.3 Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a machine learning supervised technique aimed at classification. LDA learns the projection $\mathbf{w}$ that maximizes the distance between classes and minimizes the distance between the data in each class. Over a training set where the labels corresponding to different classes are assigned, this is

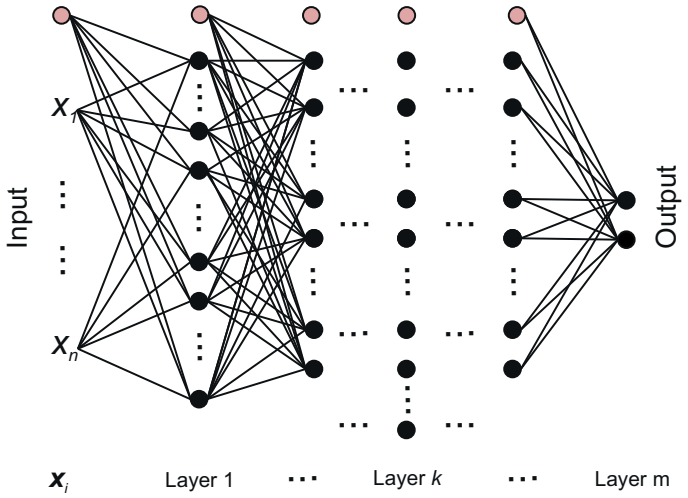


Figure 4.2 Representation of a network with multiple layers (feed-forward which turns to be feed-forward by thinking he information flowing from left to right and by noticing that nodes are not connected). Black dots represents "proper" processing nodes, while red dots represents biasing nodes (b_i^k). The n-dimensional vector $\mathbf{x}_i$ is taken as an input, while the network returns a two-dimensional output at node m

achieved by maximizing the following objective function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}} \,. \tag{4.13}$$

S_B and S_w are scatter matrices of the distances between and within the groups:

$$S_B = \sum_c (\mu_c - \bar{\mathbf{x}})(\mu_c - \bar{\mathbf{x}})^\top \,,$$
$$S_W = \sum_c \sum_{i \in c} (\mathbf{x}_i - \mu_c)(\mathbf{x}_i - \mu_c)^\top \,,$$

where in a compact notation $\sum_c$ refers to the summation over all the classes and $\sum_{i \in c}$ to the summation over all the data in a certain class c. μ_c is the mean of class c and $\bar{\mathbf{x}}$ the overall mean of all the data points (Welling, 2005).

Maximizing $J(\mathbf{w})$ therefore corresponds in finding the vector $\mathbf{w}$ such that the ratio between the distance between the groups and distance within groups is maximum, i.e. such that the class-means are well separated with respect to the variance of the data assigned to each class. Noticing that eq.(4.13) is invariant to rescalings of $\mathbf{w}$ ($\mathbf{w} \to a\mathbf{w}$),

one can chose $\mathbf{w}$ such that $\mathbf{w}^\top S_W \mathbf{w} = 1$. Therefore eq. (4.13) is equivalent to the minimization problem

$$\underset{\mathbf{w}^\top S_W \mathbf{w}=1}{\arg\min}\ \mathbf{w}^\top S_B \mathbf{w}\,,$$

whose lagrangian is:

$$\mathscr{L} = \mathbf{w}^\top S_B \mathbf{w} + \lambda\left(\mathbf{w}^\top S_w \mathbf{w} - 1\right).$$

The Karush–Kuhn–Tucker conditions for a linear mapping to be optimal tell that at the solution

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w} \tag{4.14}$$

must hold. By writing S_B as $S_B^{1/2} S_B^{1/2}$ (by using its eigenvalue decomposition, noticing that S_B is symmetric and positive definite) and definite $\mathbf{v} = S_B^{1/2}\mathbf{w}$, eq. (4.14) can be manipulated to obtain:

$$S_B^{1/2} S_W^{-1} S_B^{1/2} \mathbf{v} = \lambda \mathbf{v}\,.$$

This is a typical eigenvalue problem for the matrix $A = S_B^{1/2} S_W^{-1} S_B^{1/2}$ for which solutions λ_s, $\mathbf{v}_k$ corresponding to solutions $\mathbf{w}_k = S_B^{-1/2}\mathbf{v}_k$ can be found[4].The desired solution corresponding to the objective function (4.13) is the eigenvector $\mathbf{v}_k$ corresponding to the largest eigenvalue.

In a generalized setting of the Multilinear Discriminant Analysis (MDA) such as that of Publication II, the input data is a set of N tensors[5], $\mathscr{X}_1, ..., \mathscr{X}_N \in \mathbb{R}^{I_1 \times ... \times I_K}$, each with an associated class label c_i where $i = 1, ..., N$ and $c_i \in \{1, ..., C\}$. Analogously to the earlier case, the mean tensor of class c_i is $\mathscr{M}_i = 1/n_i \sum_{j=1}^{n_i} \mathscr{X}_{ij}$ and the mean tensor of the data $\mathscr{M} = 1/N \sum_{i=1}^{C} n_i \mathscr{M}_i$, with $\mathscr{X}_{ij}$ being the j-th sample from class c_i, and n_i the number of samples in class c_i. MDA solves for the set of projection matrices $\mathbf{W}_k \in \mathbb{R}^{I_k \times I'_k}$, $I'_k < I_k$, $k = 1, ..., K$ mapping $\mathscr{X}_{ij}$ to $\mathscr{Y}_{ij} \in \mathbb{R}^{I'_1 \times ... \times I'_K}$.

The objective function is here similar (4.13): the optimal set of matrices $\mathbf{W_k}$ is that maximizing the ratio between the inter-class distance D_B and the intra-class D_W

[4]For which the eigenvalues can be obtained as roots of the characteristic polynomial $\det(A - \lambda I)$.

[5]$\mathscr{X} \in \mathbb{R}^{I_1 \times ... \times I_K}$ is here referred to as a tensor, $i_1, ..., i_K$ are the indexes referring to each dimension. Tensor $\mathscr{X}$ is said to have K modes, where the mode-k of $\mathscr{X}$ is the I_k-dimensional vector obtained by fixing all the indexes but the i_k-th, so we call I_k the dimension of the mode-k.

distance[6]:

$$J(\mathbf{W}_1, ..., \mathbf{W}_K) = \frac{D_B}{D_W} .$$

Q. Li and Schonfeld, 2014 shows how to solve this minimization problem in an iterative manner by tensor unfolding in each mode-k by considering the trace-ratio problem:

$$J(\mathbf{W}_k) = \frac{\mathrm{tr}\left(\mathbf{W}_k{}^{\top} \mathbf{S_B}^k \mathbf{W}_k\right)}{\mathrm{tr}\left(\mathbf{W}_k{}^{\top} \mathbf{S_W}^k \mathbf{W}_k\right)} ,$$

where S_B and S_W are the scatter matrices of inter and intra -class distances for the k-th mode, which closely resembles eq.(4.13) and tr$(\cdot)$ denotes the trace operator. Similar orthogonality constraints on $\mathbf{W}_k$ as in the LDA case, allow to solve this problem in terms of I'_k eigenvectors corresponding to the largest I'_k eigenvalues of $\left(\mathbf{S}_W^k\right)\mathbf{S}_B$.

4.2 Detrended fluctuation analysis

4.2.1 The DFA algorithm

A standard approach for the detection of long-range autocorrelation, via estimation of the scaling exponent 2.4 of a time-series is the so-called Detrended Fluctuation Analysis (DFA). The method, originally introduced in (Peng, S. V. Buldyrev, Havlin et al., 1994) is nowadays widely utilized due to its robustness to non-stationarity, e.g. trends. Earlier methods, such as the Hurst's rescaled range analysis (Hurst, 1951) can lead to the false detection of long-range autocorrelation (Bryce et al., 2012). Although a number of extensions of the originally proposed DFA have been proposed (e.g. Bashan et al., 2008), the method as of (Peng, S. V. Buldyrev, Havlin et al., 1994) is widely utilized and constitutes the methodological approach of Publication III.

The following summarizes the DFA procedure for an arbitrary time-series of N observations, $\{x_t\}_{t=1,...,N}$:

I Altought not obligatory, a common practice is to consider the *profile* of the

[6]Providing here expressions for D_B and D_W would leave the notation undefined. This requires a set of mathematical definitions along with their respective notation, in the very exact way as addressed in Publication II

time-series, by considering the integrated sum:

$$y(k) = \sum_{t=1}^{k} (x_t - \bar{x}) \, .$$

Subtracting the mean $\bar{x}$ set the mean of the integrated series to zero.

II By setting a length s, the profile is divided in N/s non-overlapping windows (of equal length). In each window, a first-degree[7] polynomial approximation y_{tr}, representing the local trend of the data, is estimated (by ordinary least-squares).

III In each window m, differences (residuals) $y_m - y_{m,tr}$, $m = 1, ..., N/s$ define the m-th window detrended profile. By varying the widows of size s and considering the square root of the average variance of the residuals across the N/s windows, the fluctuation $F(s)$ is evaluated:

$$F(s) = \sqrt{\frac{1}{N/s} \sum_{m=1}^{N/s} \left[\frac{1}{s} \sum_{i=1}^{s} \left[y_m(i) - y_{m,tr}(i) \right] \right]^2} \, .$$

IV In presence of power-law scaling, $F(S) \sim s^{\alpha}$: the slope of the line approximating $F(s)$ against s in a log-log plot, estimates the scaling exponent α.

Figure 4.3 resembles the procedure.

$0.5 < \alpha < 1.5$ indicates long-range correlations, whereas $\alpha = 0.5$, $\alpha = 1$ and $\alpha = 1.5$ respectively correspond to White-noise, Brownian noise and pink-noise signals. Exponents $\alpha < 0.5$ correspond to anti-correlations (Bashan et al., 2008; Kantelhardt, 2009). In case of "crossovers" where different scaling exponents apply at different time-scales, i.e. the slope of $F(s)$ against s in the log-log plot changes, the same interpretation hold, but on a limited time-scale range. Refer to Chapter 6 for a discussion on the hypotheses underlying the DFA methodology, its applicability, and a discussion on how it relates to time-series stationarity.

[7] This corresponds to 1st order DFA, which removes linear trends. More generally a n-degree polynomial can be used.

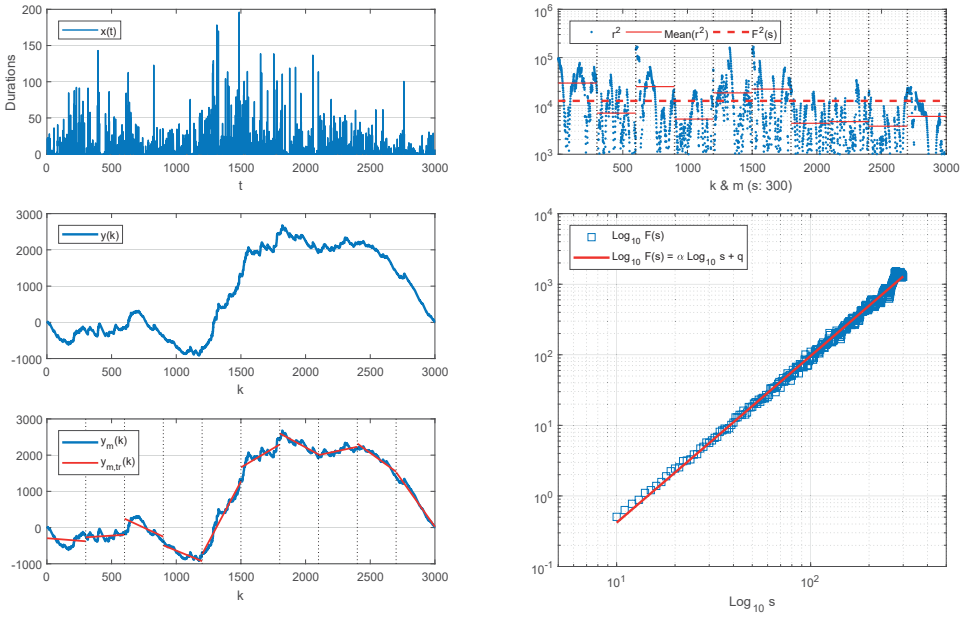


Figure 4.3 Visual representation of the DFA procedure. Order-to-order duration series for DK0010268606 (Vestas Wind Systems) on June 01, 2010 (for orders submitted at the best level on the bid side).

4.2.2 Stationarity issues in DFA

An analysis on non-stationarities in time series when applied to DFA is provided in (Z. Chen et al., 2002; Bryce et al., 2012). This are perhaps the only analysis available in this regard. For a simulated long-range correlated signal, the effect of three types on non-stationarities is analyzed in (Z. Chen et al., 2002).

(a) Signals with segments removed, i.e. non-stationarities caused by discontinuities in the signal. This is relevant in financial applications, e.g. due to the fact that markets do not trade on weekends, holidays, and at night. Z. Chen et al., 2002 find that the scaling of correlated signals is not affected by the cutting procedure, independently on the size of the cutting segment and on the number of segments removed.

(b) Signals with random spikes. In the duration series used in Publication III these correspond to seldom and *isolated* exceptionally long durations. Following Z. Chen et al., 2002, when uncorrelated spikes are added to the signal a change

in the cross-over of the scaling exponent at a characteristic scale is observed. For positively correlated signals, this is observed as small-time scales. Our data does not show any intra-day cross-over features, rather the log-log pot of fluctuation against window size is remarkably linear, suggesting that spikes-related non-stationarity are not relevant in our analyses.

(c) Signals with different local behavior, which include signals with (i) a number of segments of a certain length with different standard deviation and (ii) with different local correlation. (Z. Chen et al., 2002, Fig. 4d) shows that for correlated signals, $\alpha > 0.5$, with segments characterized by two different values of standard deviation no difference is found in the scaling exponent compared to the stationary correlated signal. Whereas the variance of durations may vary across the day, e.g. as a consequence of the U-shaped trading activity profile, DFA still constitutes a robust method.

For correlated signals, the presence of segments with different correlation can either lead to no differences in the scaling exponent wrt. to the stationary signal or to double cross-overs, with a characteristics plateau characterized by a flatter slope in the central part, which is not observed in our data. Based on the results of (Z. Chen et al., 2002) analyzing non-stationary sources of relevance also in financial time series, DFA is shown to be capable of detecting the correlation of the non-stationary signal in some circumstance, while producing crossovers in other, which are however not observed in our data. (Bryce et al., 2012) generally warn about the use of DFA in time series concluding that it does not provide any protection against non-stationarity, introduces biases, and suffers from small-sample effects. (Bryce et al., 2012) devises that explicit detrending followed by measurement of the diffusional spread of a signals' associated random walk is preferable. Note that in (Hu et al., 2001) "stationariety" is intended as "presence of trends", rather limited when compared to its broader meaning in econometrics.[8].

[8] At this point, (Z. Chen et al., 2002) and (Hu et al., 2001) provide sufficient reasons to question DFA' robustness to non-stationarity, or at least to identify of a vagueness around the term "stationarity".

4.3 Methods in volatility modeling and forecasting

This final Section presents the major methodologies adopted in Publication IV. There are two subsections accordingly. Whereas the first one has a clear connection with volatility modeling and forecasting, the second might appear and orphan in this context. That copulas are not extraneous to volatility modeling and finance has been pointed out earlier in Section 2.6, while a gap in their use in high-frequency econometrics is outlined in section 1.2. Indeed, in Publication IV, these are used for suggesting new modeling direction. This chapter is aimed at introducing methods used in Publications I-IV, while, for a general contextualization and broader discussions on their earlier use and application, pointing to Chapter 2, and to the enclosed Publications. With this perspective, Subsection 4.3.2 is not misleading, by addressing a purely statically discussion on Vine copulas, and the presentation of both parametric (ML) and non-parametric (e.g. DFA) methods in this Chapter justified. In particular, the relationship between Subsection 4.3.2, the HAR model, and the use of Vine-copulas in high-frequency volatility modelling is ascribed to Publication IV, being part of its contributions.

4.3.1 HAR model

This section discusses in detail the HAR model introduced in Section 2. Long-memory dependence in financial market volatility is long-established fact. Different models have been proposed to capture this behavior (see e.g. Section IV of Andersen, Bollerslev and Diebold, 2007, for a list of references). The HAR model is an outcome of this literature, in particular of the HARCH-class models (U. A. Müller, Dacorogna, Davé, Olsen et al., 1997), heuristically motivated by the heterogeneous market hypothesis (U. A. Müller, Dacorogna, Davé, Pictet et al., 1993): heterogeneous market participants trade on the market over different investment horizons, coexisting and interacting within the same market. E.g., high-frequency traders may be thought as participants trading at intra-day horizons, whereas large institutional traders hold their positions over longer time horizons. The typical slow-decay observed in volatility autocorrelation and stylized facts about returns' and volatility distribution can be reproduced by mixing in a simple linear model only three volatility components,

intuitively corresponding to the contribution to total daily volatility from trading on daily, weekly, and monthly horizons. Such a model, known as the Heterogeneous Autoregressive (HAR) model of (Corsi, 2009), is very attractive due to its simplicity in estimation, interpretation and in forecasting ability.

The daily latent volatility process $\tilde{\sigma}_t^{(d)}$ is modelled under a three-factor stochastic volatility specification. Factors are the past (realized) volatilities at different frequencies[9]. The three volatility terms identified in the HAR model are a daily component d, a weekly component w and a monthly component m, these are referred to as partial-volatility terms, since specific of a given time horizon. The latent volatility $\tilde{\sigma}_t^{(\cdot)}$ at any time scale is assumed to be a (linear) function of the past observed realized variance at the same time-scale [10] and of the expectation of next-period's longer-term partial volatility components[11]. The hierarchical cascade assumption reads[12]:

$$\begin{aligned}
\tilde{\sigma}_{t+1d}^{(d)} &= c^{(d)} + \phi^{(d)} RV_t^{(d)} + \gamma^{(d)} E\left[\tilde{\sigma}_{t+1w}^{(w)}\right] + \tilde{\omega}_{t+1d}^{(d)} , \\
\tilde{\sigma}_{t+1w}^{(w)} &= c^{(w)} + \phi^{(w)} RV_t^{(w)} + \gamma^{(w)} E\left[\tilde{\sigma}_{t+1m}^{(m)}\right] + \tilde{\omega}_{t+1w}^{(w)} , \\
\tilde{\sigma}_{t+1m}^{(m)} &= c^{(m)} + \phi^{(m)} RV_t^{(m)} + \tilde{\omega}_{t+1m}^{(m)} ,
\end{aligned} \tag{4.15}$$

where $RV_t^{(p)}$ is obtained by averaging p daily lagged realized variance estimates. Specifically, $RV_t^{(w)} = \frac{1}{5}\sum_{i=0}^{4} RV_{t-i}^{(d)}$ and $RV_t^{(m)} = \frac{1}{22}\sum_{i=0}^{22} RV_{t-i}^{(d)}$ are the weekly and monthly volatility components. Importantly, the error terms $\tilde{\omega}_{t+1d}^{(d)}$, $\tilde{\omega}_{t+1d}^{(w)}$ and $\tilde{\omega}_{t+1d}^{(m)}$, are serially independent, zero-mean and must guarantee positivity of the estimates.

By setting $\tilde{\sigma}_t^{(d)} = \sigma_t^{(d)}$ with $\sigma_t^{(d)}$ being the square-root of the integrated volatility[13] $\left[\int_{t-1d}^{t} \sigma_s^2 ds\right]^{\frac{1}{2}}$, by substitutions eq. (4.15) turns into:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \tilde{\omega}_{t+1d}^{(d)} , \tag{4.16}$$

[9] $RV_t^{(p)}$ denotes the realized variance estimated in t for the time-horizon p

[10] This corresponds to an AR-like structure: eq. (4.15) do not involve lagged values of $\tilde{\sigma}_t^{(\cdot)}$, but rather their respective proxies, i.e. $RV_t^{(\cdot)}$.

[11] For $\tilde{\sigma}_t^{(m)}$ only a linear function of monthly-RV remains, so the AR-like term.

[12] $t + 1d$ reads as "(end of) day t plus one day", similarly: $+1w$ and $+1w$ respectively stand for a week and a month ahead w.r.t. day t. RV are the actually observed ex-post values.

[13] As pointed out in Section 2 is the integrated volatility the usual quantity of interest, as a synthesis of the latent volatility process over an interval.

which corresponds to the three-factor representation earlier mentioned. By introducing an error term $\omega^{(d)}_{t+1d}$ that accounts for both measurement and estimation errors associated with using RV as a proxy for $\tilde{\sigma}^{(d)}_{t+1d}$ -or analogously recalling that $RV^{(d)}_{t+1d}$ is not an error-free measure for $\sigma^{(d)}_{t+1d}$, (Barndorff-Nielsen and Shephard, 2002)-, $\sigma^{(d)}_{t+1d}$ rewrites

$$\sigma^{(d)}_{t+1d} = RV^{(d)}_{t+1d} + \omega^{(d)}_{t+1d} \,. \tag{4.17}$$

By substituting eq. (4.17) into eq. (4.16) and collapsing the respective error terms, the HAR-RV model reads as:

$$RV^{(d)}_{t+1} = c^{(d)} + \beta^{(d)} RV^{(d)}_{t} + \beta^{(w)} RV^{(w)}_{t} + \beta^{(m)} RV^{(m)}_{t} + \omega_{t+1d} \,, \tag{4.18}$$

where $\omega_{t+1d} = \tilde{\omega}^{(d)}_{t+1d} - \omega^{(d)}_{t+1d}$. This corresponds to an autoregressive model with autoregressive weights taking a step-function form, restricted in a parsimonious way such that the three emerging components are economically meaningful and interpretable (Corsi, 2009).

The standard estimation of the HAR-RV model is performed via OLS. To guarantee non-negativity of the volatility estimates, eq. (4.18) can be written and estimated in the logs. To account for serial correlation a common practice is to use the Newey-West covariance correction in the estimation. Note that the HAR model can be implemented over the preferred volatility measure, e.g. over the realized kernel as in Publication IV.

As pointed out in Section 1.2, the discussion in Publication IV is based on some critical aspects of the HAR model. Here, I summarize them by referring to the above construction. (i) Linearity of equations (4.15) and thus in the linkage between the components involved in each equation. (ii) $\tilde{\omega}^{(d)}_{t+1d}$ are (a) mutually independent, (b) zero-mean, (c) left-tail truncated to guarantee positivity in the estimates. (iii) Independence of the $\beta^{(\cdot)}$ coefficients in eq. (4.16) over time. (iv) Positivity of the estimates in eq. (4.15), as (Corsi, 2009) suggest, can be achieved with an alternative specification of the model in eq. (4.18) in terms of log-RV (which is a common practice), by doing this (a) eq. (4.17) is assumed to hold in the logs as well, (b) non-log estimates are retrieved by bootstrapping, i.e. simulation (v) The HAR-RV model corresponds to a reparametrization of an AR model, with autoregressive weights

taking a step-function[14] (a) this is a step-like approximation of the typical power-law decay in volatility, (b) a limited number of volatility terms only resemble a portion of overall long-range correlation involving a continuous of time-scales. (vi) (a) Presence of autocorrelation, heteroscedasticity and general non stationarity in ω_{t+1d} require attention under OLS estimation, e.g. by using HAC standard errors, (b) although normality of the residuals is on a general level not critical for OLS applicability, confidence interval for the predictions (either for the mean response or observations) are symmetric, while e.g. volatility shows skewness. These points are further discussed in Section 3.1 of Publication IV.

Publication IV relies on some critical aspects of some of the above key-points in the HAR model construction. These are *not* to be seen as a criticism but rather as starting points for reasoning over possible limitations of the HAR model and for developing of possible extensions. Indeed, HAR's simplicity, its ability in reproducing several stylized features empirically observed in the markets and its good prediction ability, broadly motivate its widespread use.

4.3.2 Vine copulas

The pair-copula construction (PCC) for a multivariate distribution relies on sequential mixtures of bivariate conditional pair-copulas evaluated at conditional CDFs (Joe, 1996). This is based on the decomposition of multivariate distributions as products of conditional and unconditional distributions (by the law of total probability, i.e. Bayes' theorem for distributions). For a d-dimensional distribution the decomposition is however not unique: there are indeed $d(d-1)/d$ possible decompositions: *Vines* are representation specifying such decompositions and thus identifying the pair-variables and their order in the mixturing sequence leading to the construction of a multivariate density (Aas et al., 2009)[15].

In the following let F and f generically denote CDFs and densities, for univariate,

[14] Reason for which the HAR model does not belong to the class of long-memory processes (Bauwens et al., 2012, page 368).

[15] The idea of vine copulas goes back to (Joe, 1994; Joe, 1996; Cooke, 1997), while the graphical approach for their construction goes back to (Bedford et al., 2002). A comprehensive theoretical setting on which the most recent literature is based is that of (Aas et al., 2009). This is a common reference, but somewhat generic.

multivariate distributions, conditional or not. Let further c generically represent a copula density and C a copula CDF, conditional or not. The specific meaning for F, f and c is addressed by their arguments and/or subscripts, according to the following notation, referring to distinct indices $i, j, i_1, ..., i_k$ corresponding to distinct random variables $X_i, X_j, X_{i_1}, ..., X_{i_j}$ [16].

The following notation applies for unconditional distributions (uni, bi, and multi -variate respectively):

$$\begin{aligned} F_i &= F_{X_i} = F_{X_i}(x_i) = F(x_i)\,, \\ F_{ij} &= F_{X_i,X_j} = F_{X_i,X_j}\left(x_i, x_j\right) = F\left(x_i, x_j\right)\,, \\ F_{i_1,...,i_k} &= F_{X_{i_1},...,X_{i_k}} = F_{X_{i_1},...,X_{i_k}}\left(x_{i_1}, ..., x_{i_k}\right) = F\left(x_{i_1}, ..., x_{i_k}\right)\,. \end{aligned}$$

While for conditional distributions:

$$\begin{aligned} F_{i|i_1,...,i_k} &= F_{X_i|X_{i_1},...,X_{i_k}}\left(x_i|x_{i_1}, ..., x_{i_k}\right)F\left(x_i|x_{i_1}, ..., x_{i_k}\right)\,, \\ F_{ij|i_1,...,i_k} &= F_{X_i,X_j|X_{i_1},...,X_{i_k}}\left(x_i, x_j|x_{i_1}, ..., x_{i_k}\right)F\left(x_i, x_j|x_{i_1}, ..., x_{i_k}\right)\,. \end{aligned}$$

Similarly, for copulas C_{ij} capturing the dependency between variables i and j:

$$\begin{aligned} C_{ij} &= C_{ij}\left(F_i, F_j\right)\,, \\ C_{ij|i_1,...,i_k} &= C_{ij|i_1,...,i_k}\left(F_{i|i_1,...,i_k}, F_{j|i_1,...,i_k}\right)\,. \end{aligned}$$

The very same notation applies to marginal densities f and copula densities c as well.

To clarify the initial statement, first consider a $d = 3$ case. Define a random vector $\mathbf{X} = (X_1, X_2, X_3)$. From recursive conditioning (law of total probability) we obtain the following decomposition of its density:

$$f(x_1, x_2, x_3) = f(x_3|x_1, x_2)f(x_2|x_1)f_1(x_1)\,. \tag{4.19}$$

By applying the Sklar's theorem[17], conditional densities can be written in terms of

[16] This might appear somewhat not precise, but the following discussion is clear and coherent w.r.t. to this notation.

[17] For distributions, $f\left(x_i, x_j\right) = c_{ij}\left(F_i(x_i), F_j\left(x_j\right)\right)f_i(x_i)f_j\left(x_j\right)$.

copulas and marginal densities:

$$
\begin{aligned}
f(x_2|x_1) &= \frac{f(x_2,x_1)}{f_1(x_1)} = \frac{c_{12}(F_1(x_1),F_2(x_2))f_1(x_1)f_2(x_2)}{f_1(x_1)} \\
&= c_{12}(F_1(x_1),F_2(x_2))f_2(x_2)
\end{aligned}
\tag{4.20}
$$

and

$$
\begin{aligned}
f(x_3|x_1,x_2) &= \frac{f(x_2,x_3|x_1)}{f(x_2|x_1)} = \frac{c_{23|1}(F(x_2|x_1),F(x_3|x_1))f(x_2|x_1)f(x_3|x_1)}{f(x_2|x_1)} \\
&= c_{23|1}(F(x_2|x_1),F(x_3|x_1)) \times f(x_3|x_1) \\
&= c_{23|1}(F(x_2|x_1),F(x_3|x_1)) \times c_{13}(F_1(x_1),F_3(x_3))f_3(x_3)\,.
\end{aligned}
$$

These are referred to as pair-copulas, since relative to the pair of variables i and j (or $i|k$ and $j|k$). Therefore, for the joint distribution $f(x_1,x_2,x_3)$ we obtain the following decomposition:

$$
\begin{aligned}
f(x_1,x_2,x_3) &= f(x_3|x_1,x_2)f(x_2|x_1)f(x_1) \\
&= c_{23|1}\left(F_{2|1},F_{3|1}\right)f(x_3|x_1) \times c_{12}(F_1,F_2)f_2 \times f_1 \\
&= c_{23|1}c_{13}f_3 \times c_{12}f_2 \times f_1 \\
&= c_{23|1}c_{13}c_{12} \times f_1f_2f_3\,.
\end{aligned}
\tag{4.21}
$$

Equation (4.21) represents the joint density of $\mathbf{X}$ as a product of pair-copulas and marginal densities. Pair-copula construction allows to model the joint density in terms or marginals and bivariate copulas between the variables that are directly inferred from the data (in terms of pair-copula family). Therefore the joint density is easily tunable to have margins pair-copula dependencies that match those observed in the sample. For instance a direct d-dimensional fit of a Gaussian copula would imply Gaussian marginal copulas between all the pairs of variables, which might not depict the real dependence between (one or more) pairs of variables. PCC allows for a much great flexibility[18] by perfectly controlling the pair-copulas specifications and by allowing for different constructions alternatives (in terms of variables involved and their order). Here is where the PCC idea lays: by specifying the conditional copulas (and the order in which these are combined) a factorization of the joint

[18] E.g. allowing for asymmetric dependence in tail behaviors (Joe, H. Li et al., 2010)

distribution is obtained. Indeed the decomposition of $f(x_1, x_2, x_3)$ is not unique, and thus its factorization in terms of pair-copulas and marginal densities. For instance the following conditioning differs from eq. (4.19) and leads to a different pair-copula representation:

$$f(x_1, x_2, x_3) = f(x_1|x_2, x_3) f(x_2|x_3) f(x_3) = ... = c_{12|3} c_{13} c_{23} \times f_1 \times f_2 \times f_3 \,.$$

Bedford et al., 2001 introduced a graphical model called *vine* to help to organize and visualize these decompositions: a nested set of trees where the edges in the first tree are the nodes of the second tree, and so on. d-dimensional R-(regular) Vines are subset of vines such that (i) three 1 has d nodes and $d-1$ edges, (ii) three $j = 2, ..., d-1$ has $d+1-j$ nodes and $d-j$ edges, and (iii) two nodes in tree $j+1$ are joined by an edge, the corresponding edges in tree j share a node[19]. Nodes can have different degrees,

Figure 4.4 An R-Vine, $d = 4$. Numbers correspond to variables, e.g. "1" to X_1, "$c_{34|2}$" to $c_{34|2}$. Blue lines indicate edges at tree 1, orange lines edges at tree 2, green line edges at tree 3. The example corresponds to the R-Vine density decomposition $f(x_1, x_2, x_3, x_4) = c_{14|23} c_{34|12} c_{23|1} c_{12} c_{23} c_{24}$.

that is the number of nodes attaching to them. This allows to identify two important sub-classes (i) C-(canonical) Vines, where each node in the tree $j = 1, ..., d-1$ is of maximal degree, i.e. each tee T_j has a unique node of degree $j-1$ and (ii) D-(drawable) vines, where each node if of degree 1 or 2 (Bedford et al., 2001; Joe and Kurowicka, 2011).

As Figure 4.5b depicts, a C-vine has a star structure on trees $j < d-1$: in its first tree a particular variable (*root note*) is set as a note and its dependence is modelled with each other variables via bivariate copulas. Dependencies with respect to a second variable are modelled in the second tree, conditioning on the first variable. For each tree a root node is selected and all the pairwise dependencies w.r.t. this node are modelled conditioned on all the previous root nodes (E. Brechmann et al., 2013). The decomposition of a multivariate density under a C-Vine structure with root

[19] (iii) is called proximity condition and guaranteeing the feasibility of PCC.

nodes $1,...,d$ can be easily obtained by combining the recursive relation (e.g. Joe and Kurowicka, 2011):

$$f\left(x_i|x_1,...,x_{i-1}\right) = c_{(i-1)i|1,...,i-2} \times f\left(x_i|x_i,...,x_{i-2}\right),$$

with

$$f(x_1,...,x_n) = f\left(x_d|x_1,...,x_{d-1}\right) f\left(x_1,...,x_{d-1}\right) = f_1(x_1) \prod_{i=2}^{d} f\left(x_i|x_1,...,x_{i-1}\right).$$

This gives the *C-vine density*:

$$\begin{aligned} f(x_1,...,x_d) &= \prod_{i=1}^{d-1} \prod_{j=1}^{d-i} c_{i(i+j)|1,...,i-1} \times \prod_{k=1}^{d} f_k \\ &= \prod_{i=1}^{d-1} \prod_{j=1}^{d-i} c_{i(i+j)|1,...,i-1}\left(F\left(x_i|x_1,...,x_{i-1}\right), F\left(x_{i+j}|x_1,...,x_{i-1}\right)\right) \\ &\times \prod_{k=1}^{d} f_k(x_k). \end{aligned} \tag{4.22}$$

Omitted from the notation are the bivariate copula parameters $\theta_{i(i+j)|1,...,i-1}$, corresponding to each of the $c_{i(i+j)|1,...,i-1}$ copulas. The outer product runs over $d-1$ trees, while within each three there are $d-i$, $i=1,...,d-1$ pair copulas, accounted for in the inner product.

