UNIVERSITEIT VAN AMSTERDAM

MASTER THESIS

MSC. COMPUTATIONAL SCIENCE

Information-Theoretic Approach for Directed Dealer-Network Inference in the Foreign Exchange Market

Author:
Aleksander T. JANCZEWSKI

Supervisor:
Dr. Ioannis ANAGNOSTOU

Examiner:
Prof. Dr. Drona KANDHAI

Second assessor:
Dr. Rick QUAX

A thesis submitted in partial fulfilment of the requirements
for the degree of Master of Science in Computational Science.

July 19, 2022

My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.'

Claude Elwood Shannon

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science
Graduate School of Informatics

Master of Science in Computational Science

Information-Theoretic Approach for Directed Dealer-Network Inference in the Foreign Exchange Market

by Aleksander T. Janczewski

This thesis introduces econometric, and information-theoretic frameworks for inference of information flows in the dealer-network of the foreign exchange market. The microstructure approach to exchange rates and the fundamental econometric techniques for quantifying dealers' contributions to the price discovery process are thoroughly discussed. Additionally, the thesis elaborates on the information share metric proposed by Hasbrouck [1995], the notion of the common trend component established by Stock and Watson [1988], and the novel econometric approach introduced by Hagströmer and Menkveld [2019]. Limitations of the econometric approach are recognized, and a novel information-theoretic approach for network inference is proposed as an alternative method. The theoretical framework for the information-theoretic technique is introduced, and the adequacy of this approach to the inference of information flows in the dealer-network is demonstrated. Moreover, the Kraskov estimator for continuous information-theoretic metrics is explored and employed in the developed network inference algorithm. The empirical investigation is performed on two extensive high-frequency data sets containing bid and ask quotes for the EUR/USD and USD/JPY exchange rates. The results indicate that the information flows inferred with the information-theoretic network inference algorithm are, to a large extent, consistent with Hasbrouck's information share and other econometric metrics. Additionally, the investigation exposes changes in the dealer-network structure following the USD/JPY flash crash in January 2019 and the ECB's announcement of quantitative easing policy in March 2020. The findings presented in this thesis indicate that information flows inferred with information-theoretic metrics offer new insight into FX dealer-network dynamics, consistent with the literature and fundamental econometric approaches.

Acknowledgements

I want to express my gratitude to my supervisor, Dr. Ioannis Anagnostou, for his guidance and invaluable support throughout my thesis. I would also like to express my appreciation for my examiner, Prof. Drona Kandhai, for fruitful discussions and a unique opportunity to conduct my research at ING Netherlands.

I gratefully acknowledge the support given by Georgios Pierris. Our discussions on the foreign exchange market and algorithmic trading strategies allowed me to significantly extend my understanding of the dynamics in the FX dealer-network.

I am grateful for the support from my friends. In particular, the guidance provided by Gilles Magalhaes Ribeiro during my first steps with C++ and discussions on high-performance algorithms. Also, I wish to acknowledge Charel Felten for the valuable suggestions on various aspects of data visualizations techniques.

Finally, I would like to wholeheartedly thank my family for supporting my studies.

Contents

Abstract 2

Acknowledgements 3

Contents 4

List of Figures 8

List of Tables 10

1 Introduction 11

2 Literature review 16
- 2.1 Econometric approaches . . . 16
  - 2.1.1 Quantifying contribution to the price discovery process . . . 16
  - 2.1.2 Inferring information transmission in decentralized markets . . . 18
  - 2.1.3 Granger casuality . . . 21
- 2.2 Information-theoretic approach to modelling directed information flow . . . 22

3 Theory 26
- 3.1 Modelling information revelation . . . 26
  - 3.1.1 Introduction to cointegrated stochastic processes . . . 27
  - 3.1.2 Vector-error correction model . . . 32
  - 3.1.3 Vector moving average . . . 34
  - 3.1.4 Multivariate impulse response function . . . 35
  - 3.1.5 Common trend representation . . . 38
  - 3.1.6 Metrics . . . 41
    - 3.1.6.1 Information share . . . 41
    - 3.1.6.2 Price inefficiency & information speed . . . 45
    - 3.1.6.3 Bilateral connections . . . 48
- 3.2 Quantifying information flows with transfer entropy . . . 50
  - 3.2.1 Introduction to information theory . . . 50
    - 3.2.1.1 Shannon entropy . . . 50

3.2.1.2 Joint entropy . . . 54
3.2.1.3 Conditional entropy . . . 55
3.2.1.4 Kullback-Leibler divergence . . . 56
3.2.1.5 Mutual information . . . 57
3.2.1.6 Conditional mutual information . . . 58
3.2.1.7 Redundancy and synergy . . . 60
3.2.1.8 Time-delayed mutual information . . . 61
3.2.1.9 Entropy rate . . . 63
3.2.2 Transfer entropy & conditional transfer entropy . . . 65
3.2.2.1 Definition of transfer entropy . . . 65
3.2.2.2 Interpretation of transfer entropy . . . 69
3.2.2.3 Definition of conditional transfer entropy . . . 70
3.2.3 Entropy of a continuous distribution . . . 72
3.2.3.1 Differential entropy . . . 72
3.2.3.2 Transfer entropy & conditional transfer entropy . . . 73

4 Methods **75**
4.1 Data preprocessing . . . 75
4.1.1 Data description . . . 75
4.1.2 Latency adjustment . . . 77
4.1.3 Data sampling . . . 79
4.1.4 Stationarity & cointegration testing . . . 81
4.2 Econometric model . . . 83
4.2.1 VECM lag selection . . . 83
4.3 Network inference algorithm . . . 83
4.3.1 Kraskov, Stögbauer and Grassberger algorithm I . . . 83
4.3.1.1 Time-series standardization . . . 88
4.3.2 Paramater choice . . . 88
4.3.2.1 State space reconstruction . . . 89
4.3.2.2 Ragwitz criterion . . . 90
4.3.2.3 Delay reconstruction . . . 93
4.3.3 Statistical testing . . . 95
4.4 Implementation remarks . . . 97

5 Application **100**
5.1 The foreign exchange market . . . 100
5.2 Model assumptions . . . 101
5.3 Data . . . 102
5.4 Results . . . 103
5.4.1 Econometric model . . . 103
5.4.2 Information-theoretic network inference . . . 110
5.4.2.1 TE information network . . . 113
5.4.2.2 CTE information network . . . 116
5.4.3 Flash crash . . . 121
5.4.4 Quantitative easing . . . 126

6 Discussion **131**
6.1 Results . . . 131
6.1.1 Hypothesis 1 - Central dealers are more informed. . . . 131
6.1.2 Hypothesis 2 - The price discovery process is dominated by dealers who share their private information the most. . . . 132
6.1.3 Hypothesis 3 - The largest disparities in information contributions are observed between different types of dealers; i.e., banks, non-bank market makers and electronic trading platforms. . . . 133
6.1.4 Hypothesis 4 - The more a given dealer learns from other dealers, the faster is his response to the informational shocks, i.e. the faster is his price discovery process. . . . 133
6.1.5 Hypothesis 5 - There is more exchange of information between dealers when there is more to be learned. . . . 133
6.1.6 Hypothesis 6 - During a flash crash, the dealer-network deviates from a steady-state structure and eventually converges back to it. . . . 134
6.1.7 Hypothesis 7 - Public announcements lead to a decrease in the information exchange between dealers. . . . 134
6.2 Methodology . . . 135
6.2.1 Econometric model . . . 135
6.2.2 Information-theoretic model . . . 135
6.2.3 Other remarks . . . 136

7 Conclusion and future work **137**
7.1 Summary . . . 137
7.2 Contributions . . . 138
7.3 Future work . . . 139

A Example for information revelation method **141**

B Proofs of information-theoretic theorems and lemmas **148**
B.1 Proof of Lemma 3.13 - Maximum of Shannon entropy . . . 148
B.2 Proof of Lemma 3.23 - Additivity . . . 149
B.3 Proof of Lemma 3.16 - Sub-additivity . . . 149
B.4 Proof of Theorem 3.18 - Chain rule . . . 150
B.5 Proof of Corollary 3.19 - Chain rule for N random variables. . . . 150
B.6 Proof of Theorem 3.22 - Information inequality . . . 151
B.7 Proof of Lemma 3.25 - Mutual information to entropy . . . 151
B.8 Proof of Theorem 3.30 - Mutual information to entropy . . . 152
B.9 Proof of Theorem 3.46 - Differential entropy vs discrete entropy . . . 152
B.10 Proof of Theorem 3.40 - Transfer entropy is conditional time-delayed mutual information . . . 154

C Metrics validation **155**
C.1 Information Share . . . 156
C.2 Price inefficiency . . . 157

C.3 Bilateral connection 157
C.4 Transfer entropy 159

D Supplementary results for EUR/USD data set. 160
D.1 Informativeness metrics vs other types of dealer centrality. 160

E Results for USD/JPY dealer-network 161
E.1 Stationarity & Cointegration Testing 161
E.2 Average intraday quote frequency 162
E.3 Econometric baseline map 163
E.4 Scatter plots for econometric baseline 164
E.5 Summary of key information theoretic metrics 165
E.6 Scatter plots for information-theoretic baseline 165

Bibliography 167

List of Figures

3.1 An example of the empirical information revelation map. 28
3.2 EUR/USD mid prices of market maker (M1) and bank (B6) at 9h on 3rd September 2020. 31
3.3 EUR/USD mid prices of market maker (M1) and bank (B6) at 14h on 3rd September 2020. 32
3.4 Efficient price and mid prices of market 1 and market 2. 40
3.5 Shannon entropy of two-state, three-state and four-state systems as a function of probability associated with first state x_1. 52
3.6 Venn diagram illustrating mutual information common for all variables X, Y and Z. 59

4.1 Latency for the period from March 3 2020 till March 7 2020 in EUR/USD data set. 78
4.2 USD/JPY raw and resampled (RS) mid prices of bank (B2) at 22h on 2nd January 2019 - the day of flash crash. 80
4.3 Mean squared error of locally constant predictor for different embedding delay and dimension parameters. 92

5.1 Average intraday EUR/USD quote frequency for trading days in the period February 27 2020 and March 27 2020. 103
5.2 Baseline econometric map illustrating the process of information revelation in the EUR/USD dealer-network for period between 27th February and 27th March 2020. 105
5.3 Scatter plot of dealer centrality (0.2) versus price inefficiency and lower bound information share for EUR/USD baseline. 108
5.4 Scatter plot of dealer price inefficiency versus lower bound information share for EUR/USD baseline. 109
5.5 TE information map illustrating mean information flows in the EUR/USD dealer-network between 27th February and 27th March 2020. 114
5.6 Scatter plot of TE outflows versus Hasbrouck's lower and upper bound information shares for EUR/USD baseline. 115
5.7 CTE information map illustrating mean information flows in the EUR/USD dealer-network for period between 27th February and 27th March 2020. 117
5.8 Scatter plot of CTE outflows versus Hasbrouck's lower and upper bound information shares for EUR/USD baseline. 118
5.9 Scatter plot of TE and CTE inflows versus absolute change in price inefficiency from $\tau = 0$ to $\tau = 1$ for EUR/USD baseline. 119

5.10 Scatter plot of log TE inflows versus TE outflows (LHS) and log CTE inflows versus CTE outflows (RHS) for EUR/USD baseline. 120
5.11 TE (LHS) and CTE (RHS) information maps illustrating mean information flows in the USD/JPY dealer-network for period between 2nd and 17th January 2019. 121
5.12 CTE information maps illustrating changes in the USD/JPY dealer-network between 16:00 and 24:00 on the 2nd January 2019 - the day of flash crash. 124
5.13 CTE information maps illustrating changes in the USD/JPY dealer-network between 24:00 and 8:00 on the 3rd January 2019 - the next day after the flash crash. 125
5.14 CTE information maps illustrating changes in the EUR/USD dealer-network between 8:00 and 16:00 on the 12th March 2020 - the day of ECB's quantitative easing announcement. 129
5.15 CTE information maps illustrating changes in the EUR/USD dealer-network between 16:00 and 20:00 on the 12th March 2020 - the day of ECB's quantitative easing announcement. 130

A.1 Impulse response of markets to a unit shock in market 1 (M1). 142
A.2 Impulse response of markets to a unit shock in market 2 (M2). 144

C.1 Validation of Hasbrouck's information share metric and long run efficient price computed with MIRF. 156
C.2 Validation of Hägstromer and Menkveld's price inefficiency metric. 157
C.3 Validation of Hägstromer and Menkveld's bilateral connection metric. 158
C.4 Validation of the KSG algorithm implementation against the analytical solution for transfer entropy. 159

D.1 Scatter plot of dealer centrality (0.1) (top row) and (All) (bottom row) versus price inefficiency and lower bound information share for EUR/USD baseline. . . 160

E.1 Average intraday EUR/USD quote frequency for trading days in the period February 27 2020 and March 27 2020. 162
E.2 Econometric map illustrating the process of information revelation in the USD/JPY dealer-network for period between 2nd January and 17th January 2019. . 163
E.3 Scatter plot of dealer centrality (0.2) versus price inefficiency and lower bound information share for USD/JPY baseline. 164
E.4 Scatter plot of dealer price inefficiency versus lower bound information share for USD/JPY baseline. 164
E.5 Scatter plot of TE (top row) and CTE (bottom row) outflows versus Hasbrouck's lower (LHS) and upper (RHS) bound information shares for USD/JPY baseline. 165
E.6 Scatter plot of TE (LHS) and CTE (RHS) inflows versus absolute change in price inefficiency from $\tau = 0$ to $\tau = 1$ for USD/JPY baseline. 166
E.7 Scatter plot of log TE inflows versus TE outflows (LHS) and log CTE inflows versus CTE outflows (RHS) for USD/JPY baseline. 166

List of Tables

4.1 Table presents the availability of dealer's FX spot rates in different data sets. . 76
4.2 Result of augmented Dickey-Fuller tests for EUR/USD data set. 81
4.3 Result of augmented Engle-Granger two-step cointegration test for EUR/USD data set. 82

5.1 Summary of key econometric metrics characterizing the EUR/USD baseline presented in Figure 5.2. 104
5.2 Partial correlation matrix (bilateral connections) of dealers based on EUR/USD baseline at $\tau = 0$. 107
5.3 Table summarizing the mean of family-wise statistically significant at 10% level transfer entropies detected between pairs of dealers in EUR/USD data set. . . . 112
5.4 Summary of key information theoretic metrics characterizing the EUR/USD baseline presented in Figures 5.5 and 5.7. 113

A.1 Table summarizing the results of the impulse response of the system to a unit shock to market 1 (M1). 144
A.2 Table summarizing the results of the impulse response of the system to a unit shock to market 2 (M2). 145
A.3 Table summarizing the impulse response analysis of both markets. 147

E.1 Result of augmented Dickey-Fuller tests for USD/JPY data set. 161
E.2 Result of augmented Engle-Granger two-step cointegration test for USD/JPY data set. 162
E.4 Summary of key information theoretic metrics characterizing the USD/JPY TE and CTE information maps presented in Figure 5.11. 165

Chapter 1

Introduction

The Foreign Exchange (FX) market, commonly referred to as the currency market, is a worldwide, decentralized marketplace that facilitates the trading of currencies between various market participants such as banks, speculators, and investors (Garner [2011]). The FX market plays a vital role in the global economy, as it directly impacts employment through real exchange rates, inflation through the cost of imports, and international capital flows through the returns of different assets (King et al. [2011], Donnelly [2019]). According to a recent Triennial Central Bank Survey published by Bank for International Settlements (BIS), the trading in FX markets reached $6.6 trillion in average turnovers per day in April 2019, up from $540 billion as recorded in 1989 (BIS [2019a], BIS [2019b]). To understand what fueled this enormous increase in per-day turnover, it is necessary to recognize the changes in the FX market that have taken place over the past 30 years.

In the 1990s, the revolution in information and communication technology gave rise to electronic trading, which gradually dominated the foreign exchange market (Donnelly [2019], King et al. [2011]). The technological advancements led to the emergence of electronic trading and matching platforms, such as Reuters Dealing 2000 and EBS (Electronic Broking Services), which ultimately took over the market as they provided superior pricing and execution speed. These developments further facilitated improvements in the liquidity of the market (Donnelly [2019]). Eventually, the technological advancements resulted in a complete transformation of the foreign exchange market.

The electronic era gave rise to algorithmic trading; consequently, evolving the complexity and sophistication of trades. The inception of electronic trading platforms streamlined the flow of the information through the market (Record [2004]) and fundamentally changed the nature of interactions between market participants (Sager and Taylor [2006], Kissell [2013]). At the same time, the constantly improving accessibility of a high-speed internet connection simplified the entry for new electronic markets, and hence fueled further decentralization and fragmentation of the FX market (Hagströmer and Menkveld [2019], Kissell [2013]). While, today the foreign

exchange market is commonly considered as a decentralized market, characterization of modern FX market is not that straightforward.

Hagströmer and Menkveld [2019] and Sager and Taylor [2006] advocate the notion that the contemporary structure of the foreign exchange market is a hybrid one, with both centralized and decentralized market characteristics. In the hybrid structure, the decentralized trading occurs bilaterally between investors and intermediaries in quote-driven markets (Hagströmer and Menkveld [2016]), whereas the centralized trading takes place multilaterally in inter-dealer order-driven centralized brokerage platforms such as EBS (Vitale [2006], Hagströmer and Menkveld [2016]). Since there is no longer a particular centralized institution where information could be incorporated efficiently, the information flow within the dealer-network, and thus the process through which the information is incorporated into the price of the asset, has become a lot more complex (Hagströmer and Menkveld [2019]).

The process through which the information is impounded in prices over time is called the price discovery process, and it stands as the primary research focus of the branch of finance called market microstructure (Madhavan [2000]). The study of the market microstructure essentially focuses on the structure of exchanges and how investors' actions and departures from symmetric information are translated into prices and volumes (Madhavan [2000], Kissell [2013], Lyons et al. [2001]). The microstructure approach differs from traditional approaches mainly in that it recognizes that some information is not publicly available, which allows for the existence of information asymmetry (Lyons et al. [2001]). In financial markets, information asymmetry is exhibited by disparities between prices of the same asset on different markets, which can create arbitrage opportunities. Since inefficient pricing setups are eventually corrected by market participants to protect themselves from being arbitraged, it is reasonable to assume that there must exist some flow of information in the dealer-network, in which less informed counter parties learn from more informed ones (Hagströmer and Menkveld [2016]). Thus, in the market microstructure framework, the information in the decentralized markets is produced locally and then transferred between the markets (Hagströmer and Menkveld [2019], Hasbrouck [1996]).

Information in financial markets is a valuable asset as it is fundamentally linked to the price discovery process. Therefore, understanding how fundamental information is revealed and how it flows through the dealer-network is of utmost importance. Hagströmer and Menkveld [2019] list various reasons why understanding the dynamics of the information flow can be valuable. First of all, it can allow dealers to discover their position in the dealer-network. Second, it can expose market participants with superior information and thus help less informed market participants discover with whom best to interact with in order to gain new information. Gaining new information would further allow less informed market participants to readjust their position and consequently reduce exposure to the adverse selection risk. Finally, identifying which market dominates in information revelation can be helpful for regulatory purposes (Hagströmer and

Menkveld [2019], Hagströmer and Menkveld [2016]).

In the literature, there are many approaches for characterizing the price discovery process in decentralized markets. Especially noteworthy are the contributions of Gonzalo and Granger [1995], Hasbrouck [1995] and Harris et al. [1995] who were the first researchers to propose methods and metrics to quantify the extent to which each individual market contributes to the price discovery process. Hasbrouck information share (IS) and Harris-McInish-Wood component share (CS) empirical metrics rely on the assumption that the market dominates the price discovery process if it has a dominant influence on the change of the long-term cointegration equilibrium price, also known as the common efficient price (Putniņš [2013], Harris et al. [1995], Hasbrouck [1995]). On the other hand, Granger and Gonazalo's permanent-transitory (P-T) decomposition provides a different approach to the price discovery process, in which contribution is defined in terms of the market's error correction coefficients (Gonzalo and Granger [1995], Baillie et al. [2002]). Nevertheless, all of the methods use the vector error correction model and rely on the assumptions that decentralized markets are cointegrated and that they have a common trend - the standard set by Stock and Watson [1988].

A novel approach to the price discovery process has been recently proposed by Hagströmer and Menkveld [2019], who not only investigate the influence of individual markets on the changes in the long-term efficient price of the asset, but also zoom in on the short-term dynamics of information flows in the dealer-network. By accounting for the short-term dynamics, Hagströmer and Menkveld [2019] are able to infer the flow of information in the network and also characterize the process of information revelation. To do so, the authors propose a multivariate impulse-response function-based (MIRF) approach to mapping information revelation empirically. The information revelation framework proposed by Hagströmer and Menkveld [2019] relies on the ideas from the one-security-many-markets settings introduced by Hasbrouck [1995], and thus also models price dynamics using the vector error correction model. Undoubtedly, cointegration-based approaches have received a lot of attention from the research community, as they are widely used to study the dynamics of various financial instruments and markets (Booth et al. [1999], Brandvold et al. [2015], Zhang and Wei [2010], Figuerola-Ferretti and Gonzalo [2010]). However, as it will be discussed in Chapter 2, the agreement on which cointegration-based approach to the price discovery process is the best has not been reached until this day (Lehmann [2002]). Moreover, cointegration-based approaches are limiting as they can only be applied to financial instruments and markets characterized by cointegration relations.

A potential solution to this limitation can be found in the information-theoretic approaches. A relatively new information-theoretic metric called transfer entropy has been recently proposed by Schreiber [2000], building upon the concept of entropy introduced by Shannon [1948]. Transfer entropy allows for detecting and quantifying probabilistic causal relationships between variables in complex systems, which in the context of the FX market and price discovery can

be treated as quantifying information flows between dealers (Bossomaier et al. [2016]). As Schreiber states, the attractiveness of transfer entropy is rooted in the fact that there is no need to make any assumptions about the dynamics of the system that one wants to investigate, thus making it a truly generic metric (Schreiber [2000]). While employing the multivariate transfer entropy approach has recently gained much attention in neuroscience (Vicente et al. [2011]), to the best of our knowledge, there are no publications in which researchers use this approach to infer the dealer-network in the FX market. Moreover, this approach is relatively unexplored in the context of other financial markets, such as equity or fixed-income markets. Therefore, given the advantages that come with the information-theoretic approach, it would be interesting to investigate and assess whether this approach can yield results consistent with cointegration-based techniques such as the method proposed by Hagströmer and Menkveld [2019] and especially with the information share metric introduced by Hasbrouck [1995]. Furthermore, we would also like to know if the new approach provides additional insights into information flow between dealers that other methods cannot offer. These unknowns stand as the primary motivation of this thesis.

The main goal of this thesis is threefold. The first goal is to provide a theoretical framework for the concept of price discovery, cointegration-based methods proposed by Hagströmer and Menkveld [2019] and Hasbrouck [1995], and the information-theoretic approach. It is essential to present similarities and differences between both methods as well as demonstrate the adequacy of these methods to the FX market setting. The second goal is to reproduce the methodology proposed by Hagströmer and Menkveld [2019], and also develop an appropriate information-theoretic network inference algorithm. Both models will be applied to a high-frequency data set containing FX spot rates for USD/JPY and EUR/USD currency pairs from 2019 and 2020, respectively. The choice of dates used in the investigation is not arbitrary, as in the analysis, we will also take a closer look at the results provided by both methods in the vicinity of market stress times. Thus, we will use both methods to infer the flow of information in the dealer-network following the USD/JPY crash from 2$^{\text{nd}}$ January 2019 and the ECB's announcement of a quantitative easing plan from 12$^{\text{nd}}$ March 2020. The third goal is to compare the results obtained from both methods and assess if the information-theoretic approach is consistent with cointegration-based approaches.

The structure of this thesis is organized in the following manner. A brief overview of relevant literature is presented in Chapter 2. Next, in Chapter 3 the theory behind the econometric and information-theoretic approaches are covered in detail. In particular, the concept of cointegration and stationarity are introduced; additionally, information share, price inefficiency, and bilateral connection metrics are formally derived. Furthermore, the information-theoretic metrics are treated in detail. Chapter 4 introduces data processing and parameter selection methods employed in both the econometric model and network inference algorithm. Additionally, the

chapter is concluded with a discussion of the pipeline development efforts and challenges faced in the process. Finally, the results obtained from the analysis are presented in Chapter 5, discussed in Chapter 6, and conclusions and suggestions for future work can be found in Chapter 7.

Chapter 2

Literature review

This chapter provides an overview of relevant literature on econometric and information-theoretic-based approaches for inferring and quantifying information flows. The literature review is split into two main sections. Section 2.1 outlines literature on the fundamental econometric approaches for quantifying contribution to the price discovery process focusing on the work of Gonzalo and Granger [1995], Hasbrouck [1995] and Harris et al. [1995]. This section also treates the most recent developments in the relevant econometric methods on inferring information flows in dealer-network, in particular, the work of Hagströmer and Menkveld [2016] and Hagströmer and Menkveld [2019]. Section 2.2 provides an overview of various applications of information-theoretic metrics. Next, the information-theoretic network inference algorithms employing mutual information and transfer entropy are presented, including the recent developments presented in the work of Novelli et al. [2019].

2.1 Econometric approaches

2.1.1 Quantifying contribution to the price discovery process

As briefly mentioned in Chapter 1, Gonzalo and Granger [1995], Hasbrouck [1995] and Harris et al. [1995] were the first to propose methods and metrics to detect and quantify individual market's contribution to the price discovery process. The ideas and methods presented by these scholars heavily rely on the cointegration and error correction models, popularized by Granger [1981], as well as the notion of common trend component in the cointegrated systems established by Stock and Watson [1988]. Granger and Gonzalo's common factor component weight, Hasbrouck's information share, and Harris-McInish-Wood's component share are still the most widely used metrics in the empirical microstructure studies (De Jong [2002]). However, the agreement on which method and metric is superior has not been reached until this day (Lehmann [2002]).

In the literature, one can find many publications looking into the differences between the metrics and their advantages and disadvantages. For instance, Yan and Zivot [2010] argue that only the information share metric provides information on the relative informativeness of each market but advise against estimating it using high sampling frequencies data. This is because they believe that high sampling frequencies introduce transitory frictions that distort the information share estimates (Yan and Zivot [2010]).

A different perspective is provided by Putniņš [2013], who contends that both information and component share metrics are consistent with the view of price discovery only if the modeled price series have equal levels of noise. Additionally, Putniņš [2013] further proposes a new metric, the information leadership share, relying on both information and component share metrics, which he claims to be robust to differences in noise levels.

On the other hand, De Jong [2002] recognizes a very close relation between common factor component weight and information share, and reveals that the difference between these two metrics stems from the fact that Gonzalo and Granger ignore the variance in the price innovations while Hasbrouck does not. Ultimately, De Jong [2002] expresses his strong support for the variance-decomposition-based method proposed by Hasbrouck.

Another perspective is provided by Baillie et al. [2002] who reveals that the Hasbrouck's information share in fact complements Gonzalo and Granger's common factor model. Furthermore, Baillie et al. [2002] empirically shows that even though Hasbrouck, and Gonzalo and Granger define price discovery in slightly different terms, their models provide similar results if the residuals are uncorrelated and very different if a significant correlation exists (Baillie et al. [2002]). A related weakness is observed by Lehmann [2002], who recognizes that Hasbrouck's information share fails to clearly allocate the individual contributions of each market when price innovations across markets are highly correlated.

While Harris et al. [1995] are eventually persuaded by Clive Granger and follow his suggestions in their next publication Harris et al. [2002b], Hasbrouck is not convinced as he provides theoretical arguments and an empirical example against Granger and Gonzalo's approach in his following work Hasbrouck [2002] (Lehmann [2002]). Additionally, Hasbrouck further argues that the approaches proposed by Gonzalo and Granger [1995] and Harris et al. [2002b] do not measure efficient price, and are not consistent with the microstructure approach (Hasbrouck [2007]). The dispute is never resolved as later Harris et al. [2002a] replies to Hasbrouck [2002] by emphasizing the importance of statistical testing of price discovery parameters, which does not apply to the information share approach.