D-Vines differ in how the dependencies are modelled: in the first tree pair-copulas model the dependencies between the first and the second variables, the second and the third, the third and the fourth and so on. On the second tree, conditional on the second variable, the dependence between the first variable and the third is modelled; on the third three. conditional on the third variable, the dependence between the second and the fourth is modelled (Figure 4.5b). Similarly as for the C-Vines, using (Aas et al., 2009, equation 6):

$$f\left(x_j|x_{j+1},...,x_{j+i}\right) = c_{j(j+i)|j+1,...,j+i-1} \times f\left(x_i|x_{j+1},...,x_{j+i-1}\right).$$

This leads to the *D-Vine density*:

$$
\begin{aligned}
f(x_1,...,x_d) &= \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i(i+j)|(i+1,...,i+j-1)} \prod_{k=1}^{d} f_k \\
&= \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i(i+j)|(i+1,...,i+j-1)} \Big(F\Big(x_i|x_{i+1},...,x_{i+j-1}\Big), \\
&\quad F\Big(x_{i+j}|x_{i+1},...,x_{i+j-1}\Big)\Big) \times \prod_{k=1}^{d} f_k(x_k)\,.
\end{aligned}
\tag{4.23}
$$

Note that eq. (4.22) and (4.23) are special cases of the general R-Vine density [20] Bedford et al., 2001, theorem 3. The R (regular)-Vine class is much wider than the C- and R- vine classes[21], however the enormous amount of different R-vine structures to chose from is the main drawback: indeed the vast majority of applications involve C- and D- vines. A discussion on structure selection and estimation for R-Vines is provided in (e.g. Dissmann et al., 2013; Cooke et al., 2015). Importantly, conditional distributions appearing as arguments in the conditional copulas can be evaluated by the following recursive relationship (Joe, 1996):

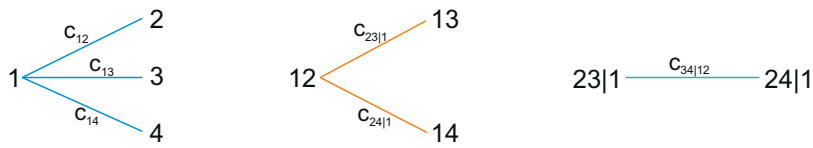
$$
F(x|\boldsymbol{v}) = \frac{\partial C_{xv_j|\boldsymbol{v}_{-j}}\Big(F\Big(x|\boldsymbol{v}_{-j}\Big), F\Big(v_j|\boldsymbol{v}_{-j}\Big)\Big)}{\partial F\Big(v_j|\boldsymbol{v}_{-j}\Big)}\,,
\tag{4.24}
$$

where $\boldsymbol{v}$ is a m-dimensional vector, v_j any of its components and $\boldsymbol{v}_{-j}$ the $(m-1)$-dimensional vector obtained by excluding v_j from $\boldsymbol{v}$. This clearly displays at a general level the sequential mixturing of conditional CDFs with copulas and how copulas on the previous tree enter as arguments in the next one.

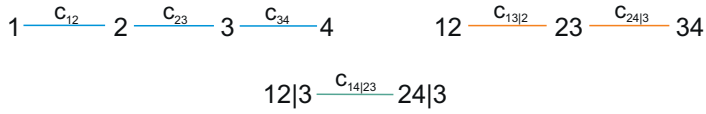
[20] Consider d variables, $\mathcal{V} = \big(T_1,...,T_{d-1}\big)$ representing a set of nested trees, $\mathbf{N} = \big(N_1,...,N_{d-1}\big)$ the node, $\mathbf{E} = \big(E_1,...,E_{d-1}\big)$ the edge set corresponding to trees in $\mathcal{V}$, and the conditions (i)-(iii) earlier mentioned defining an R-Vine. Be $c_{e_1e_2|D_e}$ bivariate copula densities assigned to each edge $e = (e_1, e_2|D_e)$ in E_j, $j = 1,...,d-1$ corresponding to the variable pairs $\Big(X_{e_1}, X_{e_2}|\mathbf{X}_{D_e} = \mathbf{x}_{D_e}\Big)$, with $\mathbf{x}_{D_e} = \{x_i : i \in D_e\}$ sub-vector of $\mathbf{x}$ with index set D_e, $\mathbf{x} = (x_1,...,x_d)$ random vector of observations. The general *R-Vine density*, (generalizing eq. (4.22) and (4.23), is given by:

$$
f(x_1,...,x_d) = \prod_{j=1}^{d-1}\prod_{e\in E_j} c_{e_1e_2|D_e}\Big(F\Big(x_{e_1}|\mathbf{X}_{D_e}\Big), F\Big(x_{e_2}|\mathbf{X}_{D_e}\Big)\Big) \times \prod_{j=1}^{d} f_j\Big(x_j\Big)\,.
$$

[21] There are $\binom{d}{2}(d-2)!2^{\binom{d-2}{2}}$ regular vines on d variables (Cooke et al., 2015).



(a) C-Vine



(b) D-Vine

Figure 4.5 C- and D- Vines. The interpretation is same as that of Figure 4.4. The C-Vine example corresponds to decomposition $f(x_1, x_2, x_3, x_4) = c_{34|12}c_{23|1}c_{24|1}c_{12}c_{13}c_{14}f_1f_2f_3f_4$, the D-Vine to $f(x_1, x_2, x_3, x_4) = c_{14|23}c_{13|2}c_{24|3}c_{12}c_{23}c_{34}f_1f_2f_3f_4$.

4.3.2.1 Estimation

From a set of possible copula families[22], the selection of given parametric copulas for the pair-copula (conditional or not) construction is performed sequentially tree by tree, since eq. (4.24). I.e. next tree's specification depends on the current tree. Given that a vine structure is decided (R, C or D), the pair-copulas therein involved are known: the estimation reduces in identifying the correct parametric family for each of them and in estimating its parameters. The unconditional pairs at the first tree can be directly estimated from the transformed data, by likelihood maximization (or minimization of AIC/BIC statistics) and with the aid of a number of graphical methods (Section 4.2.2 of Publication IV describes them). The specification of the copulas at the first tree is crucial since they affect the whole structure in virtue of eq (4.24). Given the set of estimated pair-copulas at the first tree, conditional copulas at higher trees are estimated by drawing pseudo-sampled of the conditional CDFs involved in the arguments of these copulas, and obtained from the copulas identified at earlier trees (eq. (4.24)). Higher-trees copulas are selected as well on the basis of a suitable criterion based on the likelihood evaluation. This is however achieved in a trial-and-error setting where all the copulas from different families are estimated and

[22]Standard choices are Archimedean copulas, Gaussian and t-copula, including their rotations.

only the best-performing one retained.

The log-likelihood obtained by applying a log-transform on eq. (4.22) and (4.23) reduces to a sum, where each factor corresponds to the log-likelihood of each pair-copula (conditional and not). For a C-Vine the log-likelihood reads:

$$l(\theta|\mathbf{x}) = \sum_{k=1}^{N}\sum_{i=1}^{d-1}\sum_{j=1}^{d-i}\log\Big[c_{i,i+j|1:(i-1)}\Big(F_{i|1:(i-1)}, F_{i+j|1:(i-1)}|\theta_{i,i+j|1:(i-1)}\Big)\Big]\,, \quad (4.25)$$

with $i_1 : i_k$ denoting $i_1, \dots i_k$, $\theta_{i,i+j|1:(i-1)}$ the parameter set corresponding to the copula $c_{i,i+j|1:(i-1)}$, θ their collection, and $\mathbf{x}$ collecting N observations for d variables. Thus the overall maximization of the log-likelihood for the joint density is obtained by maximizing each of the terms in the summation, i.e. through a one-by-one solving of all the pair-copulas in all the $d-1$ trees. Further details on selection of each bivariate pair can be found e.g. in Czado et al., 2012.

The above-mentioned estimation procedure and copula selection rely on the fact that there is a dependence between the variables to be modeled, so that variables correlate to some extent. The most immediate way to test for correlation is by using a correlation measure. The well known Pearson's correlation coefficient ρ is clearly affected by marginal distributions since it involves the first two moments of the random variables under investigation.

Being dependence captured by copulas only, a correlation measure that is only copula-driven and does not depend on the specific margins constitutes a better alternative. Kendall's tau rank correlation ρ_τ is not affected by the margins, since it is invariant under strictly increasing transforms of the random variables, hence only dependent on the underlying copula. For two random variables u, v and their copula C, $\rho_\tau$is defined as (Schweizer et al., 1981):

$$\rho_\tau = 4\int_0^1\int_0^1 C(u,v)\partial C(u,v) - 1\,.$$

For Gaussian, Student-t and other copulas a relation exists between ρ and ρ_τ. Importantly, for a wide range of parametric copula families a relation between ρ_τ and the copula parameter also has been established. For instance, for the Gumbel copula with parameter θ, $\rho_\tau = 1/\theta$, for a Gaussian copula with parameter ρ, $\rho = \sin(\rho_\tau \pi/2)$.

This relation (when available) paves the way for a direct parameter's estimation based no longer on maximum likelihood, but on the easier Kendall's tau-to-parameter inversion (for multivariate copulas generalization of the above definitions and relations exist, but maximum likelihood method is preferred). It is clear that this moment-like estimator with respect to the ML estimation is very attractive since is computationally straightforward and simple. Given that the efficiency of the estimator (so the standard error of the estimates) is not a primary concern, the inversion method is preferable, simply because of its simplicity. Note however that this applies to copulas driven by a single parameter (i.e. not applicable to a t-copula). Similarly, estimation is also possible based on Spearman's rank correlation coefficient or Gini margin-independent correlation.

4.3.2.2 Vine copulas in practice

The previous section presented the methodological background based on pair-copula construction and introduced the Vine-copula class. Here the procedure for the practical estimation of a Vine copula is outlined:

i Identify a suitable vine-copula structure for the problem being analyzed. I.e. within the C-Vine class select a tree. Based on the nature of the problem under analysis, perhaps some trees might not admissible and among the admissible ones, a particularly suitable tree can be identified. If this is not the case, so that tree structures are qualitatively equivalent then apply steps i-v to find the structure that e.g. maximizes the C-Vine likelihood.

ii Based on the previous step, certain pair-copulas c_{ij} need to be estimated, $i, j \in \{1, .., d\}$ and $i \neq j$. Use the data sample to conveniently model the margins F_i and F_j and reduce the sample to the $[0, 1]$ interval: obtain $u_i = F_i(x_i)$ and $u_j = F_j\left(x_j\right)$ for all the variables involved in the pair-copulas.

iii Use the sample data to estimate all the unconditional pair-copulas c_{ij}, after running an independence test to immediately identify independence copulas and, if needed, review the structure identified in step i.

iv Sequentially determine all the conditional copulas in higher trees $c_{ij|\cdot}$ conditioning on the variables in the first tree (which ones depend on step i). This is

achieved by applying a suitable criterion, such as the overall likelihood maximization or AIC/BIC minimization.

v Retrieve the Vine-copula density[23], i.e. apply (4.22).

vi A multivariate density modeling of the original sample (which is not in general of uniform margins) is obtained by applying Sklar's theorem.

Different approaches for modeling the marginal distributions in step ii are extensively discussed in Publication IV, Section 4.2.2. Graphical procedures for choosing adequate unconditional pair-copulas and likelihood-based tests are therein presented as well.

[23] Steps iii and iv can be merged and automated so that all the bivariate copulas involved at any tree are selected in such a way that in the estimated parameters the likelihood (e.g.) is maximized. In lower dimensions, however, a visual inspection and use of graphical methods for assessing the goodness-of-fit of pair-copulas on tree 1 is advisable. Indeed misspecification of the copulas at the first tree propagates to the whole structure given the sequential pair-copula construction of the joint density. For instance, w.r.t. eq. (4.21), a misspecification of c_{12} would lead to inadequate pseudo-samples of $F_{2|1}$, upon which $c_{23|1}$ is selected and estimated.

5 RESULTS

Findings of Publications I-IV are summarized in this chapter. The first sections provide a publication-specific extended and precise summary of the findings in Publications I-IV, but omitting the exclusively technical findings for which I refer to the research publications appended to this dissertation. For the sake of coherence with respect to the initial goals, each subsection terminates by recalling the corresponding research questions from Section 1.2 and summarizing the key-points aimed at addressing a concise and direct answer.

5.1 Forecasting mid-price movements with machine learning techniques

The prediction of the limit order book flow is an interesting topic both for practitioners and researchers. Among all the possible prediction tasks, Publication I and Publication II focus on the prediction of mid-price movements. In the two Publications, this is posed as a classification problem for the three labels indicating whether the mid-price movement is stale, decreasing or increasing. The prediction horizon is specified as the number of future order book events from the time at which the prediction is referred to (evaluated). Publication I analyzes the mid-price direction predictability after 1, 2, 3, 5, 10 events, Publication III for a 10-events horizon. Before addressing the finding, is important to remark that perhaps among the contributions of Publication I there are no complex applications and fine-tuning of ML methods for maximizing the performance of the above-mentioned prediction task. Rather, Publication I introduces a benchmark limit-order book dataset designed for future ML applications (and clearly, the current one). In this concern, the dataset construction,

order book processing, features extraction, and input normalization (see Section 3) is one of the major tasks. For the reproducibility of the results and for them to constitute a reliable benchmark, the experimental protocol based on increasing foldings (see Section 3.4 in Publication I) have been designed and accurately described. In this regard, the data-handling part in Publication II is minimal, and corresponds to a direct application over the very same dataset developed in Publication I (actually, on the subsample utilizing z-score normalization). But whereas Publication I utilizes the traditional approach of forming input vectors from features at a given time instance, Publication II uses tensor representations, where the time dimension is retained. In fact, the conversion between a tensor representation to a vector representation leads to the loss of temporal information.

Turning to actual results, Publication I implements two ML methods under three different data-normalization schemes and a common experimental protocol. The performance measures Publication I report show that even in the noisy high-frequency LOB data ML can effectively retrieve signals for useful for prediction. At shorter horizons, the performance is however affected by the data-noisiness, i.e. microstructure effects that prove to make the very-short-term prediction of the mid-price movement challenging. Anyhow the random mid-price dynamics, given as a noise superimposition over the efficient price, seems not to be a completely independent process, indeed the input vectors extracted from the past dynamics of the observed mid-price turns to be capable of achieving satisfactory performance measures also at 1, 2, 3 event-horizons by. However, if these prediction approaches are to be exploited by practitioners, reactions within the next-event are unfeasible (the median inter-event duration is 64 milliseconds), thus the predictability over a slightly longer horizon is more relevant from a practical perspective. Despite the data size, the out-of-sample performance (F1) is up to 43% for both methods, showing that basic machine learning techniques can effectively provide a satisfactory classification of the direction of mid-price movements. Among the normalization methods, results clearly indicate that the role played by the data-normalization is not secondary. Furthermore, we have evidence that specific combinations of ML algorithm and normalization schemes lead to considerably different performance measures than others.

RO 1 of this dissertation aims to explore the applicability of standard ML methods for the mid-price movement prediction task. Also, by addressing the role played by different normalization schemes and forecasting horizons. To pursue this goal Publi-

cation I applies two simple ML methods, under three data-normalization approaches and evaluates out-of-sample forecasting performance measures. Results confirm the feasibility of simple ML methods for this prediction task and that normalization is a key ingredient, perhaps more relevant than the classifier. The ultra-high frequency dataset developed for answering this question is made freely available and inclusive of a detailed description of the experimental protocol as well as the definition of the feature set and variables therein provided, fulfilling the second part of the research goal.

Linear discriminant analysis is applied in Publication II and proved to be an effective method of the mid-price direction prediction task. However when relying on a tensor representation of the data the corresponding multilinear discriminant analysis boosts all the performance measures up to approximately 15% and 5% with respect to the worst and best competing input-based alternatives. This indicates the importance of the contribution of the temporal information captured in tensor representation: not only current features are important for the prediction task, but interrelations in their lags generously contribute to performance improvements. Furthermore, in Publication II a regressor operating on the tensor representation is developed, and a scheme to select the best-performing model state discussed based on the algorithm's learning dynamics. This method leads to the highest F1 scores among the LDA, MDA, benchmarks of Publication I and the bag-of-features algorithms in (Passalis, Tsantekidis et al., 2017).

RQ 1 is addressed by implementing four ML methods based on time-series' tensor-representation and compared with the results from input-vector alternatives applied in the literature. Our results show that the extent to which forecasting performance measures are improved is generally widely significant for all the four measures we considered (in this regard, see Table 6.1).

5.2 Long-range correlations in limit order book markets

Publication III provides a study on the scaling behavior for duration-related variables extracted from the order book data of five securities. The scaling exponent is extracted

with the DFA method computed for, inter and cross -events durations[1]. The scaling exponents we find for order-to-order, trade-to-trade and cancel-to-cancel series are consistent with earlier analyses in the literature (e.g. Ivanov, Yuen and Perakakis, 2014; Gu et al., 2014). Power-law exponent estimates are very consistent ($\alpha \simeq 0.6$) across different stocks (and side of the book), suggesting that fractality in durations is a general phenomenon (aligned with Ivanov, Yuen, Podobnik et al., 2004), although there are some minor differences especially in the cross-events durations between the stocks traded at Helsinki, Copenhagen and Stockholm. This suggests some exchange-specific features in the long-range correlations, e.g. not all the market participants trade on multiple exchanges. Our analysis finds evidence of crossovers in the time-series we analyzed. This confirms the finding of Ivanov, Yuen, Podobnik et al., 2004 but extending it to different duration series and detecting it over several stocks. This point out that the observed fractality is complex and time-related being the compound effect of different scaling exponents characteristic of different time horizon domains. Therefore fractal properties in LOB markets seem to be indeed quite complex, reflecting the complexity of the underlying markets. The crossovers observed are interestingly placed around a day, a week and a month time scales. This analysis was possible only due to the availability of a long data period, indeed earlier studies identified two crossovers (e.g. Ivanov, Yuen, Podobnik et al., 2004; Ivanov, Yuen and Perakakis, 2014; Tiwari et al., 2017). This evidence supports the idea that there might be participants trading at different horizons o however interactions between daily, weekly and monthly goals. This aligned with the theoretical argument of (U. A. Müller, Dacorogna, Davé, Pictet et al., 1993), although Publication III provides evidence only with respect to durations. Furthermore, we find evidence of great symmetry with respect to the book side, indicating either a general and uniform behavior in market's participants, either their tendency in submitting both buy and sell limit. Importantly, this is the first analysis where different duration series are jointly analyzed within the same order book data, i.e. for a given stock and a given period we consider all the possible durations within different LOB events. Our findings unveil a true multi-level and interacting complexity: all the series and the corresponding cross-series are multifractal with a characteristic scaling exponent being very consistent on a daily level. This may indicate that e.g. trading algorithms are similar in the way the past information is processed for placing limit-orders,

[1] As a reminder, only best bid an ask levels are considered

market-orders, and cancellations.

Publication III also studies the relationship between the intra-day scaling exponent and some economic variables (like daily return and volatility). For the correlation between the scaling exponent and volatility, we have clear positive significance for all the inter-event series. Clear clusters in the correlation between α and economic variables are observed in the trade-to-trade durations, although some of them are not significant, while widespread values are observed for the cancel-to-cancel and order-order series. Very generally, this whole analysis indeed points out a true complexity in the order book dynamics, unveiling general long-range autocorrelation in the duration series ($\alpha > 0.5$) but of complex nature varying with the time-scale (crossovers).

RQ 2 is answered by analyzing the scaling exponent for different sets of duration series, not limited to single stock not to a certain side of the book. Findings from Publication III show that long-range autocorrelations are ubiquitous in all the time-series under investigations, and showing a generally remarkable homogeneity. Furthermore, the availability of long-span high-frequency data allows to unveil up to three different scaling exponents applicable at different sampling frequencies.

RQ 3 deals with the associations with economic variables and long-range autocorrelations. Publication III analyzes the association for three duration series and six economic variables, finding it to be of very heterogeneous nature across the variables, and for the order and cancellation series, across the stocks too. Indeed for certain variables and time-series, the association is either significant, largely positive, and consistent across the different factors, while for others is the opposite.

5.3 Volatility forecasting using copulas

Volatility modeling and forecasting is an important topic in econometrics. Publication IV tackles the problem of one-step-ahead forecasting of daily realized measures in a regression-like framework resembling the spirit of the HAR model. Publication IV develops an alternative model against the benchmark of Corsi, 2009. By discussing sources of misspecification and potential drawbacks in the HAR model given the complexity of volatility time-series, and alternative specification and regression approach is discussed. Publication IV suggests a novel method in linking the volatility terms

involved in the HAR model by use of the recent developments of a particular copula class, known as C-Vine copulas, for the construction multivariate distributions. On a wide high-frequency dataset of 10 stocks for 1634 trading days, realized measures are forecasted under the C-Vine HAR (CV-HAR) model. Results show improvements over the HAR model under a wide set of performance measures, both in in-sample analyses and out-of-sample analyses, using a fixed-window, an increasing-window and a rolling-window approach. Furthermore, the computation of the conditional expectation used as volatility forecast is shown to be achievable by numerical integration, in a modeling approach that involves complex mixtures of conditional distributions and copulas, thus proving the methodology of being tractable and of smooth implementation aside the theoretical complexity of vine copula construction.

Publication IV is implemented over tick-by-tick realized kernel estimates. Improvements of the C-Vine HAR model over the HAR one are shown to hold for all the measures, suggesting that besides the specific measure and sampling frequency chosen, the CV-HAR model indeed captures non-linear aspects in volatility dynamics that the HAR model misses. This holds particularly true for the non-log realized measures, where normality hypothesis in the error terms of the HAR models are clearly stretched, along with the non-negativity requirement for volatility measures. Copula construction relies on inverse CDF transforms, which are implemented with three approaches, ECDF (parametric approach), kernel smoothing of the ECDF (semi-parametric approach) and by skew-t and Inverse-Gaussian distributions CDF (parametric approach). Pair copulas involved in the C-vines construction separately involve two sets of copulas, Archimedean copulas only and Archimedeans with Gaussian and t- copulas. Results are consistent under all combinations of copula sets and inverse CDF methods. This suggests that is neither the complexity of the copulas involved in the CV-HAR model nor the method to construct the copula-samples for estimation that drives the results, i.e. there is a general gain in using a copula-based modeling approach over the HAR model. Finally, reminding hat Publication IV is inspired by the work of Sokolinskiy et al., 2011, where, in a simplified framework and with a simple and standard bivariate copula-based approach, tomorrow's volatility is forecasted based on today's with a simulation approach, the CV-HAR vine method generalizes the forecasting approach of Sokolinskiy et al., 2011 in terms of model flexibility, multivariate generalization and in this augmented setting develops an estimation method not relying on simulation. Moreover, for evaluating the performance

of the two competing models, six different measures and with formal statistical testing have been used.

RQ 4 deals with the impact that copula-based modeling for daily volatility measures has over the standard HAR model in terms of volatility forecasting. To answer this point a wide tick-by-tick dataset for ten securities covering a period of five years have been used in Publication IV. By exhaustive in-sample and out-of-sample analyses, supported by statistical tests, and a series of robustness check on the methods, indicates that the model suggested in Publication IV, methodologically relying on copulas, outperforms the HAR model under a wide set of performance measures.

6 CONCLUSIONS

In this chapter provides the final remarks for the dissertation. First, its contribution is addressed, then the reliability and validity of the research in publications I-IV are discussed. Finally, limitations, and therefore hints and suggestions for future research, are presented.

6.1 Contributions

This dissertation and specifically the publications herein presented, generally contribute to the high-frequency econometrics literature. This is achieved through three different perspectives and involve three different areas, however, linked among each other as addressed in Section 1.3. First, this dissertation looks at the problem of mid-price prediction in limit order book data, also presenting a new publicly available limit-order book dataset for machine learning applications. Second, it explores the fractal nature of different duration time-series extracted for several order book events. Third and lastly, it deals with the prediction of daily realized measures of volatility, under a copula-based approach.

Two are the contributions from Publication I, a major one and a minor one. The major contribution is not a proper literature contribution, rather a contribution to the research community involved in ML application in high-frequency finance. Indeed, a Limit Order Book (LOB) dataset is one of the outcomes of this research. Publication I relies upon a substantial heterogeneity in the datasets so far utilizes in high-frequency ML applications and the challenge of accessing free but detailed and accurate LOB data. As a consequence, comparability and reproducibility of the earlier studies are particularly challenging, also because the overall experimental protocol design is not uniformly addressed. With Publication I we disclose a publicly available

Level-II LOB dataset including the first 10 best order book levels on both the book sides along with a number of features extracted from the raw data, capturing statics and dynamical aspects of the messages inflow. As a second contribution, the problem of the mid-price prediction is addressed (for this specific data, for the specific output classification labels and estimation-forecasting scheme, and in terms of movement in the next $1, 2, 3, 5, 10$ events, but for different data normalization procedures). For this purpose, two standard ML methods (ridge regression and single layer forward-feed network) are implemented, and the respective performance measures (and errors) reported. A well-specified experimental protocol guaranteeing the replicability of the results is carefully described. In this way, the datasets along with the prediction outputs of the mid-price prediction task constitute, inclusive of performance results for standard models for future machine learning applications.

The natural tensor representation of a time-series constitutes an attractive direction for time-series modeling. Several ML algorithms exploit its classical representation in term of time-specific vector-sets of features, where the whole intertemporal connections are disregarded. In the perspective, Publication II develops and utilizes ML strategies capable of dealing with tensor inputs. Classical Fisher's Linear Discriminant Analysis (LDA) (e.g. Welling, 2005) relies on vector inputs, but its extension known as Multilinear Discriminant Analysis (MDA) is capable of accommodate tensors as inputs. LDA and MDA are applied fro the mid-price forecasting problem on the very same dataset of Publication I, and the boost in performance measures is shown to be quite noticeable when switching from LDA to MDA. However, MDA stands as a complex method for its implementation and parameter estimation. An alternative estimator based on the tensor representation of the time-series is therefore developed: Weighted Multi-channel Time-series Regressor/Regression (WMTR) [1] WMTR is of easy implementation and fast calibration. However, a number of parameters are required to be properly tuned. A method for the optimal parameter selection based on the algorithm learning rate is devised and applied. With respect to the ML methods

[1]The term "channel" is popular in signal processing language, background of the co-authors in Publication II, while "multi-channel" means multivariate inputs. "Channel" is considered as a signal that is observed/acquired with different sources. It might represent different frequency bands, or e.g. inputs from different sensors placed at different positions in space (conveying spatial information). For the dataset Publication II uses, "channel" refers to the 144 different features (characteristics, data-representations) that are perceived (extracted) from the data (Section C therein), while "multichannel" refers to the actual nature of the inputs feeding the algorithm, i.e. collections of multiple channels (vectors of features).

implemented in Publication I and the results of (Passalis, Tsantekidis et al., 2017), the WMTR leads to the best-performing F1 measure. A tensor-based representation of the time-series and appropriate ML methods capable of exploiting it seems, therefore to input vector-representations.

Publication III contributes to the literature by providing an analysis on the fractal properties of duration times series extracted from the limit order book. Earlier studies have separately analyzed inter-trade durations (Ivanov, Yuen and Perakakis, 2014, e.g.) and inter-cancellation duration (e.g. Gu et al., 2014), and detected cross-overs in the scaling exponent for inter-trades durations. Along their lines we methodologically adopt the detrended fluctuation analysis to characterize the long-range correlation for inter -order, -trade, - cancel durations as well as for the duration of the cross-events, order (submissions) to (their) cancellation, order-to-trade and order-to-cancel durations. This is done for the best book levels on both sides, for different securities all listed ad NASDAQ Nordic, but traded on different exchanges. The analyses point out a ubiquitous presence of long-range autocorrelation in all the series analyzed, consistency in the scaling exponent estimates within the single exchange and some heterogeneity between exchanges. For all the series, crossovers are identified at day, week, and month horizons, while (Ivanov, Yuen, Podobnik et al., 2004; Ivanov, Yuen and Perakakis, 2014; Tiwari et al., 2017) reported two scaling exponents for the inter-trade durations only. This indicates that fractal properties in duration series are more complex of how previously thought. Following and widely expanding (Ivanov, Yuen and Perakakis, 2014), we explore the association of some relevant but generic economic variables and the scaling exponent. The most relevant association we find is that between the scaling exponent and volatility, which is of strong financial interpretation. Furthermore, associations between the scaling exponents in the order-to-order and cancel-to-cancel series and the economic variables show complex and widespread patterns for the ten time-series involved (5 stocks, bid and ask side for each), underlying the complex nature of long-range autocorrelation in the order book, specifically for the duration series analyzed.

Publication IV presents a novel method for volatility modeling and forecasting, in particular, in modeling and forecasting daily measures of realized volatility. Several methods have been proposed as extensions of the HAR model of (Corsi, 2009). Among them, (A. J. Patton and Sheppard, 2015) separates the positive and negative contributions of intraday returns to the realized volatility, (Bollerslev et al., 2016) ex-

ploits the discrepancy between the Realized Variance (RV) measure and the Integrated Variance (IV) for finite samples, (Andersen, Bollerslev and Diebold, 2007) accounts for a jump component and, for instance (Hillebrand et al., 2007; McAleer et al., 2011; Buccheri et al., 2017) introduce non-linearities in the HAR specification. In this regard, econometric literature lacks a copula-based approach. Publication IV shows a close connection with the linear regression in the HAR's models formulation and the modeling of conditional expectations of general multivariate distribution, achievable with the so-called pair copula construction method. This approach is inspired by the work of (Sokolinskiy et al., 2011) which makes use of simple bivariate copulas for the modeling joint distributions. However, the actual framework of the HAR model, which serves as a basis and benchmark for the CV-HAR model calls for the flexible yet parsimonious multivariate copula construction method, i.e. Vine copulas. Publication IV shows how to tackle the problem of predicting tomorrows' volatility given the past information in this setting, with several different approaches in e.g. modeling the marginal CDFs or in pair-copulas selections. Importantly, it shows that the conditional expectation which serves as a predictor can be quickly and efficiently computed by numerical integration: although the complexity of the distribution (obtained as a mixture of copulas and marginal CDFs) no simulation methods are invoked. A reliable and wide dataset involving 10 stocks confirm the improvement of the CV-HAR model over the HAR model in forecasting daily realized measures. Following the practice of (e.g. Andersen, Bollerslev and Diebold, 2007; Bollerslev et al., 2016) different performance measures, in-sample and out-of-sample schemes for their evaluation and proper statistical testing are considered, confirming that the CV-HAR model is able to capture features in the volatility dynamics that the HAR model is unable of. Furthermore, the CV-HAR model has a number of smaller advantages over the HAR, e.g. does not lead to negative volatility forecasts.

6.2 Reliability and validity of the research

Trustworthiness of the results and rigor in research is important for assessing the reliability and validity of a study. Reliability relates to the accuracy of an instrument, i.e. refers to whether a thing has been done correctly (Heale et al., 2015; Siikanen, 2018). Validity, relates to the extent to which a concept is accurately measured in a

quantitative study, i.e. to how correct things have been done to answer the research questions, so to the closeness of what we believe we are measuring to what we intend to measure (Roberts et al., 2006).

Valid and reliable research requires to be developed over a solid high-quality data. All the research Publications I-IV rely on quality data, obtained from official sources. For Publication I, Publication II and Publication III the limit order book data have been directly acquired from NASDAQ Nordic. The acquired data consists of the full and complete raw data, that has been processed according to the description provided in Chapter 3. This stems for an in-depth understanding of the dataset and its processing: all data handling and in-between operations leading to a convenient workable structure has been carefully done. In particular the creation of the limit order book released with Publication I has been carefully engineered and double-checked at each step. The TAQ dataset is the most wide-spread high-frequency data used in econometric research. Files have been downloaded from Aarhus University servers, in a well-documented Matlab format, processed and cleaned with pre-existing functions, along with the rigorous guidelines of (Barndorff-Nielsen et al., 2009).

Reliability of methods in Publication I is guaranteed by the extensive literature supporting them and the numerous applications exploiting them that have already been published. ML methods in Publication I are supported by decades of literature, and by their wide application across different fields, which include high-frequency financial and econometrical domains as well (as reviewed across Chapter 2). The same applies to Publication II and Publication III, in addition, results are also coherent with earlier studies in the literature and those of closely-related applications. Concerning Publication IV, the implementation of the Vine copula constriction is made via a well-documented a widely applied R-package. The implementation of integration involved in the conditional expectation computation was verified via Monte-Carlo methods and tested over a wide number of toy-examples and not. As a robustness check, the implementation has been tested for anomalous behaviors around limiting and extreme values of its arguments as well. Different CDF construction methods, sets for copulas, volatility measures, in-sample and out-of-sample analyses (along with appropriate statistical tests and six performance measures), support the reliability of the results, which are not biased nor driven by a "lucky" sample (since involving 10 stocks over a period of five years).

Validity with respect to the research questions holds as well. All the research questions have been addressed by the use of proper methods suitable for the specific problems that each publication deals with, and the goal of developing a benchmark limit order book dataset achieved. For each publication, the decision on the methodological setup has been carefully discussed and compared with existing alternatives, to ensure it captures the exact features and quantities of interest, while sources that can potentially drive or bias the analyses have been controlled for. E.g. exclusion of the first and last 30 minutes of the regular trading hours (Publication I, Publication II, Publication III) or the implementation of the complex realized kernel, among a list of simpler alternatives, but with stricter hypothesis on the microstructure noise (Publication IV). Furthermore, Publication I,Publication II and Publication III have been peer-reviewed and updated accordingly. Moreover, several researches already made use of the dataset developed for Publication I (see the bag-of-features -related literature in Section 2.3), confirming its quality. Theoretical and methodological backgrounds for all the publications are supported by a wide literature and careful examination of the most relevant studies related to each publication and accounting for the field-specific common practices.

6.3 Limitations and suggestions for future research

6.3.1 Publication I and Publication II

Whether a research limitation is pointed out, a hint for improvement and thus future research is at the same time addressed. Research limitations and future research thus move very close to each other and are here jointly discussed.

Publication I and Publication II rely on the very same data, whereas this constitutes a solid point for result replication and comparability it also constitutes a limit, since the results therein cannot be directly abstracted toward a general level and show to globally hold. The dimensionality of the data forces to focus the attention on a predefined and limited number of securities. Theoretically, nothing prevents other securities (e.g. less liquid ones) to behave differently and to lead to even very different results for the mid-price forecasting problem. Second, the sample period is limited

to 10 trading days: results can be very different over longer periods or, for instance, during a financial crisis time. As Publication III points out, there are exchange-specific differences in the results: a limit-order book dataset involving different exchanges also for Publication I and Publication II would be interesting to analyze. Although most of the above-mentioned restrictions are because of the size of the data, indeed it would be meaningful to address how the use of a LOB dataset limited to e.g. 5 levels or augmented to deeper levels, would impact the results. I.e. understanding whether results get much better with deeper levels or much worse with fewer levels. In Publication I the prediction task relies on standard ML methods: different and more sophisticated methods could provide a better insight on mid-price predictability. On the other hand, in Publication II the MDA is quite complex and the proposed weighted-multichannel method requires a number of hyperparameters to be fine-tuned. What is a good balance between complexity and forecasting performance is an interesting point to be eventually discussed. After all, can we be sure that ML methods would outperform analytic ML methods in prediction? For instance, it would be valuable to understand whether and in which extent complex networks architectures can outperform attractive and parsimonious models such as the one of (Cont, Stoikov et al., 2010). A final remark is about the prediction task itself. The mid-price direction is classified according to three classes, can this be refined e.g. over a finer grid for price changes? Furthermore, because of operational constraints, the k-events-ahead prediction framework is of convenience. However, can we predict the mid-price movement in terms of actual time rather than event time? But how to deal with illiquid stock where transactions might have seldom occurrence? Moreover, what about other prediction tasks? Can we devise a general strategy and learning protocol that holds for different experiments? These are all interesting aspects that future works might address, and that the current research and design of Publication I, Publication II do not touch.

A limitation for the studies in Publication I and Publication II is that of not assessing statistical differences between the developed methods and the existing implementations. We report classification errors for the different models coherently with the common practice in the field, e.g. by addressing precision, recall, accuracy, and F1 measures (e.g. Kercheval et al., 2015; Tsantekidis et al., 2017b; Dixon, 2018b, wrt. financial applications), including their respective standard errors as well (though not always reported in applications, but not always applicable too). Criticism related to

this is that results do not address statistically significant differences between different implementations, and do not allow to clearly discern whereas there is an effective outperformance of a certain prediction method over another. In virtue of Chebyshev's inequality, standard errors can only outline k-standard deviations confidence intervals of worst-case probabilities. Without any connection with the distribution of the measure under analysis, point estimates and standard errors are per se little informative. t-testing would represent a feasible alternative that however Publication I and Publication II do not explore. Also, paired t-test for N-fold cross-validation (eg. comparing N accuracies obtained by training and testing N different instances of the same classifier on N equally-sized subsets) is an omission that could potentially shed light on the different performance of the different classification methods they involve.

The above-mentioned test is however addressed in Table 6.1. Based on the point estimates and standard deviations from Table I in Publication II, Table 6.1 reports p-values for the two-sided t-test for different models against the WMTR. I.e. it addresses whether there is a statistical difference in accuracy, precision, recall, and F1. Although unequal variances are taken into account, it has to be pointed out that the training-testing setup described in Publication I relying on nine foldings produces non-independent point estimates: the i-th's folding training set is entirely a subset of the $(i+1)$-th's folding training set, so that e.g. i-th and $(i+1)$-th accuracy measures are not independent. A different experimental design would allow for robust testing: this is a relevant caveat to be taken into account for future applications. With respect to Table 6.1, WMTR seems to outperform all the competing algorithms wrt. to accuracy and precision (the standard error for ridge regression is very wide, causing non-significant p-values), while no statistical differences are observed wrt. to the bag-of-features method and MDA respectively in recall and F1. Overall, the WMTR method seems the best performing within the set we considered, being not statistically worse than any other for all the four performance measures. Publication I aims at implementing simple methods and investigate their effectiveness in mid-price prediction under different normalizations and at different lags, without aiming at detecting a preferable, most performing method, i.e. tables analogous to Table 6.1 are here out of purpose and thus omitted. Similarly, cross-analyses misclassification rates from different classifiers can be approached by the McNemar's test. By excluding from the discussion implementations made by different authors but recalled in our

Publications (e.g. bag of features classifiers), this testing could have been performed for the methods we directly implemented to address the exact extent they statistically differ in misclassifying labels capturing the direction of mid-price movements. Also, the inclusion of AUC and ROC curves in future researches can help in providing clearer insights into the performance of the algorithms. Diebold-Mariano testing however in this context does not apply, since designed for regression problems rather than for assessing the forecasting performance of classification-aimed methods. Also, autocorrelation in residuals is not investigated in our applications because these do not correspond to regression problems but classification. Considering a practitioner's perspective (but an econometrician's too), point estimates of e.g. the F1 statistics are not satisfactory for adopting a method over another. Although not being a consolidated practice in ML, in possible future publications is important to strive on statistical comparisons between methods, currently limiting our analyses. Furthermore, by adopting the dataset developed with the Publication I this is not difficult to achieve since comparisons within the same dataset are not as complex as across different ones (Demšar, 2006).

	Accuracy	Precision	Recall	F1
RR	1.06E-14	**3.73E-01**	5.68E-04	2.72E-21
SLFN	4.14E-09	2.42E-02	3.60E-06	1.62E-04
LDA	4.26E-08	9.25E-04	1.50E-03	1.23E-11
MDA	1.84E-04	1.32E-02	1.93E-08	**6.73E-02**
MTR	4.64E-02	4.58E-02	1.53E-04	6.43E-04
BoF	6.70E-08	1.53E-08	**8.82E-01**	5.83E-09
N-BoF	6.25E-07	2.40E-05	2.60E-08	1.19E-06

Table 6.1 With respect to Publication II, the Table reports P-values of the two-tailed t-test for differences in forecasting performance, for different algorithms wrt. to WMTR. Differences *not* significant at 5% level are in bold.

Analyses on the stationarity of the mid-price series for the data used in Publication I and Publication II have been recently addressed and discussed in (Ntakaris, Mirone et al., 2019) using the same data source. Mid-price appears to be stationary in some cases, but not always. This underlines a complex dynamics, highly susceptible to exogenous factors driving its behavior. In classical time-series modeling, stationarity represents a strict but essential requirement: learning from the past would provide no benefit if

statistical properties of the series are not constant but changing significantly. Our publication, however, uses wide sets of 144 external variables extracted from the order book to face the mid-price prediction task. In this light, our applications use *external* variables to predict the mid-price movements and study how the dynamics of these external variables are informative of future directions of mid-price changes. All the 144 input time-series could be non-stationary and the mid-price too because the input variables are external.[2] In fact, the ML algorithms we use, look at significant patterns capable of explaining mid-price directions in any case, and the training is performed over increasing windows, always accounting for the most recent information, thus including the most recent dynamics of all the features. Overall, non-stationarity in mid-price, however, is not the main concern: First, (Ntakaris, Mirone et al., 2019) show indeed that series often exhibits stationarity. Second, whether the stationarity hypothesis does not hold, the association between the (stationary or not) external input variables and mid-price directions is uncovered by the algorithm, and good classification performance achieved.

The final two paragraphs relate to the order-flow toxicity literature and latency-related discussion from Section 2.3. Although the validity of order flow toxicity as a predictive measure for a broad number of factors is well documented in the literature, toxicity measures have generally been not considered in ML applications for different prediction tasks. This is clearly an interesting direction for future research. Also, as addressed in the earlier literature review, the non-uniform consensus about the informativeness of deep order book levels for predictions can naturally motivate extensions of Publications addressing and quantifying the exact impact that the first levels have on predictions' results wrt. the use of deep LOB data. This can be achieved by either analyzing model's coefficients (e.g. in ridge regression) or training the algorithms on subsets of features (e.g. in neural networks). It would be interesting to extend features sets to include price impact measures, analyzing their significance, and quantifying their role in prediction with respect to the remaining ones. A relevant and meaningful analysis, aimed at discerning how much of the short-term price dynamics is explainable by wide sets of deep-in-the-book features with respect to straightforward price-impact measures, and which is the time horizon over which

[2]Strictly speaking, the 144 features involve mid-price series too, this, however, introduces a very mild endogeneity factor. (i) The remaining 143 features are qualitatively different and well-differentiated from mid-price labels. (ii) Prediction is for mid-price direction labels, while the mid-price feature consists of the actually normalized mid-price thus a different series too wrt. to the forecasted one.

price impact is not negligible. E.g. if its effect is permanent, and if not, for how long it persists. In this regard, Publication I and Publication II are limited, not addressing this relevant point. With respect to the latency discussion reviewed in Section 2.3, the implementation of *simple* ML algorithms in Publication I, reads as fast and quick trainability. Complex networks architectures are clearly more demanding in terms of resources (clearly limited, although optimized), and time. The simplicity of a single hidden layer network architecture, like the one implemented in Publication I, wrt. to its satisfactory prediction ability for the mid-price movement, is attractive in this context. Furthermore, ridge regression allows for fast and robust implementation of the MP inversion (so that regularization does not cause non-invertibility issues when forcing some parameters very close to zero, and the matrix possibly close to rank deficiency). A non-negligible implementation detail that Publication I adopts, impacting the estimation speed and its overall robustness.

To potentially exploit the short-latency windows typical for LOB markets, implementations of Publication I and Publication II would need to be re-coded to fit high-level standards. The two Publications develop *potentially* fast-performing and efficient methods, however the actual implementation can be revised in this regard (e.g. adopting parallel-computing solutions): in their current well-coded but not optimal shape we cannot asses to which extent arbitrage opportunities emerging from latency can be actually exploited. The limitations of not recording run-times (and possibly reporting an averaged value) for models' training, or not utilizing a highly efficient scripting language, e.g. `C++`, and optimized computing units, are aspects of significant and immediate impact for practitioners, that future research needs to account for. Note that however the implementations have been double-checked and carefully developed, e.g. adopting the MP pseudo-inverse solution for solving Ridge Regression's parameters.

6.3.2 Publication III

The analysis of long-range correlations in the limit order book is addressed in Publication III by considering duration variables and relying on the DFA method. There are several alternative methods and improved-DFA options that might unveil different aspects on the fractal nature of the series we analyzed, e.g. a multi-fractal nature. Furthermore, do our findings apply to different markets as well? And to different

variables? The point about different variables would be relevant also for a better understanding of the heterogeneous market hypothesis (U. A. Müller, Dacorogna, Davé, Pictet et al., 1993), for which duration variables provide evidence, but how about on a market level? Very interesting would be to understand and explore the relationship between long-term memory features in duration series and in the price series, in the context of the general problem of understanding what drives price dynamics. Furthermore, we observe crossovers in the scaling exponent at day, week and moth time-scales. Besides indicating a clear complexity in the time-series, what else can be learned from them? And do they hold for other variables too? In particular, a long-range autocorrelation analysis for sub-daily horizons (e.g. 5 min) would have immediate implications in robust volatility estimation (e.g. AC-RV estimators of P. R. Hansen and Lunde, 2006 are consistent and unbiased for finite-lag autocorrelations in prices). But what is the causal relationship (if any) between fractal patterns in price and volatility? Research needs to be done in this direction since, as our research shows, even the most simple correlation measure between the scaling exponent and economic variables capture statistically significant relationships, but of complex nature and variability. Moreover, the evidence of different fractal aspects in financial markets coming from different studies call for the development of proper econometric models able to cope with them, for instance, the HAR model is an example. The following points, on the other hand, are more specific and based on results currently available in the literature.

Immediate extensions are those toward generalizations of the adopted methodology. Multi-fractal extensions of the DFA method have been proposed in the literature related financial application (Y. Wang et al., 2009; G. Cao, J. Cao et al., 2013; Niu et al., 2016, e.g.). The mono-fractality DFA addresses could represent a limited analysis, e.g. crossovers detected with DFA could be artifacts from a multi-fractal nature, i.e. scaling behaviors which cannot be accounted by a single scaling exponent. Few multifractal analyses on market-events duration have been proposed so far (e.g. Ruan et al., 2011; Gu et al., 2014, are among the few examples). This direction would represent an immediate extension towards a generalized approach expanding the current research in Publication III that we are considering to develop in the future, also by addressing the seasonality issues earlier mentioned.

Researches such as (Hopman, 2002; Bouchaud, Gefen et al., 2004) show that the order flow exhibits long-range autocorrelations in trade volume and sign, but that this does

not lead to any predictability in price changes. Our descriptive analysis detecting long-range correlations in duration can practically boost some predictive model, e.g. for market orders' arrival time, in future work it would be interesting to analyze how these findings relate to duration models such as the Hawkes process. Interestingly, (Hardiman et al., 2013) by using DFA, empirically finds a power-law decay for mid-price changes modeled with a Hawkes process, characterized by two regimes, one corresponding approximately to sub-minute time-scales, the other extending up to approximately 11 days. Mid-price dynamics is driven by best levels' dynamics: how do the long-memory features we observed for durations at best levels relate to mid-price's long-range autocorrelation, and perhaps to cross-overs observed in (Hardiman et al., 2013). This dissertation explores both modern ML methods and classical parametric approaches: how to integrate different methodological approaches merging methods from different domains, to model analyze and quantify different aspects of financial markets is an interesting point, that in future research we could aim to discuss.

In the following, I provide a reflection on the problem of time-series stationarity with respect to the DFA methodology and its limitations. I observe that under an econometric rationale there are several criticalities related to the DFA methods itself, its hypotheses and applicability range.

Recalling the discussion around non-stationarities in DFA from Section 4.2.2, based on (Z. Chen et al., 2002), the DFA implementation on our data, appears to be robust to the non-stationary sources therein addressed. Indeed, none of the typical patterns observed in (Z. Chen et al., 2002) are shown in our implementation. An analysis of the DFA methodology wrt. trends in the signal are provided in (Hu et al., 2001). A number of hypotheses for the underlying trend are therein discussed, e.g. linear, sinusoidal and power-law trends. For the duration time-series considered in Publication III sinusoidal-like trends may be of relevance. The impact of such kind of trends might lead to crossovers of complex nature. However, on an intraday level, no crossovers are observed in our data, suggesting that possible intra-day patterns are correctly addressed by the DFA method, given our sample. Aligned with this, (Ni et al., 2010) points out that the long-range correlation in inter-cancellation durations seems not affected by daily patterns. On the other hand, the mild cross-overs we observe at daily, weekly, and monthly scale could be artifacts deemed to weekly and monthly seasonality patterns. To discern this, an analysis in the inter- and cross-event duration seasonality over weekly and monthly domains, which is omitted in

the earlier closely-related research, would represent a first and relevant direction for future research. As of now, these cross-overs are believed to express intrinsic difference is the scaling exponent, motivated by an economic rationale, but actually, seasonality analyses are missing. Although results from Section 3 on unit-root testing for DFA show non-stationarity at short time-scales for time-scales from approximately 10 seconds onward, the vast majority of time-windows DFA uses for determining the characteristic typical fluctuation, are of a wider size and verified to be indeed stationary. This largely justifies the robustness of the scaling exponents' estimates obtained from our data, while accounting for the importance stationarity has, as pointed out in(Bryce et al., 2012)

Exploratory data-analyses in Publication III are aligned with the common practice and analyses typical to the closely-related application in DFA. In this concern, we have not applied any specific data processing or method-applicability analyses, guided by the widely referred quote "DFA is a well-established method for determining the scaling behavior of noisy data in the presence of trends without knowing their origin and shape" from (Kantelhardt, Koscielny-Bunde et al., 2001), based on (Peng, S. V. Buldyrev, Havlin et al., 1994; Peng, Havlin et al., 1995; S. Buldyrev et al., 1995). This is however not a reason not to address hypotheses and tests about non-stationarity of the time-series under analysis and completely disregarding the problem: we tested stationarity hypotheses motivated by the narrow literature showing that non-stationarity might lead to biased conclusions. This hypothesis-wise shallow background constitutes a criticism on the current research, however, more broadly applicable to the DFA field in general. Indeed only the above-mentioned paper of Z. Chen et al., 2002 and Hu et al., 2001 attempt at addressing these issues, while interestingly, in the original paper presenting the DFA method (Peng, S. V. Buldyrev, Havlin et al., 1994) the discussion is entirely free of hypotheses and statistical constraints for the time-series under analysis. Note that econometric applications relying on a straight use of DFA, quoting its general robustness for non-stationarity, without further empirical analyses on the data are common as well (e.g. Hardiman et al., 2013). Indeed a rigorous framework defining the exact extent and hypotheses framing the DFA applicability is not available in the literature. This is a concerning gap that future research might investigate, heavily limiting any possible attempt in e.g. unit root testing for stationarity. The literature gap does not allow to draw a conclusion of what implications of unit roots are and how to deal with them, so the interpre-

tation of any econometric analysis in this regard is penalized.[3] Indeed no research connecting DFA and econometric theory on stationarity has, so far, been proposed. This is an important direction for future research: a criticism for Publication III is perhaps that of following the mainstream, without providing extensive analyses on the data, guaranteeing the applicability of DFA. An exception is the unit-root testing described in 3, but of difficult interpretation since no references are available. On the other hand, a set of precise hypotheses outlining its applicability, and boundaries in terms of statistical properties of the series are not rigorously drawn. Based however on the analyses of Z. Chen et al., 2002; Hu et al., 2001, evidence of issues imputable to stationarity-related problems is not found in our analyses.

Stationarity issues in DFA from an econometric standpoint. From an econometrics' perspective, references to "stationarity" in the DFA context are vague and imprecise. The problem of time-series stationarity for the DFA method is challenging. There is a wide number of DFA applications from different fields that use the method without attentiveness about stationarity. These include (Ivanov, Yuen, Podobnik et al., 2004; Jiang et al., 2009; Ivanov, Yuen and Perakakis, 2014; Gu et al., 2014), closely-related to the relevant literature for Publication III. The DFA method is first proposed in (Peng, S. V. Buldyrev, Havlin et al., 1994). Surprisingly no hypotheses on the underlying time-series are there presented.

Perhaps, this constitutes a remarkable gap regarding the exact stationarity requirements upon which DFA relies. Very surprisingly there has been no published works dealing with stationarity, under a rigorous econometric perspective, and DFA. The vast majority of researches consists of straightforward applications, where stationarity is perhaps overlooked.

Reminding that a precise hypothesis on stationary requirements is not well-defined for DFA, no analyses on the implications of the results from the econometric Augmented Dickey-Fuller (ADF), (Dickey et al., 1979), Kwiatkowski–Phillips–Schmidt–Shin (KPSS) (Kwiatkowski et al., 1992) and Phillips–Perron (PP) (Phillips et al., 1988) tests on the DFA method are either available. Unit root testing and its connection

[3] Think for instance that a unit root is found in the series. The question would be: "So what?" Indeed there's no reference on how to deal with this and what to do, or not to do. There is a general consensus that DFA is robust to non-stationarity (which is however not exactly addressed, e.g. distinguishing between weak or strong stationarity, and often used interchangeably with "trend", which does not fit any rigorous econometric definition of stationarity), so that DFA is straightforwardly applied.

with DFA has been ignored, as a part of the overall disconnection between the DFA approach and econometric literature. An example: the KPSS test can constitute a feasible approach, however, its rejection would not necessarily mean DFA in-applicability: non-stationarity in higher moments appears too restrictive for DFA.

6.3.3 Publication IV

An immediate extension of Publication IV would be that of implementing D- and R- vine copula constructions under different trees. The C-Vine construction is a single possibility out of a full spectrum that has not been considered. Also, the effect of including copula rotations and different parametric copula families (e.g. BB and Twain copulas) would be interesting to analyze, in a number-of-parameter/model-complexity vs. forecasting-performance perspective too. This would, however, limit the applicability of numerical integration, requiring a simulation-based approach certainly complex. About the numerical results: do the small improvement over the HAR model (which are however typical,e.g. Bollerslev et al., 2016; Andersen, Bollerslev and Diebold, 2007) have an impact for practitioners, e.g. in option pricing, or are of little impact for practical purposes? A further limitation of the current CV-HAR model is that of not dealing with the conditional second moments and thus the variance and confidence intervals of the forecasts. A bootstrap procedure could be utilized, however, the overall model construction and estimation already involves several steps this would include a further layer of complexity. Perhaps, a direct evaluation with numerical integration could be a possibility, but will numerical integration hold for such a complex function? If not, can Monte-Carlo integration be useful? Whereas the data constitutes a very minor problem, for Publication IV the precise model specification over different possibilities and deciding how to implement any improvement seem to be major issues related to this research. Lastly, Publication IV shows that numerical integration over complex functions for retrieving conditional expectations can be indeed achieved. The very same approach could be used in a number of different application: very often the complexity of $\mathbb{E}[X|\mathbf{Y}]$ suggests a linear assumption on the multivariate conditional dependence structure as a workaround. An example could be the Granger causality test, in general involving a conditional-expectation-based predictor, turned in applications into a linear form,

either because of a normality assumption or imposed to achieve a tractable problem.

Turning to the actual implementation in Publication IV, coherently with the most-closely related literature on HAR modeling that does not report residual analyses and back-testing results on the discussed models, these are omitted from Publication IV as well. However, basic back-testing analyses have been implemented for the CV-HAR model in order to understand whether its performance is comparable or not wrt. the HAR specification. In this regard zero-mean of the residuals have been verified as well as Engle's test for residual heteroscedasticity, generally refusing the no-ARCH effects null for both the HAR and CV-HAR model. Heteroskedastic residuals in the HAR framework have already been observed in the seminal paper of (Corsi, 2009), that the CV-HAR neither is capable of removing. Fig. 6.1 further confirms this evidence. Residuals are grouped in 20-days groups, approximately corresponding to a month: quite consistent patterns are observed across the panels, pointing out heteroskedasticity, a considerable number of large residuals, and a considerable homogeneity within the two models.

Furthermore Fig. 6.2 suggests heavy right tails and non-normality in residuals (verified by histograms as well). Absence of autocorrelation in the residuals in graphically assessed in Fig. 6.2, which highlights that a general conclusion is difficult to be drawn. In particular, the top plot reports a case where both the HAR and CV-HAR are very satisfactory, leading to uncorrelated residuals over all the lags considered. The middle plot depicts a case where the CV-HAR model leas to residuals for higher correlation in the first lags wrt. the HAR model, while the opposite holds in the bottom plot. The HAR specification seems not to lead to a consistent netting out of residuals' autocorrelation: in some cases, this is accomplished remarkably well, in others, poorly. The same applies for the CV-HAR model. In general, at higher lags (≥ 10) both the HAR and the CV-HAR residuals are not found to be autocorrelated at 5% level.

Although Publication IV discusses some improvements of the CV-HAR specification in terms of measures of forecasting performance, it has to be pointed out that residuals are far from being optimal, in both the HAR and CV-HAR model, suggesting possible under-fitting issues unable to whiten the correlations and flatten the residuals towards a homoskedastic, symmetric and possibly normal behavior. This heterogeneity applies to all the different combinations of marginal modeling, copula sets

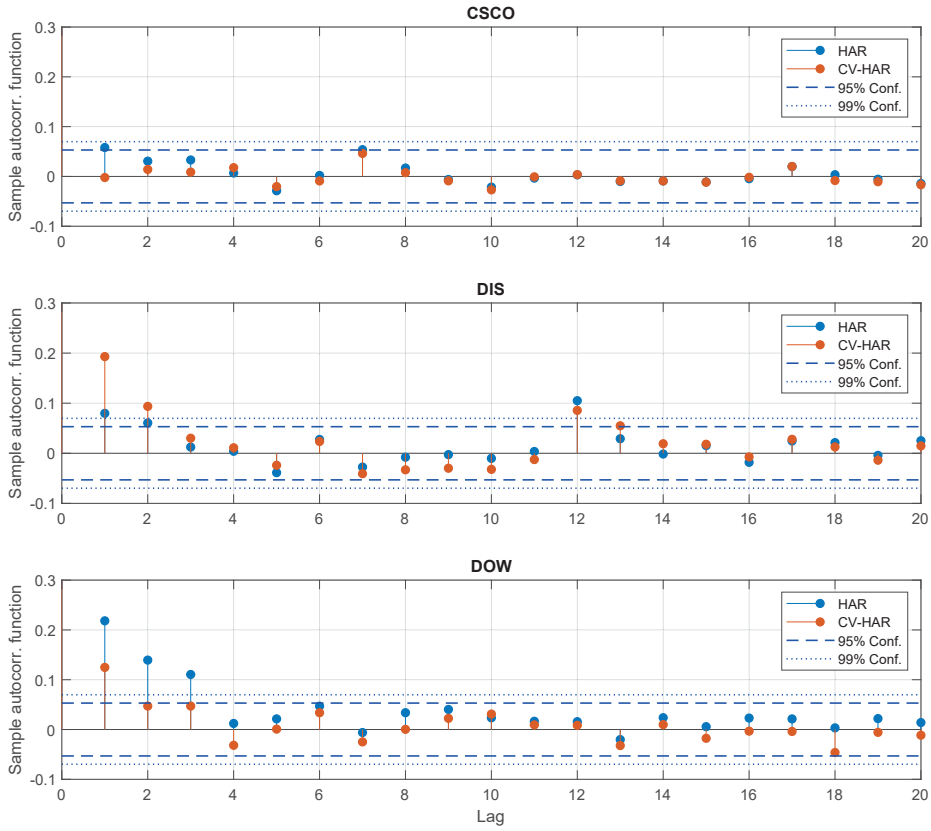


Figure 6.1 Sample autocorrelation function for the residual for the HAR and CV-HAR model. The plot depicts different situations often encountered in the data, concerning the autocorrelation in the first lags, (i) Top panel: autocorrelation in residuals for both models is largely whitened out. (ii) Middle panel: CV-HAR model leaves residuals of higher autocorrelation than those from the HAR model. (iii) Bottom panel: residuals for the HAR model are of lighter autocorrelation than those form the CV-HAR model. Fixed window modeling, estimated over 250 days with underlying Vines implemented for Empirical Cumulative Distribution Function (ECDF) margins and the reduced set of Archimedeans (perhaps the less flexible Vine specification and estimation-forecasting setting).

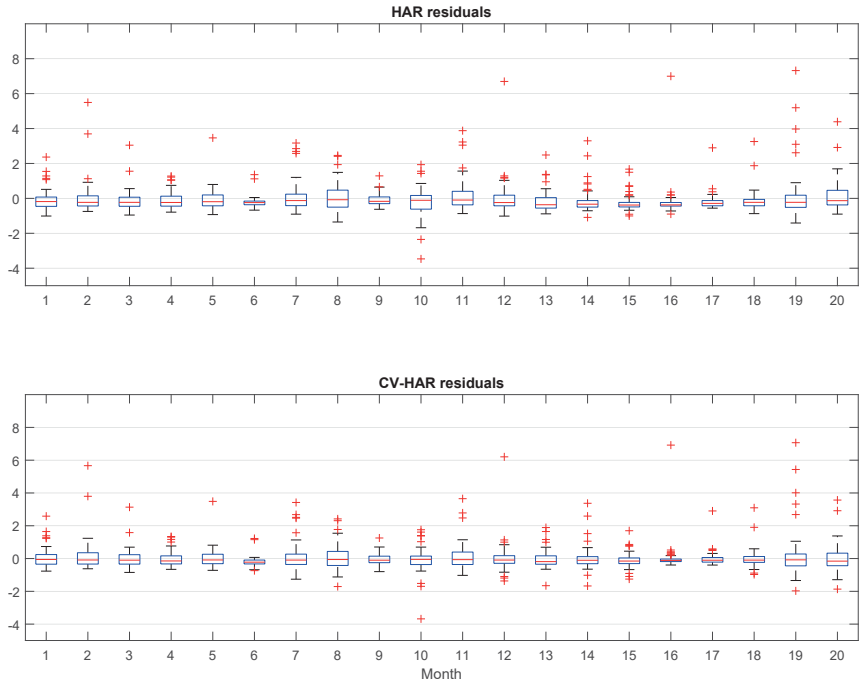


Figure 6.2 Boxplots of the residuals grouped over one-month intervals. The plotted series refers to the CSCO ticker as in the top panel of Fig. 6.1. Fat-taildness and homoskedasticity in the residuals are well-visible for both HAR and CV-HAR models.

and estimation schemes analyzed. Whereas results from Publication IV constitute statistically significant achievements in improving the forecasting of daily realized measure of volatility, the CV-HAR model (and HAR) appear to be simplistic for capturing the complex volatility dynamics. Whether additional regressor can have a positive impact in this regard, for the CV-HAR (and HAR) model it constitutes an interesting point to address in the future. Furthermore, the analyses in Publication IV, but in (e.g. Andersen, Bollerslev and Diebold, 2007; A. J. Patton and Sheppard, 2015; Andersen, Bollerslev and Diebold, 2007) as well, will definitely contribute in addressing model misspecifications and directions for its improvement by including exhaustive back-testing analyses, not limited to synthetic measures (although not simple at all, given different models' specifications, estimation approaches, e.g. rolling windows against increasing windows, and forecasting schemes, e.g. one-step-ahead rather than multi-period forecasts).

As mentioned in Section 2.6, scaling copulas approaches to high dimensions is gen-

erally known to be difficult, leading to heavily parametrized and mathematically non-tractable models. High dimensional Vine copulas are certainly among them. However the model specification as in Publication IV involving four variables is still analytically tractable, e.g. the joint density can practically be integrated, and the number of pair copulas involved is limited to six (given that there are no trivial independent variables involved). The parametric copulas used in the model specification are broadly characterized by a single parameter only, except for the t-copula which is defined over two parameters. Due to its flexibility in modeling asymmetric tail-dependence, the t-copula is broadly adopted in Vine's structure estimated in Publication IV. The CV-HAR model, therefore, involves a number of parameters between 6 and 12, most likely 8. Computations are however efficiently implemented: a full-run on the Vine copula estimation for all the different margins, sets of copulas over ten stock and all the 1636 days, takes around 6 hours on fairly good hardware (two 3.2Ghz cores and 64Gb RAM). Integration takes approximately 3 hours for all the cases[4]. From an econometric standpoint, the CV-HAR model cannot be said to be parsimonious in this regard, however, this is implicit for most of the multivariate copula applications (perhaps not for Factor copulas, but not scalable to the analyses in Publication IV). Tackling the problem addressed in Publication IV e.g. with ML, will likely lead to a higher number of parameters, and possibly of non-straightforward interpretation already by applying a simple network-based approach. However, also in high-dimensional applications, the Vine copula approach has been shown to reliable and of feasible calibration (e.g. E. C. Brechmann and Czado, 2013, a 50-dimensional appliaction). Recently (D. Müller et al., 2019) showed that in ultra-high dimensions Vine copula selection and estimation can be properly divided into sub-problems of lower complexity with important gains in estimation time and accuracy. On the other hand, alternative approaches for the non-linear regression discussed in Publication IV can be adopted. Among them, the application of e.g. a Gaussian process regression, which has desirable properties (e.g. clearly addressing uncertainty in predictors) and simple parametrization, constitutes an interesting direction for future researches connected to Publication IV. A broad number of alternatives to Vine copula for dependence modeling and non-linear regression can be applied, such as probabilistic programming methods and ML methods. This would explore the extent classical and modern method can converge or jointly used in tackling problems in

[4]Estimation of Vines is implemented in R, computation of realized measures, data cleaning and integration of conditional Vine densities in Matlab.

dependence modeling and non-linear regression. However, the use of copulas has historically been very ubiquitous and well-connected to the econometric literature: Publication IV aims at exploring their use in modeling and forecasting high-frequency measures and at covering this selected gap, acknowledging that in this context there are several alternatives, not yet widely investigated, where flexible modern methods and well-established econometric theories, generally based on parametric models specifications, can converge.

REFERENCES

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics* 44.2, 182–198.

Abergel, F., Anane, M., Chakraborti, A., Jedidi, A. and Toke, I. M. (2016). *Limit order books*. Cambridge University Press.

Abergel, F. and Jedidi, A. (2013). A mathematical approach to order book modeling. *International Journal of Theoretical and Applied Finance* 16.05, 1350025.

Aït-Sahalia, Y. and Jacod, J. (2014). *High-frequency financial econometrics*. Princeton University Press.

Ait-Sahalia, Y., Mykland, P. A. and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The review of financial studies* 18.2, 351–416.

Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics* 160.1, 160–175.

Aït-Sahalia, Y. and Yu, J. (2008). *High frequency market microstructure noise estimates and liquidity measures*. Tech. rep. National Bureau of Economic Research.

Alfarano, S. and Lux, T. (2007). A noise trader model as a generator of apparent financial power laws and long memory. *Macroeconomic Dynamics* 11.S1, 80–101.

Almgren, R. F. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied mathematical finance* 10.1, 1–18.

Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The review of economics and statistics* 89.4, 701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of financial economics* 61.1, 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2000a). *Exchange rate returns standardized by realized volatility are (nearly) Gaussian*. Tech. rep. National Bureau of Economic Research.

— (2000b). Great realizations. *Risk* 13, 105–108.

— (2003). Modeling and forecasting realized volatility. *Econometrica* 71.2, 579–625.

— (1999). Realized volatility and correlation. *LN Stern School of Finance Department Working Paper* 24.

Andersen, T. G. and Bondarenko, O. (2014). Assessing measures of order flow toxicity and early warning signals for market turbulence. *Review of Finance* 19.1, 1–54.

Andreou, E., Ghysels, E. and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics* 158.2, 246–261.

Bai, J., Ghysels, E. and Wright, J. H. (2013). State space models and MIDAS regressions. *Econometric Reviews* 32.7, 779–813.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of econometrics* 73.1, 5–59.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 74.1, 3–30.

Bandi, F. M. and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies* 75.2, 339–369.