2.1.2 Inferring information transmission in decentralized markets

The study of the transmission of information in decentralized markets through the process of trade between market participants with asymmetric information was first presented by Wolinsky [1990]. In his simple agent-based pairwise meetings model of trade, Wolinsky [1990] turns his attention to decentralized markets and introduces the idea that meetings between agents (market participants) with asymmetric information may transmit relevant information to the less informed agents. Consequently, Wolinsky [1990] proposes that the trader updates his beliefs (learns) just from the interaction with other trader, even before any deal is concluded. He focuses his efforts on the analysis of steady-state equilibria observed in a frictionless market where the number of buyers and sellers are equal. Even though Wolinsky concludes that in his simple model information is not fully revealed to less informed agents, the ideas and intuition that he introduces become fundamental for future studies of the information transmission in decentralized markets (Wolinsky [1990]).

Different variations of Wolinsky's model are studied in Serrano and Yosha [1993] and Serrano and Yosha [1996]. Blouin and Serrano [2001] also follow the line of research began by Wolinsky, but significantly depart from his work by leaving strong steady-state restrictions and focusing on one-time entry market, with non-stationary dynamics in which composition of traders changes over time. Unlike some other works, Blouin and Serrano [2001] also account for the possibility of no trade happening. By departing from strong assumptions made by Wolinsky [1990], Blouin and Serrano [2001] were able to investigate the dynamics of volumes of trades and the properties of information revelation and efficiency. They conclude that their model does not lead to information revelation at equilibrium, either, and provide further remarks on the factors they believe to facilitate the process of information revelation.

Eventually, it is Duffie and Manso [2007] who introduce the concept of information percolation. In contrast to Wolinsky [1990] and Blouin and Serrano [2001] models, Duffie and Manso [2007] propose the first agent-based model where the equilibrium in the market actually leads to full revelation of information through trading. In their model, agents also learn by observing the bids submitted by other agents. Duffie and Manso [2007] further focus their efforts on characterizing the process of information percolation through the market. Duffie builds upon his work in the following very extended seminal work, Duffie et al. [2009], in which he allows agents to change the intensity of their search for new information and introduces public information to the model. He concludes that public information reduces efforts to interact and thus gain information from other agents, which consequently reduces information sharing. Another extension is provided in Duffie et al. [2014], where information percolation in segment markets in investigated by introducing another attribute: the "connectivity" of the agent, which stands for the frequency of their bilateral trading opportunities[1].

[1] An interested reader may find many Duffie's ideas gathered in Duffie [2012].

In the literature, many other extensions to this framework can be found, such as Golosov et al. [2014] who explore the impact of private information on the information flow in the decentralized markets, Andrei and Cujean [2017], who investigate the role of information flows in generating the short-term momentum and long-term reversals, and Babus and Kondor [2018] who translate decentralization into a complex network setting to study the impact of decentralization and adverse selection on information diffusion in the dealer network. Networks found their application in the information percolation framework, after a very popular work by Billio et al. [2012], who proposed a method of mapping out securities "connectedness" in a network, and used pairwise Granger-causality to characterize its structure. Similarly, another significant contribution is the work of Diebold and Yılmaz [2014] who combine multivariate time series analysis, complex networks, and variance decomposition notions to measure financial economic risk associated with the "connectedness" of financial firms. These two contributions partially inspired the empirical method proposed by Hagströmer and Menkveld [2016].

It is Hagströmer and Menkveld [2016] who provide the first empirical framework for testing assumptions and predictions of information percolation theory presented by Duffie et al. [2009]. The novelty of work by Hagströmer and Menkveld [2016] stems from the fact that the authors depart from work focusing on the predictability of some securities returns based on an other securities' current returns such as the work by Chan [1992]. Instead, they focus on empirically identifying and outlining the process of information percolation in the dealer-network. The method proposed by Hagströmer and Menkveld [2016] quantifies the information flow between markets by first filtering out short-term pricing errors and then regressing lagged price change (with one lag only) of the source market on the price changes of the target market (Hagströmer and Menkveld [2016]). The coefficient relating these price co-movements of two markets quantifies the immediate response of the target market to the innovations in the source market, which they associate with the information flow (Hagströmer and Menkveld [2016]). The framework proposed by Hagströmer and Menkveld [2016] is an extension of widely accepted one-security-many-markets setting introduced by Hasbrouck [1995], and thus also relies on the assumption that decentralized markets are cointegrated since price responses to shocks diverge across markets in the short run but converge in a long run (Hasbrouck [1995]). Consequently, Hagströmer and Menkveld [2016] also employ the vector-error correction model to model price dynamics in the decentralized markets, and make use of Hasbrouck's information share metric. The following hypotheses are tested by Hagströmer and Menkveld [2016][2]:

1. The information about fundamental value is dispersed across all dealers in the sense that observing any dealer's quotes helps in learning fundamental value.

2. Information percolates from any dealer to any other dealer.

[2] All of the listed hypothesis are citations of Hagströmer and Menkveld [2016].

3. The release of some public information reduces the strength of information percolation between dealers.

4. The release of some public information slows down price discovery.

As previously mentioned, all of the hypotheses listed above are meant to test whether the empirical findings align with the the theoretical framework from Duffie et al. [2009]. Although the empirical findings in Hagströmer and Menkveld [2016] were found to be aligning with the findings from Duffie et al. [2009], the manuscript was later updated to significantly different version Hagströmer and Menkveld [2019].

Hagströmer and Menkveld [2016] were prompted to search for a different approach because of one particularly severe limitation of the approach presented in first publication Hagströmer and Menkveld [2016]. Namely, the fact that the method is susceptible to spurious conclusions, in case e.g. the time stamps of quotes of different dealers are unsynchronized or when latency adjustment is applied incorrectly. For example, in simple scenarios when information from dealer A arrives at dealer C via dealer B, and when information from dealer A travels directly to dealer C but is slower to arrive than when it travels to dealer B, would yield exactly the same information percolation network Hagströmer and Menkveld [2016]. While authors still believe that the method from the former manuscript Hagströmer and Menkveld [2016] has its merits, in their more recent publication, Hägstromer and Menkveld abandon the information-percolation framework and focus their efforts on characterizing the process of information revelation.

In Hagströmer and Menkveld [2019] the method of detecting and quantifying the responsiveness of markets to innovations in other markets has completely changed. The key difference between the most recent and earlier publication is that in Hagströmer and Menkveld [2019], the information flows are no longer quantified; they are inferred instead. Instead of regressing one time step lagged price changes of one market on the other market's price changes, Hagströmer and Menkveld [2019] decided to employ multivariate impulse response function (MIRF). MIRF allows them to investigate and detail how innovations in one market affect all the other markets, step by step, from an innovation until eventual convergence. The directional linkages are now replaced by edges that represent partial correlations of cumulative responses of markets to the multivariate shock at a particular time step since the introduction of the shock. The last significant change is the addition of another dimension by introducing a metric called "price inefficiency" that is interpreted as the *fraction of the information that market (...) has yet to reveal τ periods after the multivariate shock* (Hagströmer and Menkveld [2019]). The hypothesis tested by the authors have also been slightly changed, and they are now as follows:

1. The bilateral (price) connections that form the dealer network are nontrivial and stable (as opposed to dealers being matched randomly).

2. Central dealers are more informed.

3. Central dealers charge higher spreads.

4. If there is less information to be learned, then dealer connections become weaker, and information is revealed more slowly.

5. The steady-state information revelation structure grows out of a "Big Bang." Revelation starts centrally with one market dominating and gradually evolves into a steady-state decentralized structure.

Although Hagströmer and Menkveld [2019] findings support many of their hypothesis, as they say, literature on how information percolates through the dealer-network and gets revealed is still relatively young and thus limited.

2.1.3 Granger casuality

Inspired by work of Wiener [1956], and introduced into the experimental practice by Granger [1969], Granger casuality was for a long time the predominant approach for identifying direction of coupling between two related variables, objects, events (Hlaváčková-Schindler et al. [2007], Papana et al. [2011]) and was widely used in many scientific fields such as biomedicine, neuroscience, economy and finance (Gençağa [2018]). Because of the variety of applications of this statistical concept, there has been a lot of different ways of interpretation of causality. One particularly important interpretation, is that causality can be understood in terms of "flow" among processes (Hlaváčková-Schindler et al. [2007]). Even though there has been no universally accepted definition of causality, which in fact has been the topic of long-standing controversy (Hlaváčková-Schindler et al. [2007], one can find a vast literature on the use of Granger causality to detect causal relations in financial markets.

A growing interest in applying the concept of Granger causality to many different field and phenomena, quickly revealed the three major intrinsic limitations associated with this statistical concept. First of all, as Granger [1969] states himself, the traditional Granger causality application is limited to linear systems, as it can only detect linear lead-lag relations (Gencaga et al. [2015], Eichler [2012]). Secondly, it can be employed to test cause-effect relationship between two variables at a time, thus it may lead to spurious conclusions about causal relationship in a system with three or more variables at play e.g. when both variables are driven by a third variable with different delays (Gencaga et al. [2015]). Finally and most importantly, it does not quantify the strength of the causality inferred (Hlaváčková-Schindler et al. [2007], Dimpfl and Peter [2013]).

With time many of these limitations were addressed by the research community. For example, the first limitation was overcome with the nonlinear extension of the Granger causality in the field of economics proposed by Baek and Brock [1992] and Hiemstra and Jones [1994]

(Hlaváčková-Schindler et al. [2007]). A different nonlinear extension is also provided by Chen et al. [2004] who uses local linear predictions and Ancona et al. [2004] who's nonlinear predictors are based on radial basis functions. Similarly, the second limitation was also addressed by Chen et al. [2004], who propose a metric called "conditional extended Granger casualit" that can be applied to multivariate nonlinear systems and allows to determine if the detected casual relation is direct or mediated by a different process (Chen et al. [2004]). Moreover, in the literature one can also find a variety of measures that were developed using Granger causality (Papana et al. [2011]), many tailored specifically to their applications. As Granger causality does not stand as the main focus of this thesis, an interested reader is referred to a very extensive survey dedicated to inference and causality in economic time series models by Geweke [1984].

As Hlaváčková-Schindler et al. [2007] note, in many disciplines the aim it to not only detect but also quantify the strength of the relationship inferred. This is what attracted the research community towards the information-theoretic measures building upon the concept of entropy introduced by Claude Shannon in his seminal work Shannon [1948].

2.2 Information-theoretic approach to modelling directed information flow

For a long time, a common approach to cope with nonlinear systems was to use an information-theoretic metric called mutual information and introduced by Shannon [1948]. Mutual information allows measuring the information shared between random variables, broadly by quantifying the statistical distance between the joint probability distribution of random variables and the product of their marginal probability distributions (Schreiber [2000]). Mutual information found its application, for instance, in neuroscience to quantify how much information a neural response carries about the stimulus (Borst and Theunissen [1999], Theunissen et al. [1996]). It has also become a popular approach in feature selection in neural networks, allowing to assess the "information content" of features in classification tasks (Battiti [1994]) or in speech recognition Bahl et al. [1986]. Additionally, many authors employ mutual information as an alternative to correlation measure applicable to complex systems setting (Li [1990], Fiedor [2014]). However, the mutual information has one severe limitation; namely, it is a symmetric metric, and hence it does not determine the direction of relations inferred (Schreiber [2000]).

Directionality of the inferred relationships is a critical concept in many fields, such as neuroscience or genetics, where the goal is to infer causal connections between various elements of a system (D'haeseleer et al. [2000]). For example, in genetics, one would like to determine the impact of activation of a particular gene on the expression of other genes in the gene regulatory network (GRN) (Martínez et al. [2012], Meyer et al. [2007]). Therefore, a simple yet

powerful extension to the mutual information was introduced; the so-called time-delayed mutual information (TDMI).

In TDMI, the asymmetry is induced by delaying in time one of the variables with respect to the other one. This minor change provided the means to transform mutual information into a directional measure. Time-delayed mutual information found its application in neuroscience, for example quantifying information flows in corticomuscular interaction (Li et al. [2018], Jin et al. [2010]). Its superiority to Granger causality was exposed by Li et al. [2018], who showed that GC analysis fails to infer causal relationships even in minimal nonlinear systems. Consequently, they propose that in the presence of potential nonlinear interactions in the system, the TDMI metric should be used instead (Li et al. [2018]).

Inevitably, time-delayed mutual information comes with limitations. As Kaiser and Schreiber [2002] note, time-delayed mutual information does not ignore correlations due to common history or external factors (Schreiber [2000]). Therefore, Schreiber proposed a novel metric called transfer entropy, which *quantifies the exchange of information between two systems, separately for both directions, and, if desired, conditional to common input signals* (Schreiber [2000]). Broadly, for a pair of source and target processes, transfer entropy is time-delayed mutual information conditioned on the target's history. The conditioning filters out history or external driving forces that are common for the pair of processes.

Transfer entropy very quickly replaced mutual information and thus also found a wide array of applications in various scientific fields. It found its application in biology (Guo et al. [2021], Antonacci et al. [2020]), chemical engineering (Lee et al. [2020], Shu and Zhao [2013]), neuroscience (Vicente et al. [2011], Vakorin et al. [2011], Buehlmann and Deco [2010], Wibral et al. [2014a]), finance (Li et al. [2013], Sandoval [2014], Dimpfl and Peter [2013], Kwon and Yang [2008], Reddy and Sebastin [2008]) and many other fields. With the growing popularity of that metric, many investigations were performed where transfer entropy was compared to Granger causality. For example, Barnett et al. [2009] have formally shown that Granger causality and transfer entropy for Gaussian variables are equivalent up to a factor of two. Furthermore, Hlaváckova-Schindler [2011] extended upon these findings by proving equivalence of the two metrics for exponential Weinman, Gaussian mixtures, generalized normal and log-normal data distributions (Syczewska and Struzik [2015]).

Estimation of transfer entropy involves approximation of two transitional and one joint probability densities of two processes involved, usually based on a single realization of these processes. Thus, different transfer entropy estimators propose different approaches to approximating these probability densities. Hlaváčková-Schindler et al. [2007] provides a detailed account of many transfer entropy estimation techniques for both discrete and continuous random variables. Popular approaches are binning, adaptive partitioning, ranking, maximum likelihood, nearest neighbor, and neural network-based methods. Although the concept of nearest-neighbor-based

estimators for Shannon entropy was already explored over 60 years ago by Dobrushin [1958], it wasn't until the work of Kraskov et al. [2004] that this type of estimator became popular (Hlaváčková-Schindler et al. [2007]). KSG algorithm has recently received much attention, especially in neuroscience, given that it is numerically unbiased for finite samples and relatively robust in high-dimensional settings (Lizier [2014], Wibral et al. [2014b], Runge [2014])[3]. The superiority of the KSG estimator over other types of estimators is also recognized by Leisink [2019], who, in his master's thesis, conducted an in-depth comparison of various transfer entropy estimators. With the developments in KSG algorithms, for example, an extension of the algorithm to conditional mutual information by Frenzel and Pompe [2007], the application of transfer entropy to network inference algorithm has recently gained much attention, especially in bioinformatics and neuroscience (Meyer et al. [2007]).

Network inference also referred to as reverse engineering, is currently an open problem (Li et al. [2011]). It is a process of inferring the structure of interaction of complex systems components from multivariate observational data (Villaverde et al. [2014]). The goal of this approach is to uncover the underlying dynamics of complex systems in order to understand better the effect that each element has on the network as a whole (Hecker et al. [2009]). There are many different network inference algorithms, such as Boolean networks, Bayesian networks, or Association networks. Mutual information also founds its application in network inference algorithms; in particular, it is employed to weigh the network edges in two prevalent network inference algorithms: Relevance Networks proposed by Butte and Kohane [1999], and Context Likelihood of Relatedness algorithm by Margolin et al. [2006] (Hecker et al. [2009]).

The merit of the transfer entropy-based network inference algorithm is recognized by, e.g. Vicente et al. [2011], who conclude that transfer entropy is a robust measure for effective connectivity detection with neuroscience data. Similarly, Stetter et al. [2012] establish that transfer entropy-based network inference provides significantly more accurate results than Granger causality or correlation-based approaches. A lot more involved information-theoretic approach to network inference algorithm is presented by Novelli et al. [2019]. In this publication, the authors present a novel, greedy information-theoretic network inference algorithm that employs both transfer entropy and conditional transfer entropy. First, transfer entropy is used to determine statistically significant directional links between nodes in the network. Next, conditional transfer entropy is employed on the nodes with multiple statistically significant causal information contributors to filter out the insignificant ones if all significant contributions to that node are considered. Additionally, a non-uniform data embedding approach allows them to efficiently deal with high-dimensional data (Novelli et al. [2019]).

[3]Since the focus of this thesis is not to discuss different methods of transfer entropy estimation, an interested reader is referred an extensive overview of various transfer entropy estimators by Hlaváčková-Schindler et al. [2007].

In the literature, to the best of my knowledge, there are no publications on information-theoretic network inference algorithms employed in a financial setting, especially in the price discovery framework. Hence, applying state-of-the-art methods from neuroscience to infer the FX dealer-network is a unique idea on its own.

Chapter 3

Theory

The following chapter introduces the theory behind the econometric and information-theoretic approaches for the characterization of the information flow in the foreign exchange market's dealer network. To keep a clear distinction between the theory unique to each of the two approaches, the methods chapter is split into two main sections.

Section 3.1 briefly introduces fundamental intuition behind stochastic processes as well as the notions of stationarity and cointegration. The fundamental concepts introduced in Section 3.1 are based on Lütkepohl [2005] and Chatfield and Xing [2019]. Next, the theory unique to the econometric approach and the metrics proposed by Hägstromer and Menkveld are treated in detail. Finally, the section is concluded with a simple illustrative example of information revelation approach applied to two cointegrated processes.

Section 3.2 first introduces the basic concepts of information theory and provides intuition behind them. This is followed by a thorough discussion of more involved metrics, such as transfer entropy and conditional transfer entropy in Section 3.2.2. The theory introduced in Section 3.2 is based on MacKay and Mac Kay [2003], Bossomaier et al. [2016], Thomas and Joy [2006], Michalowicz et al. [2013] and Wibral et al. [2014b].

3.1 Modelling information revelation

The first approach that will be introduced is the econometric approach proposed by Hägstromer and Menkveld in their most recent publication entitled "Information Revelation in Decentralized Markets" (Hagströmer and Menkveld [2019]). In this work, the authors attempt to answer a fundamental question of how the information gets revealed in the decentralized markets by constructing, what they call, empirical maps of information revelation. An example of such information revelation network map is presented in Figure 3.1.

Hagströmer and Menkveld [2019] research is based on the foreign exchange spot rates data for a 8 dealers and one electronic communication network EBS. The map represents the FX

market's dealer-network captured in a specific time window at a particular multivariate impulse-response function's (MIRF) lag τ. The map is an undirected network graph located in a unit circle. The position and color of each node, as well as the thickness and color the edges, conveniently visualize the three key metrics; the price inefficiency, Hasbrouck's information share and the bilateral connection (partial correlation), respectively. Each metric is a nonlinear function of the parameters estimated with a vector error correction model. Without going into much details, the price inefficiency metric is used to determine the position of nodes (markets) in reference to the center of the unit circle. The distance of each market from the center of the unit circle represents the degree of inefficiency of it's quotes as compared to the long-term efficient price. The thickness and the color of network edges (bilateral connections) between pairs of markets are proportional to partial correlations of their price changes (Hagströmer and Menkveld [2019]). The color of each vertex directly corresponds to the Hasbrouck's information share, which measures the contribution of the market to the price discovery.

3.1.1 Introduction to cointegrated stochastic processes

To properly introduce the information revelation framework, it is necessary to first establish the fundamental intuition behind stochastic processes and introduce the notions of stationarity and cointegration.

Definition 3.1 (**Stochastic process**)**.** A stochastic process is a process that evolves in time with respect to some probabilistic laws (Chatfield and Xing [2019]). To illustrate further, let y_t denote a vector representing one realization of a random variable Y_t, where realization is simply a time series of observations made at discrete times t:

$$y_t = \begin{bmatrix} y_0 & y_1 & \dots & y_{T-1} & y_T \end{bmatrix}, \tag{3.1}$$

where $t \in [0, T]$, hence $t = 0, 1, \dots, T$, whereas y_i simply indicates the value associated with the process at $t = i$.

The time series y_t is just one particular realization of the stochastic process. The same stochastic process can have multiple realizations, for example, when multiple copies of the same system exist (Wibral et al. [2014b]). However, only a single realization of a random variable is usually available, especially when real data is considered. For instance, in the FX market, there is only one FX spot rate available from dealer A for currency pair B/C at a particular point in time t. Therefore, the analysis of the underlying probabilistic model of such a process is restricted to an analysis of its single realization (Chatfield and Xing [2019]). As it will become

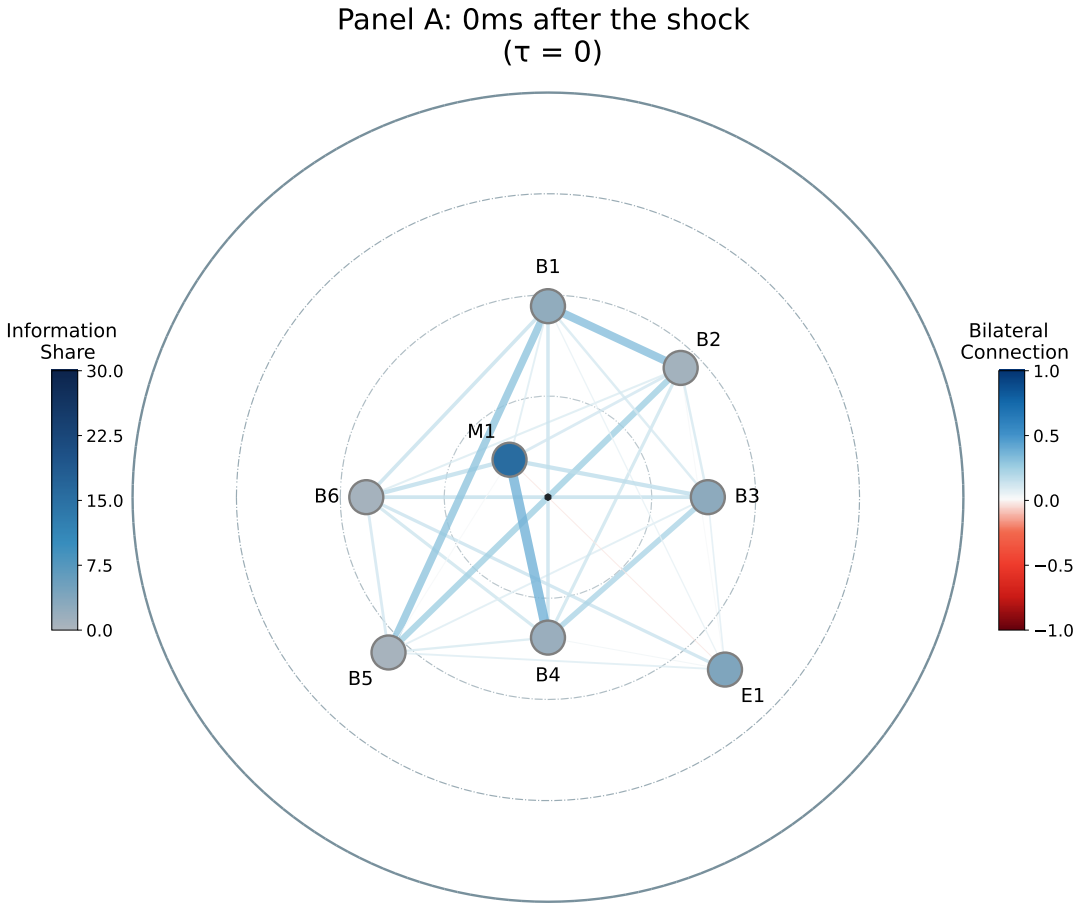

FIGURE 3.1: An example of the econometric information revelation map. The visualization technique used in this thesis is a significantly adapted visualization used by Hagströmer and Menkveld [2019].

clear later in the Section 3.2, this is a fundamental concept for information-theoretic quantities, which requires that investigated stochastic processes are stationary.

We can distinguish two fundamental classes of stochastic processes: stationary and non-stationary processes. In a broad sense, stochastic process is stationary if the statistical properties of it's one realization are the same as in any other realization of the same process i.e. the statistical properties of the stochastic process are independent of time. However, for the purposes of this thesis it is important to introduce here a proper mathematical definition of weak stationary process.

Definition 3.2 (**Weak stationary stochastic process**)**.** A (discrete) stochastic process is considered to be second-order stationary (weakly stationary) if its first and second moments are time invariant, and its autocovariance function depends only on the lag between two time

periods (Lütkepohl [2005], Chatfield and Xing [2019]). Formally,

$$\mathbb{E}[y_t] = \mu \qquad \forall\, t \in \mathbb{N}^0 \tag{3.2}$$

$$\wedge$$

$$\mathrm{Cov}[y_t, y_{t+\tau}] = \gamma(\tau) \qquad \forall\, t \wedge \tau \in \mathbb{N}^0 \tag{3.3}$$

where τ stands for a lag and $\gamma(\tau)$ indicates the autocovariance coefficient at lag τ (Chatfield and Xing [2019]). Note that since we allow $\tau = 0$, this definition also implies that the second moment (variance) needs to be constant and finite.

The concept of weak stationarity does not place any additional restrictions on any higher moments, which is the case when one considers the strict stationarity.

Definition 3.3 (**Strict stationary stochastic process**)**.** A (discrete) stochastic process is considered to be strictly stationary if all of its moments are time invariant (Chatfield and Xing [2019]). In other words, the stochastic process is strictly stationary if its underlying probability distribution is timeshift-invariant.

Note that the stationarity assumption is necessary for the econometric model and information-theoretic network inference algorithm introduced in this thesis. This, however, will become clear in Chapter 4.

Given that we have already defined stationary stochastic process, the question that naturally arises here is how do we define a non-stationary stochastic process. Simply, if a stochastic process is not stationary in the sense defined in Definition 3.2 or Definition 3.3 it is considered to be a non-stationary stochastic process (Gujarati and Porter [2003]).

Definition 3.4 (**Integrated stochastic process**)**.** A non-stationary stochastic process is considered to be integrated of order d, if it needs to be differenced d times to be transformed into a stationary series (Chatfield and Xing [2019]). Formally, a non-stationary time series integrated of order d is represented in the following manner

$$y_t \sim I(d) \tag{3.4}$$

where d stands for the order of integration. The order of integration of the series can be reduced by simply differencing the time series, as follows

$$\Delta y_t = y_t - y_{t-1} = (1 - B)y_t \qquad \sim I(d-1) \tag{3.5}$$

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = (1 - B)^2 y_t \qquad \sim I(d-2) \tag{3.6}$$

$$\Delta^d y_t = (1 - B)^d y_t \qquad \sim I(0) \tag{3.7}$$

where B is the backward shift operator, which simply shifts the data back by one period, and Δ stands for the difference operator. This simple operation (usually) allows one to remove trends from a non-stationary series, and thus transform it into a stationary one (Bisgaard and Kulahci [2011], Chatfield and Xing [2019]).

On the other hand, when multiple related time series are examined, and one suspects that there exists a long-term equilibrium relationship between them, then stochastic processes should be tested for cointegration (Chatfield and Xing [2019]).