— (2005). Realized covariation, realized beta and microstructure noise. *Unpublished paper, Graduate School of Business, University of Chicago*.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2009). Realized kernels in practice: trades and quotes. *The Econometrics Journal* 12.3, C1–C32. ISSN: 1368-4221. DOI: 10.1111/j.1368-423X.2008.00275.x. eprint: http://oup.prod.sis.lan/ectj/article-pdf/12/3/C1/27664306/ectj00c1.pdf. URL: https://doi.org/10.1111/j.1368-423X.2008.00275.x.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76.6, 1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.2, 253–280.

— (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of financial Econometrics* 4.1, 1–30.

Barucci, E. and Reno, R. (2002). On measuring volatility and the GARCH forecasting performance. *Journal of International Financial Markets, Institutions and Money* 12.3, 183–200.

Bashan, A., Bartsch, R., Kantelhardt, J. W. and Havlin, S. (2008). Comparison of detrending methods for fluctuation analysis. *Physica A: Statistical Mechanics and its Applications* 387.21, 5080–5090.

Bauwens, L. and Hafner, C. M. (2012). *Handbook of volatility models and their applications*. Vol. 3. John Wiley & Sons.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence* 32.1-4, 245–268.

Bedford, T., Cooke, R. M. et al. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics* 30.4, 1031–1068.

Beltran, H., Grammig, J. and Menkveld, A. J. (2005). *Understanding the limit order book: Conditioning on trade informativeness*. Tech. rep. CFR Working Paper.

Biais, B., Hillion, P. and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse. *the Journal of Finance* 50.5, 1655–1689.

Blasques, F., Koopman, S. J. and Lucas, A. (2014). Optimal formulations for nonlinear autoregressive processes.

Bloomfield, R., O'hara, M. and Saar, G. (2005). The "make or take" decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics* 75.1, 165–199.

Bollen, B. and Inder, B. (2002). Estimating daily volatility in financial markets utilizing intraday data. *Journal of Empirical Finance* 9.5, 551–562.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31.3, 307–327.

Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192.1, 1–18.

Bose, I. and Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & management* 39.3, 211–225.

Bouchaud, J.-P. (2010). Price impact. *Encyclopedia of quantitative finance*.

Bouchaud, J.-P., Farmer, J. D. and Lillo, F. (2009). How markets slowly digest changes in supply and demand. *Handbook of financial markets: dynamics and evolution*. Elsevier, 57–160.

Bouchaud, J.-P., Gefen, Y., Potters, M. and Wyart, M. (2004). Fluctuations and response in financial markets: the subtle nature of 'random'price changes. *Quantitative finance* 4.2, 176–190.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G. and Roncalli, T. (2000). Copulas for finance-a reading guide and some applications. *Available at SSRN 1032533*.

Brechmann, E. and Schepsmeier, U. (2013). Cdvine: Modeling dependence with c-and d-vine copulas in r. *Journal of Statistical Software* 52.3, 1–27.

Brechmann, E. C. and Czado, C. (2013). Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50. *Statistics & Risk Modeling* 30.4, 307–342.

Brechmann, E. C., Heiden, M. and Okhrin, Y. (2018). A multivariate volatility vine copula model. *Econometric Reviews* 37.4, 281–308.

Brogaard, J., Hendershott, T. and Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies* 27.8, 2267–2306.

Bryce, R. and Sprague, K. (2012). Revisiting detrended fluctuation analysis. *Scientific reports* 2, 315.

Buccheri, G. and Corsi, F. (2017). Hark the shark: Realized volatility modelling with measurement errors and nonlinear dependencies. *Available at SSRN 3089929*.

Bucci, A. et al. (2017). Forecasting realized volatility: a review. *Journal of Advanced Studies in Finance (JASF)* 8.16, 94–138.

Budish, E., Cramton, P. and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics* 130.4, 1547–1621.

Buldyrev, S., Goldberger, A., Havlin, S., Mantegna, R., Matsa, M., Peng, C.-K., Simons, M. and Stanley, H. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E* 51.5, 5084.

Cao, C., Hansch, O. and Wang, X. (2009). The information content of an open limit-order book. *Journal of Futures Markets: Futures, Options, and Other Derivative Products* 29.1, 16–41.

Cao, G., Cao, J. and Xu, L. (2013). Asymmetric multifractal scaling behavior in the Chinese stock market: Based on asymmetric MF-DFA. *Physica A: Statistical Mechanics and its Applications* 392.4, 797–807.

Cao, G., Xu, L. and Cao, J. (2012). Multifractal detrended cross-correlations between the Chinese exchange market and stock market. *Physica A: Statistical Mechanics and its Applications* 391.20, 4855–4866.

Carbone, A., Castelli, G. and Stanley, H. E. (2004). Time-dependent Hurst exponent in financial time series. *Physica A: Statistical Mechanics and its Applications* 344.1-2, 267–271.

Cenesizoglu, T., Dionne, G. and Zhou, X. (2014). *Effects of the Limit Order Book on Price Dynamics*. Tech. rep. CIRPEE.

Chen, S.-H. and Hsieh, Y.-L. (2011). Reinforcement learning in experimental asset markets. *Eastern Economic Journal* 37.1, 109–133.

Chen, Z., Ivanov, P. C., Hu, K. and Stanley, H. E. (2002). Effect of nonstationarities on detrended fluctuation analysis. *Physical review E* 65.4, 041107.

Cherubini, U. and Luciano, E. (2001). Value-at-risk Trade-off and Capital Allocation with Copulas. *Economic notes* 30.2, 235–256.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Chiriac, R. and Voev, V. (2011). Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics* 26.6, 922–947.

Chollete, L., Heinen, A. and Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime-switching copula. *Journal of financial econometrics* 7.4, 437–480.

Cizeau, P., Liu, Y., Meyer, M., Peng, C.-K. and Stanley, H. E. (1997). Volatility distribution in the S&P500 stock index. *Physica A: Statistical Mechanics and its Applications* 245.3-4, 441–445.

Cont, R. (2005). Long range dependence in financial markets. *Fractals in engineering*. Springer, 159–179.

— (2011). Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine* 28.5, 16–25.

Cont, R., Kukanov, A. and Stoikov, S. (2014). The price impact of order book events. *Journal of financial econometrics* 12.1, 47–88.

Cont, R., Stoikov, S. and Talreja, R. (2010). A stochastic model for order book dynamics. *Operations research* 58.3, 549–563.

Cooke, R. M. (1997). Markov and entropy properties of tree-and vine-dependent variables. *Proceedings of the ASA Section of Bayesian Statistical Science*. Vol. 27.

Cooke, R. M., Kurowicka, D. and Wilson, K. (2015). Sampling, conditionalizing, counting, merging, searching regular vines. *Journal of Multivariate Analysis* 138, 4–18.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7.2, 174–196.

Czado, C., Schepsmeier, U. and Min, A. (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling* 12.3, 229–255.

Czarnecki, Ł., Grech, D. and Pamuła, G. (2008). Comparison study of global and local approaches describing critical phenomena on the Polish stock exchange market. *Physica A: Statistical Mechanics and its Applications* 387.27, 6801–6811.

Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M. and Ninot, G. (2006). Fractal analyses for 'short'time series: a re-assessment of classical methods. *Journal of mathematical psychology* 50.6, 525–544.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7.Jan, 1–30.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association* 74.366a, 427–431.

Dissmann, J., Brechmann, E. C., Czado, C. and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis* 59, 52–69.

Dixon, M. (2018a). A high-frequency trade execution model for supervised learning. *High Frequency* 1.1, 32–52.

— (2018b). Sequence classification of the limit order book using recurrent neural networks. *Journal of computational science* 24, 277–286.

Dixon, M., Klabjan, D. and Bang, J. H. (2017). Classification-based financial markets prediction using deep neural networks. *Algorithmic Finance* 6.3-4, 67–77.

Dufour, A. and Engle, R. F. (2000). Time and the price impact of a trade. *The Journal of Finance* 55.6, 2467–2498.

Easley, D., De Prado, M. L. and O'Hara, M. (2011). The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. *Journal of Portfolio Management* 37.2, 118–128.

Easley, D., López de Prado, M. M. and O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies* 25.5, 1457–1493.

Embrechts, P., Höing, A. and Juri, A. (2003). Using copulae to bound the value-at-risk for functions of dependent risks. *Finance and Stochastics* 7.2, 145–167.

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond* 1, 176–223.

Ende, B., Uhle, T. and Weber, M. C. (2011). The Impact of a Millisecond: Measuring Latency Effects in Securities Trading. *Wirtschaftsinformatik*, 116.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Engle, R. F. and Ferstenberg, R. (2007). Execution risk. *The Journal of Portfolio Management* 33.2, 34–44.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* 131.1-2, 3–27.

Engle, R. F., Ghysels, E. and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics* 95.3, 776–797.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.

Erb, C. B., Harvey, C. R. and Viskanta, T. E. (1994). Forecasting international equity correlations. *Financial analysts journal* 50.6, 32–45.

Feder, J. (1988). Fractals. Plenum Press, New York. *Fractals. Plenum Press, New York.*

Flood, M. D. (2012). A brief history of financial risk and information. *Handbook of Financial Data and Risk Information* 1.

Flood, M. D., Jagadish, H., Raschid, L. et al. (2016). Big data challenges and opportunities in financial stability monitoring. *Banque de France, Financial Stability Review* 20, 129–42.

Gallo, G. M. and Otranto, E. (2015). Forecasting realized volatility with changing average levels. *International Journal of Forecasting* 31.3, 620–634.

Ghysels, E., Sinko, A. and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews* 26.1, 53–90.

Glosten, L. R. (1994). Is the electronic open limit order book inevitable?: *The Journal of Finance* 49.4, 1127–1161.

Gong, X. and Lin, B. (2018). Structural breaks and volatility forecasting in the copper futures market. *Journal of Futures Markets* 38.3, 290–339.

Gould, M. D. and Bonart, J. (2016). Queue imbalance as a one-tick-ahead price predictor in a limit order book. *Market Microstructure and Liquidity* 2.02, 1650006.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J. and Howison, S. D. (2013). Limit order books. *Quantitative Finance* 13.11, 1709–1742.

Grau-Carles, P. (2000). Empirical evidence of long-range correlations in stock returns. *Physica A: Statistical Mechanics and its Applications* 287.3-4, 396–404.

— (2001). Long-range power-law correlations in stock returns. *Physica A: Statistical Mechanics and its Applications* 299.3-4, 521–527.

Grech, D. and Mazur, Z. (2004). Can one make any crash prediction in finance using the local Hurst exponent idea?: *Physica A: Statistical Mechanics and its Applications* 336.1-2, 133–145.

Gu, G.-F., Xiong, X., Zhang, W., Zhang, Y.-J. and Zhou, W.-X. (2014). Empirical properties of inter-cancellation durations in the Chinese stock market. *Frontiers in Physics* 2, 16.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 705–730.

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24.2, 127–161.

Hansen, P. R., Huang, Z. and Shek, H. H. (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics* 27.6, 877–906.

Hansen, P. R. and Lunde, A. (2004). An unbiased measure of realized variance. *Available at SSRN 524602*.

Hansen, P., Li, Y., Lunde, A. and Patton, A. (2017). Mind the Gap: An Early Empirical Analysis of SEC's "Tick Size Pilot Program". *Unpublished manuscript*.

Hardiman, S. J., Bercot, N. and Bouchaud, J.-P. (2013). Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B* 86.10, 442.

Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. OUP USA.

Harris, L. E. and Panchapagesan, V. (2005). The information content of the limit order book: evidence from NYSE specialist trading decisions. *Journal of Financial Markets* 8.1, 25–67.

Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

— (1991). Measuring the information content of stock trades. *The Journal of Finance* 46.1, 179–207.

— (1995). One security, many markets: Determining the contributions to price discovery. *The journal of Finance* 50.4, 1175–1199.

Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets* 16.4, 646–679.

He, X., Cai, D. and Niyogi, P. (2006). Tensor subspace analysis. *Advances in neural information processing systems*, 499–506.

He, X. and Lin, S. (2019). Reinforcement Learning in Limit Order Markets. *Available at SSRN 3131347*.

Heale, R. and Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-based nursing* 18.3, 66–67.

Heneghan, C. and McDarby, G. (2000). Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes. *Physical review E* 62.5, 6103.

Hewlett, P. (2006). Clustering of order arrivals, price impact and trade path optimisation. *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 6–8.

Hillebrand, E. and Medeiros, M. C. (2007). *Forecasting realized volatility models: the benefits of bagging and nonlinear specifications*. Tech. rep. Texto para discussão.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12.1, 55–67.

Hopman, C. (2002). Are supply and demand driving stock prices. *Quantitative Finance*.

Hu, K., Ivanov, P. C., Chen, Z., Carpena, P. and Stanley, H. E. (2001). Effect of trends on detrended fluctuation analysis. *Physical Review E* 64.1, 011114.

Huang, R. D. and Stoll, H. R. (1997). The components of the bid-ask spread: A general approach. *The Review of Financial Studies* 10.4, 995–1034.

Hunt, G. (1951). Random fourier transforms. *Transactions of the American Mathematical Society* 71.1, 38–69.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.* 116, 770–799.

Ivanov, P. C., Yuen, A. and Perakakis, P. (2014). Impact of stock market structure on intertrade time and price dynamics. *PloS one* 9.4, e92885.

Ivanov, P. C., Yuen, A., Podobnik, B. and Lee, Y. (2004). Common scaling patterns in intertrade times of US stocks. *Physical Review E* 69.5, 056107.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M. and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications* 119.7, 2249–2276.

Jacod, J., Li, Y. and Zheng, X. (2017). Statistical properties of microstructure noise. *Econometrica* 85.4, 1133–1174.

Jacod, J., Podolskij, M., Vetter, M. et al. (2010). Limit theorems for moving averages of discretized processes plus noise. *The Annals of Statistics* 38.3, 1478–1545.

Jiang, Z.-Q., Chen, W. and Zhou, W.-X. (2009). Detrended fluctuation analysis of intertrade durations. *Physica A: Statistical Mechanics and its Applications* 388.4, 433–440.

— (2008). Scaling in the distribution of intertrade durations of Chinese stocks. *Physica A: Statistical Mechanics and its Applications* 387.23, 5818–5825.

Joe, H. (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

— (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, 120–141.

— (1994). Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics* 22.1, 47–64.

Joe, H. and Kurowicka, D. (2011). *Dependence modeling: vine copula handbook*. World Scientific.

Joe, H., Li, H. and Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis* 101.1, 252–270.

Jondeau, E. and Rockinger, M. (2006). The copula-garch model of conditional dependencies: An international stock market application. *Journal of international money and finance* 25.5, 827–853.

Jouanin, J.-F., Rapuch, G., Riboulet, G. and Roncalli, T. (2001). Modelling dependence for credit derivatives with copulas. *Available at SSRN 1032561*.

Kalnina, I. (2011). Subsampling high frequency data. *Journal of Econometrics* 161.2, 262–283.

Kalnina, I. and Linton, O. (2008). Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error. *Journal of econometrics* 147.1, 47–59.

Kaniel, R. and Liu, H. (2006). So what orders do informed traders use?: *The Journal of Business* 79.4, 1867–1913.

Kantelhardt, J. W. (2009). Fractal and multifractal time series. *Encyclopedia of Complexity and Systems Science*, 3754–3779.

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H., Havlin, S. and Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications* 295.3-4, 441–454.

Kantelhardt, J. W., Roman, H. E. and Greiner, M. (1995). Discrete wavelet approach to multifractality. *Physica A: Statistical Mechanics and its Applications* 220.3-4, 219–238.

Kercheval, A. N. and Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance* 15.8, 1315–1329.

Kole, E., Koedijk, K. and Verbeek, M. (2007). Selecting copulas for risk management. *Journal of Banking & Finance* 31.8, 2405–2423.

Koopman, S. J., Jungbacker, B. and Hol, E. (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance* 12.3, 445–475.

Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H. E., Goldreich, Y. and Schellnhuber, H.-J. (1998). Indication of a universal persistence law governing atmospheric variability. *Physical Review Letters* 81.3, 729.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?: *Journal of econometrics* 54.1-3, 159–178.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 1315–1335.

Laurent, J.-P. and Gregory, J. (2005). Basket default swaps, CDOs and factor copulas. *Journal of risk* 7.4, 103–122.

Lee, S. S. and Mykland, P. A. (2007). Jumps in financial markets: A new nonparametric test and jump dynamics. *The Review of Financial Studies* 21.6, 2535–2563.

Leite, A., Rocha, A. and Silva, M. (2009). Long memory and volatility in HRV: an ARFIMA-GARCH approach. *2009 36th Annual Computers in Cardiology Conference (CinC)*. IEEE, 165–168.

Leite, A., Rocha, A., Silva, M., Gouveia, S., Carvalho, J. and Costa, O. (2007). Long-range dependence in heart rate variability data: ARFIMA modelling vs detrended fluctuation analysis. *2007 Computers in Cardiology*. IEEE, 21–24.

Li, D. X. (1999). On default correlation: A copula function approach. *Available at SSRN 187289*.

Li, Q., Chen, Y., Jiang, L. L., Li, P. and Chen, H. (2016). A tensor-based information framework for predicting the stock market. *ACM Transactions on Information Systems (TOIS)* 34.2, 11.

Li, Q. and Schonfeld, D. (2014). Multilinear discriminant analysis for higher-order tensor data classification. *IEEE transactions on pattern analysis and machine intelligence* 36.12, 2524–2537.

Lillo, F. and Farmer, J. D. (2004). The long memory of the efficient market. *Studies in nonlinear dynamics & econometrics* 8.3.

Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, 6–7.

Liu, J. and Park, S. (2015). Behind stock price movement: Supply and demand in market microstructure and market influence. *The Journal of Trading* 10.3, 13–23.

Liu, Y., Gopikrishnan, P., Stanley, H. E. et al. (1999). Statistical properties of the volatility of price fluctuations. *Physical review e* 60.2, 1390.

Lux, T. and Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397.6719, 498.

Lyócsa, Š. and Molnár, P. (2016). Volatility forecasting of strategically linked commodity ETFs: gold-silver. *Quantitative Finance* 16.12, 1809–1822.

Magris, M. (2019). A C-Vine extension for the HAR model. *Unpublished manuscript*. Submitted to *Journal of Business & Economic Statistics*, May 2019.

Magris, M., Kim, J., Räsänen, E. and Kanniainen, J. (2017). Long-range auto-correlations in limit order book markets: Inter-and cross-event analysis. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/SSCI.2017.8280932.

Malevergne, Y., Sornette, D. et al. (2003). Testing the Gaussian copula hypothesis for financial assets dependences. *Quantitative Finance* 3.4, 231–250.

Malliavin, P., Mancino, M. E. et al. (2009). A Fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics* 37.4, 1983–2010.

Mandelbrot, B. B. (1997). The variation of certain speculative prices. *Fractals and scaling in finance*. Springer, 371–418.

— (1971). When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale models. *The Review of Economics and Statistics*, 225–236.

Mandelbrot, B. B. and Wallis, J. R. (1969). Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. *Water Resources Research* 5.5, 967–988.

Mantegna, R. N. and Stanley, H. E. (1995). Scaling behaviour in the dynamics of an economic index. *Nature* 376.6535, 46.

Martens, M., De Pooter, M. and Van Dijk, D. J. (2004). Modeling and forecasting S&P 500 volatility: Long memory, structural breaks and nonlinearity. *Tinbergen Institute*.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. and Barton, D. (2012). Big data: the management revolution. *Harvard business review* 90.10, 60–68.

McAleer, M. and Medeiros, M. C. (2008a). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics* 147.1, 104–119.

— (2011). Forecasting realized volatility with linear and nonlinear univariate models. *Journal of Economic Surveys* 25.1, 6–18.

— (2008b). Realized volatility: A review. *Econometric Reviews* 27.1-3, 10–45.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics* 8.4, 323–361.

Michie, D., Spiegelhalter, D. J., Taylor, C. et al. (1994). Machine learning. *Neural and Statistical Classification* 13.

Mielniczuk, J. and Wojdyłło, P. (2007). Estimation of Hurst exponent revisited. *Computational Statistics & Data Analysis* 51.9, 4510–4525.

Minsky, M. and Papert, S. A. (2017). *Perceptrons: An introduction to computational geometry*. MIT press.

Müller, D. and Czado, C. (2019). Dependence modeling in ultra high dimensions with vine copulas and the graphical lasso. *Computational Statistics & Data Analysis*.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V. and Von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance* 4.2-3, 213–239.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Pictet, O. V., Olsen, R. B. and Ward, J. R. (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*.

Muni Toke, I. and Yoshida, N. (2017). Modelling intensities of order flows in a limit order book. *Quantitative Finance* 17.5, 683–701.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nevmyvaka, Y., Feng, Y. and Kearns, M. (2006). Reinforcement learning for optimized trade execution. *Proceedings of the 23rd international conference on Machine learning*. ACM, 673–680.

Ni, X.-H., Jiang, Z.-Q., Gu, G.-F., Ren, F., Chen, W. and Zhou, W.-X. (2010). Scaling and memory in the non-Poisson process of limit order cancelation. *Physica A: Statistical Mechanics and its Applications* 389.14, 2751–2761.

Nikoloulopoulos, A. K., Joe, H. and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics & Data Analysis* 56.11, 3659–3673.

Niu, H., Wang, J. and Lu, Y. (2016). Fluctuation behaviors of financial return volatility duration. *Physica A: Statistical Mechanics and its Applications* 448, 30–40.

Nousi, P., Tsantekidis, A., Passalis, N., Ntakaris, A., Kanniainen, J., Tefas, A., Gabbouj, M. and Iosifidis, A. (2018). Machine learning for forecasting mid price movement using limit order book data. *arXiv preprint arXiv:1809.07861*.

Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting* 37.8, 852–866. DOI: 10.1002/for.2543.

Ntakaris, A., Mirone, G., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2019). Feature Engineering for Mid-Price Prediction Forecasting with Deep Learning. *arXiv preprint arXiv:1904.05384*.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics* 50.2, 379–402.

Oh, D. H. and Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics* 35.1, 139–154.

Oh, G., Um, C.-J. and Kim, S. (2006). Statistical properties of the returns of stock prices of international markets. *arXiv preprint physics/0601126*.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics* 116.2, 257–270.

O'hara, M. (1995). *Market microstructure theory*. Vol. 108. Blackwell Publishers Cambridge, MA.

Oomen, R. (2001). Using high frequency stock market index data to calculate, model and forecast realized volatility. *Department of Economics, European University Institute, Manuscript*.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business & Economic Statistics* 24.2, 219–237.

Oomen, R. C. (2005). Properties of bias-corrected realized variance under alternative sampling schemes. *Journal of Financial Econometrics* 3.4, 555–577.

Pai, P.-F. and Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 33.6, 497–505.

Palguna, D. and Pollak, I. (2016). Mid-price prediction in a limit order book. *IEEE Journal of Selected Topics in Signal Processing* 10.6, 1083–1092.

Panayi, E., Peters, G. W., Danielsson, J. and Zigrand, J.-P. (2018). Designating market maker behaviour in limit order book markets. *Econometrics and Statistics* 5, 20–44.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural computation* 3.2, 246–257.

Parlour, C. A. (1998). Price dynamics in limit order markets. *The Review of Financial Studies* 11.4, 789–816.

Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2019). Deep Adaptive Input Normalization for Price Forecasting using Limit Order Book Data. *arXiv preprint arXiv:1902.07892*.

— (2018). Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Passalis, N., Tsantekidis, A., Tefas, A., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2017). Time-series classification using neural bag-of-features. *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 301–305.

Patton, A. (2013). Copula methods for forecasting multivariate time series. *Handbook of economic forecasting*. Vol. 2. Elsevier, 899–960.

Patton, A. J. (2001a). Estimation of copula models for time series of possibly different lengths. *Economics Working Paper Series* 1.2001-17.

— (2006). Modelling asymmetric exchange rate dependence. *International economic review* 47.2, 527–556.

— (2001b). Modelling time-varying exchange rate dependence using the conditional copula. *Economics Working Paper Series*.

— (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics* 2.1, 130–168.

Patton, A. J. and Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics* 97.3, 683–697.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H. (1992). Long-range correlations in nucleotide sequences. *Nature* 356.6365, 168.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical review e* 49.2, 1685.

Peng, C.-K., Havlin, S., Stanley, H. E. and Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 5.1, 82–87.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika* 75.2, 335–346.

Podobnik, B., Horvatic, D., Ng, A. L., Stanley, H. E. and Ivanov, P. C. (2008). Modeling long-range cross-correlations in two-component ARFIMA and FIARCH processes. *Physica A: Statistical Mechanics and its Applications* 387.15, 3954–3959.

Podolskij, M., Vetter, M. et al. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli* 15.3, 634–658.

Pooter, M. d., Martens, M. and Dijk, D. v. (2008). Predicting the daily covariance matrix for s&p 100 stocks using intraday data—but which frequency to use?: *Econometric Reviews* 27.1-3, 199–229.

Qian, B. and Rasheed, K. (2004). Hurst exponent and financial market predictability. *IASTED conference on Financial Engineering and Applications*, 203–209.

Rangarajan, G. and Ding, M. (2000). Integrated approach to the assessment of long range correlation in time series data. *Physical Review E* 61.5, 4991.

Reboredo, J. C. and Ugolini, A. (2015). A vine-copula conditional value-at-risk approach to systemic sovereign debt risk for the financial sector. *The North American Journal of Economics and Finance* 32, 98–123.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica* 10.3-4, 441–451.

Roberts, P., Priest, H. and Traynor, M. (2006). Reliability and validity in research. *Nursing standard* 20.44.

Rodriguez, J. C. (2007). Measuring financial contagion: A copula approach. *Journal of empirical finance* 14.3, 401–423.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of finance* 39.4, 1127–1139.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65.6, 386.

Ruan, Y.-P. and Zhou, W.-X. (2011). Long-term correlations and multifractal nature in the intertrade durations of a liquid Chinese stock and its warrant. *Physica A: Statistical Mechanics and its Applications* 390.9, 1646–1654.

Schweizer, B., Wolff, E. F. et al. (1981). On nonparametric measures of dependence for random variables. *The annals of statistics* 9.4, 879–885.

Seppi, D. J. (1997). Liquidity provision with limit orders and a strategic specialist. *The Review of Financial Studies* 10.1, 103–150.

Sewell, M. V. and Yan, W. (2008). Ultra high frequency financial data. *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation*. ACM, 1847–1850.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shao, Y.-H., Gu, G.-F., Jiang, Z.-Q., Zhou, W.-X. and Sornette, D. (2012). Comparing the performance of FA, DFA and DMA using different synthetic long-range correlated time series. *Scientific reports* 2, 835.

Shephard, N. and Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25.2, 197–231.

Siikanen, M. (2018). Investors, Information Arrivals, and Market Liquidity: Empirical Evidence from Financial Markets. *Tampere University of Technology. Publication* 1581.

Siikanen, M., Kanniainen, J. and Luoma, A. (2017). What drives the sensitivity of limit order books to company announcement arrivals?: *Economics Letters* 159, 65–68.

Siikanen, M., Kanniainen, J. and Valli, J. (2017). Limit order books and liquidity around scheduled and non-scheduled announcements: Empirical evidence from nasdaq nordic. *Finance Research Letters* 21, 264–271.

Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance* 19.4, 549–570.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika* 9.6, 449–460.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris* 8, 229–231.

So, M. K. and Yeung, C. Y. (2014). Vine-copula GARCH model with dynamic conditional dependence. *Computational Statistics & Data Analysis* 76, 655–671.

Sokolinskiy, O. and Dijk, D. van (2011). *Forecasting volatility with copula-based time series models*. Tech. rep. Tinbergen Institute Discussion Paper.

Stošić, D., Stošić, D., Stošić, T. and Stanley, H. E. (2015). Multifractal analysis of managed and independent float exchange rates. *Physica A: Statistical Mechanics and its Applications* 428, 13–18.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Suzuki, K. (2011). *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand.

Tan, J., Wang, J., Rinprasertmeechai, D., Xing, R. and Li, Q. (2019). A Tensor-based eLSTM Model to Predict Stock Price using Financial News. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Taqqu, M. S., Teverovsky, V. and Willinger, W. (1995). Estimators for long-range dependence: an empirical study. *Fractals* 3.04, 785–798.

Tiwari, A. K., Albulescu, C. T. and Yoon, S.-M. (2017). A multifractal detrended fluctuation analysis of financial market efficiency: Comparison using Dow Jones sector ETF indices. *Physica A: Statistical Mechanics and its Applications* 483, 182–192.

Torre, K., Delignieres, D. and Lemoine, L. (2007). Detection of long-range dependence and estimation of fractal exponents through ARFIMA modelling. *British Journal of Mathematical and Statistical Psychology* 60.1, 85–106.

Tóth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J. and Bouchaud, J.-P. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X* 1.2, 021006.

Tran, D. T., Gabbouj, M. and Iosifidis, A. (2017). Multilinear class-specific discriminant analysis. *Pattern Recognition Letters* 100, 131–136.

Tran, D. T., Iosifidis, A., Kanniainen, J. and Gabbouj, M. (2018). Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE transactions on neural networks and learning systems*.

Tran, D. T., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2017). Tensor representation in high-frequency financial data for price change prediction. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/SSCI.2017.8280812.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2017a). Forecasting stock prices from the limit order book using convolutional neural networks. *2017 IEEE 19th Conference on Business Informatics (CBI)*. Vol. 1. IEEE, 7–12.

— (2017b). Using deep learning to detect price change indications in financial markets. *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2511–2515.

Van Ness, B. F., Van Ness, R. A. and Yildiz, S. (2017). The role of HFTs in order flow toxicity and stock price variance, and predicting changes in HFTs' liquidity provisions. *Journal of Economics and Finance* 41.4, 739–762.

Vandewalle, N., Ausloos, M. and Boveroux, P. (1997). Detrended fluctuation analysis of the foreign exchange market. *Econophysic Workshop, Budapest, Hungary*.

Vandewalle, N. and Ausloos, M. (1997). Coherent and random sequences in financial fluctuations. *Physica A: Statistical Mechanics and its Applications* 246.3-4, 454–459.

Vasilescu, M. A. O. and Terzopoulos, D. (2003). Multilinear subspace analysis of image ensembles. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. IEEE, II–93.

Vaz de Melo Mendes, B. and Accioly, V. B. (2014). Robust pair-copula based forecasts of realized volatility. *Applied Stochastic Models in Business and Industry* 30.2, 183–199.

Wang, G.-J. and Xie, C. (2013). Cross-correlations between Renminbi and four major currencies in the Renminbi currency basket. *Physica A: Statistical Mechanics and its Applications* 392.6, 1418–1428.

Wang, Y., Liu, L. and Gu, R. (2009). Analysis of efficiency for Shenzhen stock market based on multifractal detrended fluctuation analysis. *International Review of Financial Analysis* 18.5, 271–276.

Wei, W. C., Gerace, D. and Frino, A. (2013). Informed trading, flow toxicity and the impact on intraday trading factors. *Australasian Accounting, Business and Finance Journal* 7.2, 3–24.

Weiß, G. N. and Supper, H. (2013). Forecasting liquidity-adjusted intraday value-at-risk with vine copulas. *Journal of Banking & Finance* 37.9, 3334–3350.

Welling, M. (2005). Fisher linear discriminant analysis. *Department of computer science, University of Toronto* 3.1.

Wen, F., Gong, X. and Cai, S. (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks. *Energy Economics* 59, 400–413.

Werbos, P. (1974). Beyond Regression:" New Tools for Prediction and Analysis in the Behavioral Sciences. *Ph. D. dissertation, Harvard University*.

Werbos, P. J. et al. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE* 78.10, 1550–1560.

Yamasaki, K., Muchnik, L., Havlin, S., Bunde, A. and Stanley, H. E. (2005). Scaling and memory in volatility return intervals in financial markets. *Proceedings of the National Academy of Sciences* 102.26, 9424–9428.

Yang, T.-W. and Zhu, L. (2016). A reduced-form model for level-1 limit order books. *Market Microstructure and Liquidity* 2.02, 1650008.

Zhang, L. et al. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* 12.6, 1019–1043.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100.472, 1394–1411.

Zhang, Z., Zohren, S. and Roberts, S. (2019). DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. *IEEE Transactions on Signal Processing*.

Zheng, B., Moulines, E. and Abergel, F. (2013). Price jump prediction in a limit order book. *journal of mathematical finance* 3.2, 242–255.

— (2012). Price jump prediction in limit order book. *arXiv preprint arXiv:1204.1381*.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics* 14.1, 45–52.

Zhou, W.-X. (2009). The components of empirical multifractality in financial returns. *EPL (Europhysics Letters)* 88.2, 28004.

PUBLICATIONS

PUBLICATION

I

Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods

Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A.

Journal of Forecasting 37.8 (2018), 852–866
DOI: 10.1002/for.2543

Received: 15 May 2018 | Accepted: 7 July 2018
DOI: 10.1002/for.2543

RESEARCH ARTICLE

WILEY

Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods

Adamantios Ntakaris[1] | Martin Magris[2] | Juho Kanniainen[2] | Moncef Gabbouj[1] | Alexandros Iosifidis[3]

[1]Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland

[2]Laboratory of Industrial and Information Management, Tampere University of Technology, Tampere, Finland

[3]Department of Engineering, Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark

Correspondence
Adamantios Ntakaris, Laboratory of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, Tampere, Finland.
Email: adamantios.ntakaris@tut.fi

Funding information
H2020 Marie Sklodowska-Curie Actions, Grant/Award Number: MSCA-ITN-ETN 675044

Abstract

Managing the prediction of metrics in high-frequency financial markets is a challenging task. An efficient way is by monitoring the dynamics of a limit order book to identify the information edge. This paper describes the first publicly available benchmark dataset of high-frequency limit order markets for mid-price prediction. We extracted normalized data representations of time series data for five stocks from the Nasdaq Nordic stock market for a time period of 10 consecutive days, leading to a dataset of ∼4,000,000 time series samples in total. A day-based anchored cross-validation experimental protocol is also provided that can be used as a benchmark for comparing the performance of state-of-the-art methodologies. Performance of baseline approaches are also provided to facilitate experimental comparisons. We expect that such a large-scale dataset can serve as a testbed for devising novel solutions of expert systems for high-frequency limit order book data analysis.

KEYWORDS

high-frequency trading, limit order book, mid-price, machine learning, ridge regression, single hidden feedforward neural network

1 | INTRODUCTION

Automated trading became a reality when the majority of exchanges adopted it globally. This environment is ideal for high-frequency traders. High-frequency trading (HFT) and a centralized matching engine, referred to as a limit order book (LOB), are the main drivers for generating big data (Seddon & Currie, 2017). In this paper, we describe a new order book dataset consisting of approximately 4 million events for 10 consecutive trading days for five stocks. The data are derived from the ITCH feed provided by Nasdaq OMX Nordic and consists of the time-ordered sequences of messages that track and record all the events occurring in the specific market. It provides a complete market-wide history of 10 trading days. Additionally, we define an experimental protocol to evaluate the performance of research methods in mid-price prediction.[1]

Datasets, like the one presented here, come with challenges, including the selection of appropriate data transformation, normalization, description, and classification. This type of massive dataset requires a very good understanding of the available information that can be extracted

[1]Mid-price is the average of the best bid and best ask prices.

for further processing. We follow the information edge, as has been recently presented by Kercheval and Zhang (2015). The authors provide a detailed description of representations that can be used for a mid-price movement prediction metric. In light of this data representation, they apply nonlinear classification based on support vector machines (SVM) in order to predict the movement of this metric. Such a supervised learning model exploits class labels[2] for short- and long-term prediction. However, they train their model based on a very small (when compared to the size of the data that can be available for such applications) dataset of 4,000 samples. This is due to the limitations of many nonlinear kernel-based classification models related to their time and space complexity with respect to the training data size. On the other hand, Sirignano (2016) uses large amounts of data for nonlinear classification based on a feedforward network. The author takes advantage of the local spatial structure[3] of the data for modeling the joint distribution of the LOB's state based on its current state.

Despite the major importance of publicly available datasets for advancing research in the HFT field, there are no detailed public available benchmark datasets for method evaluation purposes. In this paper, we describe the first publicly available dataset[4] for an LOB-based HFT that has been collected in the hope of facilitating future research in the field. Based on Kercheval and Zhang (2015), we provide time series representations of approximately 4,000,000 trading events and annotations for five classification problems. Baseline results of two widely used methods—that is, linear and nonlinear regression models, are also provided. In this way, we introduce this new problem for the expert systems community and provide a testbed for facilitating future research. We hope that attracting the interest of expert systems will lead to the rapid improvement of the performance achieved in the provided dataset, thus leading to much better state-of-the-art solutions to this important problem.

The dataset described in this paper can be useful for financial expert systems in two ways. First, it can be used to identify circumstances under which markets are stable, which is very important for liquidity providers (market makers) to make the spread. Consequently, such an intelligent system would be valuable as a framework that can increase liquidity provision. Secondly, analysis of the data can be used for model selection by speculative traders, who are trading based on their predictions on market movements. In future research, this paper can be employed to identify order book spoofing—that is, situations where markets are exposed to manipulation by limit orders. In this case, spoofers could aim to move markets in certain directions by limit orders that are canceled before they are filled. Therefore, this research is relevant not only for market makers and traders but also for supervisors and regulators.

Therefore, the present work makes the following contributions: (1) To the best of our knowledge this is the first publicly available LOB-ITCH dataset for machine learning experiments on the prediction of mid-price movements. (2) We provide baselines methods based on ridge regression and a new implementation of an RBF neural network based on *k*-means algorithm. (3) The paper provides information about the prediction of mid-price movements to market makers, traders, and regulators. This paper does not suggest any trading strategies and is reliant on purely machine learning metrics prediction. Overall, this work is an empirical exploration of the challenges that come with high-frequency trading and machine learning applications.

The data from Nasdanq Helsinki Stock Exchange offers important benefits. In the USA the limit orders for a given asset are spread between several exchanges, causing fragmentation of liquidity. The fragmentation poses a problem for empirical research, because, as Gould, Porter, Williams, McDonald, Fenn, and Howison (2013) point out, the "differences between different trading platforms' matching rules and transaction costs complicate comparisons between different limit order books for the same asset." These issues related to fragmentation are not present with data obtained from less fragmented Nasdaq Nordic markets. Moreover, Helsinki Exchange is a pure limit order market, where the market makers have a limited role.

The rest of the paper is organized as follows. We provide a comprehensive literature review of the field in Section 2. Dataset and experimental protocol descriptions are provided in Section 33. Quantitative and qualitative comparisons of the new dataset, along with related data sources, are provided in Section 4. In Section 5, we describe the engineering of our baselines. Section 6 presents our empirical results and Section 7 concludes.

[2]Labels are extracted from annotations provided by experts and represent the direction of the mid-price. Three different states are defined—that is, upward, downward, and stationary movement.

[3]By local movement, the author means that the conditional movement of the future price (e.g., best ask price movement) depends, locally, on the current LOB state.

[4]The dataset can be downloaded from: https://etsin.avointiede.fi/dataset/urn-nbn-fi-csc-kata20170601153214969115https://etsin.avointiede.fi/dataset/urn-nbn-fi-csc-kata20170601153214969115.

2 | MACHINE LEARNING FOR HFT AND LOB

The complex nature of HFT and LOB spaces is suitable for interdisciplinary research. In this section, we provide a comprehensive review of recent methods exploiting

machine learning approaches. Regression models, neural networks, and several other methods have been proposed to make inferences of the stock market. Existing literature ranges from metric prediction to optimal trading strategies identification. The research community has tried to tackle the challenges of prediction and data inference from different angles. Although mid-price prediction can be considered a traditional time series prediction problem, there are several challenges that justify HFT as a unique problem.

2.1 | Regression analysis

Regression models have been widely used for HFT and LOB prediction. Zheng, Moulines, and Abergel (2012) utilize logistic regression in order to predict the inter-trade price jump. Alvim, dos Santos, and Milidiu (2010) use support vector regression (SVR) and partial least squares (PLS) for trading volume forecasting for 10 Bovespa stocks. Pai and Lin (2005) use a hybrid model for stock price prediction. They combine an autoregressive integrated moving average (ARIMA) model and an SVM classifier in order to model nonlinearities of class structure in regression estimation models. Liu and Park (2015) develop a multivariate linear model to explain short-term stock price movement where a bid–ask spread is used for classification purposes. Detollenaere and D'hondt (2017) apply an adaptive least absolute shrinkage and selection operator (LASSO)[5] for variable selection, which best explains the transaction cost of the split order. They apply an adjusted ordinal logistic method for classifying ex ante transaction costs into groups. Cenesizoglu, Dionne, and Zhou (2014) work on a similar problem. They hold that the state of the limit order can be informative for the direction of future prices and try to prove their position by using an autoregressive model.

Panayi, Peters, Danielsson, and Zigrand (2016) use generalized linear models (GLM) and generalized additive models for location, shape, and scale (GAMLSS) models in order to relate the threshold exceedance duration (TED), which measures the length of time required for liquidity replenishment, to the state of the LOB. Yu (2006) tries to extract information from order information and order submission based on the ordered probit model.[6] The author shows, in the case of Shanghai's stock market, that an LOB's information is affected by the trader's strategy, with different impacts on the bid and ask sides. Amaya, Filbien, Okou, and Roch (2015) use panel regression[7] for order imbalances and liquidity costs in LOBs so as to identify resilience in the market. Their findings show that such order imbalances cause liquidity issues that last for up to 10 minutes. Malik and Lon Ng (2014) analyze the asymmetric intra-day patterns of LOBs. They apply regression with a power transformation on the notional volume weighted average price (NVWAP) curves in order to conclude that both sides of the market behave asymmetrically to market conditions.[8] In the same direction, Ranaldo (2004) examines the relationship between trading activity and the order flow dynamics in LOBs, where the empirical investigation is based on a probit model. Cao, Hansch, and Wang (2009) examine the depth of different levels of an order book by using an autoregressive (AR) model of order 5 (the AR(5) framework). They find that levels beyond the best bid and best ask prices provide moderate information regarding the true value of an asset. Finally, Creamer (2012) suggests that the LogitBoost algorithm is ideal for selecting the right combination of technical indicators.[9]

2.2 | Neural networks

HFT is mainly a scalping[10] strategy according to which the chaotic nature of the data creates the proper framework for the application of neural networks. Levendovszky and Kia (2012) propose a multilayer feedforward neural network for predicting the price of a EUR/USD pair, trained by using the backpropagation algorithm. Sirignano (2016) proposes a new method for training deep neural networks that try to model the joint distribution of the bid and ask depth, where a focal point is the spatial nature[11] of LOB levels. Bogoev and Karam (2016) propose the use of a single hidden-layer feedforward neural (SLFN) network for the detection of quote stuffing and momentum ignition. Dixon (2016) uses a recurrent neural network (RNN) for mid-price predictions of T-bond[12] and ES futures[13] based on ultra-high-frequency data. Rehman, Khan, and

[5] Adaptive weights are used for penalizing different coefficients in the l_1 penalty term.

[6] The method is the generalization of a linear regression model when the dependent variable is discrete.

[7] Panel regression models provide information on data characteristics individually, but also across both individuals over time.

[8] Market conditions of an industry sector have an impact on sellers and buyers who are related to it. Factors to consider include the number of competitors in the sector. For example, if there is a surplus, new companies may find it difficult to enter the market and remain in business.

[9] Technical indicators are mainly used for short-term price movement predictions. They are formulas based on historical data.

[10] Scalping is a type of trading strategy according to which the trader tries to make a profit for small changes in a stock.

[11] The spatial nature of this type of neural network and its gradient can be evaluated at far fewer grid points. This makes the model less computationally expensive. Furthermore, the suggested architecture can model the entire distribution in the R^d space.

[12] Treasury bond (T-bond) is a long-term fixed interest rate debt security issued by the federal government.

[13] E-mini S&P 500 (ES futures) are electronically traded futures contracts whose value is one-fifth the size of standard S&P futures.

Mahmud (2014) apply recurrent Cartesian genetic programming evolved artificial neural network (RCGPANN) for predicting five currency rates against the Australian dollar. Galeshchuk (2016) suggests that a multilayer perceptron (MLP) architecture, with three hidden layers, is suitable for exchange rate prediction. Majhi, Panda, and Sahoo (2009) use the functional link artificial neural network (FLANN) in order to predict price movements in the DJIA[14] and S&P 500[15] stock indices.

Deep belief networks are employed by Sharang and Rao (2015) to design a medium-frequency portfolio trading strategy. Hallgren and Koski (2016) use continuous-time Bayesian networks (CTBNs) for causality detection. They apply their model on tick-by-tick high-frequency foreign exchange (FX) EUR/USD data using a Skellam process.[16] Sandoval and Hernández (2015) create a profitable trading strategy by combining hierarchical hidden Markov models (HHMM), where they consider wavelet-based LOB information filtering. In their work, they also consider a two-layer feedforward neural network in order to classify the upcoming states. They nevertheless report limitations in the neural network in terms of the volume of the input data.

2.3 | Maximum margin and reinforcement learning

Palguna and Pollak (2016) use nonparametric methods on features derived from LOB, which are incorporated into order execution strategies for mid-price prediction. In the same direction, Kercheval and Zhang (2015) employ a multi-class SVM for mid-price and price spread crossing prediction. Han et al. (2015) base their research on Kercheval and Zhang by using multi-class SVM for mid-price movement prediction. More precisely, they compare multi-class SVM (exploring linear and RBF kernels) to decision trees using bagging for variance reduction.

Kim (2001) uses input/output hidden Markov models (IOHMMs) and reinforcement learning (RL) in order to identify the order flow distribution and market-making strategies, respectively. Yang et al. (2015) apply apprenticeship learning[17] methods, like linear inverse reinforcement learning (LIRL) and Gaussian process IRL (GPIRL), to recognize traders or algorithmic trades based on the observed limit orders. Chan and Shelton (2001) use RL for market-making strategies, where experiments based on a Monte Carlo simulation and a state–action–reward–state–action (SARSA) algorithm test the efficacy of their policy. In the same vein, Kearns and Nevmyvaka (2013) implement RL for trade execution optimization in lit and dark pools. Especially in the case of dark pools, they apply a censored exploration algorithm to the problem of smart order routing (SOR). Yang, Paddrik, Hayes, Todd, Kirilenko, Beling, and Scherer (2012) examine an IRL algorithm for the separation of HFT strategies from other algorithmic trading activities. They also apply the same algorithm to the identification of manipulative HFT strategies (i.e., spoofing). Felker, Mazalov, and Watt (2014) predict changes in the price of quotes from several exchanges. They apply feature-weighted Euclidean distance to the centroid of a training cluster. They calculate this type of distance to the centroid of a training cluster where feature selection is taken into consideration because several exchanges are included in their model.

2.4 | Additional methods for HFT and LOB

HFT and LOB research activity also covers topics like the optimal submission strategies of bid and ask orders, with a focus on the inventory risk that stems from an asset's value uncertainty, as in the work of Avellaneda and Stoikov (2008). Chang (2015) models the dynamics of LOB by using a Bayesian inference of the Markov chain model class, tested on high-frequency data. An and Chan (2017) suggest a new stochastic model that is based on independent compound Poisson processes of the order flow. Talebi, Hoang, and Gavrilova (2014) try to predict trends in the FX market by employing a multivariate Gaussian classifier (MGC) combined with Bayesian voting. Fletcher, Hussain, and Shawe-Taylor (2010) examine trading opportunities for the EUR/USD where the price movement is based on multiple kernel learning (MKL). More specifically, the authors utilize SimpleMKL and the more recent LPBoost-MKL methods for training a multi-class SVM. Christensen and Woodmansey (2013) develop a classification method based on the Gaussian kernel in order to identify iceberg[18] orders for GLOBEX.

Maglaras, Moallemi, and Zheng (2015) consider the LOB as a multi-class queueing system in order to solve the problem placement of limit and market order placements. Mankad, Michailidis, and Kirilenko (2013) apply a static plaid clustering technique to synthetic data in order to

[14] The Dow Jones Industrial Average (DJIA) is the price-weighted average of the 30 largest, publicly owned US companies.

[15] S&P 500 is the index that provides a summary of the overall market by tracking some of the 500 top stocks in US stock market.

[16] A Skellam process is defined as $S(t) = N^{(1)}(t) - N^{(2}(t), \quad t \geqslant 0$, where $N^{(1)}(t)$ and $N^{(2)}(t)$ are two independent homogeneous Poisson processes.

[17] Motivation for apprenticeship learning is to use IRL techniques to learn the reward function and then use this function in order to define a Markov decision problem (MDP).

[18] Iceberg order is the conditional request made to the broker to sell or buy a larger quantity of the stock, but in smaller predefined quantities.

classify the different types of trades. Aramonte, Schindler, and Rosen (2013) show that the information asymmetry in a high-frequency environment is crucial.

Vella and Ng (2016) use higher-order fuzzy systems (i.e., an adaptive neuro-fuzzy inference system) by introducing T2 fuzzy sets, where the goal is to reduce microstructure noise in the HFT sphere. Abernethy and Kale (2013) apply market-maker strategies based on low-regr et algorithms for the stock market. Almgren and Lorenz (2006) explain price momentum by modeling Brownian motion with a drift whose distribution is updated based on Bayesian inference. Næs and Skjeltorp (2006) show that the order book slope measures the elasticity of supplied quantity as a function of asset prices related to volatility, trading activity, and an asset's dispersion beliefs.

3 | THE LOB DATASET

In this section, we describe in detail our dataset collected in order to facilitate future research in LOB-based HFT. We start by providing a detailed description of the data in Section 3.1. Data processing steps are followed in order to extract message books and LOBs, as described in Section 3.2.

3.1 | Data description

Extracting information from the ITCH flow, and without relying on third-party data providers, we analyze stocks from different industry sectors for 10 full days of ultra-high-frequency intra-day data. The data provide information regarding trades against hidden orders. Coherently, the nondisplayable hidden portions of the total volume of a so-called iceberg order are not accessible from the data. Our ITCH feed data is day specific and market wide, which means that we deal with one file per day with data over all the securities. Information (block A in Figure 1) regarding (i) messages for order submissions, (ii) trades, and (iii) cancellations is included. For each order, its type (buy/sell), price, quantity, and exact time stamp on a millisecond basis is available. In addition, (iv) administrative messages (i.e., trading halts or basic security data), (v) event controls (i.e., start and ending of trading days, states of market segments), and (vi) net order imbalance indicators are also included.

The next step is the development and implementation of a C++ converter to extract all the information relevant to a given security. We perform the same process for five stocks traded on the Nasdaq OMX Nordic at the Helsinki exchange from June 1, 2010 to June 14, 2010.[19] These data are stored in a Linux cluster. Information related to the five stocks is illustrated in Table 1. The selected stocks[20] are traded in one exchange (Helsinki) only. By choosing only one stock market exchange, the trader has the advantage of avoiding issues associated with fragmented markets. In the case of fragmented markets, the limit orders for

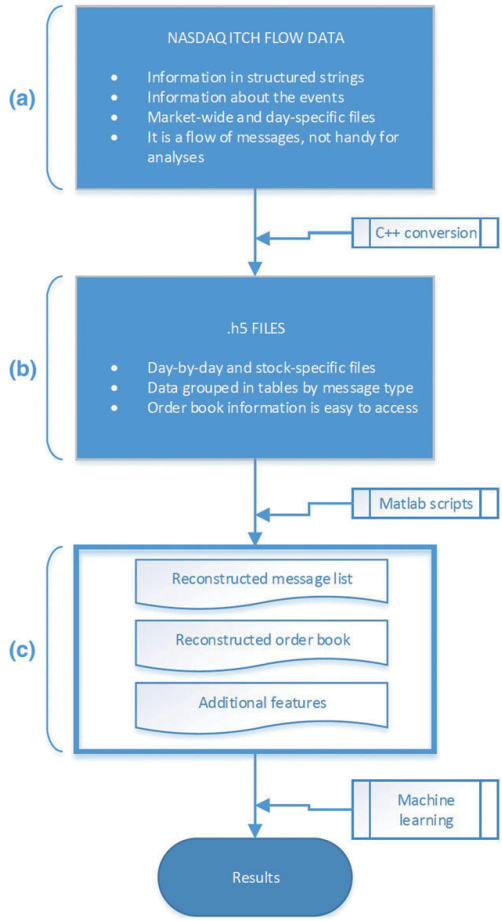


FIGURE 1 Data processing flow [Colour figure can be viewed at wileyonlinelibrary.com]

[19]There have been about 23,000 active order books, the vast majority of which are very illiquid, show sporadic activity, and correspond to little and noisy data.

[20]The choice is driven by the necessity of having a sufficient amount of data for training (this excludes illiquid stocks) while covering different industry sectors. These five selected stocks (see Table 1), which aggregate input message list and order book data for feature extraction, are about 4 GB; RTRKS was suspended from trading and delisted from the Helsinki exchange on November 20, 2014.

a given asset are spread between several exchanges, posing problems from empirical data analysis (O'Hara & Ye, 2011).

The Helsinki Stock Exchange, operated by Nasdaq Nordic, is a pure electronic limit order market. The ITCH feed keeps a record of all the events, including those that take place outside active trading hours. At the Helsinki exchange, the trading period goes from 10:00 to 18:25 (local time, UTC/GMT +2 hours). However, in the ITCH feed, we observe several records outside those trading hours. In particular, we consider the regulated auction period before 10:00, which is used to set the opening price of the day (the so-called pre-opening period) before trading begins. This is a structurally different mechanism following different rules with respect to the order book flow during trading hours. Similarly, another structural break in the order book's dynamics is due to the different regulations that are in force between 18:25 and 18:30 (the so-called post-opening period). As a result, we retain exclusively the events occurring between 10:30 and 18:00. More information related to the above-mentioned issues can be found in Siikanen, Kanniainen, and Luoma 2017 and (Siikanen, Kanniainen, & Valli, 2017). Here, the order book is expected to have comparable dynamics with no biases or exceptions caused by its proximity to the market opening and closing times.

3.2 | Limit order and message books

Message and LOBs are processed for each of the 10 days for the five stocks. More specifically, there are two types of messages that are particularly relevant here: (i) "add order messages," corresponding to order submissions; and (ii) "modify order messages," corresponding to updates on the status of existing orders through order cancellations and order executions. Example message[21] and limit order[22] books are illustrated in Tables 2 and Table 3, respectively.

LOB is a centralized trading method that is incorporated by the majority of exchanges globally. It aggregates the limit orders of both sides (i.e., the ask and bid sides) of the stock market (e.g., the Nordic stock market). LOB matches every new event type according to several characteristics. Event types and LOB characteristics describe the current state of this matching engine. Event types can be executions, order submissions, and order cancellations. Characteristics of LOB are the resolution parameters (Gould, Porter, Williams, McDonald, Fenn, & Howison, 2013), which are the tick size π (i.e., the smallest permissible price between different orders), and the lot size σ (i.e., the smallest amount of a stock that can be traded and is defined as $\{k\sigma | k = 1, 2, \dots\}$). Order inflow and resolution parameters will formulate the dynamics of the LOB, whose current state will be identified by the state variable of four elements $(s_t^b, q_t^b, s_t^a, q_t^a), t \geq 0$, where s_t^b (s_t^b) is the best bid (ask) price and q_t^b (q_t^a) is the size of the best bid (ask) level at time t.

In our data, timestamps are expressed in milliseconds based on 1 Jan 1970 format and shifted by three hours with respect to Eastern European Time (in the data, the trading day goes from 7:00 to 15:25). ITHC feed prices are recorded up to 4 decimal places and, in our data, the decimal point is removed by multiplying the price by 10,000, where currency is in euros for the Helsinki exchange. The tick size, defined as the smallest possible gap between the ask and bid prices, is 1 cent. Similarly, order quantities are constrained to integers greater than one.

3.3 | Data availability and distribution

In compliance with Nasdaq OMX agreements, the normalized feature dataset is made available to the research community.[23] The open-access version of our data has been normalized in order to prevent reconstruction of the original Nasdaq data.

3.4 | Experimental protocol

In order to make our dataset a benchmark that can be used for the evaluation of HTF methods based on LOB information, the data are accompanied by the following experimental protocol. We develop a day-based prediction framework following an anchored forward cross-validation format. More specifically, the training set is increased by 1 day in each fold and stops after $n - 1$ days (i.e., after 9 days in our case where $n = 10$). On each fold, the test set corresponds to 1 day of data, which moves in a rolling window format. The experimental setup is illustrated in Figure 2. Performance is measured by calculating the mean accuracy, recall, precision, and F1 score over all folds, as well as the corresponding standard deviation. We measure our results based on these metrics, which are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2)$$

[21] A sample from FI0009002422 on June 1, 2010.
[22] A sample from FI0009002422 on June 1, 2010.

[23] We thank Ms. Sonja Salminen at Nasdaq for her support and help.

TABLE 1 Stocks used in the analysis

ID	ISIN code	Company	Sector	Industry
KESBV	FI0009000202	Kesko Oyj	Consumer Defensive	Grocery Stores
OUT1V	FI0009002422	Outokumpu Oyj	Basic Materials	Steel
SAMPO	FI0009003305	Sampo Oyj	Financial Services	Insurance
RTRKS	FI0009003552	Rautaruukki Oyj	Basic Materials	Steel
WRT1V	FI0009000727	Wärtsilä Oyj	Industrials	Diversified Industrials

TABLE 2 Message list example

Timestamp	ID	Price	Quantity	Event	Side
1275386347944	6505727	126200	400	Cancellation	Ask
1275386347981	6505741	126500	300	Submission	Ask
1275386347981	6505741	126500	300	Cancellation	Ask
1275386348070	6511439	126100	17	Execution	Bid
1275386348070	6511439	126100	17	Submission	Bid
1275386348101	6511469	126600	300	Cancellation	Ask

TABLE 3 Order book example

			Level 1				Level 2				…
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	…
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	…
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	…
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	…
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	…
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	…

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

where TP and TF represent the true positives and true negatives, respectively, of the mid-price prediction label compared with the ground truth, where FP and FN represents the false positives and false negatives, respectively. From among the above metrics, we focus on the $F1$ score performance. The main reason that we focus on $F1$ score is based on its ability only to be affected in one direction of skew distributions, in the case of unbalanced classes like ours. On the contrary, accuracy cannot differentiate between the number of correct labels (i.e., related to mid-price movement direction prediction) of different classes where the other three metrics can separate the correct labels among different classes, with $F1$ being the harmonic mean of Precision and Recall.

We follow an event-based inflow, as used in Li, et al. (2016). This is due to the fact that events (i.e., orders, executions, and cancellations) do not follow a uniform inflow rate. Time intervals between two consecutive events can vary from milliseconds to several minutes of difference. Event-based data representation avoids issues related to such big differences in data flow. As a result, each of our representations is a vector that contains information for 10 consecutive events. Event-based data description leads to a dataset of approximately half a million representations (i.e., 394,337 representations). We represent these events using the 144-dimensional representation proposed recently by Kercheval and Zhang (2015), formed by three types of features: (a) the raw data of a 10-level limit order containing price and volume values for bid and ask orders; (b) features describing the state of the LOB, exploiting past information; and (c) features describing the information edge in the raw data by taking time into account. Derivations of time, stock price, and volume are calculated for short and long-term projections. More specifically, types in features u_7, u_8, and u_9 are: *trades, orders, cancellations, deletion, execution of a visible limit order*, and *execution of a hidden limit order*. Expressions used for calculating these features are provided in Table 4. One limitation of the adopted features

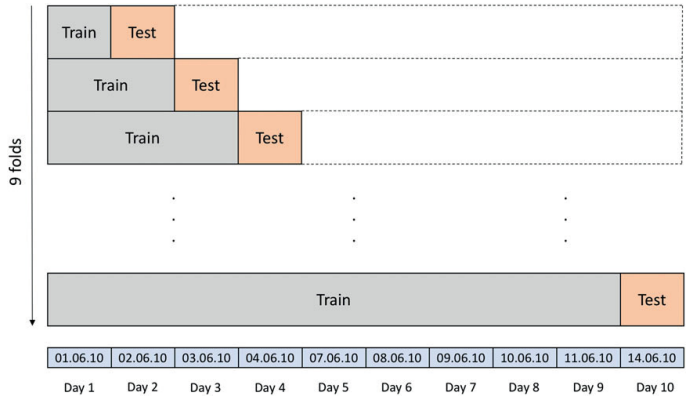


FIGURE 2 Experimental setup framework [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Feature sets

Feature set	Description	Details
Basic	$u_1 = \{P_i^{\text{ask}}, V_i^{\text{ask}}, P_i^{\text{bid}}, V_i^{\text{bid}}\}_{i=1}^{n}$	10(= n)-level LOB data
Time-insensitive	$u_2 = \{(P_i^{\text{ask}} - P_i^{\text{bid}}), (P_i^{\text{ask}} + P_i^{\text{bid}})/2\}_{i=1}^{n}$	Spread & Mid-price
	$u_3 = \{P_n^{\text{ask}} - P_1^{\text{ask}}, P_1^{\text{bid}} - P_n^{\text{bid}}, \lvert P_{i+1}^{\text{ask}} - P_i^{\text{ask}}\rvert, \lvert P_{i+1}^{\text{bid}} - P_i^{\text{bid}}\rvert\}_{i+1}^{n}$	Price differences
	$u_4 = \left\{ \frac{1}{n}\sum_{i=1}^{n} P_i^{\text{ask}}, \frac{1}{n}\sum_{i=1}^{n} P_i^{\text{bid}}, \frac{1}{n}\sum_{i=1}^{n} V_i^{\text{ask}}, \frac{1}{n}\sum_{i=1}^{n} V_i^{\text{bid}} \right\}$	Price & Volume means
	$u_5 = \left\{ \sum_{i=1}^{n} (P_i^{\text{ask}} - P_i^{\text{bid}}), \sum_{i=1}^{n} (V_i^{\text{ask}} - V_i^{\text{bid}}) \right\}$	Accumulated differences
Time-sensitive	$u_6 = \left\{ \mathrm{d}P_i^{\text{ask}}/\mathrm{d}t, \mathrm{d}P_i^{\text{bid}}/\mathrm{d}t, \mathrm{d}V_i^{\text{ask}}/\mathrm{d}t, \mathrm{d}V_i^{\text{bid}}/\mathrm{d}t \right\}_{i=1}^{n}$	Price & Volume derivation
	$u_7 = \{\lambda_{\Delta t}^1, \lambda_{\Delta t}^2, \lambda_{\Delta t}^3, \lambda_{\Delta t}^4, \lambda_{\Delta t}^5, \lambda_{\Delta t}^6\}$	Average intensity per type
	$u_8 = \left\{ \mathbf{1}_{\lambda_{\Delta_t}^1 > \lambda_{\Delta_T}^1}, \mathbf{1}_{\lambda_{\Delta_t}^2 > \lambda_{\Delta_T}^2}, \mathbf{1}_{\lambda_{\Delta_t}^3 > \lambda_{\Delta_T}^3}, \mathbf{1}_{\lambda_{\Delta_t}^4 > \lambda_{\Delta_T}^4}, \mathbf{1}_{\lambda_{\Delta_t}^5 > \lambda_{\Delta_T}^5}, \mathbf{1}_{\lambda_{\Delta_t}^6 > \lambda_{\Delta_T}^6} \right\}$	Relative intensity comparison
	$u_9 = \{\mathrm{d}\lambda^1/\mathrm{d}t, \mathrm{d}\lambda^2/\mathrm{d}t, \mathrm{d}\lambda^3/\mathrm{d}t, \mathrm{d}\lambda^4/\mathrm{d}t, \mathrm{d}\lambda^5/\mathrm{d}t, \mathrm{d}\lambda^6/\mathrm{d}t\}$	Limit activity acceleration

is the lack of information related to order flow (i.e., the sequence of order book messages). However, as can be seen in the Results Section 6, the baselines achieve relatively good performance and therefore we leave the introduction of extra features that can enhance performance to future research.

We provide three sets of data, each created by following a different data normalization strategy—that is, z-score, min–max, and decimal precision normalization—for every i data sample. Z-score, in particular, is the normalization process through which we subtract the mean from our input data for each feature separately and divide by the standard deviation of the given sample:

$$\mathbf{x}_i^{(z\text{-score})} = \frac{\mathbf{x}_i - \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j}{\sqrt{\frac{1}{N}\sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})^2}}, \tag{5}$$

where $\bar{\mathbf{x}}$ denotes the mean vector, as appears in Equation 5. On the other hand, min–max scaling, as described by

$$\mathbf{x}_i^{(\text{MM})} = \frac{\mathbf{x}_i - \mathbf{x}_{\min}}{\mathbf{x}_{\max} - \mathbf{x}_{\min}}, \tag{6}$$

is the process of subtracting the minimum value from each feature and dividing it by the difference between the maximum and minimum value of that feature sample. The third scaling setup is the decimal precision approach. This normalization method is based on moving the decimal points of each of the feature values. Calculations follow the absolute value of each feature sample:

$$\mathbf{x}_i^{(\text{DP})} = \frac{\mathbf{x}_i}{10^k}, \tag{7}$$

where k is the integer that will give us the maximum value for $|\mathbf{x}_{\text{DP}}| < 1$.

Having defined the event representations, we use five different projection horizons for our labels. Each of these

TABLE 5 HFT dataset examples

	Dataset	Public available	Unit time	Period	Asset class / No. of stocks	Size	Annotations
1	Dukascopy	✓	ms	Up to date	Various	∼20,000 events/day	×
2	truefx	✓	ms	Up to date	15 FX pairs	∼300,000 events/day	×
3	Nasdaq	AuR	ms	2008-09	Equity / 120	—	×
4	Nasdaq	AuR	ms	10/07 & 06/08	Equity / 500	∼55,000 events/day	×
5	Nasdaq	×	ms	—	Equity / 5	2,000 data points	×
6	Euronext	AuR	—	—	Several products	—	×
7	Nasdaq	×	ns	01/14-08/15	Equity / 489	50 TB	×
8	Our–Nasdaq	✓	ms	01-14/06/10	Equity / 5	4 M samples	✓

horizons portrays a different future projection interval of the mid-price movement (i.e., upward, downward, and stationary mid-price movement). More specifically, we extract labels based on short-term and long-term, event-based, relative changes for the next 1, 2, 3, 5, and 10 events for our representations dataset.

Our labels describe the percentage change of the mid-price, which is calculated as follows:

$$l_i^{(j)} = \frac{\frac{1}{k}\sum_{j=i+1}^{i+k} m_j - m_i}{m_i}, \tag{8}$$

where m_j is the future mid-price ($k = 1, 2, 3, 5$, or 10 next events in our representations) and m_i is the current mid-price. The extracted labels are based on a threshold for the percentage change of 0.002. For percentage changes equal to or greater than 0.002, we use label 1. For percentage change that varies from −0.00199 to 0.00199, we use label 2, and, for percentage change smaller or equal to −0.002, we use label 3.

4 | EXISTING DATASETS DESCRIBED IN THE LITERATURE

In this section, we list existing HFT datasets described in the literature and provide qualitative and quantitative comparisons to our dataset. The following works mainly focus on datasets that are related to machine learning methods.

There are mainly three sources of data from which a high-frequency trader can choose. The first option is the use of publicly available data (e.g., (1) Dukascopy and (2) truefx), where no prior agreement is required for data acquisition. The second option is publicly available data upon request for academic purposes, which can be found in (3) Brogaard, Hendershott, and Riordan (2014), (4) Hasbrouck and Saar (2013), (5) De Winne and D'hondt 2007, Detollenaere and D'hondt (2017), and Carrion (2013). Finally, the third and most common option is data through platforms requiring a subscription fee, like those in (6) Kercheval and Zhang (2015); Li et al. (2016), and (7) Sirignano (2016). Existing data sources and characteristics are listed in Table 5.