Definition 3.5 (**Cointegrated stochastic processes**)**.** Two time series integrated of order d are said to be co-integrated of order $CI(d, b)$ if there exists a linear combination of the two series that has an order of integration $I(d - b)$, where $b > 0$ (Engle and Granger [1987]). To better illustrate, consider the following example of two stochastic process $y_{1,t}$ and $y_{2,t}$, each integrated of order one:

$$y_{1,t} = y_{1,t-1} + \epsilon_{1,t} \qquad \sim I(1) \tag{3.8}$$

$$y_{2,t} = y_{1,t-2} + \epsilon_{2,t} \qquad \sim I(1) \tag{3.9}$$

where ϵ_{1t} and ϵ_{2t} are zero-mean, uncorrelated disturbances that are independently and identically distributed (iid.) (Hasbrouck [1995]). In the above example, each time series independently has an order of integration one. Since the difference between these two processes is a sum of stationary random variables, it clearly has an order of integration zero (Hasbrouck [1995]).

$$z_t = (y_{1,t} - y_{2,t}) = y_{1,t} - (y_{1,t-2} + \epsilon_{2,t}) = \epsilon_{1,t} + \epsilon_{1,t-1} - \epsilon_{2,t} \qquad \sim I(0) \tag{3.10}$$

Thus, the two series are co-integrated of order $CI(1, 1)$ (Chatfield and Xing [2019]).

Note that the system of stochastic cointegrated processes represented by Equations (3.8) and (3.9) will be used to illustrate various concepts in the remainder of this section. The cointegration relation from Equation (3.10) between the processes $y_{1,t}$ and $y_{2,t}$ can also be represented in the following vectorized form

$$z_t = (y_{1,t} - y_{2,t}) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \boldsymbol{\alpha}^\intercal \mathbf{y}_t \qquad \sim I(0) \tag{3.11}$$

thus revealing the cointegration vector $\boldsymbol{\alpha}$, also known as co-integrating vector (Chatfield and Xing [2019], Lütkepohl [2005], Engle and Granger [1987])). The intuition behind the cointegration relationship is as follows. The time series $y_{1,t}$ and $y_{2,t}$ treated independently drift far and widely away from zero (even if $y_{1,0} = y_{2,0} = 0$), thus their variance tends to infinity as t tends to infinity (Engle and Granger [1987]). However, the cointegration relationship suggests

that a significant part of long-term components of $y_{1,t}$ and $y_{2,t}$ cancels out, in other words they do not diverge without bound from each other (Engle and Granger [1987], Hasbrouck [1995]). Therefore, even if the processes $y_{1,t}$ and $y_{2,t}$ temporarily drift away from each other, in a long run they will converge, which suggests an existence of a long run equilibrium. An example of to cointegrated time series is presented in Figure 3.2. Finally, note that the above example represents a very basic example of cointegration. In case many processes are modelled, then multiple cointegration relations between variables may exist.

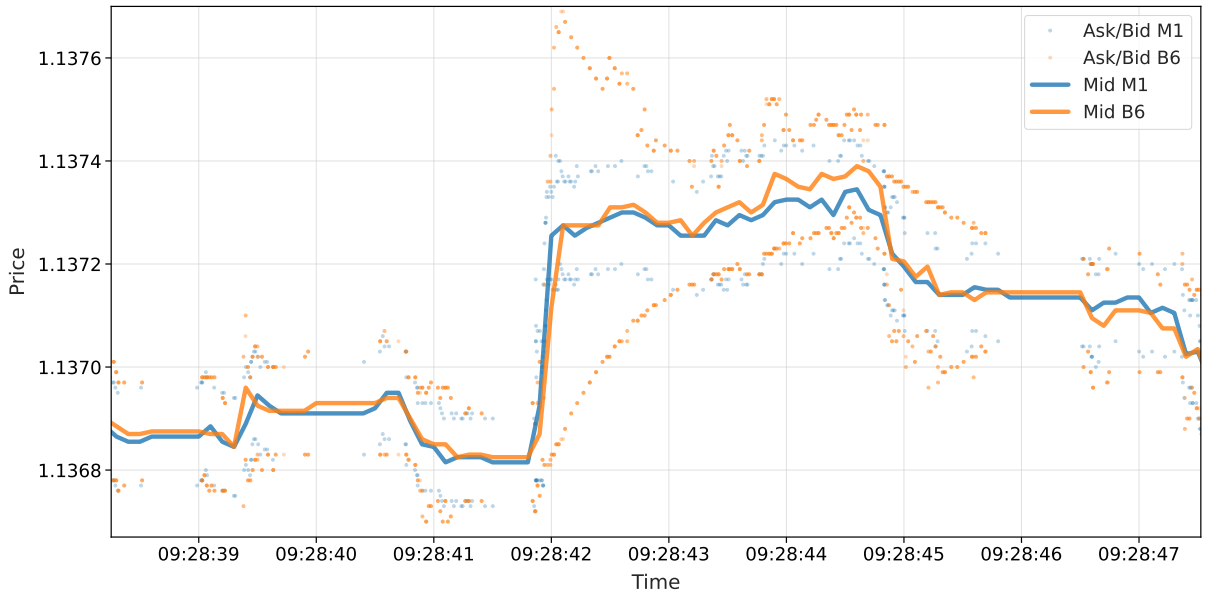

FIGURE 3.2: EUR/USD mid prices of market maker (M1) and bank (B6) at 9h on 3rd September 2020. Figure reveals the movement of mid prices of two dealers in time. The two mid prices very closely follow each other thus illustrating an example of cointegrated processes. Furthermore, one can observe that the mid prices temporarily drift away from each other at approximately 9:28:44 and converge again one second later. Ask and bid prices are presented in its raw form, whereas mid prices are resampled to 100ms. (Note that all times are presented in UTC+01:00 and the time format convention used is [hh:mm:ss.fff], where f represent milliseconds. Names of all dealers are mapped, for more details see Table 4.1.)

The notion of stationarity is fundamental for time-series analysis, as many econometric models, such as autoregressive (AR), or autoregressive–moving-average (ARMA) models, rely on the assumption that the first and second moments of the time series modelled remains constant in time. Therefore, if such models are employed, it is necessary to ensure that any non-stationary time series is transformed into stationary ones. Otherwise, the analysis may lead to spurious regression, with the estimated coefficients in the regression appearing falsely significant. To model cointegrated time series, e.g., the vector moving average model (VMA) or the vector-error correction model (VECM) can be employed.

3.1.2 Vector-error correction model

The concept of cointegration is very closely related to the error correction models. This subsection will reveal how the error correction model incorporates the cointegration relation between the modelled processes. Additionally, we will start putting the stochastic process in the context of the foreign exchange market, where each time series represents the time evolution of price quotes of a particular dealer. From now on, dealers will be referred to as markets.

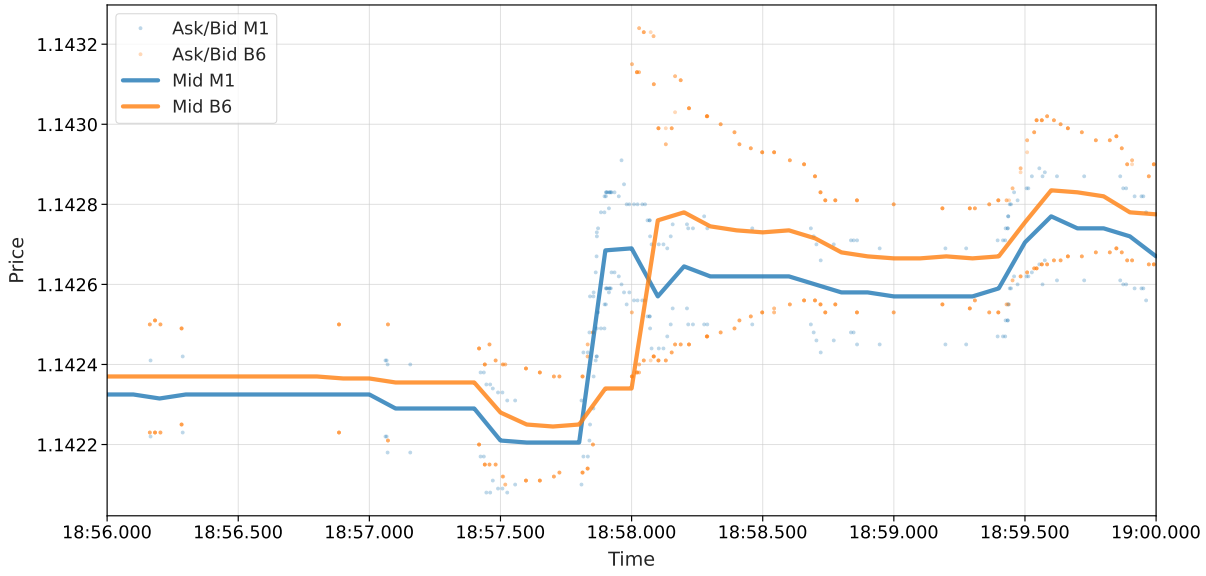

FIGURE 3.3: EUR/USD mid prices of market maker (M1) and bank (B6) at 14h on 3rd September 2020. Figure reveals how bank B6 updates its price with approximately 200ms delay as compared to the market maker. Ask and bid prices are presented in its raw form, whereas mid prices are resampled to 100ms.

Let us assume that $y_{1,t}$ represents the time-series of mid prices[1] of quotes provided by market 1 (M1)[2], whereas the $y_{2,t}$ provided by market 2 (M2) for the same currency pair, e.g., EUR/USD. For simplicity, we will further refer to mid-prices as prices or quotes. Additionally, let us assume that there exists an equilibrium relation between the two quotes $y_{1,t} = y_{2,t}$, simply because of the presence of arbitrage opportunity if $y_{1,t} \neq y_{2,t}$ at any time t. Now, if at any time t, $y_{1,t} \neq y_{2,t}$, both markets will have an incentive to correct the price disparity in their next quote, for the aforementioned reason. Thus, in the simplest scenarios, we can assume that the changes in markets' prices at time t will depend on the difference between their prices (an error) in the previous time period $t-1$. Hence, formally one could represent this relationship

[1] Mid price is the price between dealer's ask price and bid price. Thus, the mid-price represents the exact mid-point between dealer's ask and bid quotes.

[2] Note that M1 does not refer to a market on Figure 3.3.

in the following manner (Lütkepohl [2005]),

$$\Delta y_{1,t} = \alpha_1 \left(y_{1,t-1} - y_{2,t-1} \right) + \epsilon_{1,t} \tag{3.12}$$

$$\Delta y_{2,t} = \alpha_2 \left(y_{1,t-1} - y_{2,t-1} \right) + \epsilon_{2,t} \tag{3.13}$$

where α_1 and α_2 intuitively represent the "strength" of the markets' responses to the deviation from the expected equilibrium $y_{1,t} = y_{2,t}$ at the previous time step. Let us for now assume that $\epsilon_{1,t}$ and $\epsilon_{2,t}$ stand for white noise errors. The term $(y_{1,t-1} - y_{2,t-1})$ is the so called error-correction term, thus the Equations (3.12) and (3.13) represent an error correction model (Lütkepohl [2005]).

The relationship between processes presented in Equations (3.12) and (3.13) can be extended into a more general form of error correction model which also accounts for the lagged changes in each markets' prices. Additionally, let us assume a more general form of the equilibrium relation $y_{1,t} = \beta_1 y_{2,t}$,

$$\Delta y_{1,t} = \alpha_1 \left(y_{1,t-1} - \beta_1 y_{2,t-1} \right) + \gamma_{11,1} \Delta y_{1,t-1} + \gamma_{12,1} \Delta y_{2,t-1} + \epsilon_{1,t} \tag{3.14}$$

$$\Delta y_{2,t} = \alpha_2 \left(y_{1,t-1} - \beta_1 y_{2,t-1} \right) + \gamma_{21,1} \Delta y_{1,t-1} + \gamma_{22,1} \Delta y_{2,t-1} + \epsilon_{2,t} \tag{3.15}$$

where $\Delta y_{i,t-m}$ stands for the m lagged difference of i^{th} market's quotes, and γ's are simply coefficients of differenced terms.

Suppose that $y_{1,t} \wedge y_{2,t} \sim I(1)$. In this case it is apparent that all differenced terms in Equations (3.14) and (3.15) are stationary. If we further assume that $y_{1,t}$ and $y_{2,t}$ are $CI(1,1)$, then clearly the error-correction term induces the cointegration relation between the variables, and thus make these terms stationary as well (Lütkepohl [2005]). In this simple manner, we have exposed the presence of cointegration in vector-error correction model and ensured that all terms modelled are stationary. The model presented in Equations (3.14) and (3.15) can be also formulated in vector and matrix notation, which leads to the vector-error correction model representation with the following form:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^{\intercal} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_2 \Delta \boldsymbol{y}_{t-2} + \ldots + \boldsymbol{\Gamma}_M \Delta \boldsymbol{y}_{t-M} + \boldsymbol{\epsilon}_t \tag{3.16}$$

where $\boldsymbol{\alpha}$ is $(n \times n)$ loading matrix, and $\boldsymbol{\beta}$ is $(n \times n)$ cointegration matrix, and n stands for the number markets modelled. Furthermore, $\boldsymbol{\Gamma}_i$ is a $(n \times n)$ coefficient matrix of i^{th} lagged differences $\Delta \boldsymbol{y}_{t-i}$ and $\boldsymbol{\epsilon}$ is a $(n \times 1)$ column vector of residuals. Finally, one can also fully vectorize Equation (3.16) to make it convenient for programming purposes:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^{\intercal} \boldsymbol{y}_{t-1} + \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \dots & \boldsymbol{\Gamma}_M \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{y}_{t-1} \\ \Delta \boldsymbol{y}_{t-2} \\ \vdots \\ \Delta \boldsymbol{y}_{t-M} \end{bmatrix} + \boldsymbol{\epsilon}_t. \tag{3.17}$$

3.1.3 Vector moving average

As previously mentioned, the cointegrated processes can also be modelled with the vector moving average model (VMA). To demonstrate it, let us again recall an example of cointegrated markets' prices presented in Equations (3.8) and (3.9). Further, recall that the system of these two equations represents the time evolution of the cumulative price changes of two markets' quotes that can be modelled with the following equations $y_{1,t} = y_{1,t-1} + \epsilon_{1,t}$ and $y_{2,t} = y_{1,t-2} + \epsilon_{2,t}$. Note that since they are cumulative, they are integrated of order one. This time, let us consider the dynamics of price changes (differenced prices) of the two markets over time,

$$\Delta y_{1,t} = y_{1,t} - y_{1,t-1} = \epsilon_{1,t} \tag{3.18}$$

$$\Delta y_{2,t} = y_{2,t} - y_{2,t-1} = (y_{1,t-2} + \epsilon_{2,t}) - (y_{1,t-3} + \epsilon_{2,t-1}) \tag{3.19}$$

since $(y_{1,t-2} - y_{1,t-3})$ simply reduces to $\epsilon_{1,t-2}$, the system can be represent only in terms of iid. zero-mean disturbances. This leads to the following system of equations:

$$\Delta y_{1,t} = \epsilon_{1,t} \tag{3.20}$$

$$\Delta y_{2,t} = \epsilon_{2,t} - \epsilon_{2,t-1} + \epsilon_{1,t-2} \tag{3.21}$$

This representation, is also known as moving average (MA) representation. The changes of each market's prices are defined only in terms of lagged disturbances. The above system of equations can be represented in a vectorized form thus revealing the vector moving average (VMA) representation

$$\begin{aligned} &\Delta \boldsymbol{y}_t = \boldsymbol{I}\boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} \\ &\text{where } \boldsymbol{\Psi}_1 = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } \boldsymbol{\Psi}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \end{aligned} \tag{3.22}$$

where $\Delta \boldsymbol{y}_t$ is an (2×1) vector of price changes, $\boldsymbol{I}$ is an (2×2) identity matrix, $\boldsymbol{\Psi}_1$ is an (2×2) VMA coefficient matrix for disturbances at lag one, whereas $\boldsymbol{\Psi}_2$ is a VMA coefficient matrix

for disturbances at lag two. In a more general form, the VMA model can be formulated as:

$$\Delta \boldsymbol{y}_t = \boldsymbol{I}\boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t-2} + \ldots + \boldsymbol{\Psi}_{N_L}\boldsymbol{\epsilon}_{t-N_L} = \boldsymbol{\Psi}(N_L)\boldsymbol{\epsilon}_t \tag{3.23}$$

where $\Delta \boldsymbol{y_t}$ represents a $(n \times 1)$ vector of price changes where n is the number of markets modelled. Additionally, $\boldsymbol{I}$ is a $(n \times n)$ identity matrix, $\boldsymbol{\Psi}_{(\cdot)}$ is a $(n \times n)$ VMA coefficient matrix, and N_L stands for the number of lags. Finally, $\boldsymbol{\epsilon}_t$ is a $(n \times n)$ vector of disturbances.

3.1.4 Multivariate impulse response function

This subsection will illustrate how VMA representation can be utilized to determine the short-term and long-term responses of the system to shocks, where system is simply the network of markets. For this purpose, instead of treating $\boldsymbol{\epsilon}_t$ as a vector of disturbances, let us consider it as a vector of innovations in markets' prices (shocks). The shock could be a consequence of, e.g. market gathering new information about one of the currencies that affects its fundamental value, or simply just changing mind about the fair exchange rate for particular currency pair.

To gain this new perspective provided by the VMA representation, again let us consider the system of stochastic processes that was presented in Equations (3.8) and (3.9), i.e. $y_{1,t} = y_{1,t-1} + \epsilon_{1,t}$ and $y_{2,t} = y_{1,t-2} + \epsilon_{2,t}$. Let us further assume that at $t = 0$ the markets agree on the EUR/USD exchange rate, formally $y_{1,0} = y_{2,0} = 0$ (markets' quotes are initially at equilibrium). Furthermore, suppose that at $t = 0$ market 1 learns new information and changes its price by one unit value. On the other hand, market 2 doesn't learn any new information and thus its quote remains the same. In other words, we introduce a unit shock to the exchange rate of market 1, i.e. $\epsilon_{1,0} = 1$, while the same shock is not introduced to price of market 2, i.e. $\epsilon_{2,0} = 0$. Let us further assume that there are no other shocks in this system for the rest of the time, i.e. for $t \neq 0$ $\epsilon_{1,t} = 0 \wedge \epsilon_{2,t} = 0$. Now, let us track how the introduced shock propagates through the system at each time step of the simulation. At $t = 0$,

$$\Delta \boldsymbol{y}_{0,M1} = \boldsymbol{I}\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.24}$$

where $M1$ subscript indicates that the changes that we are tracking are a consequence of the shock being introduced to the market 1 (M1) price. Equation (3.24) reveals the contemporaneous price change in the zeroth time step, i.e. the new information is directly impounded into the quotes of the first market, while the second market maintains the equilibrium price.

$$\Delta \boldsymbol{y}_{1,M1} = \boldsymbol{I}\boldsymbol{\epsilon}_1 + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{3.25}$$

In the first time step, we do not observe any responses from any of the markets.

$$\Delta \boldsymbol{y}_{2,M1} = \boldsymbol{I}\boldsymbol{\epsilon}_2 + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_1 + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{3.26}$$

In the second time step, the second market updates its quote, as the market impounds the entire shock into its price. Notice that this response happens two time steps after the shock was introduced. Since, $\boldsymbol{\Psi}_i = \boldsymbol{O}_{2\times 2}$ for $i > 2$, and $\boldsymbol{\epsilon}_t = \boldsymbol{0}$ for $t \neq 0$, hence $\Delta \boldsymbol{y}_j = \boldsymbol{0}$ for $j > 2$. Thus, there is no need to iterate over more time steps as the system has reached the new equilibrium.

We can also compute the cumulative price changes of each market for all time step of the simulation by summing up the price changes, hence

$$\boldsymbol{y}_{1,M1} - \boldsymbol{y}_0 = \Delta \boldsymbol{y}_{0,M1} + \Delta \boldsymbol{y}_{1,M1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.27}$$

$$\boldsymbol{y}_{2,M1} - \boldsymbol{y}_0 = \Delta \boldsymbol{y}_{0,M1}, + \Delta \boldsymbol{y}_{1,M1} + \Delta \boldsymbol{y}_{2,M2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.28}$$

Equation (3.27) exposes the short-term (instant) cumulative price changes, whereas Equation (3.28) uncovers the long-term cumulative price changes. Since cumulative price changes of both markets are the same, we can conclude that a new equilibrium (long-term) price has been reached. Thus, as a result of a unit shock in the first market's price, the system transitioned to new equilibrium state with a new common price.

Now, let us consider the same simulation, but this time we introduce a unit shock to the second market, as follows

$$\Delta \boldsymbol{y}_{0,M2} = \boldsymbol{I}\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{3.29}$$

The zeroth time step is analogous to the one observed in the previous simulation, however

$$\Delta \boldsymbol{y}_{1,M2} = \boldsymbol{I}\boldsymbol{\epsilon}_1 + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \tag{3.30}$$

the first time step it can be observed, that the second market instantly corrects price differential and reduces the price back to initial equilibrium price.

$$\Delta \boldsymbol{y}_{2,M2} = \boldsymbol{I}\boldsymbol{\epsilon}_2 + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_1 + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{3.31}$$

The above simulation did not show any response of the first market to the shock introduced to

the second market price. This is not surprising, simply because there is no term in Equation (3.8) that relates the first market's price change to the second market's price change. Accumulating the price changes from all of the time steps,

$$\boldsymbol{y}_{1,M2} - \boldsymbol{y}_0 = \Delta\boldsymbol{y}_{0,M2} + \Delta\boldsymbol{y}_{1,M2} = \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}0\\-1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \tag{3.32}$$

$$\boldsymbol{y}_{2,M2} - \boldsymbol{y}_0 = \Delta\boldsymbol{y}_{0,M2} + \Delta\boldsymbol{y}_{1,M2} + \Delta\boldsymbol{y}_{2,M2} = \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}0\\-1\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \tag{3.33}$$

It appears, that after both one and two time steps the cumulative change of both markets' prices is simply zero, thus there is no change in the equilibrium state of the system.

Finally, let us consider cumulative sum of VMA coefficient matrices,

$$\tilde{\boldsymbol{\Psi}}_1 = \boldsymbol{I} + \boldsymbol{\Psi}_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & -1\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} \tag{3.34}$$

$$\tilde{\boldsymbol{\Psi}}_2 = \boldsymbol{I} + \boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & -1\end{bmatrix} + \begin{bmatrix}0 & 0\\1 & 0\end{bmatrix} = \begin{bmatrix}1 & 0\\1 & 0\end{bmatrix} \tag{3.35}$$

where $\tilde{\boldsymbol{\Psi}}_\tau$ stands for the cumulative sum of VMA coefficients after τ time steps from the shock. An observant reader will notice that the cumulative sum of VMA coefficient retains information about the short-term and long-term responses of both markets. Notice that

$$\tilde{\boldsymbol{\Psi}}_1 = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}\boldsymbol{y}_{1,M1} - \boldsymbol{y}_0 & \boldsymbol{y}_{1,M2} - \boldsymbol{y}_0\end{bmatrix} \tag{3.36}$$

$$\tilde{\boldsymbol{\Psi}}_2 = \begin{bmatrix}1 & 0\\1 & 0\end{bmatrix} = \begin{bmatrix}\boldsymbol{y}_{2,M1} - \boldsymbol{y}_0 & \boldsymbol{y}_{2,M2} - \boldsymbol{y}_0\end{bmatrix} \tag{3.37}$$

That is, the first and second row of the first column of $\tilde{\boldsymbol{\Psi}}_1$ matrix corresponds to a short-term cumulative price changes of the first and second market, respectively, to a unit shock in the first market's quote. Analogously, the second column of $\tilde{\boldsymbol{\Psi}}_1$ corresponds to a short-term cumulative price changes of the first and second market, to a unit shock in the second market's quote (Pesaran and Shin [1998]). In the same manner, columns of $\tilde{\boldsymbol{\Psi}}_2$ matrix correspond to long-term cumulative price changes for both markets. Additionally, one can notice that each row of the $\tilde{\boldsymbol{\Psi}}_2$ is exactly the same, which aligns with cointegration relationship that ensures that the long-run price changes are the same for all markets (Hasbrouck [1995]). Let us formally define ψ as the common row vector of $\lim_{\tau\to\infty} \tilde{\boldsymbol{\Psi}}_\tau$,

$$\lim_{\tau \to \infty} \tilde{\boldsymbol{\Psi}}_\tau = \begin{bmatrix} \boldsymbol{\psi} \\ \vdots \\ \boldsymbol{\psi} \end{bmatrix} \tag{3.38}$$

where $\boldsymbol{\psi}$ is an $1 \times n$ vector representing long-run cumulative price changes common for all of the markets. The above detailed study offers some important insights about the properties of VMA coefficient matrices. The three central properties of the VMA coefficient matrices that need to be stated more clearly:

1. VMA coefficient matrix ($\boldsymbol{\Psi}_i$) contains information about the strength and the direction of the response of all variables at i^{th} time step after the shock. Therefore, it intrinsically stores the information about the responsiveness of the variables and the dependencies between variables.

2. The cumulative sum of VMA coefficient matrices ($\tilde{\boldsymbol{\Psi}}_\tau$) reveals the cumulative short-term response of variables to the shock. Namely, it exposes by how much each market adjusted its prices until τ time steps from the shock.

3. The $\lim_{\tau \to \infty} \tilde{\boldsymbol{\Psi}}_\tau$, and $\boldsymbol{\psi}$ provide the information about the long-term response of variables to a unit shock. Additionally, it provides a measure of how much a shock in one market's quotes impacts the long-term common price in all of the markets.

In this subsection, we were able to observe how in systems with cointegrated processes change in one variable can impact other variables over time. The last step necessary to fully comprehend Hasbrouck's framework is to connect the notions of short-term and long-term responses to the common trend representation proposed by Stock and Watson [1988].

3.1.5 Common trend representation

To discuss the common trend representation of cointegrated processes, let us again consider the processes from Equations (3.8) and (3.9). This time however, let us consider the system in terms of price change dynamic, previously presented in Equations (3.20) and (3.21) that is, $\Delta y_{1,t} = \epsilon_{1,t}$ and $\Delta y_{2,t} = \epsilon_{2,t} - \epsilon_{2,t-1} + \epsilon_{1,t-2}$. In order to expose the common trend representation, consider integrating the equations with respect to time, as follows

$$\sum_{x=0}^{t} \Delta y_{1,x} = \sum_{x=0}^{t} \epsilon_{1,x} \tag{3.39}$$

$$\sum_{x=0}^{t} \Delta y_{2,x} = \sum_{x=0}^{t} \epsilon_{1,x-2} + \sum_{x=0}^{t} (\epsilon_{2,x} - \epsilon_{2,x-1}) \tag{3.40}$$

A basic transformation of the above system of equations, leads to the following representation (Stock and Watson [1988]),

$$y_{1,t} = y_{1,0} + \left(\sum_{x=0}^{t} \epsilon_{1,x}\right) \tag{3.41}$$

$$y_{2,t} = y_{2,0} + \left(\sum_{x=0}^{t} \epsilon_{1,x}\right) + (-\epsilon_{1,t} - \epsilon_{1,t-1} + \epsilon_{2,t} - \epsilon_{2,0}) \tag{3.42}$$

which reveals the common trend component $\left(\sum_{x=0}^{t} \epsilon_{1,x}\right)$. Further vectorization of the equations results in the following

$$\boldsymbol{y}_t = \boldsymbol{y}_0 + \mathbf{1}\left(\sum_{x=0}^{t} \begin{bmatrix} 1 & 0 \end{bmatrix} \boldsymbol{\epsilon}_x\right) + \boldsymbol{\Psi}^*(N_L)\boldsymbol{\epsilon}_t \tag{3.43}$$

where $\boldsymbol{\Psi}^*(N_L)$ stands for a matrix polynomial in the lag operator, and accounts for so called transitory effects (Hasbrouck [1995]). Recall that in the long-term response investigation Equation (3.37) we determined that for this same system of equations the long-term response of variables to shocks is $\boldsymbol{\psi} = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Thus Equation (3.43) reduces to the following common trend representation

$$\boldsymbol{y}_t = \boldsymbol{y}_0 + \mathbf{1}\sum_{i=1}^{t} \boldsymbol{\psi}\boldsymbol{\epsilon}_i + \boldsymbol{\Psi}^*(N_L)\boldsymbol{\epsilon}_t \tag{3.44}$$

The above representation reveals a very important insight. The price series of cointegrated markets can be represented as a sum of three independent components (Stock and Watson [1988], Lütkepohl [2005], Warne [1993]).