In particular, the datasets are at a millisecond resolution, except for number 6 in the table. Access to various asset classes including FX, commodities, indices, and stocks is also provided. To the best of our knowledge, there is no available literature based on this type of dataset for equities. Another source of free tick-by-tick historical data is the truefx.com site, but the site provides data only for the FX market for several pairs of currencies at a millisecond resolution. The data contain information regarding timestamps (in millisecond resolution) and bid and ask prices. Each of these .csv files contains approximately 200,000 events per day. This type of data is used in a mean-reverting jump-diffusion model, as presented in Suwanpetai (2016).

There is a second category of datasets available upon request (AuR), as seen in Hasbrouck and Saar (2013). In this paper, the authors use the Nasdaq OMX ITCH for two periods: October 2007 and June 2008. For that period, they run samples at 10-minute intervals for each day where they set a cutoff mechanism for available messages per period.[24] The main disadvantage of uniformly sampling HFT data is that the trader loses vital information. Events come randomly, with inactive periods varying from a few milliseconds to several minutes or hours. In our work, we overcome this challenge by considering the information based on event inflow, rather than equal time sampling. Another example of data that is available only for academic purposes is Brogaard et al. (2014). The dataset contains information regarding timestamps, price, and buy–sell side prices but no other details related to daily events or feature vectors. Hasbrouck and Saar provide a detailed description of their Nasdaq OMX ITCH data, which is not directly accessible for testing and comparison with their

[24]The authors provide a threshold, which is based on 250 events per 10-minute sample interval.

baselines. They use these data to applying low-latency strategies based on measures that capture links between submissions, cancellations, and executions. De Winne and D'hondt (2007) and Detollenaere and D'hondt (2017) use similar datasets from Euronext for LOB construction. They specify that their dataset is available upon request from the provider. What is more, the data provider supplies details regarding the LOB construction by the user. Our work fills that gap since our dataset provides the full LOB depth and it is ready for use and comparison with our baselines.

The last category of dataset has dissemination restrictions. An example is the paper by Kercheval and Zhang (2015), where the authors are trying to predict the mid-price movement by using machine learning (i.e., SVM). They train their model with a very small number of samples (i.e., 4,000 samples). The HFT activity can produce a huge volume of trading events daily, as our database does with 100,000 daily events for only one stock. Moreover, the datasets in Kercheval and Zhang and in Sirignano (2016) are not publicly available, which makes comparison with other methods impossible. In the same direction, we also add works such as Hasbrouck (2009), Kalay, Sade, and Wohl (2004), and Kalay, Wei, and Wohl (2002), which utilize TAQ and Tel Aviv stock exchange datasets (not for machine learning methods), and require subscription.

5 | BASELINES

In order to provide performance baselines for our new dataset of HFT with LOB data, we conducted experiments with two regression models using the data representations described in Section 3.4. Details on the models used are provided in Sections 5.1 and 5.2. The baseline performances are provided in Section 6.

5.1 | Ridge regression (RR)

Ridge regression defines a linear mapping, expressed by the matrix $\mathbf{W} \in \mathbb{R}^{D\times C}$, that optimally maps a set of vectors $\mathbf{x}_i \in \mathbb{R}^D,\ i = 1, \cdots, N$ to another set of vectors (noted as target vectors) $\mathbf{t}_i \in \mathbb{R}^C,\ i = 1, \cdots, N$, by optimizing the following criterion:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{t}_i\|_2^2 + \lambda\|\mathbf{W}\|_F^2, \tag{9}$$

or using a matrix notation:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \|\mathbf{W}^T\mathbf{X} - \mathbf{T}\|_F^2 + \lambda\|\mathbf{W}\|_F^2. \tag{10}$$

In the above, $\mathbf{X} = [\mathbf{x}_i, \ldots, \mathbf{x}_N]$ and $\mathbf{T} = [\mathbf{t}_i, \ldots, \mathbf{t}_N]$ are matrices formed by the samples $\mathbf{x}_i$ and $\mathbf{t}_i$ as columns, respectively.

In our case, each sample $\mathbf{x}_i$ corresponds to an event, represented by a vector (with $D = 144$), as described in Section 3.4. For the three-class classification problems in our dataset, the elements of vectors $\mathbf{t}_i \in \mathbb{R}^C$ ($C = 3$ in our case) take values equal to $t_{ik} = 1$, if $\mathbf{x}_i$ belongs to class k, and if $t_{ik} = -1$ otherwise. The solution of Equation 10 is given by

$$\mathbf{W} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{T}^T, \tag{11}$$

or

$$\mathbf{W} = \left(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{X}\mathbf{T}^T, \tag{12}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimensions. Here, we should note that, in our case, where the size of the data is large, $\mathbf{W}$ should be computed using Equation 12, since the calculation of Equation 11 is computationally very expensive.

After the calculation of $\mathbf{W}$, a new (test) sample $\mathbf{x} \in \mathbb{R}^D$ is mapped on its corresponding representation in space $\mathbb{R}^C$—that is, $\mathbf{o} = \mathbf{W}^T\mathbf{x}$—and is classified according to the maximum value of its projection:

$$l_{\mathbf{x}} = \arg\max_k o_k. \tag{13}$$

5.2 | SLFN network-based nonlinear regression

We also test the performance of a nonlinear regression model. Since the application of kernel-based regression is computationally too intensive for the size of our data, we use an SLFN (Figure 3) network-based regression model. Such a model is formed as follows.

For fast network training, we train our network based on the algorithm proposed in Huang, Zhou, Ding, and Zhang (2012), Zhang, Kwok, and Parvin (2009), and Iosifidis, Tefas, and Pitas (2017). This algorithm is formed by

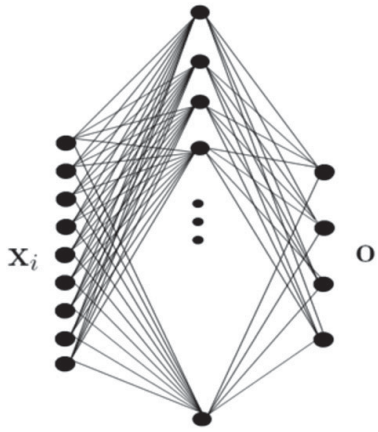


FIGURE 3 SLFN

two processing steps. In the first step, the network's hidden layer weights are determined either randomly (Huang, Zhou, Ding, & Zhang, 2012) or by applying clustering on the training data. We apply K-means clustering in order to determine K prototype vectors, which are subsequently used as the network's hidden layer weights.

Having determined the network's hidden layer weights $\mathbf{V} \in \mathbb{R}^{D\times K}$, the input data $\mathbf{x}_i, i = 1, \dots, N$ are nonlinearly mapped to vectors $\mathbf{h}_i \in \mathbb{R}^K$, expressing the data representations in the feature space determined by the network's hidden layer outputs $\mathbb{R}^K$. We use the radial basis function—that is, $\mathbf{h}_i = \varphi_{\text{RBF}}(\mathbf{x}_i)$—calculated in an element-wise manner, as follows:

$$h_{ik} = \exp\left(\frac{\|\mathbf{x}_i - \mathbf{v}_k\|_2^2}{2\sigma^2}\right), \quad k = 1, \cdots, K, \tag{14}$$

where σ is a hyperparameter denoting the spread of the RBF neuron and $\mathbf{v}_k$ corresponds to the kth column of $\mathbf{V}$.

The network's output weights $\mathbf{W} \in \mathbb{R}^{K\times C}$ are subsequently determined by solving for

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \|\mathbf{W}^T\mathbf{H} - \mathbf{T}\|_F^2 + \lambda\|\mathbf{W}\|_F^2, \tag{15}$$

where $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N]$ is a matrix formed by the network's hidden layer outputs for the training data and $\mathbf{T}$ is a matrix formed by the network's target vectors $\mathbf{t}_i, i = 1, \dots, N$ as defined in Section 5.1. The network's output weights are given by

$$\mathbf{W} = \left(\mathbf{H}\mathbf{H}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{H}\mathbf{T}^T. \tag{16}$$

After calculation of the network parameters $\mathbf{V}$ and $\mathbf{W}$, a new (test) sample $\mathbf{x} \in \mathbb{R}^D$ is mapped on its corresponding representations in spaces $\mathbb{R}^K$ and $\mathbb{R}^C$; that is, $\mathbf{h} = \phi_{\text{RBF}}(\mathbf{x})$ and $\mathbf{o} = \mathbf{W}^T\mathbf{h}$, respectively. It is classified according to the maximal network output:

$$l_{\mathbf{x}} = \arg\max_k o_k. \tag{17}$$

6 | RESULTS

In our first set of experiments, we have applied two supervised machine learning methods, as described in Sections 5.1 and 5.2, on a dataset that does not include the auction period. Results with the auction period will also be available. Since there is not a widely adopted experimental protocol for these datasets, we provide information for the five different label scenarios under the three normalization setups.

The tables in this section provide details regarding the results of experiments conducted on raw data and three different normalization setups. We present these results, for our baseline models, in order to give insight into the preprocessing step for a dataset like ours, to examine the strength of the predictability of the projected time horizon, and to understand the implications of the suggested methods. Data normalization can significantly improve the metric's performance in combination with the use of the right classifier. More specifically, we measure the predictability power of our models via the performance of the metrics of accuracy, precision, recall, and $F1$ score. For instance, Table 6 presents the results based on raw data (i.e., no data decoding), and in the case of the linear classifier RR and label 5 (i.e., the 5th mid-price event as predicted horizon), we achieve an $F1$ score of 40%, where as in Table 7 (i.e., the Z-score data decoding method), Table 8 (i.e., min–max data decoding method), and Table 9 (i.e., the decimal precision decoding method), we achieve 43%, 42%, and 40%, respectively. This shows

TABLE 6 Results based on unfiltered representations

Label	$\text{RR}_{\text{Accuracy}}$	$\text{RR}_{\text{Precision}}$	$\text{RR}_{\text{Recall}}$	RR_{F1}
1	0.637 ± 0.055	0.505 ± 0.145	0.337 ± 0.003	0.268 ± 0.014
2	0.555 ± 0.064	0.504 ± 0.131	0.376 ± 0.023	0.320 ± 0.050
3	0.489 ± 0.061	0.423 ± 0.109	0.397 ± 0.031	0.356 ± 0.070
5	0.429 ± 0.049	0.402 ± 0.113	0.425 ± 0.038	0.400 ± 0.093
10	0.453 ± 0.054	0.400 ± 0.105	0.400 ± 0.030	0.347 ± 0.066
Label	$\text{SLFN}_{\text{Accuracy}}$	$\text{SLFN}_{\text{Precision}}$	$\text{SLFN}_{\text{Recall}}$	SLFN_{F1}
1	0.636 ± 0.055	0.299 ± 0.075	0.335 ± 0.002	0.262 ± 0.015
2	0.536 ± 0.069	0.387 ± 0.132	0.345 ± 0.009	0.260 ± 0.035
3	0.473 ± 0.074	0.334 ± 0.080	0.357 ± 0.005	0.270 ± 0.021
5	0.381 ± 0.038	0.342 ± 0.058	0.370 ± 0.020	0.327 ± 0.043
10	0.401 ± 0.039	0.284 ± 0.102	0.356 ± 0.020	0.290 ± 0.070

TABLE 7 Results based on *Z*-score normalization

Label	$RR_{Accuracy}$	$RR_{Precision}$	RR_{Recall}	RR_{F1}
1	0.480 ± 0.040	0.418 ± 0.021	0.435 ± 0.029	0.410 ± 0.022
2	0.498 ± 0.052	0.444 ± 0.025	0.443 ± 0.031	0.440 ± 0.031
3	0.463 ± 0.045	0.438 ± 0.027	0.437 ± 0.033	0.433 ± 0.034
5	0.439 ± 0.042	0.436 ± 0.028	0.433 ± 0.028	0.427 ± 0.041
10	0.429 ± 0.046	0.429 ± 0.028	0.429 ± 0.043	0.416 ± 0.044
Label	**$SLFN_{Accuracy}$**	**$SLFN_{Precision}$**	**$SLFN_{Recall}$**	**$SLFN_{F1}$**
1	0.643 ± 0.056	0.512 ± 0.037	0.366 ± 0.019	0.327 ± 0.046
2	0.556 ± 0.066	0.550 ± 0.029	0.378 ± 0.011	0.327 ± 0.030
3	0.512 ± 0.069	0.497 ± 0.024	0.424 ± 0.047	0.389 ± 0.082
5	0.473 ± 0.036	0.468 ± 0.024	0.464 ± 0.028	0.459 ± 0.031
10	0.477 ± 0.048	0.453 ± 0.056	0.432 ± 0.025	0.410 ± 0.040

TABLE 8 Results Based on min–max normalization

Label	$RR_{Accuracy}$	$RR_{Precision}$	RR_{Recall}	RR_{F1}
1	0.637 ± 0.054	0.499 ± 0.118	0.339 ± 0.005	0.272 ± 0.015
2	0.561 ± 0.063	0.467 ± 0.117	0.400 ± 0.028	0.368 ± 0.060
3	0.492 ± 0.070	0.428 ± 0.111	0.400 ± 0.030	0.357 ± 0.072
5	0.437 ± 0.048	0.419 ± 0.078	0.429 ± 0.043	0.417 ± 0.063
10	0.452 ± 0.054	0.421 ± 0.110	0.399 ± 0.028	0.348 ± 0.066
Label	**$SLFN_{Accuracy}$**	**$SLFN_{Precision}$**	**$SLFN_{Recall}$**	**$SLFN_{F1}$**
1	0.640 ± 0.055	0.488 ± 0.104	0.348 ± 0.007	0.291 ± 0.022
2	0.558 ± 0.065	0.469 ± 0.066	0.399 ± 0.023	0.367 ± 0.050
3	0.499 ± 0.063	0.447 ± 0.068	0.410 ± 0.032	0.370 ± 0.063
5	0.453 ± 0.038	0.441 ± 0.041	0.444 ± 0.030	0.432 ± 0.050
10	0.450 ± 0.048	0.432 ± 0.070	0.406 ± 0.037	0.377 ± 0.062

TABLE 9 Results based on decimal precision normalization

Label	$RR_{Accuracy}$	$RR_{Precision}$	RR_{Recall}	RR_{F1}
1	0.638 ± 0.054	0.518 ± 0.132	0.341 ± 0.007	0.277 ± 0.018
2	0.551 ± 0.066	0.473 ± 0.118	0.372 ± 0.018	0.315 ± 0.045
3	0.490 ± 0.069	0.432 ± 0.113	0.386 ± 0.023	0.330 ± 0.059
5	0.435 ± 0.051	0.406 ± 0.115	0.430 ± 0.039	0.405 ± 0.095
10	0.451 ± 0.052	0.417 ± 0.108	0.399 ± 0.029	0.349 ± 0.067
Label	**$SLFN_{Accuracy}$**	**$SLFN_{Precision}$**	**$SLFN_{Recall}$**	**$SLFN_{F1}$**
1	0.641 ± 0.055	0.512 ± 0.027	0.351 ± 0.007	0.297 ± 0.024
2	0.565 ± 0.063	0.505 ± 0.020	0.410 ± 0.026	0.385 ± 0.054
3	0.504 ± 0.061	0.465 ± 0.032	0.421 ± 0.040	0.393 ± 0.073
5	0.457 ± 0.038	0.451 ± 0.029	0.449 ± 0.031	0.438 ± 0.046
10	0.461 ± 0.053	0.453 ± 0.036	0.420 ± 0.035	0.399 ± 0.053

that in the case of the linear classifier the suggested decoding methods did not offer any significant improvements, since the variability of the performance range is approximately 3%. On the other hand, our nonlinear classifier (i.e., SLFN) for the same projected time horizon (i.e., label 5) reacted more efficiently in the decoding process. SLFN achieves 33% for the *F*1 score for nonnormalized data, while the *Z*-score, min–max and decimal precision methods achieve 46%, 43%, and 43%, respectively. As a result, normalization improves the *F*1 score performance by almost 10%.

Normalization and model selection can also affect the predictability of mid-price movements over the projected time horizon. Very interesting results come to light if we try to compare the *F*1 performance over different time horizons. For instance, we can see that, regardless of the decoding method, the *F*1 score is always better for label 5 than 1, meaning that 'our models'

predictions are better further in the future. This result is significant, especially with unfiltered data and min–max and decimal precision normalizations, when $F1$ score is approximately 27%, in the case of the one-step prediction problem (label 1), and 43% in the case of the five-step problem (label 5).

Another aspect of the experimental results above stems from the pros and cons of linear and nonlinear classifiers. More specifically, the RR linear classifier performed better on the raw dataset and for the Z-score decoding method in terms of $F1$ when compared to the SLFN (i.e., nonlinear classifier). This is not the case for the last decoding methods (i.e., min–max and decimal precision), where our nonlinear classifier presents similar or better results than RR. An explanation for this $F1$ performance discrepancy is due to each of these methods' engineering has. The RR classifier tends to be very efficient in high-dimensional problems, and these types of problems are linearly separable, in most cases. Another reason that RR can perform better when compared to a nonlinear classifier is that RR can control the complexity by penalizing the bias, via cross-validation, using the ridge parameter. On the other hand, a nonlinear classifier is prone to overfitting, which means that in some cases it offers a better degree of freedom for class separation.

7 | CONCLUSION

This paper described a new benchmark dataset formed by the Nasdaq ITCH feed data for five stocks for 10 consecutive trading days. Data representations that were exploited by order flow features were made available. We formulated five classification tasks based on mid-price movement predictions for 1, 2, 3, 5, and 10 predicted horizons. Baseline performances of two regression models were also provided in order to facilitate future research in the field. Despite the data size, we achieved an average out-of-sample performance ($F1$) of approximately 46% for both methods. These very promising results show that machine learning can effectively predict mid-price movement.

Potential avenues of research that can benefit from exploiting the provided data include: (a) prediction of the stability of the market, which is very important for liquidity providers (market makers) to make the spread, as well as for traders to increase liquidity provision (when markets can be predicted to be stable); (b) prediction on market movements, which is important for expert systems used by speculative traders; (c) identification of order book spoofing—that is, situations where markets are manipulated by limit orders. Although there is no spoofing activity information available for the provided data, the exploitation of such a large corpus of data can be used in order to identify patterns in stock markets that can be further analyzed as normal or abnormal.

ACKNOWLEDGMENT

This work was supported by H2020 Project BigDataFinance MSCA-ITN-ETN 675044 (http://bigdatafinance.eu), Training for Big Data in Financial Research and Risk Management.

ORCID

Adamantios Ntakaris http://orcid.org/0000-0001-6949-5337

REFERENCES

Abernethy, J., & Kale, S. (2013). Adaptive market making via online learning, *Advances in Neural Information Processing Systems* (pp. 2058–2066). Cambridge, MA: MIT Press.

Almgren, R., & Lorenz, J. (2006). Bayesian adaptive trading with a daily cycle. *Journal of Trading*, *1*(4), 38–46.

Alvim, L. G., dos Santos, C. N., & Milidiu, R. L. (2010). Daily volume forecasting using high frequency predictors. In *Proceedings of the 10Th IASTED International Conference*, Acta Press, Calgary, Canada, Vol. 674, pp. 248.

Amaya, D., Filbien, J.-Y, Okou, C., & Roch, A. F. (2015). Distilling liquidity costs from limit order books. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2660226.

An, Y., & Chan, N. H. (2017). Short-term stock price prediction based on limit order book dynamics. *Journal of Forecasting*, *36*(5), 541–556.

Aramonte, S., Schindler, J. W., & Rosen, S. (2013). Assessing and combining financial conditions indexes. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2976840.

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224.

Bogoev, D., & Karam, A. (2016). An Empirical Detection of High Frequency Trading Strategies. (*Working Paper*). Durham, UK: Durham University.

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, *27*(8), 2267–2306.

Cao, C., Hansch, O., & Wang, X. (2009). The information content of an open limit-order book. *Journal of Futures Markets*, *29*(1), 16–41.

Carrion, A. (2013). Very fast money: High-frequency trading on the NASDAQ. *Journal of Financial Markets*, *16*(4), 680–711.

Cenesizoglu, T., Dionne, G., & Zhou, X. (2014). Effects of the limit order book on price dynamics. Retrieved from https://depot.erudit.org/bitstream/003996dd/1/CIRPEE14-26.pdf.

Chan, N. T., & Shelton, C. (2001). An electronic market-maker. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/7220/AIM-2001-005.pdf?sequence=2.

Chang, Y. L. (2015). Inferring Markov chain for modeling order book dynamics in high frequency environment. *International Journal of Machine Learning and Computing*, *5*(3), 247–251.

Christensen, H. L., & Woodmansey, R. (2013). Prediction of hidden liquidity in the limit order book of globex futures. *Journal of Trading*, *8*(3), 68–95.

Creamer, G. (2012). Model calibration and automated trading agent for euro futures. *Quantitative Finance*, *12*(4), 531–545.

De Winne, R., & D'hondt, C. (2007). Hide-and-seek in the market: placing and detecting hidden orders. *Review of Finance*, *11*(4), 663–692.

Detollenaere, B., & D'hondt, C. (2017). Identifying expensive trades by monitoring the limit order book. *Journal of Forecasting*, *36*(3), 273–290.

Dixon, M. (2016). High frequency market making with machine learning. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2868473.

Felker, T., Mazalov, V., & Watt, S. M. (2014). Distance-based high-frequency trading. *Procedia Computer Science*, *29*, 2055–2064.

Fletcher, T., Hussain, Z., & Shawe-Taylor, J. (2010). Multiple kernel learning on the limit order book. In *Proceedings of the First Workshop on Applications of Pattern Analysis*, Vol. *11*, pp. 167–174.

Galeshchuk, S. (2016). Neural networks performance in exchange rate prediction. *Neurocomputing*, *172*, 446–452.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., & Howison, S. D. (2013). Limit order books. *Quantitative Finance*, *13*(11), 1709–1742.

Hallgren, J., & Koski, T. (2016). Testing for causality in continuous time Bayesian network models of high-frequency data. arXiv preprint retrieved from https://arxiv.org/abs/1601.06651.

Han, J., Hong, J., Sutardja, N., & Wong, S. F. (2015). Machine Learning Techniques for Price Change Forecast Using the Limit Order Book Data. (*Working Paper*). Berkeley, CA: University of California, Berkeley.

Hasbrouck, J. (2009). Trading costs and returns for US equities: Estimating effective costs from daily data. *Journal of Finance*, *64*(3), 1445–1477.

Hasbrouck, J., & Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, *16*(4), 646–679.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *42*(2), 513–529.

Iosifidis, A., Tefas, A., & Pitas, I. (2017). Approximate kernel extreme learning machine for large scale data classification. *Neurocomputing*, *219*, 210–220.

Kalay, A., Sade, O., & Wohl, A. (2004). Measuring stock illiquidity: An investigation of the demand and supply schedules at the TASE. *Journal of Financial Economics*, *74*(3), 461–486.

Kalay, A., Wei, L., & Wohl, A. (2002). Continuous trading or call auctions: Revealed preferences of investors at the Tel Aviv stock exchange. *Journal of Finance*, *57*(1), 523–542.

Kearns, M., & Nevmyvaka, Y. (2013). Machine Learning for Market Microstructure and High Frequency Trading. In D. Easley, M. López De Prado, & M. O'Hara (Eds.), *High Frequency Trading: New Realities for Traders, Markets and Regulators*. London, UK: Risk Books.

Kercheval, A. N., & Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance*, *15*(8), 1315–1329.

Kim, A. J. (2001). Input/Output Hidden Markov Models for Modeling Stock Order Flows. (*Technical Report No. 1370*). Cambridge, MA: MITAI Laboratory.

Levendovszky, J., & Kia, F. (2012). Prediction based-high frequency trading on financial time series. *Periodica Polytechnica: Electrical Engineering and Computer Science*, *56*(1), 29–34.

Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., Min, H., & Deng, F. (2016). Empirical analysis: Stock market prediction via extreme learning machine. *Neural Computing and Applications*, *27*(1), 67–78.

Liu, J., & Park, S. (2015). Behind stock price movement: Supply and demand in market microstructure and market influence. *Journal of Trading*, *10*(3), 13–23.

Maglaras, C., Moallemi, C. C., & Zheng, H. (2015). Optimal execution in a limit order book and an associated microstructure market impact model. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2610808.

Majhi, R., Panda, G., & Sahoo, G. (2009). Development and performance evaluation of FLANN based model for forecasting of stock markets. *Expert Systems with Applications*, *36*(3), 6800–6808.

Malik, A., & Lon Ng, W. (2014). Intraday liquidity patterns in limit order books. *Studies in Economics and Finance*, *31*(1), 46–71.

Mankad, S., Michailidis, G., & Kirilenko, A. (2013). Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method. *Algorithmic Finance*, *2*(2), 151–165.

Næs, R., & Skjeltorp, J. A. (2006). Order book characteristics and the volume–volatility relation: Empirical evidence from a limit order market. *Journal of Financial Markets*, *9*(4), 408–432.

O'Hara, M., & Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics*, *100*(3), 459–474.

Pai, P.-F., & Lin, C.-S (2005). A hybrid Arima and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497–505.

Palguna, D., & Pollak, I. (2016). Mid-price prediction in a limit order book. *IEEE Journal of Selected Topics in Signal Processing*, *10*(6), 1083–1092.

Panayi, E., Peters, G. W., Danielsson, J., & Zigrand, J.-P. (2016). Designating market maker behaviour in limit order book markets. *Econometrics and Statistics*, *5*, 20–44.

Ranaldo, A. (2004). Order aggressiveness in limit order book markets. *Journal of Financial Markets*, *7*(1), 53–74.

Rehman, M., Khan, G. M., & Mahmud, S. A. (2014). Foreign currency exchange rates prediction using CGP and recurrent neural network. *IERI Procedia*, *10*, 239–244.

Sandoval, J., & Hernández, G. (2015). Computational visual analysis of the order book dynamics for creating high-frequency foreign exchange trading strategies. *Procedia Computer Science*, *51*, 1593–1602.

Seddon, J. J., & Currie, W. L. (2017). A model for unpacking big data analytics in high-frequency trading. *Journal of Business Research*, *70*, 300–307.

Sharang, A., & Rao, C. (2015). Using machine learning for medium frequency derivative portfolio trading. arXiv preprint retrieved from https://arxiv.org/abs/1512.06228

Siikanen, M., Kanniainen, J., & Luoma, A. (2017). What drives the sensitivity of limit order books to company announcement arrivals? *Economics Letters*, *159*, 65–68.

Siikanen, M., Kanniainen, J., & Valli, J. (2017). Limit order books and liquidity around scheduled and non-scheduled announcements: Empirical evidence from NASDAQ Nordic. *Finance Research Letters*, *21*, 264–271.

Sirignano, J. (2016). Deep learning for limit order books. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2710331.

Suwanpetai, P. (2016). Estimation of exchange rate models after news announcement. In *AP16Thai Conference 2016: Sixth Asia–Pacific Conference on Global Business, Economics, Finance and Social Sciences*.

Talebi, H., Hoang, W., & Gavrilova, M. L. (2014). Multi-scale foreign exchange rates ensemble for classification of trends in FOREX market. *Procedia Computer Science*, *29*, 2065–2075.

Vella, V., & Ng, W. L. (2016). Improving risk-adjusted performance in high frequency trading using interval type-2 fuzzy logic. *Expert Systems with Applications*, *55*, 70–86.

Yang, S., Paddrik, M., Hayes, R., Todd, A., Kirilenko, A., Beling, P., & Scherer, W. (2012). Behavior Based Learning in Identifying High Frequency Trading Strategies. In *2012 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, IEEE, Piscataway, NJ, pp. 1–8.

Yang, S. Y., Qiao, Q., Beling, P. A., Scherer, W. T., & Kirilenko, A. A. (2015). Gaussian process-based algorithmic trading strategy identification. *Quantitative Finance*, *15*(10), 1683–1703.

Yu, Y. (2006). The Limit Order Book Information and the Order Submission Strategy: a Model Explanation. In *2006 International Conference on Service Systems and Service Management*, IEEE, Piscataway, NJ, Vol. *1*, pp. 687–691.

Zhang, K., Kwok, J. T., & Parvin, B. (2009). Prototype Vector Machine for Large Scale Semi-Supervised Learning. In *Proceedings of the 26Th Annual International Conference on Machine Learning*, ACM, New York, NY, pp. 1233–1240.

Zheng, B., Moulines, E., & Abergel, F. (2012). Price jump prediction in limit order book. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2026454.

Adamantios Ntakaris is an ESR within the Marie Curie BigDataFinance project in the Dept. of Signal Processing at Tampere University of Technology. He received a B.Sc. in Mathematics in 2009 from the Aristotle University of Thessaloniki and an M.Sc. in Financial Modelling and Optimization in 2014 from the University of Edinburgh. In 2014 Adamantios completed an industrial placement at Standard Life Investments in Edinburgh. Before commencing his PhD, he worked as an Effective Interest Rate Analyst at CitiGroup investment bank in Edinburgh, and as a Maths Olympiad Coach in Thessaloniki.

Martin Magris is an Early Stage Researcher within the Marie Curie BigDataFinance training network in the Laboratory of Industrial and Information Management at Tampere University of Technology (Finland) since April 2016. He received a B.Sc. in Statistics and Mathematics in 2013 and a M.Sc in Statistical and Actuarial Sciences in 2015 from Universitá degli studi di Trieste, Italy. As a part of his master studies, Martin visited Aarhus university for seven months in 2014. In the years 2015-2016, before commencing his PhD, Martin worked as actuarial analyst for a non-life insurance company, specifically in the car-insurance pricing and in the development, profit-testing and pricing of multiple-peril non-life insurance products.

Juho Kanniainen is a Professor of Financial Engineering at the Tampere University of Technology, Finland. His research agenda is focused on quantitative finance with emphasis on big data problems. Dr. Kanniainen has published in many journals in Finance and Engineering, including Review of Finance, Journal of Banking and Finance, and Digital Signal Processing. He has been coordinating two international EU projects, BigDataFinance (www.bigdatafinance.eu) and HPCFinance (www.hpcfinance.eu).

Moncef Gabbouj is a Professor of Signal Processing at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Academy of Finland Professor during 2011-2015. He held several visiting professorships at different universities. Dr. Gabbouj is currently the TUT-Site Director of the NSF IUCRC funded Center for Visual and Decision Informatics. His research interests include Big Data analytics, multimedia content-based analysis, indexing and retrieval, artificial intelligence, machine learning, pattern recognition, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Alexandros Iosifidis is currently an Assistant Professor of Machine Learning and Computer Vision in the Department of Engineering, at Aarhus University, Denmark. He has held Postdoctoral Researcher positions in Tampere University of Technology, Finland and Aristotle University of Thessaloniki, Greece. He has participated in many R&D projects financed by EU, Greek, Finnish, and Danish funding agencies and companies. He has co-authored more than 120 papers in international journals and conferences proposing novel Machine Learning techniques and their application in a variety of problems.

How to cite this article: Ntakaris A, Magris M, Kanniainen J, Gabbouj M, Iosifidis A. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*. 2018;37:852–866. https://doi.org/10.1002/for.2543

PUBLICATION

II

Tensor representation in high-frequency financial data for price change prediction

Tran, D. T., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A.

2017 IEEE Symposium Series on Computational Intelligence (SSCI). ed. by 2017
DOI: 10.1109/SSCI.2017.8280812

Tensor Representation in High-Frequency Financial Data for Price Change Prediction

Dat Thanh Tran*, Martin Magris†, Juho Kanniainen†, Moncef Gabbouj* & Alexandros Iosifidis‡
*Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
†Laboratory of Industrial and Information Management, Tampere University of Technology, Tampere, Finland
‡Department of Engineering, Electrical & Computer Engineering, Aarhus University, Aarhus, Denmark
Email:{dat.tranthanh,martin.magris,juho.kanniainen,moncef.gabbouj}@tut.fi, alexandros.iosifidis@eng.au.dk

***Abstract*—Nowadays, with the availability of massive amount of trade data collected, the dynamics of the financial markets pose both a challenge and an opportunity for high frequency traders. In order to take advantage of the rapid, subtle movement of assets in High Frequency Trading (HFT), an automatic algorithm to analyze and detect patterns of price change based on transaction records must be available. The multichannel, time-series representation of financial data naturally suggests tensor-based learning algorithms. In this work, we investigate the effectiveness of two multilinear methods for the mid-price prediction problem against other existing methods. The experiments in a large scale dataset which contains more than 4 millions limit orders show that by utilizing tensor representation, multilinear models outperform vector-based approaches and other competing ones.**

I. Introduction

High Frequency Trading (HFT) is a form of automated trading that relies on the rapid, subtle changes of the markets to buy or sell assets. The main characteristic of HFT is high speed and short-term investment horizon. Different from a long-term investors, high frequency traders profit from a low margin of the price changes with large volume within a relatively short time. This requires the ability to observe the dynamics of the market to predict prospective changes and act accordingly. In quantitative analysis, mathematical models have been employed to simulate certain aspects of the financial market in order to make a prediction of the potential asset price, stock trends, etc. The performance of traditional mathematical models relies heavily on hand-crafted features. With recent advances in computational power, more and more machine learning models have been introduced to to predict financial market behaviours. Popular machine learning methods in HFT include regression analysis [1], [2], [3], [4], [5], multilayer feed forward network [6], [7], [8], convolutional neural network [9] , recurrent neural network [10], [11], [12].

With large volume of data and the erratic behaviours of the market, neural network-based solutions have been widely adopted to learn both the suitable representation of the data and the corresponding classifiers. This resolves the limitation in hand-crafted models. All kinds of deep architectures have been proposed, ranging from traditional multilayer feed-forward models [6], [7], [8] to Convolutional Neural Network (CNN) [9], Recurrent Neural Network (RNN) [10], [11], [12], Deep Belief Networks [13], [14], [15]. For example, in [9] a CNN with both 2D and 1D convolution masks was trained to predict stock price movements. On a similar benchmark HFT dataset, a RNN with Long Short-Term Memory Units (LSTM) [12] and or a Neural Bag-of-Features (N-BoF) [16] network generalizing the (discriminant) Bag-of-Feature model (BoF) [17] were proposed to perform the same prediction task.