The first component on the right-hand-side of Equation (3.44) is a vector of initial values that account for any non-stochastic initial disparities between prices in each market. For example, if the price vector involves both bid and ask quotes, the disparities between these terms in $\boldsymbol{y}_0$ will reflect the average spread (Hasbrouck [1995]).

The second component, is a product of a scalar random walk and a unit vector. This term represents a random-walk component that is shared by all markets prices (Stock and Watson [1988]). Recall that $\boldsymbol{\psi}$ contains information about the long-term common response of all market prices to a unit innovation in each market. Clearly, $\boldsymbol{\psi}$ weights innovations $\boldsymbol{\epsilon}_t$, by their respective impacts on the long-term common response. Thus, the product $\boldsymbol{\psi}\boldsymbol{\epsilon}_t$ simply provides information about the overall long-term response of market prices to all shocks captured in $\boldsymbol{\epsilon}_t$ at time t. Hasbrouck refers to the term $\boldsymbol{\psi}\boldsymbol{\epsilon}_t$ as *the component of the price change that is permanently impounded into the security price and presumably due to new information* (Hasbrouck [1995]). This common random-walk component captures changes to, what Hasbrouck's calls, an implicit unobservable efficient price common to all markets (Hasbrouck [1995]). The efficient price is the

price to which the market's would converge to, if one would isolate all the future disturbances from the system and let it reach an equilibrium.

Additionally, note that if we consider the fact that $\boldsymbol{\epsilon_t}$ are zero-mean uncorrelated disturbances, thus the common trend component is a martingale, that forecasts the efficient price using the information available until time t (Hirsa and Neftci [2013]).

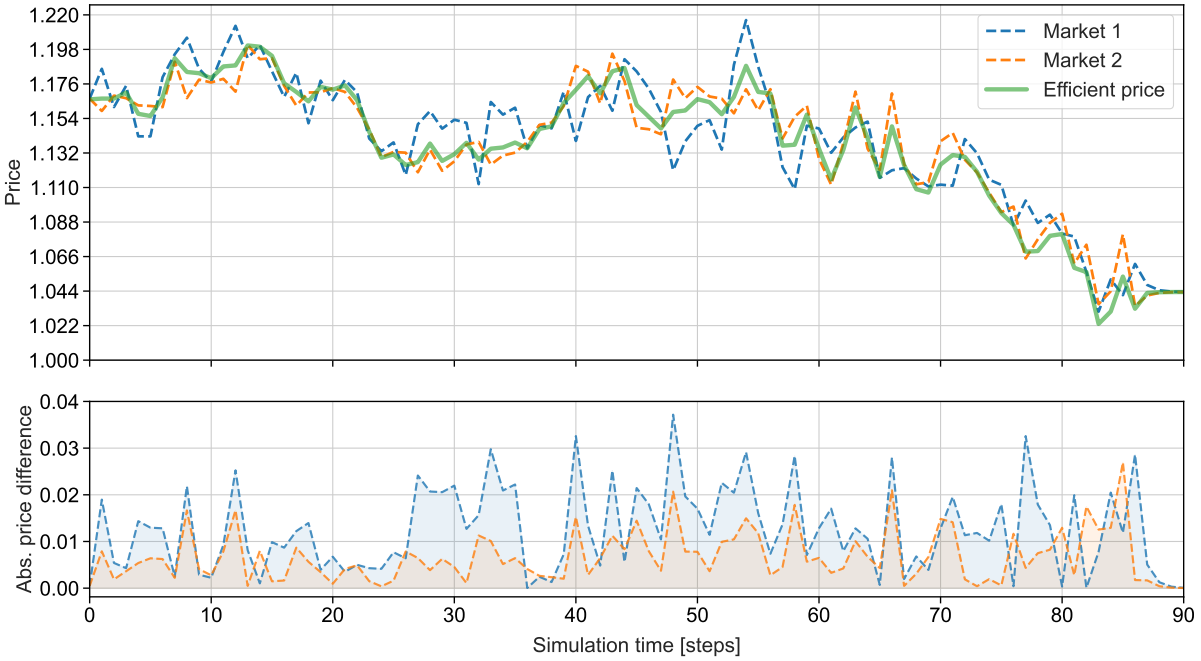

FIGURE 3.4: Efficient price and mid prices of market 1 and market 2. The top figure illustrates the movements of the implicit efficient price common to all markets. The bottom figure presents the absolute price difference between markets' prices and efficient price. From both figures it is clear that market 2 has larger impact on the changes in the efficient price, as efficient price much closely follows its price changes. This observation suggests that market 2 has higher information share - metric that will be introduced in Section 3.1.6.1. Markets in this figure are simulated with Equations (A.3) and (A.4).

The third component is a zero-mean covariance stationary term which reflects transitory effects unique to each market (Lütkepohl [2005], Hasbrouck [1995]). This transitory effects explain the deviations between the observed price and the efficient price (Hasbrouck [1995]), and can be attributed to, for example, bid-ask bounces and inventory adjustments (Baillie et al. [2002]).

Although we have already shown that a cointegrated processes can be represented in the form of VECM model and the common trend representation, it is also important to formally show that these two representation are equivalent. According to to a Theorem proposed by Soren Johansen (Johansen et al. [1995], Lütkepohl [2005]), also known as Granger Representation Theorem, the variables that are modelled with VECM model 3.16, can also be represented in

the following form:

$$\boldsymbol{y}_t = \boldsymbol{y}_0 + \lim_{\tau\to\infty} \tilde{\boldsymbol{\Psi}}_\tau \sum_{i=1}^{t} \boldsymbol{\epsilon}_i + \boldsymbol{\Psi}^*(N_L)\boldsymbol{\epsilon}_t \tag{3.45}$$

if one recalls that $\boldsymbol{\psi}$ is a common vector of $\lim_{\tau\to\infty} \tilde{\boldsymbol{\Psi}}_\tau$, it becomes clear that this representation is identical to Equation (3.44). Furthermore, (Johansen [1991]) shows that the matrix of the long-term response can be calculated in one step using matrices obtained from VECM model, with the following formula

$$\lim_{\tau\to\infty} \tilde{\boldsymbol{\Psi}}_\tau = \boldsymbol{\beta}_\perp \left[\boldsymbol{\alpha}_\perp^\intercal \left(\boldsymbol{I}_n - \sum_{i=1}^{M} \boldsymbol{\Gamma}_i \right) \boldsymbol{\beta}_\perp \right]^{-1} \boldsymbol{\alpha}_\perp^\intercal \tag{3.46}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}_i$, are the same matrices that were presented in the VECM representation Equation (3.16) and M stands for the highest number of lagged differences modelled. Also, $\perp$ stands for the orthogonal complement of a matrix (Lütkepohl [2005]).

3.1.6 Metrics

The metrics subsection is split into four smaller parts, each individually treating all metrics used in the Information Revelation approach proposed by Hagströmer and Menkveld [2019].

3.1.6.1 Information share

In Section 3.1.5, the notions of efficient price and the innovation in the efficient price were introduced. We determined that innovations in a single market can impact the innovations in the efficient price common to all markets. In this framework, the innovations in market prices are associated with incorporation of new information; the innovations in the efficient price are what Hasbrouck considers the price discovery process – *the impounding of new information into the security price* (Hasbrouck [1995]). Hasbrouck further proposes a variance decomposition-based metric, which he calls information share, to quantify the contribution of each market to the innovations in the efficient price.

Before we discuss information share, let us first define the covariance matrix of innovations in the following manner,

$$\boldsymbol{\Omega} = \mathrm{Cov}[\boldsymbol{\epsilon}_t] \tag{3.47}$$

where $\boldsymbol{\epsilon}_t$ is a $(n \times 1)$ vector of innovations in each markets quotes at time t. Notice until this point we have assumed that innovations in the markets are uncorrelated, thus $\boldsymbol{\Omega}$ would be a diagonal matrix, which made the analysis somewhat simplistic. In the context of real foreign exchange market, the innovation in one market can be accompanied by an innovation in other markets, i.e. they may be correlated (Hasbrouck [1995], Lütkepohl [2005]). Therefore, in

order to appropriately attribute contributions of each market to the innovations in the efficient price, i.e. to disentangle relations between variables, Hasbrouck proposes to orthogonalize the covariance matrix by employing Cholesky decomposition - the traditional approach suggested by Sims [1980] (Hasbrouck [1995]). The decomposition provides means to assess the maximal and minimal explanatory power of each market (Hasbrouck [2002]). Thus, the correlated innovations vector $\boldsymbol{\epsilon}_t$ is decomposed into instantaneously uncorrelated innovations $\boldsymbol{u}_t$ (Durlauf and Blume [2016], Hamilton [1994]).

For any real symmetric positive-definite matrix $\boldsymbol{\Omega}$ there is a unique lower triangular matrix $\boldsymbol{F}$, such that

$$\boldsymbol{\Omega} = \boldsymbol{F}\boldsymbol{F}^{\intercal} \tag{3.48}$$

where $\boldsymbol{F}$ is the Cholesky factorization of $\boldsymbol{\Omega}$ (Lay [2016]). Now, consider $\boldsymbol{\epsilon}_t$ given by a factor structure,

$$\boldsymbol{\epsilon}_t = \boldsymbol{F}\boldsymbol{u}_t \tag{3.49}$$

where $\boldsymbol{u}_t$ is an $(n \times 1)$ vector of serially uncorrelated random variables with $\mathbb{E}[\boldsymbol{u}_t] = 0$ and $\text{Var}[\boldsymbol{u}_t] = \boldsymbol{I}$ - $(n \times n)$ identity matrix (Hasbrouck [1995]). More explicitly,

$$\begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \\ \vdots \\ \epsilon_{n,t} \end{bmatrix} = \begin{bmatrix} f_{11} & 0 & 0 & \dots & 0 \\ f_{21} & f_{22} & 0 & \dots & 0 \\ f_{31} & f_{32} & f_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ f_{n1} & f_{n2} & f_{n3} & \dots & f_{nn} \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ \vdots \\ u_{n,t} \end{bmatrix} \tag{3.50}$$

where f_{ij} corresponds to the element in the i^{th} row and j^{th} column of lower triangular matrix $\boldsymbol{F}$. The above presentation is crucial to show what Hägstromer and Menkveld call an *unattractive feature of Cholesky factorization* (Hagströmer and Menkveld [2019]). Notice that with this setup, we are able to represent innovation in the first market $\epsilon_{1,t}$ only in terms of $u_{1,t}$,

$$\epsilon_{1,t} = f_{11}u_{1,t} \tag{3.51}$$

However,

$$\epsilon_{2,t} = f_{21}u_{1,t} + f_{22}u_{2,t} \tag{3.52}$$

$$\epsilon_{3,t} = f_{31}u_{1,t} + f_{32}u_{2,t} + f_{33}u_{3,t} \tag{3.53}$$

which accounts for the fact that innovation in the first market has potentially instantaneous effect on the innovation in the second market (Durlauf and Blume [2016]). Next, the first and second market may have instantaneous effect on the innovation in the third market etc. Thus

this results in a system in recursive form where the order of the variables plays an important role (Durlauf and Blume [2016], Hasbrouck [1995]). The importance of this setup and its impact on information share metric will become clearer once we derived the variance of efficient price that can be uniquely attributed to each market. Before that, it is necessary to first determine the total variance of the efficient price innovations.

Recall from Section 3.1.5 that $\boldsymbol{\psi}\boldsymbol{\epsilon}_t$ is the increment that is impounded into the efficient price at time t, due to the new information - in other words the efficient price innovation. The variance of the efficient price innovation is

$$\begin{aligned}
\text{Var}[\boldsymbol{\psi}\boldsymbol{\epsilon}_t] &= \text{Var}[\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t] && (3.54)\\
&= \mathbb{E}\left[(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t])\,(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t])^\intercal\right] && (3.55)\\
&= \mathbb{E}\left[(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \boldsymbol{\psi}\boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t])\,(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \boldsymbol{\psi}\boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t])^\intercal\right] && (3.56)\\
&= \mathbb{E}\left[(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t)\,(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t)^\intercal\right] && (3.57)
\end{aligned}$$

since $\boldsymbol{\psi}$ and $\boldsymbol{F}$ are constant and given that $\mathbb{E}[\boldsymbol{u}_t] = 0$. Also, note that $(ABC)^\intercal = C^\intercal B^\intercal A^\intercal$ and that $\text{Cov}[\boldsymbol{u}_t] = \boldsymbol{I}$, hence

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t)\,(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t)^\intercal\right] &= \mathbb{E}\left[(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t)\,(\boldsymbol{u}_t^\intercal\boldsymbol{F}^\intercal\boldsymbol{\psi}^\intercal)\right] && (3.58)\\
&= \boldsymbol{\psi}\boldsymbol{F}\;\mathbb{E}\left[\boldsymbol{u}_t\boldsymbol{u}_t^\intercal\right]\boldsymbol{F}^\intercal\boldsymbol{\psi}^\intercal && (3.59)\\
&= \boldsymbol{\psi}\boldsymbol{F}\boldsymbol{I}\boldsymbol{F}^\intercal\boldsymbol{\psi}^\intercal && (3.60)\\
&= \boldsymbol{\psi}\boldsymbol{F}\boldsymbol{F}^\intercal\boldsymbol{\psi}^\intercal && (3.61)\\
&= \boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal && (3.62)
\end{aligned}$$

Thus, the total variance of the efficient price innovations is

$$\text{Var}[\boldsymbol{\psi}\boldsymbol{\epsilon}_t] = \boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal \tag{3.63}$$

Now, let us attempt to quantify each market's contribution to the variance of the efficient price innovation. Recall that $\boldsymbol{\psi}$ is an $(1 \times n)$ row vector, in which value i quantifies the change in the efficient price due to a unit shock (innovation) in market i. Thus clearly, the product of i^{th} element of $\boldsymbol{\psi}$ and i^{th} element of $\boldsymbol{\epsilon}_t$, i.e. $[\boldsymbol{\psi}]_i[\boldsymbol{\epsilon}_t]_i$ is the contribution of i^{th} market to the efficient price innovation at time t. Where $[\cdot]_i$ indicates the i^{th} element of a vector $\cdot$. Thus, using the factor structure from Equation (3.49) the variance of the efficient price innovation attributable to i^{th} market is,

$$\text{Var}\left[[\boldsymbol{\psi}\boldsymbol{\epsilon}_t]_i\right] = \text{Var}\left[[\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t]_i\right] = \left([\boldsymbol{\psi}\boldsymbol{F}]_i\right)^2 \tag{3.64}$$

the reduction takes advantage of the fact that $\text{Var}[\boldsymbol{u}_t] = 0$. Thus, if we normalize the efficient price innovation attributable to i^{th} market by the total variance of efficient price innovation, we obtain Hasbrouck's Information Share metric

$$S_i = \frac{([\boldsymbol{\psi F}]_i)^2}{\boldsymbol{\psi \Omega \psi}^\intercal} \tag{3.65}$$

where S_i stands for the information share of i^{th} market.

Recall, that earlier in this section we have identified that the employment of Cholesky decomposition imposes a specific hierarchy, in which first (order-wise) variables are allowed to have an impact on the following variables in the sequence (Hasbrouck [1995]). Thus, the information share of the first market is maximized whereas, the last market's information share is minimized. This becomes much more clear when one considers a more explicit form of $(\boldsymbol{\psi F})^2$

$$\left(\begin{bmatrix} \psi_1 & \psi_2 & \dots & \psi_n \end{bmatrix} \begin{bmatrix} f_{11} & 0 & 0 & \dots & 0 \\ f_{21} & f_{22} & 0 & \dots & 0 \\ f_{31} & f_{32} & f_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ f_{n1} & f_{n2} & f_{n3} & \dots & f_{nn} \end{bmatrix} \right)^2 = \begin{bmatrix} (\psi_1 f_{11} + \psi_2 f_{21} + \dots + \psi_n f_{n1})^2 \\ (\psi_2 f_{22} + \psi_3 f_{32} + \dots + \psi_n f_{n2})^2 \\ (\psi_3 f_{33} + \psi_4 f_{43} + \dots + \psi_n f_{n3})^2 \\ \vdots \\ (\psi_n f_{nn})^2 \end{bmatrix}^\intercal \tag{3.66}$$

Notice, that for the first market's variance contribution we account for all the possible instantaneous impacts in all the other markets, whereas the n^{th} market's information share is not allowed to have instantaneous effects on other markets (Durlauf and Blume [2016]). In order to cope with this problem, Hasbrouck proposes to use upper and lower bounds of information share. That is, the upper bound of market's information share is obtained when one permutes $\boldsymbol{\psi}$ and $\boldsymbol{\Omega}$ to place market as first (Hasbrouck [1995]). The lower bound of market's information share is obtained when one places the market as last (Hasbrouck [1995], Baillie et al. [2002]). In this setup, the upper bound accounts for not only market's own contribution but also its correlation with other markets, whereas the lower bound only accounts for the pure contribution that is uncorrelated with any other markets (Baillie et al. [2002]).

The publication entitled "Price discovery and common factor analysis" by Baillie et al. [2002] provides an in-depth insight on the impact of the correlation between markets on the size of the gap between the lower and the upper bound of Hasbrouck's information share metric. Baillie et al. [2002] show that the higher the correlation between variables, the greater (smaller) the upper (lower) bound. Furthermore, they analytically demonstrate that if the markets are uncorrelated, then upper bound and lower bound are equivalent. Thus, the correlation between markets directly impacts the uncertainty about market's unique contribution to the efficient

price innovation.

Finally, Baillie et al. [2002] also prove that in fact the largest (smallest) information share occurs when the variables is first (last) in the sequence, but only if one assumes that the cross correlation between variables is positive. Since, we cannot assume that the cross correlation between variables is always positive, Hägstromer and Menkveld take even more cautious approach to determine the lower and upper bounds. Namely, they define the lower bound information share as the minimum information share across all possible permutations of market sequence (Hagströmer and Menkveld [2019]),

$$\text{InfoShare}_i = \min_{x \in \text{Aut}(X)} \frac{\left([\boldsymbol{\psi} \boldsymbol{F}_x]_{x(i)} \right)^2}{\boldsymbol{\psi} \boldsymbol{\Omega} \boldsymbol{\psi}^\intercal} \tag{3.67}$$

where X is the sequence of markets used in VECM modelling, $\text{Aut}(X)$ stands for the automorphism with all possible combinations of market sequences (Hagströmer and Menkveld [2019]). Hence, x simply stands for one particular combination of market sequence, the $\boldsymbol{F}_x$ is the Cholesky factor of $\boldsymbol{\Omega}$ which was resequenced according to x and $x(i)$ stands for the position of market i in the sequence x (Hagströmer and Menkveld [2019]).

In conclusion, it is important to note that Hasbrouck emphasizes the fact that upper and lower bound of information share should not in any case be associated with the confidence intervals, as the number of samples does not affect the gap between the upper and lower bound (Hasbrouck [2002]). Hasbrouck further points out that information share is only a relative measure that quantifies relative information share but does not in any way quantify in the absolute sense the total information that is impounded into the efficient price (Hasbrouck [1995]). Thus, intuitively *information share measures "who moves first" in the process of price adjustment* (Hasbrouck [1995]), whereas the uncertainty about this stems from the correlation between markets.

3.1.6.2 Price inefficiency & information speed

In this section, our main focus will be deriving and developing an intuition behind the price inefficiency metric. This will be followed by a brief explanation of the information speed measure, which is a derived from price inefficiency metric.

To begin with, recall from Section 3.1.4, the observations made about the cumulative sum of VMA coefficient matrices until τ time steps after unit shock ($\tilde{\boldsymbol{\Psi}}_\tau$). As we have already established, the ($\tilde{\boldsymbol{\Psi}}_\tau$) matrix retains the information about the cumulative short-term responses of each market to a multivariate shocks (unit shock in all of the markets). In other words, ($\tilde{\boldsymbol{\Psi}}_\tau$) matrix reveals by how much each markets adjusted its prices until τ time steps from a unit multivariate shock. Now, for each market we would like to quantify how close to efficient price

innovations are the price adjustments that are made by market until τ steps after the shock introduction to all of the markets. Although, the idea behind this method of characterization the short-term responses may seem quiet convoluted at this moment, it should become a lot more clear in the remainder of this section.

Hägstromer and Menkveld decided to consider covariance of cumulative short-term responses of markets to innovations (shocks) $\boldsymbol{\epsilon}_t$ at time τ with the long-term response of the markets to the same shocks. The derivation of covariance employs the factor structure previously presented in Equation (3.49), and the assumptions that $\mathbb{E}[\boldsymbol{u}_t] = 0$ and $\text{Cov}[\boldsymbol{u}_t] = \boldsymbol{I}$ about the vector of serially uncorrelated random variables $\boldsymbol{u}_t$,

$$\begin{aligned}
\text{Cov}[\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\epsilon}_t, \boldsymbol{\psi}\boldsymbol{\epsilon}_t] &= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t]\right)\left(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t]\right)^\intercal\right] && (3.68)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t]\right)\left(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t - \boldsymbol{\psi}\boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t]\right)^\intercal\right] && (3.69)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t\right)\left(\boldsymbol{\psi}\boldsymbol{F}\boldsymbol{u}_t\right)^\intercal\right] && (3.70)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t\right)\left(\boldsymbol{u}_t^\intercal \boldsymbol{F}^\intercal \boldsymbol{\psi}^\intercal\right)\right] && (3.71)\\
&= \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\intercal\right]\boldsymbol{F}^\intercal \boldsymbol{\psi}^\intercal && (3.72)\\
&= \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\Omega}\boldsymbol{\psi}^\intercal && (3.73)
\end{aligned}$$

In this simple manner we have derived a formula for the covariance between the short-term and the long-term responses of all markets. From Equations (3.36) and (3.37) recall that i^{th} element of j^{th} row of matrix $\tilde{\boldsymbol{\Psi}}_\tau$ corresponds to cumulative response of the j^{th} market to a shock in i^{th} market. Thus, the j^{th} row of matrix $\tilde{\boldsymbol{\Psi}}_\tau$, denoted as $\tilde{\boldsymbol{\Psi}}_{j,\tau}$, represents the cumulative responses of j^{th} market to independent shocks in all of the markets. Consequently, the covariance of j^{th} market responses to its respective long-term responses is

$$\text{Cov}[\tilde{\boldsymbol{\Psi}}_{j,\tau} \boldsymbol{\epsilon}_t, \boldsymbol{\psi}\boldsymbol{\epsilon}_t] = \tilde{\boldsymbol{\Psi}}_{j,\tau} \boldsymbol{\Omega}\boldsymbol{\psi}^\intercal \tag{3.74}$$

Hägstromer and Menkveld proposes to normalize this covariance by the total variance of efficient price innovations, which yields the beta coefficient metric associated with with j^{th} market:

$$\beta_{j,\tau} = \frac{\text{Cov}[\tilde{\boldsymbol{\Psi}}_{j,\tau} \boldsymbol{\epsilon}_t, \boldsymbol{\psi}\boldsymbol{\epsilon}_t]}{\text{Var}(\boldsymbol{\psi}\boldsymbol{\epsilon}_t)} = \frac{\tilde{\boldsymbol{\Psi}}_{j,\tau} \boldsymbol{\Omega}\boldsymbol{\psi}^\intercal}{\boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal} \tag{3.75}$$

Notice that the beta coefficient is analogous to the Pearson correlation coefficient. The only difference is that the beta coefficient is a ratio of covariance to a variance of only one variable. The normalization factor proposed by Hägstromer and Menkveld is motivated by the fact that $\lim_{\tau \to \infty} \tilde{\boldsymbol{\Psi}}_{j,\tau} = \boldsymbol{\psi}$, hence beta coefficient of j^{th} market converges to 1 as τ tends to infinity (Hagströmer and Menkveld [2019]). Let us consider a simple example to better illustrate

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tilde{\boldsymbol{\Psi}}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tilde{\boldsymbol{\Psi}}_1 = \begin{bmatrix} 1.5 & 1 \\ 0.5 & 1.5 \end{bmatrix}, \tilde{\boldsymbol{\Psi}}_2 = \begin{bmatrix} 1.5 & 2 \\ 1.5 & 2 \end{bmatrix} = \lim_{\tau \to \infty} \tilde{\boldsymbol{\Psi}}_{j,\tau} \tag{3.76}$$

First, let us consider the beta coefficient for $\tau = 0, 1, 2$ for the 1^{st} market i.e. $\tilde{\boldsymbol{\Psi}}_{1,\tau}$.

$$\beta_{1,0} = \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{1.5}{1.5^2 + 2^2} = 0.24 \tag{3.77}$$

Notice that beta coefficient metric takes into account the short-term responses of the market when the shock is introduced to that same market, and also when the shock is introduced to the other market. In this manner, beta coefficient provides a collective measure of responsiveness of the market to shocks in all of the markets.

$$\beta_{1,1} = \frac{\begin{bmatrix} 1.5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{\begin{bmatrix} 1.5 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{1.5^2 + 2}{1.5^2 + 2^2} = 0.68 \tag{3.78}$$

At $\tau = 1$ we can notice that even though 1^{st} market's response to the shock in its own market is already equivalent to the long-term response, the beta coefficient is still not equal to 1. This is because, 1^{st} market's cumulative response until $\tau = 1$ to the shock in the 2^{nd} one is only 1, whereas the long term response is 2. Thus, the beta coefficient of j^{th} market will be equal to 1 for a τ at which j^{th} market cumulative responses to any shock is equivalent to the efficient price innovation. As follows, at $\tau = 2$

$$\beta_{1,2} = \frac{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}}{\begin{bmatrix} 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}} = \frac{1.5^2 + 2^2}{1.5^2 + 2^2} = 1 \tag{3.79}$$

It appears that after two time steps since introduction of a multivariate shock, the 1^{st} market reveals all the information associated with shocks as beta coefficient reaches 1. Thus, in this regards, $\beta_{j,\tau}$ can be treated as a fraction of the information content that market j has revealed until τ time steps since the shock (Hagströmer and Menkveld [2019]). Furthermore, $\beta_{j,\tau}$ intuitively, is a normalized measure of the responsiveness of the market to multivariate shocks, where responsiveness is measured in terms of the proportion of short-term adjustment of prices

until τ relative to the expected long-term price adjustment (efficient price innovation). Thus, Hägstromer and Menkveld note that the speed with which $\beta_{j,\tau}$ approaches 1 captures the speed of information revelation in j^{th} market (Hagströmer and Menkveld [2019]), in other words how fast does a market impound into its prices the entire information associated with the shock at $\tau = 0$. Alternatively, one can also see beta coefficient as a measure of collective price efficiency of the responses to a multivariate shocks.

Hägstromer and Menkveld decide to focus on the exact opposite of the beta coefficient measure. Namely, they define the price inefficiency metric as follows:

$$PI_{j,\tau} =\mid 1 - \beta_{j,\tau} \mid \tag{3.80}$$

where $PI_{j,\tau}$ stands for price inefficiency of j^{th} market at τ time steps after the shock. They further propose that price inefficiency measure should be interpreted *as the fraction of the information that market j has yet to reveal τ periods after the shock $\boldsymbol{\epsilon}_t$* (Hagströmer and Menkveld [2019]).

Information speed

Hägstromer and Menkveld take one step further and propose a metric that is derived from price inefficiency, which they call Information Speed. They define Information Speed as *the average time it takes for a market to incorporate 80% of the information contained in a price shock* (Hagströmer and Menkveld [2019]). This definition translates to minimum time τ at which market's price inefficiency is below 20%. Finally, the measure is multiplied by minus one.

3.1.6.3 Bilateral connections

Another metric that is essential for characterization of the short-term responses of markets and generating the empirical information revelation maps is the partial correlation matrix. Partial correlation matrix of cumulative responses of markets is used to define what Hägtromer and Menkveld call the strength of bilateral connections between a pair of markets (Hagströmer and Menkveld [2019]). The partial correlation matrix is based on the covariance of the cumulative response matrices $\tilde{\boldsymbol{\Psi}}_\tau$ and can be computed for each τ time steps since the multivariate shock. The strength of bilateral connections between a pair of markets at τ is defined as the partial correlation of their cumulative response τ time steps since the introduction of multivariate shock (Hagströmer and Menkveld [2019]). This method of computing the partial correlation matrix is motivated by the work of (Hero and Rajaratnam [2012]).

The covariance matrix of market's cumulative responses until τ, i.e. $\tilde{\boldsymbol{\Psi}}_\tau$, to a multivariate shock $\boldsymbol{\epsilon}_t$ can be determined in the following manner

$$\begin{aligned}
\text{Cov}[\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\epsilon}_t, \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\epsilon}_t] &= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t]\right)\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \mathbb{E}[\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t]\right)^\intercal\right] && (3.81)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t]\right)\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t - \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\mathbb{E}[\boldsymbol{u}_t]\right)^\intercal\right] && (3.82)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t\right)\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t\right)^\intercal\right] && (3.83)\\
&= \mathbb{E}\left[\left(\tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\boldsymbol{u}_t\right)\left(\boldsymbol{u}_t^\intercal \boldsymbol{F}^\intercal \tilde{\boldsymbol{\Psi}}_\tau^\intercal\right)\right] && (3.84)\\
&= \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{F}\mathbb{E}\left[\boldsymbol{u}_t \boldsymbol{u}_t^\intercal\right] \boldsymbol{F}^\intercal \tilde{\boldsymbol{\Psi}}_\tau^\intercal && (3.85)\\
&= \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\Omega} \tilde{\boldsymbol{\Psi}}_\tau^\intercal && (3.86)
\end{aligned}$$

For clarity purposes, let us denote the covariance matrix of cumulative responses until time τ as follows:

$$\boldsymbol{\Sigma}_\tau = \tilde{\boldsymbol{\Psi}}_\tau \boldsymbol{\Omega} \tilde{\boldsymbol{\Psi}}_\tau^\intercal \tag{3.87}$$

where $\boldsymbol{\Sigma}_\tau$ stands for the covariance matrix of cumulative responses of markets until time τ from the multivariate shock $\boldsymbol{\epsilon}_t$. The partial correlation matrix associated with the covariance matrix of cumulative responses of markets is computed with the following formula:

$$\boldsymbol{R}_\tau = \boldsymbol{D}_{\boldsymbol{\Sigma}_\tau^{-1}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_\tau^{-1} \boldsymbol{D}_{\boldsymbol{\Sigma}_\tau^{-1}}^{-\frac{1}{2}} \circ \boldsymbol{K} \tag{3.88}$$

where $\boldsymbol{R}_\tau$ stands for the partial correlation matrix (adjacency matrix), $\boldsymbol{D}_{\boldsymbol{\Sigma}_\tau^{-1}}$ denotes the diagonal matrix obtained from $\boldsymbol{\Sigma}_\tau^{-1}$ by zeroing out all non-diagonal entries (Hero and Rajaratnam [2012]). The $\circ$ is the element-wise product (also known as Hadamard product) and $\boldsymbol{K}$ is an $(n \times n)$ matrix whose off-diagonal entries are set to negative one, whereas the diagonal entries are set to one (Hagströmer and Menkveld [2019]). Let us denote $\rho_{i,j,\tau}$ element in i^{th} row and j^{th} column of $\boldsymbol{R}_\tau$ matrix, which represents the partial correlation between i^{th} and j^{th} markets. Also note that $\rho_{i,j,\tau} = \rho_{j,i,\tau}$.

The use of partial correlation matrix to determine the strength of bilateral connections between markets, is motivated by the fact that partial correlation allows to eliminate the potential influence of other variables on the correlation, the so called confounding variables (Hero and Rajaratnam [2012], Baba et al. [2004]). Thus the partial correlation exposes the correlation in innovations of two markets that is uncorrelated with other markets, and hence private to the pair (Hagströmer and Menkveld [2019]).

A detailed example for the information revelation method is presented in Appendix A.

3.2 Quantifying information flows with transfer entropy

In the following section, the fundamental information-theoretic metrics, such as Shannon entropy, Kullback-Leibler divergence and mutual information, are formally introduced in a discrete setting. Afterwards, the conditional mutual information is treated in detail, with a thorough discussion of the concept of multivariate interaction as well as the associated notions of redundancy elimination and synergy. Next, the transfer entropy and conditional transfer entropy are defined and their interpretations clarified. Finally, the metrics are presented in their continuous analogs .

3.2.1 Introduction to information theory

3.2.1.1 Shannon entropy

One of the most elementary quantities in the Information Theory is the Shannon information content, which is a measure of the amount of information one gains if an event with some associated probability happens (MacKay and Mac Kay [2003]).

Definition 3.6 (**Shannon information content**)**.** The Shannon information content in base b, i.e. $h_b(x)$ (also called surprisal) of an outcome x is defined as (MacKay and Mac Kay [2003], Bossomaier et al. [2016]):

$$h_b(x) \triangleq \log_b \left(\frac{1}{p_X(x)} \right) \tag{3.89}$$

where $p_X(x) = \Pr\{X = x\}$ denotes the probability that a random variable X gets the value x. Variable b represents the base of the logarithm. The base of the logarithm used for the computation establishes the units in which entropy is expressed. For example if $b = 2$, the entropy is expressed in "bits", whereas if $b = e$ (euler constant), the entropy is expressed in "nats" (Shannon [1948]).

The idea behind this metric is as follows: if the probability associated with particular outcome $X = x$ is close to one $p_X(x) \to 1^-$, then $h(x) \to 0$, which translates to nearly no surprise about the outcome. On the other hand, if $p_X(x) \to 0^+$, then $h(x) \to +\infty$, which means that the outcome is unexpected.

The Shannon information content is in fact the key ingredient for the most fundamental information theoretic metric called Shannon entropy[3]. Shannon entropy was first introduced by Claude Shannon in his landmark paper Shannon [1948].

[3] *I thought of calling it "information," but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."* (Baniel [2009])

Definition 3.7 (**Shannon entropy**)**.** Let $\mathcal{X}$ denote the alphabet[4] of random variable X. Furthermore, let $|\mathcal{X}| < \infty$, i.e. the alphabet is finite. Then, the Shannon entropy in base b, i.e. $H_b(X)$, of a random variable X with a finite alphabet $\mathcal{X}$ governed by a probability mass function $p_X(x)$, where $p_X(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$ is defined as (Thomas and Joy [2006]):

$$H_b(X) \triangleq -\sum_{x\in\mathcal{X}} p_X(x) \log_b \left(p_X(x)\right) \tag{3.90}$$

$$= \sum_{x\in\mathcal{X}} p_X(x) \log_b \left(\frac{1}{p_X(x)}\right) \tag{3.91}$$

$$= \mathbb{E}\left[\log_b \left(\frac{1}{p_X(X)}\right)\right] \tag{3.92}$$

Equation (3.92) reveals how Shannon entropy is simply the expectation of the Shannon information content of all possible outcomes. It is usually interpreted as a measure of average uncertainty about the random variable X. This intuition is further demonstrated in the following examples.

Remark 3.8 (Convention). If $p_X(x) = 0 \implies 0 \log(0) = 0$. This follows from $\lim_{x\to 0^+} x \log(x) = 0$

Example 3.1 (Entropy of a fair coin)**.** *Let random variable Y represent a fair coin with two faces $\mathcal{Y} = \{1, 2\}$ and $p_Y(y) = Pr\{Y = y\}$. The Shannon entropy of the fair coin is:*

$$\begin{aligned} H_2(Y) &= -\sum_{y=1}^{2} p(y) \log_2 \left(p(y)\right) \\ &= -\sum_{y=1}^{2} \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \\ &= \log_2(2) \end{aligned}$$

Example 3.2 (Entropy of a fair dice)**.** *Now, consider a random variable X representing a fair dice with six faces $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and $p_X(x) = Pr\{X = x\}$. The Shannon entropy of a fair dice is:*

$$\begin{aligned} H_2(X) &= -\sum_{x=1}^{6} p(x) \log_2 \left(p(x)\right) \\ &= -\sum_{x=1}^{6} \frac{1}{6} \log_2 \left(\frac{1}{6}\right) \\ &= \log_2(6) \end{aligned}$$

[4] Alphabet is the set of all possible states or outcomes of a random variable.

Examples 3.1 and 3.2 demonstrate that the Shannon entropy of a fair dice is clearly higher than the entropy of a fair coin. Without doubt, there is more uncertainty about the outcome from tossing a die than there is from tossing a coin.

Remark 3.9 (Notation). For the rest of this thesis, we will be referring to Shannon entropy as entropy. Also, we will be using entropy in "nats" units, i.e. $b = e$ (euler constant). Thus, the notation $\log(\cdot)$ will represent a natural logarithm, and $H(X)$ will be used to symbolize $H_e(X)$. Finally, instead of denoting the probability mass function by $p_X(x)$, it will be represented with $p(x)$.

To better illustrate the concept of entropy, consider systems (variables) with two, three and four possible states (outcomes). For a two-state-system, let us denote the probability associated with the first state as $p(x_1)$. Next, for a two-state-system $p(x_2) = 1 - p(x_1)$, whereas for three and four-state-systems let $p(x_i) = \frac{1-p(x_1)}{n-1}$, for $i \neq 1$ where i is the state index and n is the cardinality of the system $|\mathcal{X}|$. In other words, let all the remaining probability $1 - p(x_1)$ be uniformly distributed among all the states other than state 1. Figure 3.5 visualizes how the

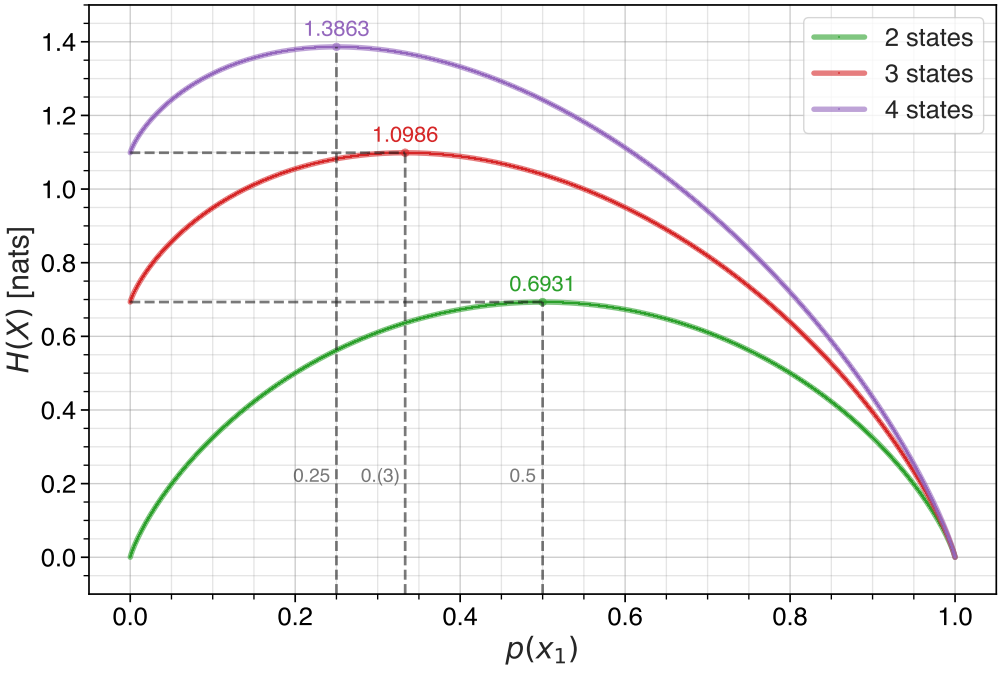

FIGURE 3.5: Shannon entropy of two-state, three-state and four-state systems as a function of probability associated with first state x_1. Figure illustrates when the maximum Shannon entropy of a variable (system) is reached, and also how systems with more possible states have higher entropy. Furthermore, figure shows that Shannon entropy is a strictly concave function.

entropy of the system changes as a function of probability associated with first state x_1. From Figure 3.5, it should be clear that maximal entropies are reached when probabilities associated with the states are uniformly distributed among all the possible states. This aligns with the

intuition that we developed, as the largest uncertainty about the state of the system is reached when the probability of occurrence of each state is as likely as occurrence of any other state. Furthermore, as the $p(x_1) \to 1$, the $H(X) \to 0$, which makes sense, as there is nearly no uncertainty about the state of the system.

Another important insight presented in Figure 3.5 is that as the number of states increases, so does the entropy of the system (with an exception when probability associated with one of the states tends to 1). Clearly, if there are more states a system can occupy, then there is higher uncertainty about the state that the system can be found in. Finally, from Figure 3.5 one can observe that Shannon entropy is a strictly concave function. This property will be proved later in this section.

One should also note that information-theoretic definition of entropy is nearly identical to the statistical thermodynamics entropy formulation established by Ludwig Boltzmann in 1877 (Boltzmann [2015]). While there is a significant difference in theory and the interpretation of two formulations, the metrics and intuition behind them are the same. As Shannon [1948] puts it, the Boltzmann constant present in the statistical thermodynamics entropy formulation merely amounts to a different units of entropy measure.

Having provided a detailed account on the intuition behind Shannon entropy, it is time to mention here four fundamental properties of Shannon entropy metric (Shannon [1948], Thomas and Joy [2006], Floudas and Pardalos [2008]):

Lemma 3.10 (Shannon entropy is nonnegative)**.**

$$H(X) \geq 0 \tag{3.93}$$

Proof. $0 \leq p(x) \leq 1 \implies \log\left(\frac{1}{p(x)}\right) \geq 0$. □

Lemma 3.11 (The base of the Shannon entropy measure and thus its units can be changed by multiplying it with a factor)**.**

$$H_b(X) = \log_b(a) H_a(X) \tag{3.94}$$

Proof. $\log_b(p(x)) = \log_b(a) \log_a(p(x))$ □

Lemma 3.12 (The Shannon entropy of a probability distribution with certaint outcome is 0)**.**

$$H(X) = 0 \iff p(x_i) = 1 \quad \text{for any } x_i \in \mathcal{X} \tag{3.95}$$

Proof. $p(x) = 1 \implies \log\left(\frac{1}{p(x)}\right) = 0$ □

Lemma 3.13 (Maximum entropy and strict concavity)**.**

$$H_{\max}(X) = \log\left(|\mathcal{X}|\right) \iff p(x_i) = \frac{1}{|\mathcal{X}|} \quad \forall\, x_i \in \mathcal{X} \tag{3.96}$$

Proof. See Appendix B, Section B.1. □

3.2.1.2 Joint entropy

Up to this point we have considered the uncertainty of only one random variable. The notion of entropy can be further extended to a pair of discrete random variables, say X and Y. For multiple random variables one can consider for example *joint entropy* and *conditional entropy*.

Definition 3.14 (**Joint entropy**)**.** Consider two discrete random variables X and Y with marginal probability mass functions $p(x)$ and $p(y)$, respectively. A joint entropy $H(X,Y)$ of a pair of (discrete) variables X and Y with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, is defined as:

$$H(X,Y) \triangleq -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x,y)) \tag{3.97}$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{1}{p(x,y)}\right) \tag{3.98}$$

$$= \mathbb{E}\left[\log\left(\frac{1}{p(X,Y)}\right)\right] \tag{3.99}$$

where $p(x,y)$ is the joint distribution of X and Y. The joint entropy can be interpreted as the expected uncertainty about states X and Y treated together as one random variable (X,Y). In this case, the pair (X,Y) has alphabet $\mathcal{X}\times\mathcal{Y}$. There are many properties of the joint entropy. The two most important ones are treated in the following lemmas.

Lemma 3.15 (Entropy is additive for independent random variables)**.**

$$H(X,Y) = H(X) + H(Y) \iff p(x,y) = p(x)p(y) \tag{3.100}$$

Proof. See Appendix B, Section B.2. □

Lemma 3.16 (Entropy is sub-additive for dependent random variables)**.**

$$H(X,Y) \leq H(X) + H(Y) \tag{3.101}$$

Proof. See Appendix B, Section B.3. □

3.2.1.3 Conditional entropy

Definition 3.17 (**Conditional entropy**)**.** Consider two discrete random variables X and Y with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, as well as marginal probability mass functions $p(x)$ and $p(y)$. The entropy of Y conditioned on X, i.e. the conditional entropy $H(Y|X)$, is defined as:

$$H(Y|X) \triangleq -\sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \tag{3.102}$$

$$= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\left(p(y|x)\right) \tag{3.103}$$

$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(p(y|x)\right) \tag{3.104}$$

$$= \mathbb{E}\left[\log\left(\frac{1}{p(Y|X)}\right)\right] \tag{3.105}$$

The conditional entropy is the average entropy of variable Y for each value of X. Intuitively, conditional entropy can be interpreted as the uncertainty about a variable Y left after considering some additional context provided by variable X (concept analogous to conditional probability) (Bossomaier et al. [2016]) In the following theorem, the relationship between joint entropy and conditional entropy will be presented.

Theorem 3.18 (Chain rule)**.** *The joint entropy of random variables X and Y is equal to the sum of the marginal entropy of X and entropy of Y conditioned on X, as follows:*

$$H(X,Y) = H(X) + H(Y|X) \tag{3.106}$$

$$H(X,Y) = H(Y) + H(X|Y) \tag{3.107}$$

Proof. See Appendix B, Section B.4 □

Corollary 3.19 (Chain rule for joint entropy of N random variables)**.** *For discrete random variables $X_1, X_2, \ldots, X_n$ based on $p(x_1, x_2, \ldots, x_n)$*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) \tag{3.108}$$

Proof. See Appendix B, Section B.5 □

3.2.1.4 Kullback-Leibler divergence

Another very important metric is the Kullback-Leibler (KL) divergence, also known as cross-entropy and relative entropy (Jumarie [1990]). The Kullback-Leibler divergence is a measure of "distanc" between two probability mass functions (Thomas and Joy [2006]).

Definition 3.20 (**Kullback-Leibler divergence**)**.** The KL divergence between probability mass functions $p(x)$ and $q(x)$, with a finite alphabet $\mathcal{X}$ is defined as:

$$D(p(x)||q(x)) \triangleq \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right) \tag{3.109}$$

$$= \mathbb{E} \left[\log \left(\frac{p(X)}{q(X)} \right) \right] \tag{3.110}$$

KL divergence can be thought of as a measure of amount of information lost when using one probability mass function $q(x)$ to express other one $p(x)$ (Bossomaier et al. [2016]). It should be pointed out that KL divergence is not strictly a distance measure, as it does not satisfy the triangle inequality, and it is not symmetric (Bossomaier et al. [2016]). As it will become clear in the next subsections, KL divergence is used to define mutual information and transfer entropy measures.

Remark 3.21 (Conventions). The definition of KL divergences uses the following three conventions: $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. As a result, in case there is any even $x \in X$, such that $p(x) > 0$ and $q(x) = 0$ then $D(p||q) = \infty$ (Thomas and Joy [2006]).

Theorem 3.22 (Information inequality)**.**

$$D(p(x)||q(x)) \geq 0 \tag{3.111}$$

$$D(p(x)||q(x)) = 0 \iff p(x) = q(x) \quad \forall\, x \in \mathcal{X} \tag{3.112}$$

Proof. See Appendix B, Section B.6. □

Lemma 3.23 (Non-symmetricity)**.**

$$D(p(x)||q(x)) \neq D(q(x)||p(x)) \tag{3.113}$$

Proof. This lemma can be easily proved by simply choosing $p(x) \neq q(x)$ and showing that the relation does not hold. □

3.2.1.5 Mutual information

Another very important metric is the so called mutual information. Intuitively, mutual information is the amount of information shared between random variables X and Y (Bossomaier et al. [2016]). It may as well be interpreted as the measure of information that one random variable contains about the other random variable (Thomas and Joy [2006]).

Definition 3.24 (**Mutual information**)**.** Let X and Y be two discrete random variables defined on a finite spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, with a joint probability mass function $p(x, y)$. Moreover, let $p(x)$ and $p(y)$ be the marginal probability mass functions of X and Y, respectively. Hence, the mutual information of random variables X and Y is the KL divergence between their joint probability distribution and the product of their marginal probability distributions (Bossomaier et al. [2016]). Following the Definition 3.20,

$$I(X;Y) \triangleq \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3.114}$$

$$= D\left(p(x,y)||p(x)p(y)\right) \tag{3.115}$$

$$= \mathbb{E}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right] \tag{3.116}$$

Mutual information has a few very important and convenient properties which are presented in the following lemmas. Especially important is the fact that mutual information is a symmetric metric.

Lemma 3.25 (From Mutual information to Entropy)**.**

$$I(X;Y) = H(X) - H(X|Y) \tag{3.117}$$

Proof. See Appendix B, Section B.7. □

Lemma 3.26 (Mutual information symmetry)**.**

$$I(X;Y) = I(Y;X) \tag{3.118}$$

Proof. This lemma can be proved by simply using Definition 3.24 and recognizing that $p(x, y) = p(y, x)$ and $p(x)p(y) = p(y)p(x)$. □

Lemma 3.27 (Mutual information of a variable with itself)**.**

$$I(X;X) = H(X) \tag{3.119}$$

Proof. Using Lemma 3.25, $I(X,X) = H(X) - H(X|X)$. From Definition 3.17, it is clear that $H(X|X) = 0$, hence $I(X,X) = H(X)$. □

Lemma 3.28 (Mutual information for independent processes)**.**

$$I(X;Y) = 0 \iff p(x,y) = p(x)p(y) \tag{3.120}$$

Proof. The above follows from Equation (3.116), as $p(x,y) = p(x)p(y) \implies \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \log(1) = 0$. □

3.2.1.6 Conditional mutual information

When more than two random variables are considered, one may be interested in determining the mutual information between a pair of processes conditioned on a third process. For this purpose, the *conditional mutual information* can be used.

Definition 3.29 (**Conditional mutual information**)**.** Let X, Y and Z be discrete random variables with finite alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. The mutual information between X and Y conditioned on Z is defined as (Wyner [1978]):

$$I(X;Y|Z) \triangleq \sum_{z\in\mathcal{Z}} p(z) \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y|z) \log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) \tag{3.121}$$

$$= \sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y,z) \log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) \tag{3.122}$$

$$= \mathbb{E}\left[\log\left(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right)\right] \tag{3.123}$$

$$\tag{3.124}$$

Using, Lemma 3.25 we can also express mutual information in terms of conditional entropies, as follows:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{3.125}$$

Just like entropy, mutual information also satisfies a chain rule.

Theorem 3.30 (Mutual information chain rule)**.**

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I\left(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1\right) \tag{3.126}$$

Proof. See Appendix B, Section B.8, □

All the metrics introduces up to this point are conveniently presented in Figure 3.6.

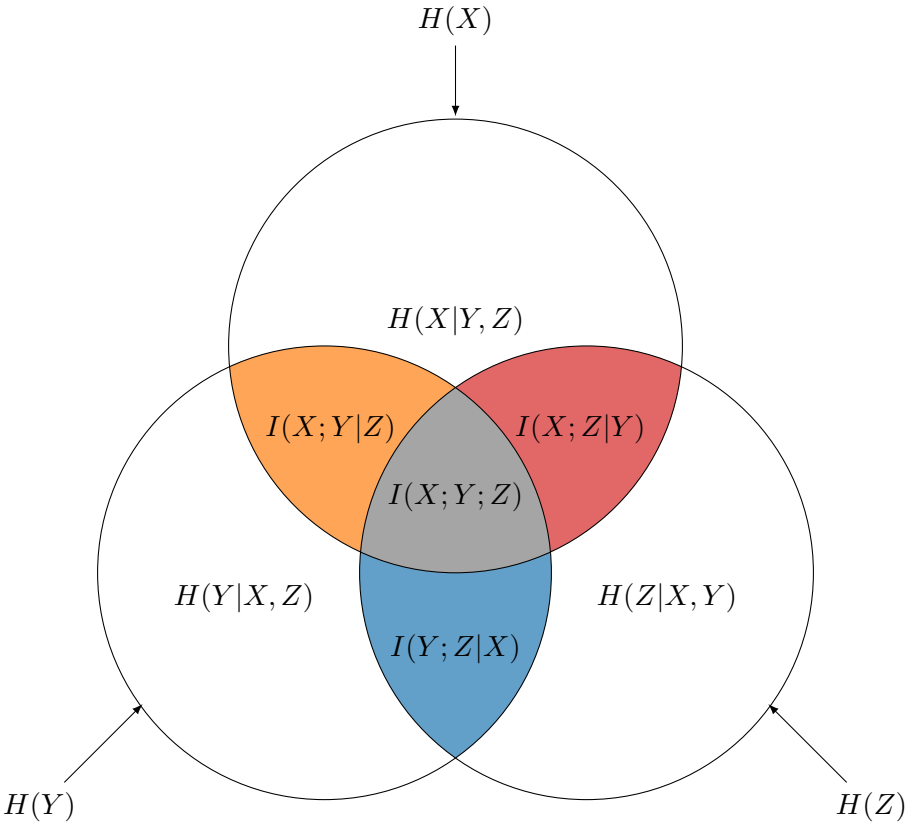

FIGURE 3.6: Venn diagram illustrating mutual information common for all variables X, Y and Z. The diagram also illustrates the entropy, joint entropy, conditional entropy and conditional mutual information. The arrows pointing towards the perimeters refer to the area inside the entire circle. The text inside the region refers to the area of that section.

3.2.1.7 Redundancy and synergy

Before moving on to the time-delayed mutual information, it is necessary to break down a common misconception about the conditional mutual information, which has key implications on the interpretation of this metric.

It may appear that mutual information between variables X and Y conditioned on Z, i.e. $I(X;Y|Z)$, should provide a measure of unique information shared between variables X and Y, given information provided by Z (Bossomaier et al. [2016]). Thus, an expectation is that the conditioning should reduce the mutual information between X and Y variables, i.e. $I(X;Y|Z) \leq I(X;Y)$, just like it was the case for the entropies and conditional entropies. However, this is not exactly the case for conditional mutual information. While conditioning in mutual information does eliminate redundancy, it also incorporates synergistic contributions from multivariate interactions between the conditional variables and the variables for which the mutual information is estimated.

To illustrate the concept of redundancy, consider three fair coins represented by random variables X, Y and Z, hence clearly $X = Y = Z$. The mutual information between two fair coins is $I(X;Y) \approx 1.443$ nats. However, if we condition the mutual information between X and Y on Z, then $I(X;Y|Z) = I(X;Z|Y) = 0$, which indicates that Y does not provide any unique information that Z already has. This is expected simply because Z and Y hold the same information about X (Bossomaier et al. [2016]). Thus, in this case its clear that conditioning eliminates redundancy, and consequently reduces the mutual information.

However, now let us consider a Boolean exclusive-OR or XOR operation between Y and Z, i.e. $X = Y \oplus Z$ (Lizier et al. [2014]). If we let Y and Z be independent and randomized, then $I(X;Y) = I(X;Z) = 0$ (Bossomaier et al. [2016]). On the other hand, if we condition the mutual information on the other variable from XOR relation, then $I(X;Y|Z) = I(X;Z|Y) \approx 1.443$ nats. In this example, conditioning reveals the information contributions that are possible to uncover only if we know Y and Z. This type of information contribution is the synergistic contributions from interaction of Y with Z (Bossomaier et al. [2016]). Synergistic information is a property of a set of random variables that can better predict a single target while cooperating than their sum of single-source predictions (Quax et al. [2017], Griffith and Koch [2014]).