Tensor representation offers a natural representation of the time-series data, where time corresponds to one of the tensor orders. Therefore, it is intuitive to investigate machine learning models that utilize tensor representations. In traditional vector-based models, the features are extracted from the time-series representation and form an input vector to the model. The pre-processing step to convert a tensor representation to a vector representation might lead to the loss of temporal information. That is, the learned classifiers might fail to capture the the interactions between spatio-temporal information due to vectorization. Because many neural network-based solutions, such as CNN or RNN, learn the data directly in the tensor form, this could explain why many neural network implementations outperform traditional vector-based models with hand-crafted features. With advances in mathematical tools and algorithms dealing with tensor input, many multilinear discriminant techniques as well as tensor regression have been proposed for image and video classification problems such as [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. However, there are few works investigating the performance of the tensor-based multilinear methods in financial problems [28]. Different from neural network methodology which requires heavy tuning of network topology and parameters, the beauty of tensor-based multilinear techniques is that the objective function is straightforward to interpret and very few parameters are required to tune the model. In this work, we propose to use two multilinear techniques based on the tensor representation of time-series financial data to predict the mid price movement based on information obtained from Limit Order Book (LOB) data. Specifically, the contribution of this paper is as follows

- We investigate the effectiveness of tensor-based discriminant techniques, particularly Multilinear Discriminant Analysis (MDA) in a large scale prediction problem

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

of mid-price movement with high-frequency limit order book data.
- We propose a simple regression classifier that operates on the tensor representation, utilizing both the current and past information of the stock limit order book to boost the performance of the vector-based regression technique. Based on the observation of the learning dynamics of the proposed algorithm, efficient scheme to select the best model's state is also discussed.

The rest of the paper is organized as follows. Section 2 reviews the mid-price movement prediction problem given the information collected from LOB as well as related methods that were proposed to tackle this problem. In Section 3, MDA and our proposed tensor regression scheme are presented. Section 4 shows the experimental analysis of the proposed methods compared with existing results on a large-scale dataset. Finally, conclusions are drawn in Section 5.

II. High Frequency Limit Order Data

In finance, a limit order placed with a bank or a brokerage is a type of trade order to buy or sell a set amount of assets with a specified price. There are two types of limit order: a buy limit order and a sell limit order. In a sell limit order (ask), a minimum sell price and the corresponding volume of assets are specified. For example, a sell limit order of 1000 shares with minimum prize of \$20 per share indicates that the investors wish to sell the share with maximum prize of \$20 only. Similarly, in a buy limit order (bid), a maximum buy price and its respective volume must be specified. The two types of limit orders consequently form two sides of the LOB, the bid and the ask side. LOB aggregates and sorts the order from both sides based on the given price. The best bid price $p_b^{(1)}(t)$ at the time instance t is defined as the highest available price that a buyer is willing pay per share. The best ask price $p_a^{(1)}(t)$ is in turn the lowest available price at a time instance t that a seller is willing to sell per share. The LOB is sorted so that best bid and ask price is on top of the book. The trading happens through a matching mechanism based on several conditions. Generally, when a bid price exceeds an ask price, i.e. $p_b^{(i)}(t) > p_a^{(j)}(t)$, the trading happens between the two investors. In addition to executions, the order can disappear from the order book by cancellations.

Given the availability of LOB data, several problems can be formulated, such as price trend prediction, order flow distribution estimation or detection of anomalous events that cause turbulence in price change. One of the popular tasks given the availability of LOB data is to predict the mid-price movements, i.e. to classify whether the mid-price increases, decreases or remains stable based on a set of measurements. The mid-price is a quantity defined as the mean between the best bid price and the best ask price at a given time, i.e.

$$p_t = \frac{p_a^{(1)}(t) + p_b^{(1)}(t)}{2} \tag{1}$$

which gives a good estimate of the price trend.

The LOB dataset [29] used in this paper, referred as FI-2010, was collected from 5 different Finnish stocks (Kesko, Outokumpu, Sampo, Rautaruukki and Wartsila) in 5 different industrial sectors. The collection period is from 1st of June to 14th of June 2010, producing order data of 10 working days. The provided data was extracted based on event inflow [30] which aggregates to approximately 4.5 million events. Each event contains information from the top 10 orders from each side of the LOB. Since each order consists of a price (bid or ask) and a corresponding volume, each order event is represented by a 40-dimensional vector. In [29], a 144-dimensional feature vector was extracted for every 10 events, leading to $453,975$ feature vector samples. For each feature vector, FI-2010 includes an associated label which indicates the movement of mid-price (increasing, decreasing, stationary) in the next 10 order events. In order to avoid the effect of different scales from each dimension, the data was standardized using z-score normalization

$$\mathbf{x}_{norm} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma_{\mathbf{x}}} \tag{2}$$

Given the large scale of FI-2010, many neural network solutions have been proposed to predict the prospective movement of the mid-price. In [9], a CNN that operates on the raw data was proposed. The network consists of 8 layers with an input layer of size 100×40, which contains 40-dimensional vector representation of 100 consecutive events. The hidden layers contain both 2D and 1D convolution layers as well as max pooling layer. In [12], a RNN architecture with LSTM units that also operates on a similar raw data representation was proposed with separate normalization schemes for order prices and volumes. Beside conventional deep architecture, a N-BoF classifier [16] was proposed for the problem of the mid-price prediction. The N-BoF network in [16] was trained on 15 consecutive 144-dimensional feature vectors which contain order information from 150 most recent order events and predicted the movements in the next $k = \{10, 50, 100\}$ order events.

It should be noted that all of the above mentioned neural network solutions utilized not only information from the current order events but also information from the recent past. We believe that the information of the recent order events plays a significant role in modelling the dynamics of the mid-price. The next section presents MDA classifier and our proposed regression model that take into account the contribution of past order information.

III. Tensor-based Multilinear Methods for Financial Data

Before introducing the classifiers to tackle mid-price prediction problem, we will start with notations and concepts used in multilinear algebra.

A. Multilinear Algebra Concepts

In this paper, we denote scalar values by either low-case or upper-case characters $(x, y, X, Y \dots)$, vectors by lower-case

bold-face characters $(\mathbf{x}, \mathbf{y}, \dots)$, matrices by upper-case bold-face characters $(\mathbf{A}, \mathbf{B}, \dots)$ and tensor as calligraphic capitals $(\mathcal{X}, \mathcal{Y}, \dots)$. A tensor with K modes and dimension I_k in the mode-k is represented as $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$. The entry in the i_kth index in mode-k for $k = 1, \dots, K$ is denoted as $\mathcal{X}_{i_1, i_2, \dots, i_K}$.

Definition 1 (Mode-k Fiber and Mode-k Unfolding): The mode-k fiber of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ is a vector of I_k-dimensional, given by fixing every index but i_k. The mode-k unfolding of $\mathcal{X}$, also known as mode-k matricization, transforms the tensor $\mathcal{X}$ to matrix $\mathbf{X}_{(k)}$, which is formed by arranging the mode-k fibers as columns. The shape of $\mathbf{X}_{(k)}$ is $\mathbb{R}^{I_k \times I_{\bar{k}}}$ with $I_{\bar{k}} = \prod_{i=1, i \neq k}^{K} I_i$.

Definition 2 (Mode-k Product): The mode-k product between a tensor $\mathcal{X} = [x_{i_1}, \dots, x_{i_K}] \in \mathbb{R}^{I_1 \times \dots I_K}$ and a matrix $\mathbf{W} \in \mathbb{R}^{J_k \times I_k}$ is another tensor of size $I_1 \times \cdots \times J_k \times \cdots \times I_K$ and denoted by $\mathcal{X} \times_k \mathbf{W}$. The element of $\mathcal{X} \times_k \mathbf{W}$ is defined as $[\mathcal{X} \times_k \mathbf{W}]_{i_1, \dots, i_{k-1}, j_k, i_{k+1}, \dots, i_K} = \sum_{i_k=1}^{I_K} [\mathcal{X}]_{i_1, \dots, i_{k-1}, i_k, \dots, i_K} [\mathbf{W}]_{j_k, i_k}$.

With the definition of mode-k product and mode-k unfolding, the following equation holds

$$(\mathcal{X} \times_k \mathbf{W}^T)_{(k)} = \mathbf{W}^T \mathbf{X}_{(k)} \tag{3}$$

For convenience, we denote $\mathcal{X} \times_1 \mathbf{W}_1 \times \cdots \times_K \mathbf{W}_K$ by $\mathcal{X} \prod_{k=1}^{K} \times_k \mathbf{W}_k$.

B. Multilinear Discriminant Analysis

MDA is the extended version of the Linear Discriminant Analysis (LDA) which utilizes the Fisher criterion [31] as the optimal criterion of the learnt subspace. Instead of seeking an optimal vector subspace, MDA learns a tensor subspace in which data from different classes are separated by maximizing the interclass distances and minimizing the intraclass distances. The objective function is thus maximizing the ratio between interclass distances and intraclass distances in the projected space. Formally, let us denote the set of N tensor samples as $\mathcal{X}_1, \dots, \mathcal{X}_N \in \mathbb{R}^{I_1 \times \cdots \times I_K}, i = 1, \dots, N$, each with an associated class label $c_i, i = 1, \dots, C$. In addition, $\mathcal{X}_{i,j}$ denotes the jth sample from class c_i and n_i denotes the number of samples in class c_i. The mean tensor of class c_i is calculated as $\mathcal{M}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{X}_{i,j}$ and the total mean tensor is $\mathcal{M} = \frac{1}{N} \sum_{i}^{C} \sum_{j=1}^{n_i} \mathcal{X}_{i,j} = \frac{1}{N} \sum_{i=1}^{C} n_i \mathcal{M}_i$.

MDA seeks a set of projection matrices $\mathbf{W}_k \in \mathbb{R}^{I_k \times I_k^{'}}, I_k^{'} < I_k, k = 1, \dots, K$ that map $\mathcal{X}_{i,j}$ to $\mathcal{Y}_{i,j} \in \mathbb{R}^{I_1^{'} \times \cdots \times I_K^{'}}$, with the subspace projection defined as

$$\mathcal{Y}_{i,j} = \mathcal{X}_{i,j} \prod_{k=1}^{K} \times_k \mathbf{W}_k^T \tag{4}$$

The set of optimal projection matrices are obtained by maximizing the ratio between interclass and intraclass distances, measured in the tensor subspace $\mathbb{R}^{I_1^{'} \times \cdots \times I_K^{'}}$. Particularly, MDA maximizes the following criterion

$$J(\mathbf{W}_1, \dots, \mathbf{W}_K) = \frac{D_b}{D_w} \tag{5}$$

where

$$D_b = \sum_{i=1}^{C} n_i \| \mathcal{M}_i \prod_{k=1}^{K} \times_k \mathbf{W}_k - \mathcal{M} \prod_{k=1}^{K} \times_k \mathbf{W}_k \|_F^2 \tag{6}$$

and

$$D_w = \sum_{i=1}^{C} \sum_{j=1}^{n_i} \| \mathcal{X}_{i,j} \prod_{k=1}^{K} \times_k \mathbf{W}_k - \mathcal{M}_i \prod_{k=1}^{K} \times_k \mathbf{W}_k \|_F^2 \tag{7}$$

are respectively interclass distance and intraclass distance. The subscript F in (6) and (7) denotes the Frobenius norm. D_b measures the total square distances between each class mean $\mathcal{M}_i$ and the global mean $\mathcal{M}$ after the projection while D_w measures the total square distances between each sample and its respective mean tensor. By maximizing (5), we are seeking a tensor subspace in which the dispersion of data in the same class is minimum while the dispersion between each class is maximum. Subsequently, the classification can then be performed by simply selecting the class with minimum distance between a test sample to each class mean in the discriminant subspace. Since the projection in (4) exposes a dependancy between each mode-k, each $\mathbf{W}_k$ cannot be optimized independently. An iterative approach is usually employed to solve the optimization in (5) [[27], [26], [32]. In this work, we propose to use the CMDA algorithm [32] that assumes orthogonal constraints on each projection matrix $\mathbf{W}_k^T \mathbf{W}_k = \mathbf{I}, k = 1, \dots, K$ and solves (5) by iteratively solving a trace ratio problem for each mode-k. Specifically, D_b and D_w can be calculated by unfolding the tensors in mode-k as follows

$$\begin{aligned} D_b = & tr\Big(\sum_{i=1}^{C} n_i \Big[(\mathcal{M}_i - \mathcal{M}) \prod_{p=1}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)} \\ & \Big[(\mathcal{M}_i - \mathcal{M}) \prod_{p=1}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)}^T \Big) \end{aligned} \tag{8}$$

and

$$\begin{aligned} D_w = & tr\Big(\sum_{i=1}^{C} \sum_{j=1}^{n_i} \Big[(\mathcal{X}_{i,j} - \mathcal{M}_i) \prod_{p=1}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)} \\ & \Big[(\mathcal{X}_{i,j} - \mathcal{M}_i) \prod_{p=1}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)}^T \Big) \end{aligned} \tag{9}$$

where $tr()$ in (8) and (9) denotes the trace operator. By utilizing the identity in (3), D_b and D_w are further expressed as

$$\begin{aligned} D_b &= tr\Big(\mathbf{W}_k^T \Big(\sum_{i=1}^{C} n_i \Big[(\mathcal{M}_i - \mathcal{M}) \prod_{p=1, p \neq k}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)} \\ &\quad \Big[(\mathcal{M}_i - \mathcal{M}) \prod_{p=1, p \neq k}^{K} \times_p \mathbf{W}_p^T \Big]_{(k)}^T \Big) \mathbf{W}_k \Big) \\ &= tr\big(\mathbf{W}_k^T \mathbf{S}_b^k \mathbf{W}_k \big) \end{aligned} \tag{10}$$

and

$$D_w = tr\Bigg(\mathbf{W}_k^T\Bigg(\sum_{i=1}^{C}\sum_{j=1}^{n_i}\Big[(\mathcal{X}_{i,j}-\mathcal{M}_i)\prod_{p=1,p\neq k}^{K}\times_p\mathbf{W}_p^T\Big]_{(k)}\Big[(\mathcal{X}_{i,j}-\mathcal{M}_i)\prod_{p=1,p\neq k}^{K}\times_p\mathbf{W}_p^T\Big]_{(k)}^T\Bigg)\mathbf{W}_k\Bigg) = tr\big(\mathbf{W}_k^T\mathbf{S}_w^k\mathbf{W}_k\big) \tag{11}$$

where $\mathbf{S}_b^k$ and $\mathbf{S}_w^k$ in (10) and (11) denote the interclass and intraclass scatter matrices in mode-k. The criterion in (5) can then be converted to a trace ratio problem with respect to $\mathbf{W}_k$ while keeping other projection matrices fixed as

$$\mathbf{J}\big(\mathbf{W}_k\big) = \frac{tr\big(\mathbf{W}_k^T\mathbf{S}_b^k\mathbf{W}_k\big)}{tr\big(\mathbf{W}_k^T\mathbf{S}_w^k\mathbf{W}_k\big)} \tag{12}$$

With the orthogonality constraint of $\mathbf{W}_k$, the solution of (12) is given by $I_k^{'}$ eigenvectors corresponding to $I_k^{'}$ largest eigenvalues of $(\mathbf{S}_w^k)^{-1}\mathbf{S}_b$. Usually, a positive λ is added to the diagonal of $\mathbf{S}_w^k$ as a regularization, which also enables stable computation in case $\mathbf{S}_w^k$ is not a full rank matrix. In the training phase, after randomly initializes $\mathbf{W}_k$, CMDA algorithm iteratively goes through each mode k, optimizes the Fisher ratio with respect to $\mathbf{W}_k$ while keeping other projection matrices fixed. The algorithm terminates when the changes in $\mathbf{W}_k$ below a threshold or the specified maximum iteration reached. In the test phase, the class with minimum distance between the class mean and the test sample in the tensor subspace is assigned to the test sample.

C. Weighted Multichannel Time-series Regression

For the FI-2010 dataset, in order to take into account the past information one could concatenate T 144-dimensional feature vectors corresponding to the $10T$ most recent order events to form a 2-mode tensor sample, i.e. a matrix $\mathcal{X}_i \in \mathbb{R}^{144\times T}, i = 1,\ldots,N$. For example, a training tensor sample of size 144×10 contains information of 100 most recent order events in the FI-2010 dataset. 10 columns represent information at 10 time-instances with the 10th column contains the latest order information. Each of the 144 rows encode the temporal evolution of the 144 features (or channels) through time. Generally, given N 2-mode tensor $\mathcal{X}_i \in \mathbb{R}^{D\times T}, i = 1,\ldots,N$ that belong to C classes indicated by the class label $c_i = 1,\ldots,C$, the proposed Weighted Multichannel Time-series Regression (WMTR) learns the following mapping function

$$f\big(\mathcal{X}_i\big) = \mathbf{W}_1^T\mathcal{X}_i\mathbf{w}_2 \tag{13}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D\times C}$ and $\mathbf{w}_2 \in \mathbb{R}^T$ are learnable parameters. The function f in (13) maps each input tensor to a C-dimensional (target) vector. One way to interpret f is that $\mathbf{W}_1$ maps D-dimensional representation of each time-instance to a C-dimensional (sub)space while $\mathbf{w}_2$ combines the contribution of each time-instance into a single single vector, by using a weighted average approach. In order to deal with unbalanced datasets, such as FI-2010, the parameters $\mathbf{W}_1$, $\mathbf{w}_2$ of the WMTR model are determined by minimizing the following weighted least square criterion

$$J\big(\mathbf{W}_1,\mathbf{w}_2\big) = \sum_{i=1}^{N} s_i\|\mathbf{W}_1^T\mathcal{X}_i\mathbf{w}_2 - \mathbf{y}_i\|_F^2 + \lambda_1\|\mathbf{W}_1\|_F^2 + \lambda_2\|\mathbf{w}_2\|_F^2 \tag{14}$$

where $\mathbf{y}_i \in \mathbb{R}^C$ is the corresponding target of the ith sample with all elements equal to -1 except the c_ith element, which is set equal to 1. λ_1 and λ_2 are predefined regularization parameters associated with $\mathbf{W}_1$ and $\mathbf{w}_2$. We set the value of the predefined weight s_i equal to $1/\sqrt[r]{N_{c_i}}, r > 0$, i.e. inversely proportional to the number of training samples belonging to the class of sample i, so that errors in smaller classes contribute more to the loss. The weight of each class is controlled by parameter r: the smaller r, the more contribution of the minor classes in the loss. The unweighted least square criterion is a special case of (14) when $r \to +\infty$, i.e. $s_i = 1, \forall i$.

We solve (14) by applying an iterative optimization process that alternatively keeps one parameter fixed while optimizing the other. Specifically, by fixing $\mathbf{w}_2$ we have the following minimization problem

$$J_2\big(\mathbf{W}_1\big) = \|\big(\mathbf{W}_1^T\mathbf{X}_2 - \mathbf{Y}_2\big)\mathbf{S}_2\|_F^2 + \lambda_1\|\mathbf{W}_1\|_F^2 \tag{15}$$

where $\mathbf{X}_2 = \big[\mathcal{X}_1\mathbf{w}_2,\ldots,\mathcal{X}_N\mathbf{w}_2\big] \in \mathbb{R}^{D\times N}$, $\mathbf{Y}_2 = [\mathbf{y}_1,\ldots,\mathbf{y}_N] \in \mathbb{R}^{C\times N}$ and $\mathbf{S}_2 \in \mathbb{R}^{N\times N}$ is a diagonal matrix with the $\mathbf{S}_{2_{i,i}} = \sqrt{s_i}, i = 1,\ldots,N$. By solving $\frac{\partial J_2}{\partial \mathbf{W}_1} = 0$, we obtain the solution of (15) as

$$\mathbf{W}_1^* = \big(\mathbf{X}_2\mathbf{S}_2\mathbf{S}_2^T\mathbf{X}_2^T + \lambda_1\mathbf{I}\big)^{-1}\mathbf{X}_2\mathbf{S}_2\mathbf{S}_2^T\mathbf{Y}_2^T \tag{16}$$

where $\mathbf{I}$ is the identity matrix of the appropriate size.

Similarly, by fixing $\mathbf{W}_1$, we have the following regression problem with respect to $\mathbf{w}_2$

$$J_1(\mathbf{w}_2) = \|\mathbf{S}_1\big(\mathbf{X}_1\mathbf{w}_2 - \mathbf{Y}_1\big)\|_F^2 + \lambda_2\|\mathbf{w}_2\|_F^2 \tag{17}$$

where $\mathbf{X}_1 = \big[\mathcal{X}_1^T\mathbf{W}_1,\ldots,\mathcal{X}_N^T\mathbf{W}_1\big]^T \in \mathbb{R}^{CN\times T}$, $\mathbf{Y}_{(1)} = [\mathbf{y}_1^T,\ldots,\mathbf{y}_N^T]^T \in \mathbb{R}^{CN}$ and $\mathbf{S}_1 \in \mathbb{R}^{CN\times CN}$ is a diagonal matrix with $\mathbf{S}_{1_{C(i-1)+k,C(i-1)+k}} = \sqrt{s_i}; k = 1,\ldots,C; i = 1,\ldots,N$. Similar to $\mathbf{W}_1$, optimal $\mathbf{w}_2$ is obtained by solving for the stationary point of (17), which is given as

$$\mathbf{w}_2^* = \big(\mathbf{X}_1^T\mathbf{S}_1^T\mathbf{S}_1\mathbf{X}_1 + \lambda_2\mathbf{I}\big)^{-1}\mathbf{X}_1^T\mathbf{S}_1^T\mathbf{S}_1\mathbf{Y}_1 \tag{18}$$

The above process is formed by two convex problems, for which each processing step obtains the global optimum solution. Thus, the overall process is guaranteed to reach a local optimum for the combined regression criterion. The algorithm terminates when the changes in $\mathbf{W}_1$ and $\mathbf{w}_2$ are below a threshold or the maximum number of iterations is reached. In the test phase, f in (13) maps a test sample to the feature space and the class label is inferred by the index of the maximum element of the projected test sample.

Usually, multilinear methods (including multilinear regression ones) are are randomly initialized. This means that, in our case, one would randomly initialize the parameters in

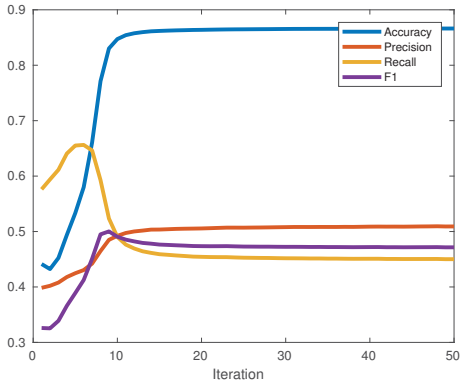


Fig. 1. Performance measure of WMTR on training data

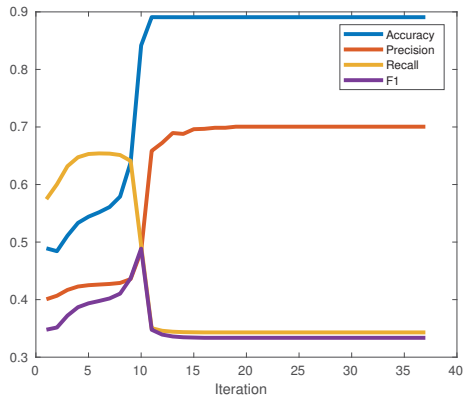


Fig. 2. Performance measure of MTR on training data

$\mathbf{w}_2$ in order to define the optimal regression values stored in $\mathbf{W}_1$ on the first iteration. However, since for WMTR when applied to LOB data, the values of $\mathbf{w}_2$ encode the contribution of each time-instance in the overall regression, we chose to initialize it as $\mathbf{w}_2 = [0\, 0 \ldots 1]^T$. That is, the first iteration of WMTR corresponds to the vector-based regression using only the representation for the current time-instance. After obtaining this mapping, the optimal weighted average of all time-instances is determined by solving for $\mathbf{w}_2$.

IV. Experiments

A. Experiment Setting

We conducted extensive experiments on the FI-2010 dataset to compare the performance of the multilinear methods, i.e. MDA and and the proposed WMTR, with that of the other existing methods including LDA, Ridge Regression (RR), Single-hidden Layer Feed Forward Network (SLFN), BoF and N-BoF. In addition, we also compared WMTR with its unweighted version, denoted by MTR, to illustrate the effect of weighting in the learnt function. Regarding the train/test evaluation protocol, we followed the anchored forward cross-validation splits provided by the database [29]. Specifically, there are 9 folds for cross-validation based on the day basis; the training set increases by one day for each consecutive fold and the day following the last day used for training is used for testing, i.e. for the first fold, data from the first day is used for training and data in the second day is used for testing; for the second fold, data in the first and second day is used as for training and data in the third day used for testing; for the last fold, data in the first 9 days is used for training and the 10th day is used for testing.

Regarding the input representation of the proposed multilinear techniques, MDA and WMTR both accept input tensor of size $\mathbb{R}^{144\times 10}$, which contains information from 100 consecutive order events with the last column contains information from the last 10 order events. For LDA, RR and SLFN, each input vector is of size $\mathbb{R}^{144}$, which is the last column of the input from MDA and WMTR, representing the most current information of the stock. The label of both tensor input and vector input is the movement of the mid-price in the next 10 order events, representing the future movement that we would like to predict. Since we followed the same experimental protocol as in [29] and [16], we directly report the result of RR, SLFN, BoF, N-BoF in this paper.

The parameter settings of each model are as follows. For WMTR, we set maximum number of iterations to 50, the terminating threshold to $1e{-}6$; $\lambda_1, \lambda_2 \in \{0.01, 0.1, 1, 10, 100\}$ and $s_i = n_{c_i}^{-1/r}$ with $r \in \{2, 3, 4\}$. For MTR, all paramter settings were similar to WMTR except $s_i = 1$, $\forall i$. For MDA, the number of maximum iterations and terminating threshold were set similar to WMTR, the projected dimensions of the first mode is from 5 to 60 with a step of 5 while for the second mode from 1 to 8 with a step of 1. In addition, a regularization amount $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ was added to the diagonal of $\mathbf{S}_w^k$.

B. Performance Evaluation

It should be noted that FI-2010 is a highly unbalanced dataset with most samples having a stationary mid-price.

TABLE I
Performance on FI-2010

	Accuracy	Precision	Recall	F1
RR	46.00 ± 2.85	43.30 ± 9.9	43.54 ± 5.2	42.52 ± 1.22
SLFN	53.22 ± 7.04	49.60 ± 3.81	41.28 ± 4.04	38.24 ± 5.66
LDA	63.82 ± 4.98	37.93 ± 6.00	45.80 ± 4.07	36.28 ± 1.02
MDA	71.92 ± 5.46	44.21 ± 1.35	60.07 ± 2.10	46.06 ± 2.22
MTR	86.08 ± 4.99	51.68 ± 7.54	40.81 ± 6.18	40.14 ± 5.26
WMTR	81.89 ± 3.65	46.25 ± 1.90	51.29 ± 1.88	$\mathbf{47.87 \pm 1.91}$
BoF	57.59 ± 7.34	39.26 ± 0.94	51.44 ± 2.53	36.28 ± 2.85
N-BoF	62.70 ± 6.73	42.28 ± 0.87	61.41 ± 3.68	41.63 ± 1.90

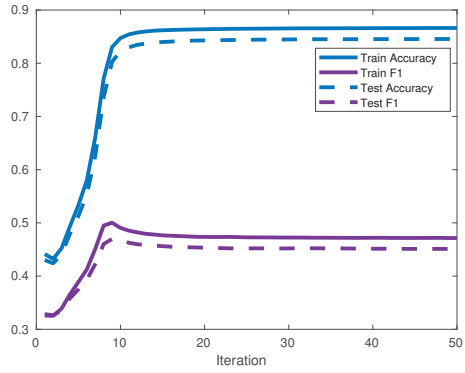


Fig. 3. Performance measure of WMTR on both train and test set

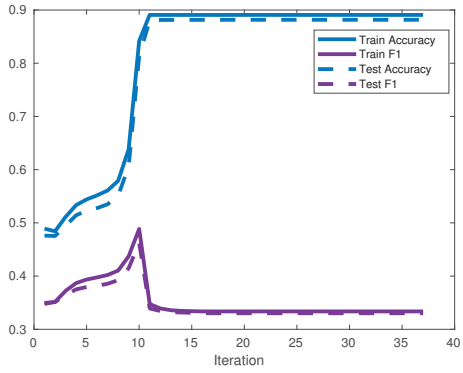


Fig. 4. Performance measure of MTR on both train and test set

Therefore we use average $f1$ score per class [33] as a performance measure to select model parameters since $f1$ expresses a trade-off between precision and recall. More specifically, for each cross-validation fold, the competing methods are trained with all combinations of the above mentioned parameter settings on the training data. We selected the learnt model that achieved the highest $f1$ score on the training set and reported the performance on the test set. In addition to $f1$, the corresponding average precision per class, average recall per class and accuracy are also reported. Accuracy measures the percentage the predicted labels that match the ground truth. Precision is the ratio between true positives over the number of samples predicted as positive and recall is the ratio between true positive over the total number of true positives and false negatives. $f1$ is the harmonic mean between precision and recall. For all measures, higher values indicate better performance.

Table 1 shows the average performance with standard deviation over all 9 folds of the competing methods. Comparing two discriminant methods, i.e. LDA and MDA, it is clear that MDA significantly outperforms LDA on all performance measures. This is due to the fact that MDA operates on the tensor input, which could hold both current and past information as well as the temporal structure of the data. The improvement of tensor-based approaches over vector-based approach is consistent also in case of regression (WMTR vs RR). Comparing multilinear techniques with N-BoF, MDA and WMTR perform much better than N-BoF in terms of $f1$, accuracy and precision while recall scores nearly match. WMTR outperforming MTR in large margin suggests that weighting is important for the highly unbalanced dataset such as FI-2010. Overall, MDA and WMTR are the leading methods among the competing methods in this mid-price prediction problem.

C. WMTR analysis

Figure 1 shows the dynamic of the learning process of WMTR on the training data of the first fold. There is one interesting phenomenon that can be observed from the training process. In the first 10 iterations, all performance measures improve consistently. After the 10th iteration, $f1$ score drops a little then remains stable while accuracy continues to improve. This phenomenon can be observed at every parameter setting. Since WMTR minimizes the squared error between the target label and the predicted label, constant improvement before converging observed from the training accuracy is expected. The drop in $f1$ score after some k iterations can be explained as follows: in the first k iterations, WMTR truly learns the generating process behind the training samples; however, at a certain point, WMTR starts to overfit the data and becomes bias towards the dominant class. The same phenomenon was observed from MTR with more significant drop in $f1$ since without weight MTR overfits the dominant class severely. Figure 2 shows the training dynamic of MTR with similar parameter setting except the class weight in the loss function. Due to this behaviour, in order to select the best learnt state of WMTR and MTR, we measured $f1$ score on the training data at each iteration and selected the model's state which produced the best $f1$. The question is whether the selected model performs well on the test data? Figure 3 and Figure 4 plots accuracy and $f1$ of WMTR and MTR measured on the training set and the test set at each iteration. It is clear that the learnt model that produced best $f1$ during training also performed best on the test data. The margin between training and testing performance is relatively small for both WMTR and MTR which shows that our proposed algorithm did not suffer from overfitting. Although the behaviours of WMTR and MTR are the similar, the best model learnt from MTR is biased towards the dominant class, resulting in inferior performance as shown in the experimental result.

V. Conclusions

In this work we have investigated the effectiveness of multilinear discriminant analysis in dealing with financial prediction based on Limit Order Book data. In addition, we proposed a

simple bilinear regression algorithm that utilizes both current and past information of a stock to boost the performance of traditional vector-based regression. Experimental results showed that the proposed methods outperform their counterpart exploiting vectorial representations, and outperform existing solutions utilizing (possibly deep) neural network architectures.

VI. Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 675044 BigDataFinance.

References

[1] B. Zheng, E. Moulines, and F. Abergel, "Price jump prediction in limit order book," 2012.

[2] L. G. Alvim, C. N. dos Santos, and R. L. Milidiu, "Daily volume forecasting using high frequency predictors," in *Proceedings of the 10th IASTED International Conference*, vol. 674, p. 248, 2010.

[3] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[4] B. Detollenaere and C. D'hondt, "Identifying expensive trades by monitoring the limit order book," *Journal of Forecasting*, vol. 36, no. 3, pp. 273–290, 2017.

[5] E. Panayi, G. W. Peters, J. Danielsson, and J.-P. Zigrand, "Designating market maker behaviour in limit order book markets," *Econometrics and Statistics*, 2016.

[6] J. Levendovszky and F. Kia, "Prediction based-high frequency trading on financial time series," *Periodica Polytechnica. Electrical Engineering and Computer Science*, vol. 56, no. 1, p. 29, 2012.

[7] J. Sirignano, "Deep learning for limit order books," 2016.

[8] S. Galeshchuk, "Neural networks performance in exchange rate prediction," *Neurocomputing*, vol. 172, pp. 446–452, 2016.

[9] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting stock prices from the limit order book using convolutional neural networks," in *IEEE Conference on Business Informatics (CBI), Thessaloniki, Greece*, 2017.

[10] M. Dixon, "High frequency market making with machine learning," 2016.

[11] M. Rehman, G. M. Khan, and S. A. Mahmud, "Foreign currency exchange rates prediction using cgp and recurrent neural network," *IERI Procedia*, vol. 10, pp. 239–244, 2014.

[12] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Using deep learning to detect price change indications in financial markets," in *European Signal Processing Conference (EUSIPCO), Kos, Greece*, 2017.

[13] A. Sharang and C. Rao, "Using machine learning for medium frequency derivative portfolio trading," *arXiv preprint arXiv:1512.06228*, 2015.

[14] J. Hallgren and T. Koski, "Testing for causality in continuous time bayesian network models of high-frequency data," *arXiv preprint arXiv:1601.06651*, 2016.