While in the above examples redundant and synergistic relationships are present when they occur separately and independently, they may in fact occur simultaneously. A standard example used to illustrate this are the OR and AND logic gates (Bossomaier et al. [2016], Griffith and Koch [2014]). The fact that synergistic relationships as well as redundancy elimination may occur simultaneously poses a significant problem for the information-theoretic research community. Some researchers would like to determine unique and independent information contributions from each variable involved in the system (including us). Others, are interested in decomposing the information in a multivariate system in terms of a redundant (shared),

unique, and synergistic information contributions from each variable and actually quantifying each one of them (Williams and Beer [2010], Gutknecht et al. [2021]). The problem is that these components and dependencies cannot be decomposed and quantified with classic information-theoretic measures (Bossomaier et al. [2016]).

The idea of systematically quantifying synergies has in fact perplexed researchers now for many years, and it continues to be an open problem in information theory (Quax et al. [2017], Gutknecht et al. [2021]). It is a particularly important problem because synergy is a fundamental phenomenon for many fields such as biology, neuroscience, genetics, or physics, where various phenomena are a consequence of multivariate interactions (Quax et al. [2017], Williams and Beer [2010]). While the most common approach appears to be the Partial Information Decomposition framework formalized by Williams and Beer [2010], the common agreement on the proper approach is yet to be reached.

While this is a very interesting topic, further discussion of this problem is very complex and thus definitely out of the scope of this thesis. However, it is important to note the consequences of the above-presented issue. Namely, conditional mutual information may be either greater or smaller than the associated "unconditioned" mutual information measure, due to the multivariate interactions that are not explicitly captured by classic information-theoretic metrics (Lizier et al. [2014], Bossomaier et al. [2016]). Given the above remarks, one should also note that the areas depicted in Venn diagram in Figure 3.6 may in fact represent negative quantities (MacKay and Mac Kay [2003]). As it will become clear later in this thesis, the above-presented problem is directly impacting the transfer entropy and conditional transfer entropy metrics.

3.2.1.8 Time-delayed mutual information

In Lemma 3.26 we demonstrated that mutual information is a symmetric metric. However, the research community came up with a very simple yet powerful extension to the mutual information, which allows to induce the asymmetry and hence transform it into a directional measure. In mutual information, the asymmetry can be induced by delaying in time one of the variables, which leads to so called time-delayed mutual information. For this purpose, it is necessary to introduce another subscript t, which will denote the time index. As we now shall consider the mutual information between two time series.

Since we are transitioning into considering information-theoretic metrics for stochastic processes, the notion of stationarity and its importance needs to be clarified. As already mentioned in Section 3.1.1, the notion of stationarity is also fundamental for information-theoretic metrics that consider stochastic processes. When for example, the mutual information between two time series is considered, the underlying assumption is that the two time series are stationary. This assumption is essential, in particular when only a single realization of stochastic processes

is available for the analysis. This is because information-theoretic metrics are defined using probability distributions, and the only time we may evaluate underlying probabilistic distribution based on a single realization of the process is when that distribution does not change over time, i.e., the process is stationary (Wibral et al. [2014b]). With that being said, from this point moving forward, we will assume that all stochastic processes considered are stationary unless stated otherwise.

Definition 3.31 (**Time-delayed mutual information**)**.** Let X_t and Y_t represent two discrete stationary stochastic processes. Assume that we introduce a time lag L in variable Y_t, i.e. Y_{t-L}. Now, let $p(x_t)$ and $p(y_{t-L})$ be the marginal probability mass functions associated with X_t and Y_{t-L}, respectively (Li et al. [2018], Albers and Hripcsak [2012a]). Also, let $p(x_t, y_{t-L})$ denote the joined probability mass function. Using mutual information definition 3.24, the mutual information between the variable X_t and time-delayed variable Y_{t-L} is defined as

$$I(X_t; Y_{t-L}) \triangleq \sum_{x_t} \sum_{y_{t-L}} p(x_t, y_{t-L}) \log \left(\frac{p(x_t, y_{t-L})}{p(x_t)p(y_{t-L})} \right) \tag{3.127}$$

$$= D\left(p(x_t, y_{t-L}) || p(x_t)p(y_{t-L})\right) \tag{3.128}$$

$$= \mathbb{E}\left[\log \left(\frac{p(X_t, Y_{t-L})}{p(X_t)p(Y_{t-L})} \right) \right] \tag{3.129}$$

While the metric is identical to the basic mutual information, the introduction of the time lag parameter provides the means to induce directional sense to this metric (Schreiber [2000]).

Lemma 3.32 (Time-delayed mutual information asymmetry)**.**

$$Y_t \neq X_t \implies I(X_t; Y_{t-L}) \neq I(Y_t; X_{t-L}) \tag{3.130}$$

Proof. This is a consequence of the asymmetry that is introduced by the time lag parameter. □

Time-delayed mutual information is not a perfect metric, as it fails to account for the shared history as well as the common external driving forces (Bossomaier et al. [2016], Albers and Hripcsak [2012b]). This is in fact a quite significant limitation, especially when one considers a Markov process in which variable's past states contribute significant information about its potential future state (Bossomaier et al. [2016]). Therefore, it is necessary to filter out these contributions to avoid spurious conclusions about the flow of information from one variable to another. A solution to this limitation was proposed by Schreiber [2000] who introduced the metric called transfer entropy. Transfer entropy does account for the past of the target variable when the flow of the information between variables is estimated. In order to properly introduce the idea behind this concept, it is necessary to clarify the notions of dynamical entropies, and

explain how one can account for the uncertainty reduction due to the information about the past states of the target variable (Schreiber [2000]).

3.2.1.9 Entropy rate

Shannon entropy introduced in Definition 3.7 can also be used to characterize the dynamical structure of a random process (Kaiser and Schreiber [2002], Jumarie [1990]). Using entropy, one can easily determine how much uncertainty about the future state of the random variable is reduced by the past states of the variable.

Definition 3.33 (**Entropy rate**)**.** An entropy rate measures the stochastic process' intrinsic randomness, which is the uncertainty about the present measurement, given the information embedded in the infinitely long history (Jurgens and Crutchfield [2021]). In other words, the entropy rate is a measure of unpredictability of a stochastic process (Dulek and Schaffner [2017]). The entropy rate of stochastic process X_t is defined in the following manner

$$H\left(\{X_t\}\right) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{3.131}$$

where X_n denotes the state of the process at time $t = n$. Entropy rate in a way also measures how the entropy of a sequence grows as n grows large (Dulek and Schaffner [2017]). Alternatively, entropy rate can be interpreted as the average number of nats (or equivalently bits for log base 2) needed to encode one additional state of the process, given the knowledge about all the previous states of the system (Prokopenko et al. [2013])

With entropy rate being introduced, we can now also consider the measure of uncertainty about the next state state given some fixed-length history.

Definition 3.34 (**Entropy of the future state given the past d_x states**)**.** Consider a discrete stochastic process X_t, and let X_n denote the state at time $t = n$. Let $p(x_{n+1}|\boldsymbol{x_n}^{(d_x)})$ denote the transition probability, where $\boldsymbol{x_n}^{(d_x)}$ is a vector of past d_x states (as will be discussed later in this thesis d_x is called embedding dimension) (Kaiser and Schreiber [2002], Prokopenko et al. [2013]). The Shannon entropy of the future state given past d_x states is defined as (Kaiser and Schreiber [2002]):

$$H(X_{n+1}|\boldsymbol{X_n}^{(d_x)}) = \sum_{\boldsymbol{x_n}^{(d_x)}\in\mathcal{X}^{d_x}} p(\boldsymbol{x_n}^{(d_x)}) \sum_{x_{n+1}\in\mathcal{X}} p(x_{n+1}|\boldsymbol{x_n}^{(d_x)}) \log\left(\frac{1}{p(x_{n+1}|\boldsymbol{x_n}^{(d_x)})}\right) \tag{3.132}$$

$$= \sum_{\boldsymbol{x_n}^{(d_x)}\in\mathcal{X}^{d_x}} \sum_{x_{n+1}\in\mathcal{X}} p(x_{n+1}, \boldsymbol{x_n}^{(d_x)}) \log\left(\frac{1}{p(x_{n+1}|\boldsymbol{x_n}^{(d_x)})}\right) \tag{3.133}$$

If one has more information about the underlying process of the stochastic process, then more properties of the conditional entropy can be derived. As presented below, we can consider a stationary process or a stationary Markov process.

Theorem 3.35 (Entropy of stationary stochastic process)**.** *Consider a stationary stochastic process X_t, and let X_n denote the state at time $t = n$. The entropy of a stationary stochastic process is not increasing with n (Thomas and Joy [2006])*

$$H(X_{n+1}|X_1, X_2, \ldots, X_n) \leq H(X_{n+1}|X_n, \ldots, X_2) = H(X_n|X_{n-1}, \ldots, X_1) \tag{3.134}$$

Proof. The above follows from the properties of stationary stochastic process discussed in Section 3.1.1. □

Theorem 3.36 (Entropy of stationary Markov process)**.** *Consider a stationary Markov stochastic process X_t, and let X_n denote the state at time $t = n$. The entropy of a stationary Markov stochastic process is increasing with n (Thomas and Joy [2006])*

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) \tag{3.135}$$

$$H(X_n|X_1, X_2) = H(X_n|X_2) \tag{3.136}$$

$$H(X_n|X_2) = H(X_{n-1}|X_1) \tag{3.137}$$

Where inequality in Equation (3.135) follows from the fact that conditioning can only reduce the entropy. Equality in Equation (3.136) is by Markovity, and equality in Equation (3.137) is per property of a stationary process.

3.2.2 Transfer entropy & conditional transfer entropy

3.2.2.1 Definition of transfer entropy

Schreiber [2000] introduces the *transfer entropy*, also referred to as the *apparent transfer entropy*, to overcome the challenges posed by mutual information metric (Lizier and Rubinov [2012], Lizier and Prokopenko [2010]).

The two main challenges were the fact that mutual information is a symmetric metric, and that it does not account for the shared history, nor for the common external driving forces (Bossomaier et al. [2016], Albers and Hripcsak [2012b]). As discussed in the previous subsection, this is especially important when one considers Markov processes, as the previous state of the process may in fact provide a lot of information about the next state. With that in mind, Schreiber [2000] proposes to use Kullback-Leibler divergence to measure the deviation from the generalized Markov property, i.e.:

$$p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}) = p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \tag{3.138}$$

As Schreiber [2000] states, the dynamics of the processes is embedded in the transition probabilities. In case there is no flow of information from process x_t (source) to process y_t (target), then x_t will not have any impact on the transitional probabilities of process y_t, and then the generalized Markov property will hold (Schreiber [2000], Bossomaier et al. [2016]). On the other hand, if the source process provides any unique information and consequently improves the prediction of the future state of target process, or in other words reduces the uncertainty about the future state of target process, then one could say that the next state of target process is not independent of the source process (Wibral et al. [2014b]). Therefore, to quantify the potential reduction of the uncertainty about the future state of y_t given the information about source process x_t, Schreiber [2000] proposes to quantify any deviations from the generalized Markov property assumption by computing the Kullback-Leibler divergence between the transitional probabilities from Equation (3.138).

Definition 3.37 (**Transfer entropy**)**.** Consider two discrete stochastic process X_t and Y_t with marginal probability mass functions $p(x_t)$ and $p(y_t)$. Let $p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})$ denote the transition probability, where $\boldsymbol{y}_t^{(d_y)}$ is a vector of past d_y states of process y_t. The transfer entropy from x_t to y_t with source-target time-lag $L = 1$ is defined as:

$$\mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)} \triangleq D\left(p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})||p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})\right) \tag{3.139}$$

$$= \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log\left(\frac{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})}{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})}\right) \tag{3.140}$$

$$= \mathbb{E}\left[\log\left(\frac{p(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)})}{p(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)})}\right)\right] \tag{3.141}$$

where $p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})$ denotes the joined probability mass function. At this point, it is necessary to emphasize the underlying assumption behind the history length of the target process, i.e. d_y. History length is also called an embedding dimension, however this will be further discussed in Section 4.3.2.1. By choosing some finite history length of target process Y_t, we are making an assumption that Y_t is a Markov process of order d_y. Formally, the following assumption is made: $p(y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = p(y_{t+1}|y_t, y_{t-1}, \ldots, y_{t+1-d_y})$. Thus, the assumption is that the process is conditionally independent from the states of history further than the past d_y states. Also, bear in mind that for the sake of simplicity, we choose time-lag $L = 1$, however generally any source-target time-lag can be used. Note that source-target time-lag will be discussed in detail in Section 4.3.2.3.

Thus, transfer entropy is simply mapping the difference between the hypotheses that the transitional probability of Y_t is independent of X_t, in a comparable manner to how mutual information measures the deviation from the hypothesis that the processes are independent (Michalowicz et al. [2013]).

Finally, it also important to stress that transfer entropy does not quantify true causal relationship, but it quantifies a probabilistic account of causality based on observational causality just like Wiener-Granger causality (Wibral et al. [2012]). Moreover, it has been determined that Granger causality and transfer entropy are equivalent for exponential Wienmann, log-normal and Gaussian distributions up to factor of two (Barnett et al. [2009], Hlaváckova-Schindler [2011])

Basic properties of transfer entropy such as non-negativity and non-symmetricity are presented in the following lemmas.

Lemma 3.38 (Transfer entropy non-negativity)**.**

$$\mathrm{TE}^{(d_y,d_x)}_{X_t\to Y_t} \geq 0 \tag{3.142}$$

$$\mathrm{TE}^{(d_y,d_x)}_{X_t\to Y_t} = 0 \iff p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}) = p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \quad \forall\, y_t \in Y_t \tag{3.143}$$

Proof. The properties follow from definition and previously proofed properties of Kullback-Leibler divergence (Lemma 3.22). □

Lemma 3.39 (Transfer entropy non-symmetricity)**.**

$$Y_t \neq X_t \implies \mathrm{TE}^{(d_y,d_x)}_{X_t\to Y_t} \neq \mathrm{TE}^{(d_y,d_x)}_{Y_t\to X_t} \tag{3.144}$$

Proof. The argument is analogous to argument used for time-delayed mutual information (Lemma 3.32). □

Theorem 3.40 (Transfer entropy is conditional time-delayed mutual information)**.**

$$\mathrm{TE}^{(d_y,d_x)}_{X_t\to Y_t} = I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)}|\boldsymbol{Y}_t^{(d_y)}) \tag{3.145}$$

Proof. See Appendix B, Section B.10 □

Theorem 3.40 and Definition 3.37 reveal a close relation between information-theoretic metrics. The transfer entropy is defined with Kullback-Leibler divergence measure, but it can also be represented as the conditional time-delayed mutual information. Furthermore, from the previous subsections we know that conditional mutual information can be decomposed into more granular fashion in terms of conditional entropies, which can be further simplified to joint entropies and individual entropies. With that in mind, we can granularize the transfer entropy metric in order to reveal the underlying intuitition behind this metric using more fundamental metrics. In fact, the recipe for this treatment is already quiet strongly ingrained in the scientific community that zooms in on the so called local transfer entropy, the research being led, among other, by Professor Joseph Lizier (Lizier and Prokopenko [2010], Lizier et al. [2012], Lizier et al. [2014]).

Theorem 3.41 (Decomposition of the uncertainty about the future state)**.** Consider two stochastic processes X_t and Y_t. Let us further assume that these are the only processes that exist in the system, and that the information flows from X_t to Y_t. Then, the uncertainty about the future state Y_{t+1} of target process Y can be decomposed into three components: active information storage, transfer entropy from X_t to Y_t, and the remaining intrinsic uncertainty about the future state of target process (Lizier and Prokopenko [2010]).

$$H(Y_{t+1}) = \underbrace{I(Y_{t+1}, \boldsymbol{Y}_t^{(d_y)})}_{AIS(Y^{(d_y)})} + \mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)} + \underbrace{H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)})}_{IU(Y_{t+1})} \tag{3.146}$$

where $\mathrm{AIS}(Y^k)$ stands for the active information storage and $IU(Y_{t+1})$ denotes the remaining intrinsic uncertainty in the future state of Y given its history and the source X (Lizier and Prokopenko [2010]).

Proof. By Theorem 3.40, transfer entropy can be represented as a conditional mutual information. The conditional mutual information can be decomposed into two conditional entropies by Equation (3.125). Hence, the following decomposition of transfer entropy can be performed

$$\begin{aligned}
\mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)} &= I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)}|\boldsymbol{Y}_t^{(d_y)}) && (3.147)\\
&= \underbrace{H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)})}_{H(Y_{t+1},\boldsymbol{Y}_t^{(d_y)})-H(\boldsymbol{Y}_t^{(d_y)})} - H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) && (3.148)\\
&= \underbrace{H(Y_{t+1}, \boldsymbol{Y}_t^{(d_y)})}_{H(Y_{t+1})+H(\boldsymbol{Y}_t^{(d_y)}|Y_{t+1})} - H(\boldsymbol{Y}_t^{(d_y)}) - H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) && (3.149)\\
&= H(Y_{t+1}) + \underbrace{H(\boldsymbol{Y}_t^{(d_y)}|Y_{t+1}) - H(\boldsymbol{Y}_t^{(d_y)})}_{-I(Y_{t+1},\boldsymbol{Y}_t^{(d_y)})} - H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) && (3.150)\\
&= H(Y_{t+1}) - I(Y_{t+1}, \boldsymbol{Y}_t^{(d_y)}) - H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) && (3.151)
\end{aligned}$$

Let us consider rearranging the terms to define $H(Y_{t+1})$ i.e. the uncertainty about the future state Y_{t+1}. Simple terms rearrangement leads to

$$H(Y_{t+1}) = \underbrace{I(Y_{t+1}; \boldsymbol{Y}_t^{(d_y)})}_{AIS(Y^{(d_y)})} + \mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)} + \underbrace{H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)})}_{IU(Y_{t+1})} \tag{3.152}$$

□

The above-presented decomposition shows a different and more granular perspective on the uncertainty of the future state of a process. Active information storage is a relatively new metric introduced by Lizier et al. [2012]. Lizier defined active information storage as *the average mutual*

information between the semi-infinite past of the process and its next state (Lizier et al. [2012]). Thus, active information storage simply quantifies the contribution of the past states $\boldsymbol{Y_t}^{d_y}$ of the target process Y_t to the reduction of the uncertainty about its future state Y_{t+1}. In other words, it represents the information gained from the past of the target that is directly used to compute the next state of the process (Lizier and Prokopenko [2010], Lizier et al. [2012], Lizier et al. [2014]).

The second term $\mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)}$ represents the information transferred from the source process about the future state of the target process. And finally, $IU(Y_{t+1})$ is the remaining uncertainty about target's future state after considering the transferred information from the source and active storage information from target's past states (Lizier et al. [2014]). The notion of active storage information is very involved concept, and a more detailed account on this metric is out of the scope of this thesis.

This representation closely aligns with the transfer entropy interpretation that was developed until this point. To sum up, transfer entropy quantifies the unique predictive in the probabilistic sense information provided by the source process about the future state of the target process.

3.2.2.2 Interpretation of transfer entropy

Before moving to conditional transfer entropy, it is necessary to elaborate on a few particularly important aspects regarding the interpretation of the transfer entropy measure. MacKay and Mac Kay [2003] provide broad account on recent studies that demonstrate why transfer entropy should be interpreted as predictive information transfer. In similar fashion Lizier and Rubinov [2012], as well as Williams and Beer [2011] clearly highlight various misconceptions about transfer entropy and conditional transfer entropy metrics.

First of all, the absolute value of the transfer entropy does not measure "causal effect size" (MacKay and Mac Kay [2003], Chicharro and Ledberg [2012]). As Chicharro and Ledberg [2012] state, in complex systems, subsystems are usually connected bidirectionally, and hence the interactions between the subsystems should not always be considered as simple cause-and-effect interaction. Moreover, as Williams and Beer [2011] further clarify, while transfer entropy's conditioning on the target's history remove the information shared by the source and target due to common histories, it also adds in higher-order synergistic information due to the interactions of the source's and target's history (if such exists) (Williams and Beer [2010], Williams and Beer [2011]). In a more recent publication Wibral et al. [2017] elaborate on these potential interactions and refers to them as the "state dependent transfer entropy."

Secondly, it is very important to realize that transfer entropy is not a measure of coupling strength. While for example Kaiser and Schreiber [2002] provide a detailed account on how coupling between two linear processes with Gaussian distributions translates to transfer entropy,

it should not be interpreted in this manner. As discussed, there are much more higher-order interactions at play, especially in the systems with many variables involved, and hence transfer entropy cannot be used to predict and estimate potential coupling between just a pair of variables (Wibral et al. [2014b]). Moreover, one should note a simple example when transfer entropy fails to identify potential coupling between variables. If we increase coupling between two variables up to the point when variables are pretty much completely synchronized, the transfer entropy measure will not detect any information flow (Wibral et al. [2014b]). This limitation is observed and clearly described by Gourévitch et al. [2006] based on variables modelled with the ARX model.

3.2.2.3 Definition of conditional transfer entropy

The transfer entropy is a sufficient metric to quantify probabilistic causal relationships when two processes are considered. However, the ability of transfer entropy to identify unique information contributions becomes a lot more tricky when more than two processes are jointly distributed (Bossomaier et al. [2016]). If we consider three processes, i.e. X_t, Y_t and Z_t, there are many possible types and combinations of interactions that can take place between these processes. For example, the so called common driver effect where process Z_t has an impact on both processes X_t and Y_t, but a different lags (say $X_{t-1} = Z_t$ and $Y_{t-2} = Z_t$), then the analysis with the use of transfer entropy could lead to spurious conclusions that suggest that process X_t has an impact on the future states of Y_t process,i.e. $X_t = Y_{t-1}$ while in fact the process Z_t is the driving force (Bossomaier et al. [2016]).

Therefore, it is necessary to filter out the influences from other sources, in order to directly quantify the influence of individual source process on the future state of the target process. As in the case of many other metrics presented up to this point, there is a conditional analog for transfer entropy called conditional transfer entropy.

Definition 3.42 (**Conditional transfer entropy**)**.** Consider three discrete stochastic process X_t, Y_t and Z_t with marginal probability mass functions $p(x_t)$, $p(y_t)$ and $p(z_t)$ and joint probability mass function $p(x_t, y_t, z_t)$. Furthermore, let $\boldsymbol{y}_t^{(d_y)}$, $\boldsymbol{x}_t^{(d_x)}$ and $\boldsymbol{z}_t^{(d_z)}$ denote vectors of past d_y, d_x and d_z states of process Y_t, X_t and Z_t, respectively. The conditional transfer entropy, i.e. the transfer entropy from x_t to y_t conditioned on z_t is defined as:

$$\mathrm{CTE}_{X_t \to Y_t | Z_t}^{(d_y, d_x, d_z)} \triangleq D\left(p(y_{t+1} | \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)}) || p(y_{t+1} | \boldsymbol{y}_t^{(d_y)}, \boldsymbol{z}_t^{(d_z)})\right) \tag{3.153}$$

$$= \sum_{\boldsymbol{z}_t^{(d_z)}} \sum_{\boldsymbol{x}_t^{(d_x)}} \sum_{\boldsymbol{y}_t^{(d_y)}} \sum_{y_{t+1}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)}) \log \left(\frac{p(y_{t+1} | \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)})}{p(y_{t+1} | \boldsymbol{y}_t^{(d_y)}, \boldsymbol{z}_t^{(d_z)})} \right) \tag{3.154}$$

$$= \mathbb{E}\left[\log \left(\frac{p(Y_{t+1} | \boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}, \boldsymbol{Z}_t^{(d_z)})}{p(Y_{t+1} | \boldsymbol{Y}_t^{(d_y)}, \boldsymbol{Z}_t^{(d_z)})} \right) \right] \tag{3.155}$$

Note that the subscript of summation has been omitted, however each respective variable is summed over its own alphabet, i.e. $\boldsymbol{z}_t^{(d_z)} \in \mathcal{Z}^{d_z}, \boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}, \boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}, y_{t+1} \in \mathcal{Y}$. From the above definition, it is clear that conditional transfer entropy is nothing but transfer entropy conditioned on another causal information contributor (Lizier and Rubinov [2012], Lizier et al. [2010]). Hence, it can also be represented as conditional time-delayed mutual information.

Theorem 3.43 (Conditional Transfer entropy is conditional time-delayed mutual information)**.**

$$\mathrm{CTE}_{X_t \to Y_t | Z_t}^{(d_y, d_x, d_z)} = I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)} | \boldsymbol{Y}_t^{(d_y)}, \boldsymbol{Z}_t^{(d_z)}) \tag{3.156}$$

Recall the remarks made in Section 3.2.1.7 about the consequences of conditioning mutual information on other causal information contributors. While the conditional transfer entropy eliminates the transfer of information from X_t to Y_t which is redundant given information about Z_t, it also captures potential synergistic contributions from multisource interactions of X_t and Z_t, as well as their interactions with the source's history vector (Lizier and Rubinov [2012]).

The conditional transfer entropy can be used to condition on any number of possible sources. One can distinguish a special case of transfer entropy, commonly known as the complete transfer entropy (Lizier et al. [2010]).

Definition 3.44 (**Complete transfer entropy**)**.** Consider two discrete stochastic processes: source X_t, and target Y_t and a set of all the other causal information contributors to Y_t, i.e. $\mathbf{S}_{\mathcal{Z}} = \{Z_{1,t}, Z_{2,t}, \ldots, Z_{n,t}\}$, where $Z_{i,t} \neq Y_t, X_t \quad \forall i$.

$$\mathrm{CTE}_{X_t \to Y_t | \mathbf{S}_{\mathcal{Z}} t}^{(d_y, d_x, d_z)} = I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)} | \boldsymbol{Y}_t^{(d_y)}, \mathbf{S}_{\mathcal{Z}} t^{(\boldsymbol{d_z})}) \tag{3.157}$$

where $\boldsymbol{d_z}$ is the vector of history lengths considered for each contributor in set $\mathbf{S}_{\mathcal{Z}}$. As Lizier et al. [2010] notes, since complete transfer entropy it accounts for all the interactions of the sources in stochastic systems the complete transfer entropy can only occasionally be negative, due to statistical fluctuations.

There is also the collective transfer entropy; however, the discussion of this does not add any value to this thesis. Hence, I refer interested readers to an extensive work by Professor Joseph Lizier, especially the following publication Lizier et al. [2010].

3.2.3 Entropy of a continuous distribution

In Section 3.2.1, the information-theoretic measures are discussed in the context of discrete random variables and probability mass functions associated with them. In a similar manner, all the metrics presented earlier can be defined for a continuous random variables with associated probability density functions (Shannon [1948]). Following the steps of Claude Shannon, we define the Shannon entropy of a continuous random variable, nowadays referred to as *differential entropy*.

3.2.3.1 Differential entropy

Definition 3.45 (**Differential entropy**)**.** A differential entropy of a continuous random variable X with probability density function $f(x)$ with the set-theoretic support $S = \{x \in X : f(x) \neq 0\}$ is defined in the following manner (Thomas and Joy [2006], Kolmogorov [1993]):

$$h(X) \triangleq -\int_S f(x) \log\left(f(x)\right) \, dx \tag{3.158}$$

While differential entropy retains most of the properties presented for the discrete entropy, there are a few ways in which it actually departs from them (Shannon [1948], Michalowicz et al. [2013]). First of all, differential entropy can take negative values, i.e. it is no longer non-negative as previously stated in Lemma 3.10. To illustrate this, consider the following example in which differential entropy of a uniform distribution is computed.

Example 3.3 (Differential entropy of uniform distribution)**.** *Let X denote a continuous random variable distributed uniformly in range $[0, a]$. Thus, the probability density function associated with X is $f(x) = 1/a \quad \forall x \in [0, a]$. The differential entropy of this uniform distribution is*

$$h(X) = -\int_0^a \frac{1}{a} \log\left(\frac{1}{a}\right) dx = \log\left(a\right) \tag{3.159}$$

From the above result it should be clear that for $a < 1$, $\log(a) < 0 \implies h(X) < 0$. Thus, clearly differential entropy can take negative values, as opposed to discrete entropy.