[15] J. Sandoval and G. Hernández, "Computational visual analysis of the order book dynamics for creating high-frequency foreign exchange trading strategies," *Procedia Computer Science*, vol. 51, pp. 1593–1602, 2015.

[16] N. Passalis, A. Tsantekidis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Time-series classification using neural bag-of-features," in *European Signal Processing Conference (EUSIPCO), Kos, Greece*, 2017.

[17] A. Iosifidis, A. Tefas, and I. Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 185–192, 2014.

[18] A. Shashua and A. Levin, "Linear image coding for regression and classification using the tensor-rank principle," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–I, IEEE, 2001.

[19] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[20] K. Liu, Y.-Q. Cheng, and J.-Y. Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion," *Pattern Recognition*, vol. 26, no. 6, pp. 903–911, 1993.

[21] H. Kong, E. K. Teoh, J. G. Wang, and R. Venkateswarlu, "Two-dimensional fisher discriminant analysis: forget about small sample size problem [face recognition applications]," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2, pp. ii–761, IEEE, 2005.

[22] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in neural information processing systems*, pp. 1569–1576, 2005.

[23] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Advances in neural information processing systems*, pp. 499–506, 2006.

[24] D. Cai, X. He, and J. Han, "Subspace learning based on tensor analysis," tech. rep., 2005.

[25] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–93, IEEE, 2003.

[26] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Discriminant analysis with tensor representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 526–532, IEEE, 2005.

[27] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, 2007.

[28] Q. Li, Y. Chen, L. L. Jiang, P. Li, and H. Chen, "A tensor-based information framework for predicting the stock market," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 2, p. 11, 2016.

[29] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price prediction of limit order book data," *arXiv preprint arXiv:1705.03233*, 2017.

[30] X. Li, H. Xie, R. Wang, Y. Cai, J. Cao, F. Wang, H. Min, and X. Deng, "Empirical analysis: stock market prediction via extreme learning machine," *Neural Computing and Applications*, vol. 27, no. 1, pp. 67–78, 2016.

[31] M. Welling, "Fisher linear discriminant analysis," *Department of Computer Science, University of Toronto*, vol. 3, no. 1, 2005.

[32] Q. Li and D. Schonfeld, "Multilinear discriminant analysis for higher-order tensor data classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2524–2537, 2014.

[33] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

PUBLICATION III

Long-range auto-correlations in limit order book markets: Inter-and cross-event analysis

Magris, M., Kim, J., Räsänen, E. and Kanniainen, J.

2017 IEEE Symposium Series on Computational Intelligence (SSCI). ed. by 2017
DOI: 10.1109/SSCI.2017.8280932

Long-range Auto-correlations in Limit Order Book Markets: Inter- and Cross-event Analysis

Martin Magris*,[1], Jiyeong Kim†, Esa Räsänen†, Juho Kanniainen*
*Laboratory of Industrial and Information Management, Tampere University of Technology, Tampere, Finland
†Laboratory of Physics, Tampere University of Technology, Tampere, Finland
[1]E-mail: martin.magris@tut.fi

***Abstract*—Long-range correlation in financial time series reflects the complex dynamics of the stock markets driven by algorithms and human decisions. Our analysis exploits ultra-high frequency order book data from NASDAQ Nordic over a period of three years to numerically estimate the power-law scaling exponents using detrended fluctuation analysis (DFA). We address inter-event durations (order to order, trade to trade, cancel to cancel) as well as cross-event durations (time from order submission to its trade or cancel). We find strong evidence of long-range correlation, which is consistent across different stocks and variables. However, given the crossovers in the DFA fluctuation functions, our results indicate that the long-range correlation in inter-event durations becomes stronger over a longer time scale, i.e., when moving from a range of hours to days and further to months. We also observe interesting associations between the scaling exponent and a number of economic variables, in particular, in the inter-trade time series.**

I. Introduction

Many natural and economic time series exhibit long-range power decaying correlations. The dynamics of the times series depicting a complex system is often characterized by *scaling laws*, over a continuous range of *time scales* and frequencies [1]. A system characterized by self-similar structures over different time-scales is called a *fractal*. Financial and economic systems are highly complex and stochastic, characterized by numerous degrees of freedom and highly susceptible to exogenous factors. This complexity emerges as time-dependent properties (e.g., *trends*), or more generally, non-stationariety in the time-series. A reliable method for detecting long-range correlations, known as *deterended fluctuation analysis* (DFA), which is robust for non-stationariety, has been developed in [2]. Because of its simplicity and wide applicability, DFA has been extensively used in the analyses of the long-range correlations in natural, social and economic data (see section II-B). However, in the existing literature DFA with intra-day high-frequency financial data has not been extensive. The closest works related to our research are, among others, [3], [4], [5], which studied the long memory and multifractal nature of inter-trade durations, and [6], which analyzed the long-range properties of inter-cancel durations and characterized their distribution. Set apart from previous studies, we do not limit our analysis to one type of order book events. We compare the correlation properties in each type of order book event, i.e., order, trade or cancellation, and make cross-analyses between different events (e.g. order submission and its cancellation) as well.

In this paper, we study the fractal properties of order flow data over different time scales, providing new insights into understanding the complex dynamics of the order book, also in relation to selected economic variables. In order to do so, we utilize the ultra-high frequency order book data, with five securities traded in NASDAQ Nordic over three years. We apply DFA on inter-event and cross-event time intervals of all message types (order submissions, transactions, and cancellations); this has not been done in the extant literature, even though message types are likely be interconnected. Our main finding is that fractal properties are ubiquitous in the time series of the order book, but of complex nature. The scaling properties show crossovers and peculiar relationships with a number of variables of economic interest (such as mean inter-event duration, daily return and volatility).

II. Method and data

A. Detrended fluctuation analysis

Due to the non-stationary nature of financial time series, conventional methods, e.g., Hurst's rescaled range analysis [7], can lead to a false detection of long-range correlations (see e.g., [8]), as they assume stationary time series. Detrended fluctuation analysis (DFA), originally introduced in [2], incorporates a detrending scheme in the fluctuation analysis, making the algorithm robust to non-stationarities like trends. Therefore, DFA has been established as a reliable method to analyze long-range correlation in financial time series. We provide a brief description of DFA as it is our main analysis method. A detailed and thorough description of the algorithm can be found in [9].

For a time series of length N, whose observations are $\{x_t\}_{t=1,\dots,N}$, the DFA procedure can be summarized in four steps:

i. We define the profile of the time series by taking an integrated sum of the series:

$$y(k) = \sum_{t=1}^{k} (x_t - \langle x \rangle)$$

The subtraction by the mean $\langle x \rangle$ sets the global mean to zero; however, it is not obligatory.

ii. The profile is divided into N/s non-overlapping windows of equal length s. In each window, an n-degree polynomial approximation y_{tr}, representing local trend,

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

is computed by a least-squares fit. In our analysis we use 1st order DFA, in which a linear trend is eliminated from each window.

iii. We compute the variance of the residuals, or the detrended profile, $(y_m - y_{m,tr})$, $m = 1, ..., N/s$ for the m-th window and then average the variances. By taking the square root of the average variance, we obtain the DFA fluctuation F as a function of window size s, as we repeat the procedure for all the window sizes:

$$F(s) = \sqrt{\frac{1}{N/s} \sum_{m=1}^{N/s} \left[\frac{1}{s} \sum_{i=1}^{s} [y_m(i) - y_{tr,m}(i)]^2 \right]} \quad (1)$$

iv. Since we have $F(s) \sim s^{\alpha}$ in presence of power-law scaling, we plot $F(s)$ against s in log-log scale and calculate the slope of the linear fit in the log-log plot to obtain the *scaling exponent* α. Note that time series may require more than one scaling exponent to describe different correlation behaviors at different time scales. This "crossover" can be detected as change in slope in the log-log plot of $F(s)$ against s.

The scaling exponent α describes the nature of the correlation present in the data. White noise (uncorrelated signal) and Brownian noise are characterized by $\alpha = 0.5$ and $\alpha = 1.5$, respectively. Values $0.5 < \alpha < 1.5$ indicate long-term correlations, i.e., fractality, with $\alpha = 1$ corresponding to perfect $1/f$ fractal behavior (pink noise). Values $\alpha < 0.5$ correspond to anticorrelations [10], [1].

B. DFA in financial literature

The nature of a power-law describing the long-range correlation implies self-similar patterns in the time series over a long period, which is of particular interest in economic and financial problems. The existence of long memory behavior in asset returns was first considered in [11], from which a rich literature on the fractal properties of financial time series followed. DFA analysis in finance is indeed ubiquitous and applied to various time series; we provide a number of examples. In the earliest studies, [12] showed that the scaling in the distribution of the S&P 500 index can be described by a non-gaussian process and that the scaling exponent is constant over the period of six years analyzed. The first use of DFA on the evolution of currency rate exchange was proposed in [13], where the authors found a close association between the scaling exponent and economic events, economic policies, and the information propagation among economic systems. These analyses have been expanded to a wider number of exchange rates in [14]. The use of the long-range correlation analysis to identify different phases in the evolution of financial markets and predict forthcoming changes (e.g. financial crashes) is discussed in, e.g., [15], [16]. Examples of long-range correlation analyses applied to financial returns can be found in, among the others, [17], [18], [19], [20]. Among the latest studies, [21] compared the results implied by DFA with those from standard time series models, while [22] provides a very interesting application of DFA to investigate market efficiency of Dow Jones ETF.

Previous studies related to the present work include [3], which analyzed the distribution and fractal behavior of the inter-trade durations over a four-years period for thirty stocks listed at NYSE. Time intervals between consecutive trades were also considered in [4] for stocks traded at NYSE and NASDAQ, finding that power-law correlations in inter-trade times are influenced by the market structure and coupled with the power-law correlations of absolute returns and volatility. Inter-trade times for ultra-high frequency order book data were analyzed in [23] and [5]; in the latter work, strong evidence of crossovers between two different power-scaling regimes from 23 stocks was found. Fractal properties of inter-cancellations duration have been analyzed for 18 stocks in Shenzhen exchange in [6], using three variants of the standard DFA.

The above mentioned studies, though dealing with a large number of stocks, consider a limited number of variables, e.g., inter-trade or inter-cancellation times only. Joint analyses of the fractal structure of several variables (as proposed in [6]) are definitely interesting and need to be expanded to a larger number of time series. The full order book data contain orders, trades, cancellations and cross-events for each day down to ultra-short time scales, thus allowing a wide perspective to the dynamics, especially when combined with standard economic variables (e.g. daily return or volatility).

C. Data

We processed the raw ITCH order book flow and reconstructed the order book for a selected number of securities; for a detailed description of the procedure, we refer to [24]. The ITCH flow contains the full information about all the events related to a particular security; therefore, our order book data is complete and has accurate timestamps with millisecond precision. For our analysis, we have selected five stocks from NASDAQ Nordic (see Table I), based on their liquidity, which directly determines the number of events in the order book. More information on NASDAQ Nordic order book market can be found, e.g., in [25], [26]. We consider 752 trading days ranging from June 1, 2010 to May 31, 2013. In order to avoid bias in the data, non-regular trading hours are excluded as well as the events occurred in the first and last 30 minutes of the trading day, thus considering the time window spanning from 7:30 a.m. to 3:00 p.m. (2:30 p.m. for the stock traded in Copenhagen). The variables we use in our analysis are listed in Table II. We also differentiate the variables based on the side of the order book, while focusing on the events occurring at the best levels only.

III. Results and Discussion

A. Estimation of the power-law scaling exponent α

Using the DFA method described in section II-A, we compute the power-law scaling exponent α for each variable (Table II) that is recorded on a trading day. Fig. 1 shows a sample series of α values over 752 trading days. The mean

TABLE I: List of five stocks of our selection at NASDAQ Nordic. The average number of records in the order book for each variable is computed over 752 trading days and over bid and ask sides.

ID	ISIN code	Company	Exchange	Avg. number of records				
				or-or	tr-tr	ca-ca	or-tr	or-ca
DK_1	DK0010268606	Vestas Wind Systems	Copenhagen	4295	1730	2943	1020	3276
FI_1	FI0009005318	Nokian Renkaat Oyj	Helsinki	6405	1462	4812	903	5503
FI_2	FI0009007835	Metso Oyj	Helsinki	7480	1698	5628	1029	6452
SE_1	SE0000101032	Atlas Copco A	Stockholm	13031	2499	12330	1530	11502
SE_2	SE0000115446	Volvo B	Stockholm	15851	3883	14327	2406	13446

TABLE II: Time-variables and their description. For the cancel-cancel durations, only consecutive cancellations *occurring at the best level* are taken into account.

Variable	Description
or-or	Time to the next order (inter-order duration)
tr-tr	Time to the next trade (inter-trade duration)
ca-ca	Time to the next cancellation (inter-cancel duration)
or-tr	Lifetime of orders that led to a trade (time from order submission to its trade)
or-ca	Lifetime of orders that have been canceled (time from order submission to its cancellation)

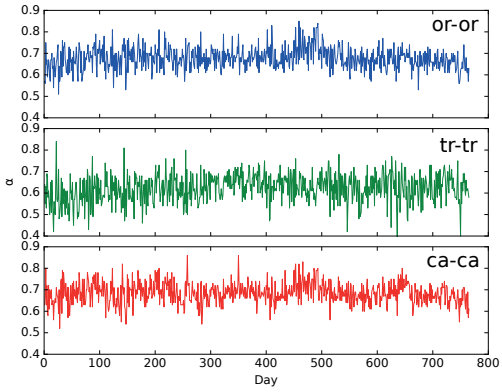


Fig. 1: Series of α values computed for inter-events durations for the bid side of stock FI_1.

and standard deviation of the α values over all the available days are found in Table III.

Across the multiple stocks we analyzed, the inter-event (or-or, tr-tr, and ca-ca) durations have the mean α values around 0.65, which are well above the threshold of random noise ($\alpha \approx 0.5$), thus indicating the presence of (weak) long-range correlations in the waiting times within a day. The cross-events variables (or-tr and or-ca) also yield similar results, but with clearly larger day-to-day variance around the mean α, compared to that of inter-event variables.

The remarkable consistency of the results between the ask and bid sides suggests a symmetry of the correlation properties between the sides. This may imply that the way trading algorithms are implemented to look upon the past events on the bid and ask sides is very similar.

While mean α values are noticeably consistent between the stocks within a single exchange, there is slight heterogeneity in the mean alphas between the exchanges, especially for the cross-events variables. This may suggest that there are exchange-specific differences in the nature of the long-range correlation of stock market, since, e.g., not all the market participants (at NASDAQ Nordic) may trade at multiple exchanges.

TABLE III: Means and standard deviations of the daily scaling exponent α, for different stocks and order book sides.

Stock	Variable	Ask	Bid
		Mean α	Mean α
DK_1 Vestas Wind System	or-or	0.68 $\pm$0.047	0.68 $\pm$0.050
	tr-tr	0.65 $\pm$0.066	0.65 $\pm$0.060
	ca-ca	0.69 $\pm$0.057	0.68 $\pm$0.055
	or-tr	0.57 $\pm$0.090	0.57 $\pm$0.098
	or-ca	0.58 $\pm$0.104	0.59 $\pm$0.104
FI_1 Nokian Renkaat Oyj	or-or	0.68 $\pm$0.050	0.68 $\pm$0.050
	tr-tr	0.63 $\pm$0.064	0.62 $\pm$0.065
	ca-ca	0.69 $\pm$0.051	0.68 $\pm$0.050
	or-tr	0.58 $\pm$0.112	0.58 $\pm$0.114
	or-ca	0.63 $\pm$0.077	0.63 $\pm$0.076
FI_2 Metso Oyj	or-or	0.68 $\pm$0.046	0.68 $\pm$0.045
	tr-tr	0.63 $\pm$0.065	0.63 $\pm$0.064
	ca-ca	0.69 $\pm$0.047	0.69 $\pm$0.046
	or-tr	0.58 $\pm$0.094	0.57 $\pm$0.091
	or-ca	0.64 $\pm$0.082	0.63 $\pm$0.084
SE_1 Atlas Copco A	or-or	0.68 $\pm$0.039	0.68 $\pm$0.041
	tr-tr	0.65 $\pm$0.060	0.65 $\pm$0.060
	ca-ca	0.68 $\pm$0.040	0.68 $\pm$0.041
	or-tr	0.64 $\pm$0.098	0.64 $\pm$0.095
	or-ca	0.71 $\pm$0.067	0.71 $\pm$0.069
SE_2 Volvo B	or-or	0.69 $\pm$0.038	0.69 $\pm$0.038
	tr-tr	0.66 $\pm$0.050	0.66 $\pm$0.050
	ca-ca	0.68 $\pm$0.043	0.68 $\pm$0.045
	or-tr	0.64 $\pm$0.088	0.64 $\pm$0.090
	or-ca	0.71 $\pm$0.065	0.71 $\pm$0.064

B. Crossovers in correlation behaviors

By aggregating the daily recordings of a variable over all the available days, we observe crossover phenomena, i.e., changes in the scaling behavior at different time scales. Such crossovers in inter-event durations have been reported in earlier studies, e.g., [3], [22], in which two scaling exponents characterizing short- and long-range correlation are studied. In our analysis, however, motivated by the long time span of 3 years and

the accuracy up to millisecond precision of the ultra-high frequency data, we compute three scaling exponents, α_1, α_2, and α_3, characterizing the correlation properties at intra-day, day, and month time scales respectively (Fig. 2).

In DFA, the estimates of α_1, α_2, and α_3 are obtained by calculating the slopes locally within relevant time scale ranges in the log-log plot of the fluctuation $F(\tilde{s})$ against $\tilde{s}$, as shown in Fig. 2. Note that we use s – the number of events – as our time variable. The time scale $\tilde{s}$ is normalized by the average number of the events (order, trade, or cancel) during a trading day. Intra-day α_1 is calculated in the range 0.003-0.1 ($\times$ average daily activity), α_2 in the range 0.3-3 and α_3 in the range 10-100. Dotted lines in Fig. 2, represent the time scales of a trading day ($\log_{10} 1 = 0$) and a month ($\log_{10} 30 = 1.48$) respectively. Coherent with [3], we also observe a "bump" around the time scale of a trading day, indicating a clear crossover.

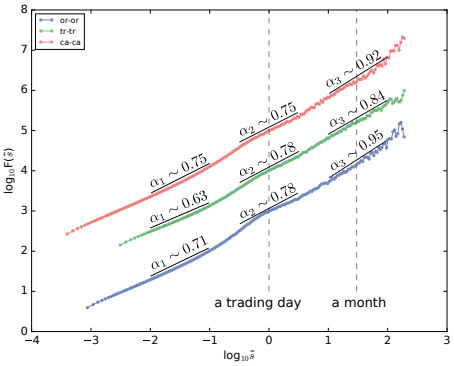


Fig. 2: Log-log plot of fluctuation (1) as a function of normalized scale $\tilde{s}$ for the inter-event duration series on the bid side for the stock SE_1. For each series we denote the local regressions and the corresponding slopes, i.e., α.

The final result is summarized in Fig. 3. For all inter-event variables, a significant increase in α is observed from intra-day to a day scale. Difference between α_2 and α_3 are not as prominent due to much bigger variance of α_3; however, Fig. 3 suggests that there is an increase in α from a day to a month scale for or-or and ca-ca durations. Increase in α at larger time scale signifies that the long-range correlation in inter-event durations become strong over longer time. This agrees with the economic theory (see e.g. [27], [28]) and supports the heterogeneous markets hypotheses of [29].

C. Correlation between the scaling exponent and economic variables

Utilizing the complete order book data available to us, we perform further analyses on the relationship between the power-law scaling exponent α and a selected number of market variables of interest. From financial point of view, it is interesting to understand whether there are some market variables that are closely correlated with the scaling behavior of the market flow. We investigate the correlation between the mean intra-day α of inter-event durations and the following economic variables: i) avg. duration: mean inter-event duration of a trading day, ii) activity: number of orders, trades, or cancels during a trading day, i.e., length of the inter-event duration time series, iii) avg. quantity: mean quantity (volume) of orders, trades, or cancels in a trading day, iv) daily (log) return, v) volatility[1]. The first four variables are directly related to the very same process, that generates the inter-event time variables, while daily return and volatility are determined by the (mid-) price process.

In Fig. 4, we present the results from the correlation analysis by plotting the (Pearson) correlation coefficient between α and X, the economic variables, as mentioned above. Stronger correlation is characterized by a larger magnitude of the correlation coefficient. Significance level of 0.01 is marked by dashed lines.

For the time series of tr-tr durations, the correlation coefficients of all the stocks are clustered together for each variable. The clustering pattern among the stocks suggests that the correlation properties in tr-tr durations are not stock-specific, but may apply at a more general level. Consistency in the correlation measures between stocks within a same exchange is also not observed. On the other hand, the correlation coefficients of bid and ask sides are often found near to each other, confirming again the symmetry of the order book between the sides.

Average inter-trade duration and α are negatively correlated, while activity and volatility are positively correlated to α. Negative correlation between average inter-trade duration and α is in agreement with the previous study with NYSE and NASDAQ stocks [4]. The significant correlation found between volatility and α is of financial interpretation. In periods of high volatility, trades tend to adjust their positions more often and accordingly become more reactive to other participants' submissions and cancellations. Furthermore, high volatility forces the participants to closely track the market evolution and make consequent decisions based on the market history from the past up to the most recent events, thus resulting in long-range dependence in tr-tr durations, i.e., larger α. High volatility also infers high activity and shorter tr-tr duration; therefore, the positive and negative correlation respectively with α provides a cross-check in the reasoning.

On the other hand, very weak (but not statistically significant) or no correlation is observed between α and average quantity or daily return. Indeed, market returns are known to be difficult to predict and either weakly dependent or not related to most of the economic variables [31]. The lack of correlation indicates that the dynamics, which lead to long-range correlation in the inter-event durations, has little impact on the price evolution and its level at the end of the day. Similarly, our results also shows that the (daily mean) quantity

[1]As a proxy for the daily volatility, the realized variance (RV) is used. We implement the sub-sampling and averaging estimator of [30], estimating and averaging the RV over different sub-sampling grids of five minutes.

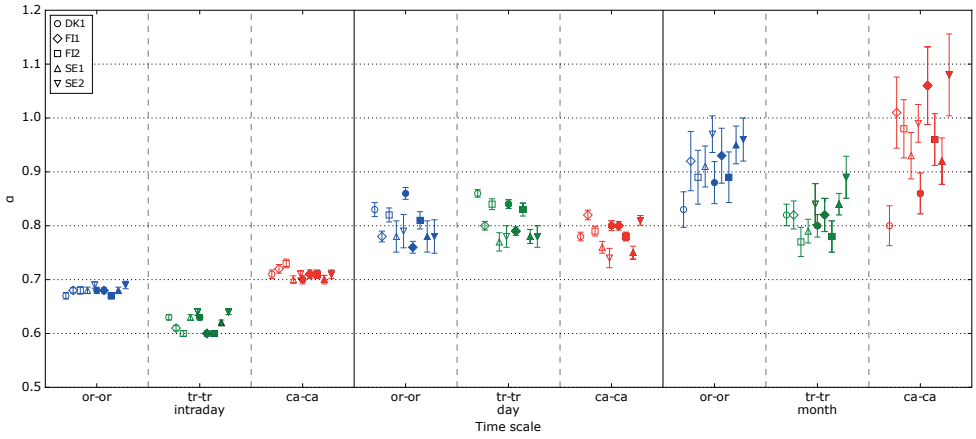


Fig. 3: Local α values computed at different time scales for each inter-event variable and stock. We report the 95% confidence intervals. Filled (empty) markers represent bid (ask) side data.

of orders, trades, or cancels is mostly uncorrelated with time-related variables, as it has no significant impact on α.

Interestingly, the clustering patterns in tr-tr duration is absent in or-or and ca-ca durations. The correlation coefficients of or-or and especially ca-ca are widely spread out, sometimes even across both positive and negative range. Therefore, it is more challenging to draw a strong conclusion from them.

IV. Conclusion

In this paper we provide an extensive empirical analysis on the scaling behavior of the order book flow for selected stocks traded at NASDAQ Nordic. We perform DFA, a reliable analysis method for detecting long-range correlation in a time series, on the ultra-high frequency order book data with millisecond precision. We calculate the DFA scaling exponent α, which characterizes the nature of the correlation present in the time series of inter-trade, -order, -cancel durations, yielding consistent results with previous works, e.g., [4], [5], [6]. We do not limit our analysis to inter-event durations, but extend the analysis to consider cross-events durations, such as lifetimes from order submission to its trade or cancellation. We find an outstanding consistency in the estimates of α values across different stocks and trading days, providing evidence that the time-related variables in the order book data are long-range correlated, i.e., fractal.

Taking the advantage of a complete and accurate order book data of 752 days, we investigate over different time scales, finding a significant change in correlation properties. This "crossover" phenomenon, also addressed in the past by, e.g., [3], [22], shows that the long-range correlation in intra-day inter-event durations becomes stronger at time scales beyond a day. The crossover analysis sheds light on the complexity of the market flow, as it takes more than one α to fully characterize the long-range correlation in the inter-event durations.

We also study the relationship between intra-day α and simple economic explanatory variables, such as mean inter-event duration, activity, mean quantity, daily return, and volatility. We find that, for tr-tr durations, the correlation coefficients of different stocks of both bid and ask sides form a cluster for each economic variable, while the clustering patterns are absent in or-or and ca-ca durations. We explain from a financial point of view, the negative correlation between the average tr-tr duration over a day and α, the positive correlation that activity and volatility have with α, and the little to no correlation shown by average trade quantity and daily return with α.

Our analysis reveals the complexity of the dynamics of stock markets, driven by automatic algorithms and human decision. The work presented in this paper will be expanded to include a wider number of stocks to further investigate other variables that may be relevant to the long-range dependence in the order book flow, addressing the dependence of the scaling exponent on the time scale. We will consider not only time intervals but also include a fractal analysis for the volumes of orders, trades and cancellations. Moreover, in order to gain even better insights into the market dynamics, the present work will be enriched with more sophisticated analyses, i.e., implementing other statistical methods and variations of DFA (such as multifractal DFA and multiscale entropy analysis).

Acknowledgment

The research leading to these results received funding from the Academy of Finland, the Finnish Academy of Science and Letters, and from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No. 675044.

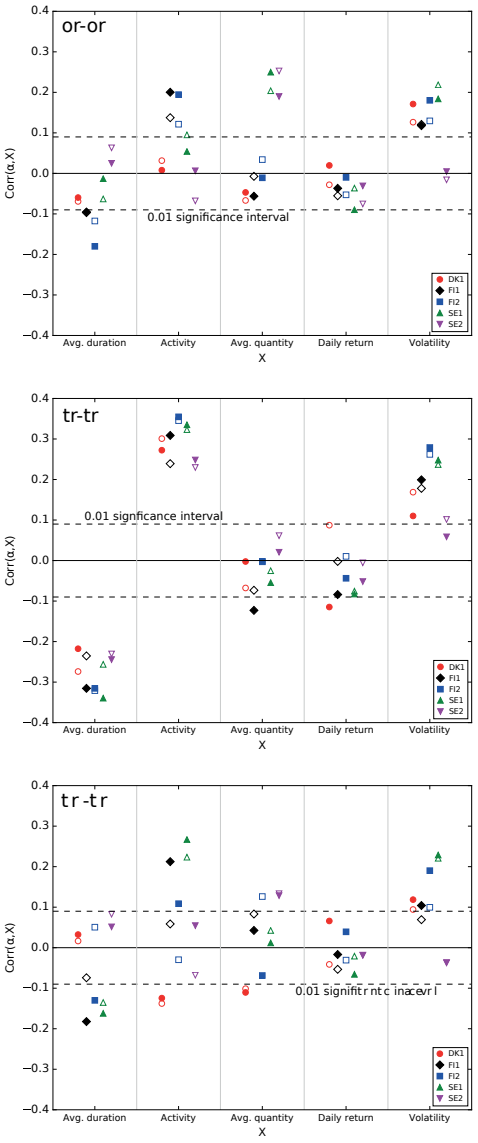


Fig. 4: Correlations coefficients between the daily α values and market variables, for the inter-event series. Filled and empty markers respectively denote bid and ask side data. Dashed lines mark the significance threshold, beyond which correlation is statistically significant at 99%.

REFERENCES

[1] J. W. Kantelhardt, "Fractal and multifractal time series," in *Mathematics of complexity and dynamical systems*. Springer, 2012, pp. 463–487.

[2] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of dna nucleotides," *Physical review E*, vol. 49, no. 2, p. 1685, 1994.

[3] P. C. Ivanov, A. Yuen, B. Podobnik, and Y. Lee, "Common scaling patterns in intertrade times of us stocks," *Physical Review E*, vol. 69, no. 5, p. 056107, 2004.

[4] P. C. Ivanov, A. Yuen, and P. Perakakis, "Impact of stock market structure on intertrade time and price dynamics," *PloS one*, vol. 9, no. 4, p. e92885, 2014.

[5] Z.-Q. Jiang, W. Chen, and W.-X. Zhou, "Detrended fluctuation analysis of intertrade durations," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 4, pp. 433–440, 2009.

[6] G.-F. Gu, X. Xiong, W. Zhang, Y.-J. Zhang, and W.-X. Zhou, "Empirical properties of inter-cancellation durations in the chinese stock market," *Frontiers in Physics*, vol. 2, p. 16, 2014.

[7] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, vol. 116, pp. 770–808, 1951.

[8] R. Bryce and K. Sprague, "Revisiting detrended fluctuation analysis," *Scientific reports*, vol. 2, p. 315, 2012.

[9] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 295, no. 3, pp. 441–454, 2001.

[10] A. Bashan, R. Bartsch, J. W. Kantelhardt, and S. Havlin, "Comparison of detrending methods for fluctuation analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5080–5090, 2008.

[11] B. B. Mandelbrot, "When can price be arbitraged efficiently? a limit to the validity of the random walk and martingale models," *The Review of Economics and Statistics*, pp. 225–236, 1971.

[12] R. N. Mantegna and H. E. Stanley, "Scaling behaviour in the dynamics of an economic index," *Nature*, vol. 376, no. 6535, pp. 46–49, 1995.

[13] N. Vandewalle and M. Ausloos, "Coherent and random sequences in financial fluctuations," *Physica A: Statistical Mechanics and its Applications*, vol. 246, no. 3-4, pp. 454–459, 1997.

[14] N. Vandewalle, M. Ausloos, and P. Boveroux, "Detrended fluctuation analysis of the foreign exchange market," in *Econophysic Workshop, Budapest, Hungary*, 1997.

[15] D. Grech and Z. Mazur, "Can one make any crash prediction in finance using the local hurst exponent idea?" *Physica A: Statistical Mechanics and its Applications*, vol. 336, no. 1, pp. 133–145, 2004.

[16] Ł. Czarnecki, D. Grech, and G. Pamuła, "Comparison study of global and local approaches describing critical phenomena on the polish stock exchange market," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6801–6811, 2008.

[17] S. Benbachir and M. El Alaoui, "A multifractal detrended fluctuation analysis of the moroccan stock exchange," *International Research Journal of Finance and Economics*, no. 78, pp. 6–17, 2011.

[18] A. Carbone, G. Castelli, and H. Stanley, "Time-dependent hurst exponent in financial time series," *Physica A: Statistical Mechanics and its Applications*, vol. 344, no. 1, pp. 267–271, 2004.

[19] P. Oświe, J. Kwapień, S. Drożdż *et al.*, "Multifractality in the stock market: price increments versus waiting times," *Physica A: Statistical Mechanics and its Applications*, vol. 347, pp. 626–638, 2005.

[20] J. Alvarez-Ramirez, J. Alvarez, E. Rodriguez, and G. Fernandez-Anaya, "Time-varying hurst exponent for us stock markets," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 24, pp. 6159–6169, 2008.

[21] J. R. Thompson and J. R. Wilson, "Multifractal detrended fluctuation analysis: Practical applications to financial time series," *Mathematics and Computers in Simulation*, vol. 126, pp. 63–88, 2016.

[22] A. K. Tiwari, C. T. Albulescu, and S.-M. Yoon, "A multifractal detrended fluctuation analysis of financial market efficiency: Comparison using dow jones sector etf indices," *Physica A: Statistical Mechanics and its Applications*, vol. 483, pp. 182–192, 2017.

[23] Z.-Q. Jiang, W. Chen, and W.-X. Zhou, "Scaling in the distribution of intertrade durations of chinese stocks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 23, pp. 5818–5825, 2008.

[24] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price prediction of limit order book data," *arXiv preprint arXiv:1705.03233*, 2017.

[25] M. Siikanen, J. Kanniainen, and J. Valli, "Limit order books and liquidity around scheduled and non-scheduled announcements: Empirical evidence from nasdaq nordic," *Finance Research Letters*, vol. 21, pp. 264–271, 2017.

[26] M. Siikanen, J. Kanniainen, and A. Luoma, "What drives the sensitivity of limit order books to company announcement arrivals?" *Economics Letters*, vol. 159, pp. 65–68, 2017.
[27] E. F. Fama and K. R. French, "Dividend yields and expected stock returns," *Journal of Financial Economics*, vol. 22, no. 1, pp. 3–25, 1988.
[28] J. Y. Campbell, A. W.-C. Lo, and A. C. MacKinlay, *The econometrics of financial markets*. Princeton University Press, 1997.
[29] U. A. Müller, M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward, "Fractals and intrinsic time: A challenge to econometricians," *Unpublished manuscript, Olsen & Associates, Zürich. International AEA Conference on Real Time Econometrics in Luxembourg*, 1993.
[30] L. Zhang, P. A. Mykland, and Y. Aït-Sahalia, "A tale of two time scales: Determining integrated volatility with noisy high-frequency data," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1394–1411, 2005.
[31] R. Cont, "Long range dependence in financial markets," in *Fractals in Engineering*. Springer, 2005, pp. 159–179.

PUBLICATION
IV

A C-Vine extension for the HAR model

Magris, M.

Ed. by 2019. *Unpublished manuscript*. Submitted to *Journal of Business & Economic Statistics*, May 2019