Second of all, the continuous entropy (differential entropy) is not a "natural" extension of discrete entropy (Michalowicz et al. [2013]). To illustrate this, consider a continuous random variable X with Riemann integrable probability density $f(x)$. Now, let us suppose that we divide the space of X into equisized bins of length Δ, thus X^{Δ} is simply a discretized analog of X (Michalowicz et al. [2013]). If differential entropy was a "natural" extension of a discrete entropy, then the following relation should hold $\lim_{\Delta\to 0} h(X^{\Delta}) \to h(X)$, which in fact is not the case (Thomas and Joy [2006]).

Theorem 3.46 (Differential entropy and Discrete entropy)**.**

$$\lim_{\Delta\to 0} h(X) - h(X^{\Delta}) = \log(\Delta) \tag{3.160}$$

Proof. See Appendix B, Section B.9. □

As proofed above, the differential entropy is not a "natural" extension of a discrete entropy. Additionally, as presented in Equation (3.159), the differential entropy cannot be interpreted as a measure of uncertainty, since $f(x)$ now represents the probability density, which would need to be integrated over some finite interval to acquire the meaning of probability (Michalowicz et al. [2013]). Thus, differential entropy is a functions that describes the associated probability distribution and can be interpreted as a measure of relative uncertainty (Michalowicz et al. [2013]). While the above-presented concept is important to mention for the completeness, the rest of the metrics presented in the previous section can be naturally extended into continuous random variables without any consequences. Furthermore, the continuous forms of mutual information, transfer entropy and conditional transfer entropy retain interpretations from their respective discrete analogs (Michalowicz et al. [2013]).

As the thesis will be focused on estimating the continuous apparent and conditional transfer entropies, it is necessary to briefly reintroduce these measures in their continuous forms.

3.2.3.2 Transfer entropy & conditional transfer entropy

Definition 3.47 (**Transfer entropy**)**.** Consider two continuous processes X_t and Y_t with marginal probability density functions $f(x_t)$ and $f(y_t)$. Let $f(y_{t+1}|\boldsymbol{y}_t^{(d_y)})$ denote the transition probability, where $\boldsymbol{y}_t^{(d_y)}$ is a vector of past d_y states of process y_t. The transfer entropy from X_t to Y_t with time-lag $\tau = 1$ is defined as:

$$\mathrm{TE}_{X_t\to Y_t}^{(d_y,d_x)} \triangleq D\left(f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})||f(y_{t+1}|\boldsymbol{y}_t^{(d_y)})\right) \tag{3.161}$$

$$= \int_{\mathbb{R}^{d_x}}\int_{\mathbb{R}^{d_y}}\int_{\mathbb{R}} f(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log\left(\frac{f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})}{f(y_{t+1}|\boldsymbol{y}_t^{(d_y)})}\right) dy_{t+1} d\boldsymbol{y}_t^{(d_y)} d\boldsymbol{x}_t^{(d_x)} \tag{3.162}$$

where $f(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})$ denotes the joined probability density function. Additionally, the $\int_{\mathbb{R}^D}$ denotes the D-dimensional integral over the support of the variable (Michalowicz et al. [2013]). Similarly to the discrete form of transfer entropy, d_y and d_x represent the history lengths of processes Y_t and X_t, respectively.

Analogously to Definition 3.47, the definition of conditional transfer entropy is presented.

Definition 3.48 (**Conditional transfer entropy**)**.** Consider three continuous stochastic process X_t, Y_t and Z_t with marginal probability density functions $f(x_t)$, $f(y_t)$ and $f(z_t)$ and joint probability density function $f(x_t, y_t, z_t)$. Furthermore, let $\boldsymbol{y}_t^{(d_y)}$, $\boldsymbol{x}_t^{(d_x)}$ and $\boldsymbol{z}_t^{(d_z)}$ denote vectors of past d_y, d_x and d_z states of process Y_t, X_t and Z_t, respectively. The conditional transfer entropy, i.e. the transfer entropy from x_t to y_t conditioned on z_t is defined as:

$$\mathrm{CTE}_{X_t \to Y_t | Z_t}^{(d_y, d_x, d_z)} \triangleq D\left(f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)}) || f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{z}_t^{(d_z)})\right) \tag{3.163}$$

$$= \int_{\mathbb{R}^{d_z}} \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \int_{\mathbb{R}} f(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)}) \times \tag{3.164}$$

$$\log\left(\frac{f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}, \boldsymbol{z}_t^{(d_z)})}{f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{z}_t^{(d_z)})}\right) dy_{t+1} d\boldsymbol{y}_t^{(d_y)} d\boldsymbol{x}_t^{(d_x)} d\boldsymbol{z}_t^{(d_z)}$$

Chapter 4

Methods

The following chapter introduces data processing and parameter selection methods employed in both the econometric model and network inference algorithm. First, data exploratory and preprocessing efforts are described. In particular, the notion of time stamp synchronization is addressed, and the data sampling procedure is explained. Next, the Kraskov algorithm used to estimate transfer entropies is treated in detail. Additionally, the optimal delay embedding is introduced. The Ragwitz criterion and its role in determining the optimal parameter settings for the Kraskov algorithm is also addressed. Finally, the network inference algorithm implementation efforts and the main challenges faced are outlined.

4.1 Data preprocessing

4.1.1 Data description

In this thesis, we use three independent high-frequency FX spot rate data sets; EUR/USD, USD/JPY and EUR/CHF. Model development and validation is carried out using the EUR/CHF data set, whereas the quantitative analysis is performed on EUR/USD and USD/JPY data sets. All data sets are comprised of privately collected, irregularly spaced, temporal data of FX rates of different dealers, provided to us by courtesy of ING Netherlands.

The EUR/USD data set is comprised of 109,449,980 high-frequency observations of the best[1] bid and ask prices quoted by eight parties between February 27 2020, and March 27 2020. This time period encompasses 22 trading days and 8 non-trading days (weekends)[2].

[1] At a particular point in time a, dealer can provide multiple ask and bit quotes for different volumes at different price points. The best bid and ask prices are the ones that represent the narrowest spread, or in other words, the most competitive ask and bid prices. The best ask and bid prices are used to closely follow the data preprocessing steps described by Hagströmer and Menkveld [2019]. Further discussion of this choice can be found in Chapter 6.

[2] While there is no trading on Saturday, on Sunday trading starts at 23:00 UTC+1. The reason for this is thoroughly explained in Section 5.3.

The USD/JPY data set is comprised of 14,284,491 observations of bid and ask prices quoted by six parties between January 1 2019, and January 17 2019. The data set is significantly shorter than the one with EUR/USD price quotes due to the data availability issues. In this period, there are 12 trading days and 4 non-trading days.

The EUR/CHF data set is comprised of 761,012 high-frequency observations of the best bid and ask prices quoted by eight parties on the day of the Swiss franc crash on January 15 2015. Given that access to this data was available at the beginning of this thesis project, it was used mainly for model development and validation purposes. While this data set was considered to be used in the quantitative investigation, unfortunately, the data from the days following the Swiss franc crash were unavailable. Hence the investigation in its full scope could not be performed on this data set.

In our quantitative investigation, we are going to distinguish three types of trading parties; the market makers (M)[3], banks (B), and electronic trading platforms (E). Dealers' presence in particular data set is outlined in Table 4.1. While the data will be described in more detail later in Chapter 5, in this section the primary focus is on the considerations of data preprocessing.

Each data point is marked with two different timestamps, hence, we need to decide which one should be used for price discovery investigation. Moreover, we also have to address the possibility that the temporal data are not correctly synchronized. This is very important, especially when one recalls the remarks made about latency adjustment and the fact that it was the main limitation of the methodology from Hagströmer and Menkveld [2016].

Data sets		Dealers										
Currency pair	Time period	B1	B2	B3	B4	B5	B6	M1	E1	E2	E3	E4
EUR/USD	2020/02/27 - 2020/03/27	x	x	x	x	x	x	x	x			
USD/JPY	2019/01/01 - 2019/01/17	x	x	x	x		x		x			
EUR/CHF	2015/01/15	x			x		x		x	x	x	x

TABLE 4.1: Table presents the availability of dealer's FX spot rates in different data sets.

[3]Banks usually act as market makers as well, however, here, we are making a distinction between a bank and a non-bank market maker.

4.1.2 Latency adjustment

In the context of the FX market, the latency is the time it takes for an electronic signal to travel from its origin to its destination (Addison et al. [2019], Hagströmer and Menkveld [2016]). Thus, latency captures the time between the action of a dealer updating their quote and the moment when other dealers observe that particular update. Nowadays, the average latencies have been reduced to a fraction of a millisecond with one millisecond being the upper bound (Menkveld [2013], Hasbrouck and Saar [2013], Chen et al. [2018]).

Latency is a crucial concept that needs to be addressed here since it may distort our perception of the actions and reactions in the dealer-network. For example, the quote update from bank B1 may travel for 10 milliseconds to the ING's server, whereas the signal from bank B2 may travel for 100 milliseconds. Consequently, based on the ING's quote reception timestamps, one could incorrectly infer that B1 made a quote update before B2, even if their quote updates were simultaneous.

As already mentioned, each data point in both EUR/USD and USD/JPY data sets come with two different time steps; the original timestamp that is assigned at the origin by the dealer who posts the quote (Timestamp 1) and the timestamp that ING sets at the moment when the quote update is received in the system (Timestamp 2). Both timestamps are collected and stored at a precision of 1 millisecond.

Since our goal in this thesis is to accurately uncover the flow of the information and describe the price discovery process, the best choice is to use Timestamp 1. This is because Timestamp 1 reflects the quote updates from the perspective of the market participant posting that particular quote without any latency[4]. On the other hand, Timestamp 2 would present the process of price discovery from ING's perspective, as it would include the latencies mentioned above and hence distort our perception of dealers interactions.

The problem with using Timestamp 1 is that there is no way of verifying if the timestamps were assigned correctly. This is a concern since it is not uncommon for a dealer clock to be not synchronized with other clocks. Hence, given that we would like to use the timestamps assigned by other market participants at the origin, we must ensure that there are no anomalies present in the data. Since we do not have access to any other data than the one provided by ING, the only viable way to ensure latency consistency is to compare Timestamp 1 and Timestamp 2. Thus, the underlying assumption is that the ING's clock has been adequately synchronized and remained so over the entire time. Hence it will be used as a point of reference. With that established, the comparison of timestamps can be made to ensure that the difference between the two timestamps is not drastically changing over time, i.e. the latencies stay approximately constant.

[4]Note that Hagströmer and Menkveld [2019] did not have a choice to use their timestamps. The authors had to work with the timestamps as they were assigned by OLSEN data.

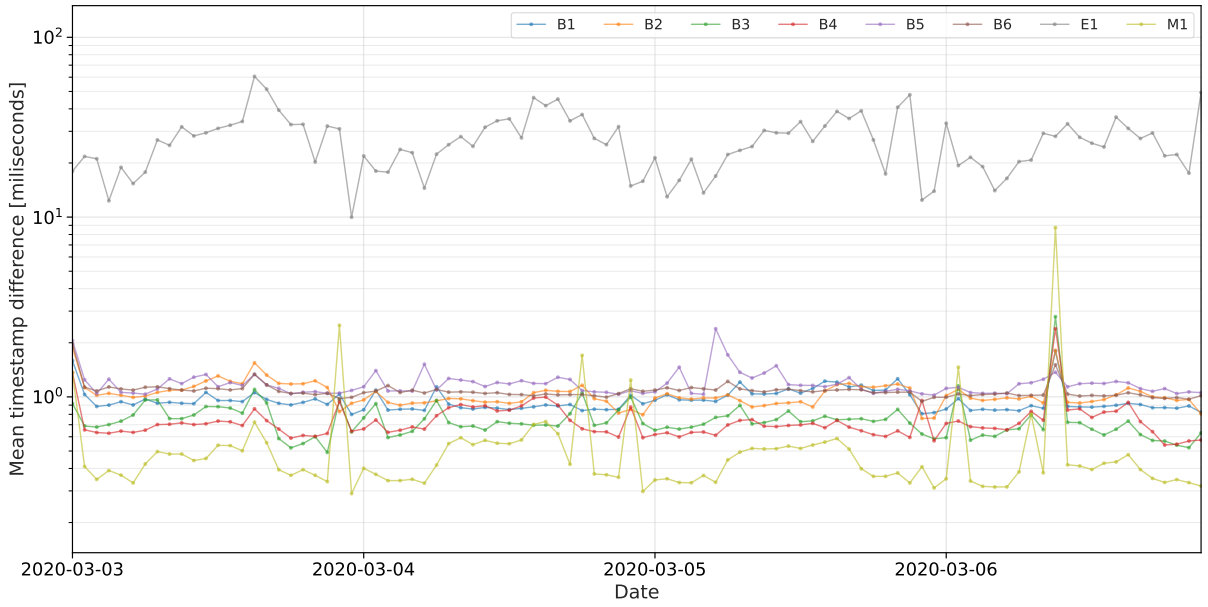

FIGURE 4.1: Mean difference between Timestamp 1 and Timestamp 2 aggregated per hour for the period from March 3 2020 till March 7 2020 in EUR/USD data set. Figure reveals how mean signal latency for various dealers fluctuates over time. The y-axis is in logarithmic scale.

In Figure 4.1 one can observe mean differences per hour between the ING's timestamp and timestamps assigned by different counter-parties in the EUR/USD data set. Thus, in essence we observe an average signal latency from ING's perspective. From the figure, it is clear that the mean difference between Timestamp 1 and Timestamp 2 is generally smaller than 100 milliseconds. The highest latency reaching close to 100 milliseconds is observed for electronic trading platform E1. Additionally, it can be observed that on March 6 2020, there is a spike in latency that is observed for all counter parties. While this spike is very clear on the figure, it's amplitude is not larger than 10 milliseconds as compared to the "base" latencies observed for each counter-party. Given that the data will be further resampled into 100 millisecond intervals, the spike will not have any impact on the results, and thus it can be disregarded.

While in this figure, we only take a closer look at one week of mean disparities between timestamps, the investigation is performed on the entire EUR/USD and USD/JPY data sets. This exploration does not reveal any anomalies in either EUR/USD or USD/JPY data sets. For all the counter-parties the latency remains approximately constant. Consequently, no latency adjustment is needed.

4.1.3 Data sampling

As mentioned in the previous section, data with FX spot rates from each dealer are irregularly spaced in time simply because quote updates from dealers are posted at different times and in various time intervals between consecutive updates. Each market participant has the freedom to update its quotes whenever deemed necessary. However, to employ both the econometric and information-theoretic models, the temporal data must be transformed into regularly spaced data with the same sampling frequency applied to all time series.

There are multiple reasons for that, in particular the fact that we need to somehow put all quote prices from different dealers on equal footing (Hagströmer and Menkveld [2016]). Only after unifying the time-space for quotes of all dealers can we perform an analysis on them. Moreover, both econometric and information-theoretic models rely on the data with an evenly spaced time grid. In essence, regularly spaced data amount to having a consistent number of sample points for each time series.

Many factors need to be accounted for to choose the sampling period correctly. First of all, note that the precision of the time stamps is one millisecond; hence sampling to a higher sampling period, e.g., 20 milliseconds, essentially comes down to aggregating various observations from those 20 milliseconds into one data point. Thus by undersampling, we condense the informational content of a particular series of FX rates and therefore potentially lose relevant information.

To maintain all the information that could potentially be extracted from the data, it would be best to resample the data to the lowest time interval between two quote updates of any two dealers, i.e. the precision of the clock. Only then would we be able to map all of the observations on an evenly spaced time grid. However, this approach would undoubtedly lead to a significant increase in the overall number of data points. And as it will be established in Section 4.3.1, the computational effort of the Kraskov algorithm that we will use to estimate the entropies increases as the length of the time series increases. Hence, oversampling is simply an unfeasible approach given the scope and the time constraints of this thesis, and the computational limitations.

We also need to consider the conclusions from our latency investigation. Since we are hoping to observe dealers' responses to changes in other dealers' quotes, we need to account for how the signal latency affects the reaction time of dealers to the changes in other dealers' quotes. This is especially important for the econometric model, which inherently assumes that the reaction of dealers to disparities in quotes happens at the first lag $(t-1)$. This is a smaller concern for the information-theoretic approach, for which, as discussed later in this section, we will scan through different delays and determine the true delay between the signal and the response. Given that the highest mean latency observed on Figure 4.1 is close to 100 milliseconds, it is reasonable to use it as our sampling period. Assuming that other dealers do not experience significantly higher latencies than the ones we observed, 100 milliseconds is sufficient time for all dealers to

observe any other dealers' updates and take action. Additionally, this choice also aligns with the sampling period employed in both Hagströmer and Menkveld [2016] and Hagströmer and Menkveld [2019].

Finally, the so-called forward-fill approach is applied to resample the data. In simple terms, in the forward approach we fill the missing values with the closest value directly prior the missing one. This is approach is employed to ensure that no information from the future is propagated backwards in time.

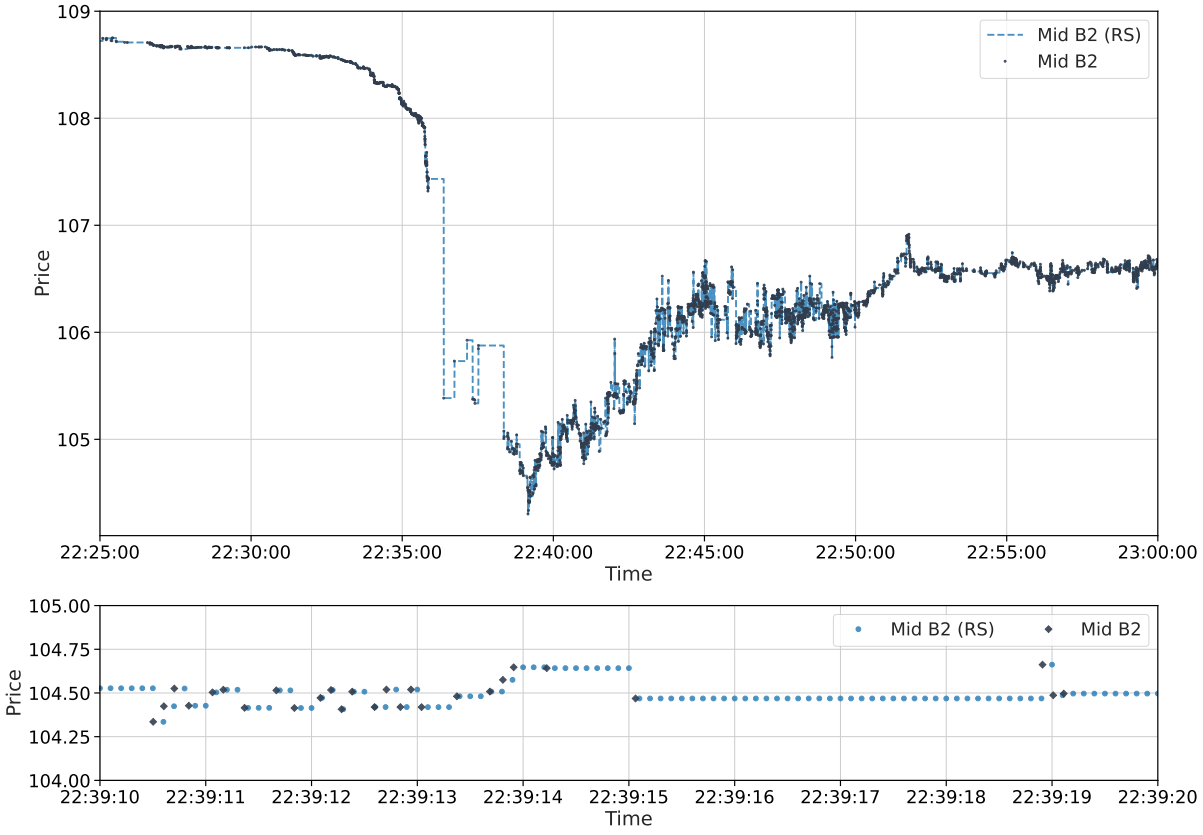

FIGURE 4.2: USD/JPY raw and resampled (RS) mid prices of bank (B2) at 22h on 2nd January 2019 - the day of flash crash. The top figure presents the evolution of price quotes during the USD/JPY flash crash. The bottom figure zooms on the data points presenting the both raw and mid prices resampled to 100ms time intervals. Time is expressed in London local time, i.e. UTC+1.

4.1.4 Stationarity & cointegration testing

As discussed in Section 3.1, the econometric model relies on the assumption that the first and second moments of the time series modelled remain constant in time, i.e., the time series is stationary. Moreover, recall that stationarity assumption is also made in the definition of transfer entropy (Wibral et al. [2014b]). Therefore, it is necessary to ensure all non-stationary time series are transformed into stationary ones before the time series is used for the analysis.

Dates	Dealers							
	B1	B2	B3	B4	B5	B6	M1	E1
2020-02-27	1	2	1	1	1	0	0	1
2020-02-28	1	1	1	1	1	2	1	1
2020-03-02	0	0	0	0	0	0	0	0
2020-03-03	2	0	1	0	1	2	1	1
2020-03-04	1	0	0	0	0	0	0	0
2020-03-05	0	0	0	0	0	0	0	0
2020-03-06	2	1	1	1	1	1	2	1
2020-03-09	0	0	0	0	0	0	0	0
2020-03-10	0	0	0	0	0	0	0	0
2020-03-11	0	1	0	1	0	0	0	0
2020-03-12	0	0	0	0	0	0	0	0
2020-03-13	3	4	5	5	4	5	4	5
2020-03-16	1	0	0	0	0	0	0	0
2020-03-17	1	1	1	0	1	0	0	1
2020-03-18	0	0	0	0	0	0	0	0
2020-03-19	1	1	0	1	1	1	1	1
2020-03-20	1	0	0	0	0	2	2	2
2020-03-23	0	0	0	0	0	0	0	0
2020-03-24	0	0	0	0	0	0	0	0
2020-03-25	0	0	1	0	0	0	0	0
2020-03-26	0	0	0	0	0	0	0	0
2020-03-27	0	0	0	0	1	0	0	0

TABLE 4.2: Result of augmented Dickey-Fuller tests for EUR/USD data set. The table reveals the number of subsamples that were determined to be stationary for a significance level $\alpha = 0.05/96$. For each dealer-date combination there are 96 subsamples tested, hence the Bonferroni correction is applied. The subsamples used for testing are 5-minute-long time-series resampled to 100ms time intervals.

Another assumption of the econometric model is that the time-series of price quotes of different dealers are cointegrated. Therefore, it is necessary to verify the assumptions mentioned

above by employing the augmented Dickey-Fuller and Engle-Granger stationarity and cointegration tests[5].

As it will be discussed later in Chapter 5, both econometric and information-theoretic models will be used on 5-minute-long time series. Thus, for each non-trading day, the time between 8:00 and 16:00 (i.e., 8 hours) will be divided into 5-minute-long time windows, yielding 96 subsamples per trading day. We must ensure that our stationarity and cointegration assumptions apply to each subsample used for the analysis. Thus, the stationarity and cointegration tests are performed for each time series from each time window included in the quantitative analysis. The testing results are presented in Table 4.2.

The null-hypothesis of augmented Dickey-Fuller test is that a unit root is present in a time series, i.e. the time series is integrated of order one. Table 4.2 reveals that for only 94 out of 16896 subsamples we reject the null-hypothesis. In particular, on the day following the quantitative easing announcement many subsamples were found to be stationary. For the remaining 99.4% of subsamples we failed to reject the null-hypothesis, thus with 5% confidence we can conclude that they are integrated of order one.

Dealers	Dealers						
	B2	B3	B4	B5	B6	M1	E1
B1	118	180	130	116	141	36	83
B2		117	54	70	59	44	49
B3			139	122	149	76	139
B4				47	46	35	34
B5					48	44	44
B6						29	33
M1							30

TABLE 4.3: Result of augmented Engle-Granger two-step cointegration test for EUR/USD data set. The test is performed for each dealer-dealer pair. The table presents the number of subsamples that were determined to be not cointegrated at significance level $\alpha = 0.05/96$. For each dealer-dealer combination there are $22 \cdot 96 = 2112$ subsamples tested. The subsamples used for testing are 5-minute-long time-series resampled to 100ms time intervals.

The null-hypothesis of augmented Engle-Granger two-step cointegration test is that the two time-series are cointegrated. Table 4.3 reveals that for only 2112 out of 59136 dealer pairs we reject the null-hypothesis. For the remaining 96.4% of subsamples we failed to reject the null-hypothesis. The stationarity and cointegration tests are also performed for the USD/JPY data set. The results for this currency pair are presented in Appendix E.1.

[5] `Scipy`'s implementation of these tests is used (Virtanen et al. [2020]).

4.2 Econometric model

4.2.1 VECM lag selection

In the econometric model, VECM lag is the only parameter that needs to be determined. For the VECM lag selection, we follow the methodology outlined in Hagströmer and Menkveld [2019]. The authors propose to use the Bayesian Information Criterion (BIC) to determine the number of lags used in the VECM model. The maximum lag that is investigated is lag 20. Thus, the optimal VECM lag is selected for each subsample based on the BIC criterion. Since BIC is a popular criterion, an interested reader is referred to Lütkepohl [2005] where this criterion is extensively treated.

In addition to the lag selection, for each subsample, each market is required to have at least 10 changes in their quote midpoints, otherwise, it is excluded from the subsample.

4.3 Network inference algorithm

In the network inference algorithm, both transfer entropy and conditional transfer entropy will be employed to infer information flows in the FX dealer-network. First, transfer entropy will be computed to identify dealers that are causal information contributors and quantify their information contributions for each unique dealer pair in the FX dealer-network. Next, the conditional transfer entropy will be employed to filter out the information flows that do not contribute any unique information to the target dealer, when other information contributions are considered. While the theory section may create an impression that estimating apparent and conditional transfer entropies can be easily accomplished, it is a pretty complex task, especially for continuous analogs of these metrics (Kraskov et al. [2004]). In this thesis, the Kraskov algorithm estimator is employed, and hence it will be thoroughly introduced in the following subsection (Wibral et al. [2014b]).

4.3.1 Kraskov, Stögbauer and Grassberger algorithm I

The concept of nearest-neighbor-based estimators for Shannon entropy was already explored a long time ago; first by Dobrushin [1958], and also later by Vasicek [1976] (Hlaváčková-Schindler et al. [2007], Kraskov et al. [2004]). However, the estimators proposed by these scholars could not be extended for the estimation of mutual information (Hlaváčková-Schindler et al. [2007]). Fairly recently, Kraskov et al. [2004] introduced a new estimator, known as Kraskov, Stögbauer and Grassberger (KSG) estimator, which has quickly gained wide acceptance in the information-theoretic community (Lizier [2014]). KSG estimator is an improved adaptation of a naive Kozaczenko-Leonenko nearest-neighbor estimator for continuous data that was proposed by Kozachenko and Leonenko [1987]. KSG is currently considered to be the top of the class transfer

entropy estimator extensively used in neuroscience (Lizier [2014], Lord et al. [2018]). This is mainly because, it is a relatively simple to implement, parameter-free tool, which can handle estimating mutual information in complex multi-dimensional systems (Lord et al. [2018]). The KSG algorithm was extended for conditional mutual information estimation by Frenzel and Pompe [2007].

To be precise, Kraskov et al. [2004] propose two very similar algorithms. The theory behind these two algorithm is quiet elaborate, and is not easy to comprehend. While it is not necessary to delve into the mathematical formulation and derivation of this estimators, we will try to develop a broad understanding of these algorithms.

Let us begin with recalling the definition of the differential entropy (definition 3.45 in section 3.2.3.1). The differential entropy of a continuous random $\mathbb{R}$-valued variable X whose probability density function $f(x)$ is:

$$h(X) \triangleq -\int_{\mathbb{R}} f(x) \log\left(f(x)\right)\, dx$$

when the integral exists (Devroye and Györfi [2021]). The integral can also be expressed in the following manner:

$$h(X) \triangleq \int_0^1 \log\left(\frac{d}{dp}F^{-1}(p)\right)\, dp \tag{4.1}$$

Now, following the steps of Vasicek [1976], let $x_1, x_2, x_3, \ldots, x_N$ be a sample from the probability density distribution $F(x)$. Vasicek [1976] further proposes that the distribution function F can be approximated by the empirical distribution F_n using the distances between the neighboring points. To do so, the difference operator is used instead of differential operator (Vasicek [1976]). Now, if we reorder the sample points from the smallest to the largest so that $x_1 \leq x_2 \leq \ldots \leq x_N$, then the differential entropy can be approximated by:

$$\widehat{h}(X) = \frac{1}{N}\sum_{i=1}^{N} \log\left(N\frac{x_{i+m} - x_{i-m}}{2m}\right) \quad \forall \frac{i-1}{N} < p \leq \frac{1}{N} \tag{4.2}$$

where $i = m+1, m+2, \ldots, N-m$ and $m < \frac{N}{2}$. If $p \leq \frac{m}{N}$ then $\frac{x_{i+m}-x_1}{2m}$ and if $p > \frac{N-m}{N}$ then $\frac{x_N - x_{i-m}}{2m}$ are used instead (Vasicek [1976]). Following this line of thought, the simplest estimator based on distances for one-dimension x was formulated as:

$$\widehat{h}(X) = \frac{1}{N-1}\sum_{i=1}^{N-1} \log\left(x_{i+1} - x_i\right) - \psi(1) + \psi(N) \tag{4.3}$$

where, $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function[6] - the derivative of the log of the gamma function (Kraskov et al. [2004], Bossomaier et al. [2016]). The problem with this formulation is that it can not be easily generalized to higher dimensions, and therefore this method is not applicable for the estimation of mutual information (Kraskov et al. [2004]). Kozachenko and Leonenko [1987] build upon this idea and address the dimensional limitations of the naive formulation by replacing the distances between sorted points with the K-th nearest-neighbor distances in d-dimensional space (Bossomaier et al. [2016]). The idea is that this allows us to approximate the probability density, since given the distance $\frac{\epsilon}{2}$ to the K-th nearest-neighbor, this implies that there are $N-K-1$ points distance further than $\frac{\epsilon}{2}$ and $K-1$ points that are at a distance smaller then $\frac{\epsilon}{2}$ (Kraskov et al. [2004]). Consequently, the probability density function can be approximated with the trinomial formula (Kraskov et al. [2004]). The estimator of differential entropy proposed by Kozachenko and Leonenko [1987] is defined as:

$$\widehat{h}(X) = \frac{d}{N}\sum_{i=1}^{N}\log(\epsilon_i) + \log(c_d) - \psi(K) + \psi(N) \tag{4.4}$$

where, d is the dimension of variable x, ϵ is twice the distance to the K-th nearest-neighbor, and c_d denotes the volume of the d-dimensional unit ball (Kraskov et al. [2004], Berrett et al. [2019]). For example for the Chebyshev distance (maximum norm) $c_d = 1$, whereas for Euclidean distance $c_d = \pi/(1+d/2)/2^d$ (Kraskov et al. [2004]). It is important to note that the estimator proposed by Kozachenko and Leonenko [1987] comes with some bias, as the underlying assumption is that the density remains constant in the region captured within distance ϵ (Kraskov et al. [2004]). The bias is further controlled by K, as it effectively determines the region within distance ϵ, which is assumed to maintain a constant density. The larger the K, the further neighbor is considered, the longer the distance, hence the larger the region that is assumed to have constant density. However, as presented by Kraskov et al. [2004], the estimator comes with relatively low error scaling with $\sim K/N$ or $\sim K/N\log(N/K)$.

The bias, however, becomes a problem if one considers estimating, for example, mutual information by estimating each differential entropy separately. Recall that mutual information can be expressed as the sum of the differential entropies of the marginal and joint densities, as follows:

$$I(X;Y) = h(X) + h(Y) - h(X,Y)$$

The problem is that using Kozachenko and Leonenko [1987] estimator to compute $h(X)$, $h(Y)$ $h(X,Y)$ will results in different distance scales being used for each estimate (Kraskov et al.

[6] Digamma function is the derivative of the log of the gamma function - $\Gamma(x)$ (Bossomaier et al. [2016])

[2004]). In particular, the distances in the joint space will be larger than the distances in the marginal spaces (Kraskov et al. [2004]). This problem has been resolved by Kraskov et al. [2004].

Kraskov et al. [2004] propose the following solution. First, to find the K_{XY}-th nearest neighbor distance ϵ_{XY} in the joint distribution (X, Y). Next, in each marginal space determine the number of nearest neighbors that are within the respective K-th nearest neighbor distance found in the joint space. Thus, K is different for each marginal space, but the distance scales are the same for all spaces (Bossomaier et al. [2016]). Moreover, the bias terms are consequently of the same order and likely to cancel each other out (Bossomaier et al. [2016]). This approach is possible, since Kraskov et al. [2004] observe that Kozachenko and Leonenko [1987] formula holds for any K, hence it does not need to be fixed for the estimation of differential entropies in the marginal space (Kraskov et al. [2004]).

Definition 4.1 (KSG Algorithm I - Mutual Information)**.** Consider two continuous processes X_t and Y_t with marginal probability density functions $p(x_t)$ and $p(y_t)$. Furthermore, let us denote the joint density pf X_t and Y_t with $p(x_t, y_t)$. Then, the KSG estimate of mutual information is computed with the following formula (Kraskov et al. [2004], Lord et al. [2018]):

$$\widehat{I}(X;Y)_{KSG(1)} = \psi\left(K_{XY}\right) + \psi\left(N\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\psi\left(n_X + 1\right) + \psi\left(n_Y + 1\right)\right) \tag{4.5}$$

$$= \psi\left(K_{XY}\right) + \psi\left(N\right) - \left\langle\psi\left(n_X + 1\right) + \psi\left(n_Y + 1\right)\right\rangle \tag{4.6}$$

where, N represents the number of points (or simply data points) that are used to estimate mutual information. Moreover, $\langle\cdots\rangle$ denotes the average over all data points, and K_{XY} represents K-th nearest-neighbor that is used to determine the distances in the joint space (X, Y) (Vejmelka and Paluš [2008]). Finally, n_X and n_Y denote the number of nearest neighbors found within the distance of ϵ_{XY} in X-space and Y-space, respectively.

In conclusion, Kraskov et al. [2004] proposed a robust estimator that can be scaled up to virtually any number of random variables. While it is a standard to use Chebyshev distance (maximum norm) to compute the distances, (as it can be easy to work with in large dimensions), in fact any distance metric can be used (Kraskov et al. [2004]). In this thesis, the max-norm KSG algorithm is employed.

Definition 4.2 (KSG Algorithm I - Conditional Mutual Information)**.** As mentioned in the introduction to this section, KSG algorithm for mutual information estimator was further extended to conditional mutual information by Frenzel and Pompe [2007] and adapted to transfer entropy by Gómez-Herrero et al. [2015] (Lindner et al. [2011]); the estimator takes the following form:

$$\widehat{I}(X;Y|Z)_{KSG(1)} = \psi\left(K_{XYZ}\right) - \left\langle\psi\left(n_Z + 1\right) - \psi\left(n_{XZ} + 1\right) - \psi\left(n_{YZ} + 1\right)\right\rangle \tag{4.7}$$

This estimator can be used to estimate the transfer entropy. The KSG algorithm can be easy extended to a case where more conditionals need to be accounted for. For example the conditional transfer entropy can be computed in the following manner:

$$\widehat{I}(X;Y|Z_1,\ldots,Z_a)_{KSG(1)} = \psi\left(K_{XYZ_1\ldots Z_a}\right) - \tag{4.8}$$

$$\left\langle \psi\left(n_{Z_1\ldots Z_a}+1\right) - \psi\left(n_{XZ_1\ldots Z_a}+1\right) - \psi\left(n_{YZ_1\ldots Z_a}+1\right)\right\rangle \tag{4.9}$$

where a denotes the number of information contributors that transfer entropy is conditioned on. Thus, it becomes clear how easy it is to extend the estimator to any number of conditionals.

The only parameter that one needs to specify for the KSG estimator is the K, i.e. which K-th nearest neighbor is chosen to determine the distances in the joint space that will be mapped to the marginal spaces. As mentioned earlier, K does effectively control the bias of the estimation, which here scales with a factor K/N (Kraskov et al. [2004]). However, numerical experiments strongly suggest that the estimator is relatively stable to the choice of K (of course given that $N >> K$) (Lizier [2014]). Accordingly, one must ensure that there is enough data that can be used to approximate the probability density functions, and hence estimate information-theoretic metrics.

It appears that typically K is set to 4 as a default parameter value, however it is recommended for each application to independently scan through K space to determine the appropriate choice of that parameter given the underlying distribution of the data (Lizier [2014], Wibral et al. [2014b], Hlaváčková-Schindler et al. [2007]).

For example, Kraskov et al. [2004] propose to use $K = 2 - 4$, the numerical experiments presented by Frenzel and Pompe [2007] reveal that it is necessary to determine an optimal balance between systematic and statistical errors. The numerical experiments performed by Frenzel and Pompe [2007] provide a lot of important insights, in particular the fact that the highest standard error reduction is observed between $K = 2$ and $K = 8$, while the bias essentially remains the same. Similar results are presented by Runge [2014] who concludes that as we increase K the decrease in variance is much more significant than the increase in bias. The robustness of the estimator to choice of parameter K is also presented by Kraskov et al. [2004], Frenzel and Pompe [2007] and Vejmelka and Paluš [2008].

Based on the results from the literature and numerical experiments with the FX data, the decision was made to choose $K = 8$ which very significantly reduces the variance of the estimate, while maintaining negligible systematic errors. This choice was motivated by large variance of transfer entropy estimates for $K = 4$ observed for the time windows that included EUR/CHF and USD/JPY flash crashes.

Finally, it is necessary to note that the choice of K parameter value also impacts the computational complexity of the nearest-neighbor searching algorithm. Higher values of K significantly

extend the run time. Hence, it is necessary to balance the systematic and statistical errors, given the computational limitations. Additionally, the computational effort of the KSG algorithm also increases as the length of the time-series increases. As we have more data points, there are more data points that we need to find the nearest-neighbors for. Thus, longer time-series (N) also considerably extends the run time of the KSG algorithm.

4.3.1.1 Time-series standardization

Apart from transforming data into stationary time-series, Kraskov et al. [2004] suggest to standardize the time series to zero mean and unit variance; this approach is widely accepted by the scientific community (Lizier [2014], Vejmelka and Paluš [2008]). After the time series is standardized, it is also recommended to introduce very low-amplitude noise to the data. For double-precision floating-point operations, Kraskov et al. [2004] proposes to add noise of order 10^{-10}. This treatment is essential when one works with empirical data with limited precision, potentially resulting in many points having equal values. Consequently, this would lead to breaking the assumption of continuously distributed points and result in spurious estimates (Kraskov et al. [2004]). Adding noise to the time series introduces very slight stochasticity to the estimator. Generally, if the information transfer is significant, it will remain significant after the addition of the noise (Lizier [2015]). The noise does not affect the estimation unless many data points have the same value. In this case, however, this would suggest that the data are not continuously distributed, and hence KSG algorithm should not be used.

4.3.2 Paramater choice

While the KSG estimator is practically parameter-free (besides the K parameter), it is still necessary to consider the history lengths and the time-delay embedding of the time-series used with the KSG estimator. Recall a few critical factors discussed in the theory section: the Markov order of target process d_y and the history length of the source process d_x. As previously discussed, the past states of the stationary Markov processes can contribute significant information about their potential future state. Hence, ideally one would like to account for the entire history of the target process to properly reflect the notion of information transfer (Lizier and Rubinov [2012]). Unfortunately, in practice this is not possible with the KSG algorithm because each extra length of history introduces one more dimension that needs to be taken into account in the nearest-neighbor search. Not only does this considerably increase the computational effort, but it also exhausts the statistical power of the KSG algorithm (Lizier [2022]).

4.3.2.1 State space reconstruction

To determine appropriate history lengths for both source and target processes and their respective time-delay embeddings, the notion of state-space reconstruction and Takens' delay embedding theorem need to be introduced.

To illustrate the idea of state-space reconstruction, consider an example of a simple pendulum, with horizontal position $x = 0$ being the center when the pendulum is at rest, $x < 0$ left-hand-side and $x > 0$ right-hand-side. Let us further assume that we do not know the length of the pendulum rod or the angle at which the pendulum was initiated. Now, if we know the position value of the pendulum at one point in time t, say $x_t = -1$, while we know that the pendulum is positioned on the left-hand side, it is impossible to tell if, at this moment of time, the pendulum is swinging to the right or left (Wibral et al. [2014b]). Consequently, we have no information that could help predict its future position x_{t+1}, other than the fact that it will be in some undefined vicinity of $x \sim -1$. However, if we learn that at $t = t - 1$ the pendulum is at $x_{t-1} = -0.9$, then with that knowledge, we can reason that the pendulum is most likely still swinging towards the left. Thus, it is likely that the future position of the pendulum will be $x_{t+1} < -1$. In this manner, with a more extended history of the pendulum's position, we can determine the state that the pendulum is currently found at, e.g. broadly it is in the state of swinging to the left-hand side or on the left-hand side but swinging to the right-hand side. Thus, with a longer history of the past states, we can come up with a much better prediction of the future position of the pendulum.

However, the entire history of the pendulum may not necessarily be essential to account for. In a simple case scenario, with no gravity or any other external force in place, the pendulum's motion would be perfectly periodic. Hence, having the information collected in one full swing would provide all the information about the system. However, in a more complex scenario, the position value from time point $t = t - 100$ may be entirely irrelevant for the current dynamics, for example if some extra force was applied at $t = t - 10$. This is important to note, especially when one considers complex non-linear systems, for which we usually do not know the underlying system dynamics. The problem becomes even more involved when one needs to reconstruct the system dynamics based on just a single time-series (Bossomaier et al. [2016]).

In a nutshell, Takens [1981] proposed a theorem that the past observations of the process can be used to reconstruct the state of a d-dimensional nonlinear dynamical system (Bossomaier et al. [2016]). Let y_t denote the dependent variable observed at some discrete time points $t \in \{1, 2, \ldots, T\}$. Now, let us further assume that y_t is a function of some d-dimensional state space $\boldsymbol{x}_t = \{x_t^1, x_t^2, \ldots, x_t^d\}$, thus

$$y_t = f\left(\boldsymbol{x}_t\right) \tag{4.10}$$

As already mentioned, while working with empirical data that represent the time-evolution of some nonlinear complex system, the underlying state-space may be unknown, or very difficult to reconstruct (Bossomaier et al. [2016]). By Takens' theorem we can reconstruct the state space from the observations of y_t, as follows:

$$\widehat{y_t} = \{y_{t-d_y\tau_{\text{emb}}}, y_{t-(d_y-1)\tau_{\text{emb}}}, y_{t-(d_y-2)\tau_{\text{emb}}} \ldots, y_{t-1}\} \tag{4.11}$$

where τ_{emb} is called the embedding delay, which simply determines the time lag between the consecutive past states that are accounted for in the vector of past states. Also, d_y denotes the embedding dimension which is simply the length of the history y_t that is accounted for. Recall the definition of transfer entropy:

$$\text{TE}_{X_t \to Y_t}^{(d_y,d_x)} \triangleq D\left(f(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})||f(y_{t+1}|\boldsymbol{y}_t^{(d_y)})\right) \tag{4.12}$$

thus here,

$$\boldsymbol{y}_t^{(d_y)} = \{y_t, y_{t-1\tau_{\text{emb}}}, y_{t-2\tau_{\text{emb}}}, \ldots, y_{t-(d_y-1)\tau_{\text{emb}}}\} \tag{4.13}$$

$$\boldsymbol{x}_t^{(d_x)} = \{x_t, x_{t-1\tau_{\text{emb}}}, x_{t-2\tau_{\text{emb}}}, \ldots, x_{t-(d_x-1)\tau_{\text{emb}}}\} \tag{4.14}$$

Given the time series of length N, the number of embedding vectors that can be constructed is $N - (d_y - 1)\tau_{\text{emb}}$ (Hegger et al. [1999]). Thus, the history length and the embedding delay have a direct impact on the length of the time series that can be used for estimation with KSG algorithm.

Having introduced the Takens' theorem and elaborated on the intuition behind state-space reconstruction, we can now look into optimizing the history length and embedding delay for a give time series.

4.3.2.2 Ragwitz criterion

Various methods exist to determine an optimal choice of embedding dimension and embedding delays. Broadly, there are two approaches: the uniform (UE) and non-uniform embedding (NUE) (Montalto et al. [2014]). The uniform embedding amounts to selecting a particular combination of embedding dimension and delay parameters for the entire time series before the process of estimating the transfer entropy between variables (Montalto et al. [2014]). On the other hand, the non-uniform embedding approach consists of progressively and individually, for each time step, selecting the most relevant past states (given some maximal lag considered) that maximizes the amount of information about the target variable (Montalto et al. [2014]). The superiority of the NUE approach over the UE lies in the fact that the NUE approach allows to minimize the dimensionality of the embedding vectors and hence improve the performance

of the KSG estimator (Shahsavari Baboukani et al. [2020], Lizier [2022]). Consequently, the NUE approach becomes much more applicable when the transfer of information between a large number of processes is investigated (Novelli et al. [2019]. Since uniform embedding is a lot easier to implement, it is currently the most widely used approach (Lindner et al. [2011]). However, nowadays, much attention is given to non-uniform methods that can be tailored toward particular applications.

Given the time limitations of my thesis, the decision was made to focus on the most common uniform embedding approaches, the Cao criterion Cao [1997] and Ragwitz criterion Ragwitz and Kantz [2002] (Lindner et al. [2011]). The Cao criterion is a preferred method when the optimal embedding dimension and delay need to be determined for a deterministic (chaotic) system (Lindner et al. [2011]). The Cao criterion is considered to be rather a more heuristic approach, in which the embedding dimension is determined using a false neighbor criterion, whereas the embedding delay is set to the first zero of the auto-correlation function (ACF) or the auto-information (Lindner et al. [2011], Guo et al. [2021]). As Cao [1997] states himself, his criterion is meant to provide a practical method to determine the minimum embedding dimension that is computationally efficient, does not involve any subjective parameters, and is suitable to use for time-series with high-dimensional attractors.

On the other hand, for stochastically driven systems, the Ragwitz criterion should be used instead (Lindner et al. [2011]). Ragwitz criterion provides means to jointly optimizes both the embedding dimension and delay parameters and was designed to be used for both deterministic and stochastic data from Markovian processes (Lindner et al. [2011]). Ragwitz criterion employs a locally constant predictor of the future state (w_{t+1}) of embedding vector $\boldsymbol{w}_t^k$. The prediction of the future state of the variable is estimated using the future states of the nearest neighbors of the variable after embedding is applied (Lindner et al. [2011]).

Definition 4.3 (Ragwitz's locally constant predictor)**.** Let ϵ denote the neighborhood diameter and $\mathcal{U}_n$ be the neighborhood (subscript n) of $\boldsymbol{w}_t^k$, where $\mathcal{U}_n = \{\boldsymbol{w}_n^k : ||\boldsymbol{w}_n^k - \boldsymbol{w}_t^k|| \leq \epsilon\}$. Ragwitz's locally constant estimate $\widehat{w}_{t+1}$ of the future state w_{t+1} is defined as:

$$\widehat{w}_{t+1} = \frac{1}{|\mathcal{U}_n|} \sum_{\boldsymbol{w}_n^k \in \mathcal{U}_n} y_{n+1} \tag{4.15}$$

which simply amounts to taking the mean value of the future states (w_{n+1}) of the nearest neighbors of the embedding vector $\boldsymbol{w}_t^k$. In practice, ϵ is replaced by a specific number of K nearest neighbors taken into account for the estimation, which is a "natural" substitute when the KSG algorithm is employed (Lizier [2014]). Next, the squared error of the local predictor is computed for each time index in the embedding vector, and the mean squared error is determined.

$$e^2 = \frac{1}{|t|} \sum_t (\widehat{w}_{t+1} - w_{t+1})^2 \tag{4.16}$$

This procedure is adopted for each combination of parameters k and τ_{emb} that one chooses to investigate. Thus, for each combination of k, and τ_{emb} the mean squared errors are determined. Based on Ragwitz's criterion, the parameters that yield the smallest mean squared error should be used.

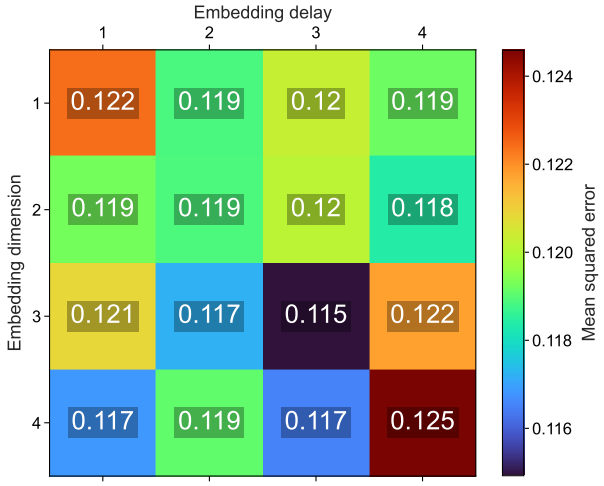

FIGURE 4.3: Mean squared error of locally constant predictor for different embedding delay and dimension parameters. The figure illustrates the parameters space that is investigated during the embedding optimization process. In this example the lowest mean squared error of 0.115 is observed for embedding delay and dimension set to 3.

Ragwitz's criterion is applied to each pair of source (X_t) and target (Y_t) processes investigated. Therefore, it is also necessary to emphasize slight but notable differences when the optimal embedding parameters are determined for source and target processes. Namely, when one investigates the optimal embedding settings for a target process Y_t, the target's future states y_{t+1} of the nearest neighbors are used for the locally constant prediction. In this setting, the embedding vector is meant to capture the underlying state of the target process for the Markov process of order d_y - hence the idea is to maximize the self-prediction (Lizier [2014]). Whereas, in the case when we want to determine the optimal embedding for the source process X_t, the target's future states y_{t+1} that correspond to the source's nearest neighbors "current" states are used instead. In other words, for each source's nearest neighbor, its corresponding future state of the target process is used in the prediction.

The reason why this is the case is very simple, namely, since we are interested in determining the transfer entropy from source to target, we are interested in the predictive power of the

sources past states on the future states of the target. Thus, naturally we are interested in finding the optimal embedding that minimizes the prediction of the future states of the target process.

Moreover, it is necessary to emphasize the importance of the proper state-space reconstruction. As already discussed when the notion of transfer entropy was introduced, the history of the target process, especially when one considers Markovian process, may in fact provide a lot of significant information about its future state (Lizier [2014]). Hence, given that we cannot include the entire history of the target process, and thus we are making some approximations, it is necessary to perform the state-space reconstruction with utmost attention. As insufficient state-space reconstruction may lead to spurious conclusions about the flow of information from one variable to another, when in fact it is the history of the target process that contributes the information (Wibral et al. [2014b]).

Finally, it should be noted that the state space reconstruction and the Ragwitz criterion, which we use for the nonlinear models, is an approach analogous to using the Bayesian information criterion for the linear models (Pan and Duraisamy [2020]). In essence, both criterion are used to determine the length of the history of the variable that needs to be accounted for.

In our information-theoretic network inference algorithm, the maximal embedding dimension and delays are set to 4. Higher embedding dimensions would be computationally unfeasible given the scope of this thesis and computational limitations of the KSG algorithm, which we elaborated on in this section.

4.3.2.3 Delay reconstruction

After the embedding dimension and delays of both source and target processes are determined, it is necessary to determine the "true" delay for the source-target interaction that is investigated Wibral et al. [2012]. Up to this point we have not explicitly considered the possibility of the information being transferred at any other delay than delay of 1. In practice, when we work on complex systems, we do not have much or any knowledge about the underlying system dynamics. Thus, it may not always be possible to know what is the "true" delay between the interaction of source and target processes. Recall, that econometric model assumes that the interaction between source process and target process occurs at the first lead-lag. While this assumption is not necessarily wrong, it makes the model heavily dependent on the correct latency adjustment of the data and the accuracy of data collection and data processing. Further, recall that the latency adjustment was also the problem that Hagströmer and Menkveld [2016] considered as the major limitation of the model from the first manuscript.

The situation is different when we consider the information-theoretic approach. Wibral et al. [2013] not only provides the method to determine an optimal delay that should be used for the

estimation of transfer entropy, but also they prove that the optimal delay is in fact the true delay. Thus, even if one works with the data that was not correctly adjusted for the latency, the transfer entropy will still be detected, just at a different delay.

The method for delay reconstruction as proposed by Wibral et al. [2013] and also suggested by Lizier [2022] is in fact very simple. In order to determine the optimal delay between source and target processes it is necessary to estimate transfer entropy for each delay in the delay-space that one considers. The true source-target delay is the one that yields the largest transfer entropy estimate. Recall that in the section 3.2 the transfer entropy was defined as:

$$\mathrm{TE}_{X_t \to Y_t}^{(d_y, d_x)} = I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)} | \boldsymbol{Y}_t^{(d_y)}) \tag{4.17}$$

for the purpose of discussion of the source-target delay note that we can also represent it slightly differently, as follows:

$$\mathrm{TE}_{X_t \to Y_t}^{(d_y, d_x)} = I(Y_t; \boldsymbol{X}_{t-u}^{(d_x)} | \boldsymbol{Y}_{t-1}^{(d_y)}) \tag{4.18}$$

by simply shifting future and history data points of Y_t process by one backwards, and introducing the delay parameter u for the source process X_t. In this formulation u represents source-target delay. Given this formulation, the optimal source-target delay is determind to be:

$$\delta = \underset{u \in \mathcal{U}}{\operatorname{argmax}} \left(I(Y_t; \boldsymbol{X}_{t-u}^{(d_x)} | \boldsymbol{Y}_{t-1}^{(d_y)}) \right) \tag{4.19}$$

where $\mathcal{U}$ is the investigated delay-space, i.e. $\mathcal{U} = \{1, 2, 3, \ldots, u_{\max}\}$, and δ represent the true delay between source and target processes.

Note that we only consider different delays between the source and target processes. The delay between the target's history and the target's future states needs to remain equal to one. Changing the delay between the target's future states and the target's history would violate Wiener's principle of causality (Wibral et al. [2013])[7].

Finally, Wibral et al. [2013] extends upon the notion of "true" delay by providing the framework for detection of multiple delays and feedback loops. While this is very interesting, it needs to remain as potential future extension of this thesis.

In our information-theoretic network inference algorithm, we set the maximal source-target delay to 4 (which, given the resampling interval, is equivalent to 400 ms). From our latency investigation and the literature, it is clear that 400ms is more than enough time for dealers to react to quote updates of other dealers. Furthermore, the analysis of the distribution of true source-target delays revealed that most of the true delays recovered were equal to 1.

[7] An interested reader is referred to Wibral et al. [2013], where the reason behind not delaying target's history and the notion of self prediction optimality are clearly explained.

4.3.3 Statistical testing

Having estimated the transfer entropy between the source and target processes, it is necessary to assess whether the estimate is statistically significant. While in theory, the estimate of transfer entropy should be zero if the future state of the target process is conditionally independent from its own history and the history of the source process, this is not exactly the case when the KSG algorithm is used.

As previously stated, not only does the KSG algorithm introduce bias, but also we are working with finite-size sample and we introduce low-amplitude noise to the data. Hence, the transfer entropy estimate will not necessarily be equal to zero, even if the source's history does not provide unique information about the target's future state. Therefore, it is imperative to assess the statistical significance of transfer entropy estimates.

Since we can not make any assumptions about the underlying distribution of transfer entropy estimates, we need to employ alternative techniques such as permutation testing or bootstrapping (Bossomaier et al. [2016], Wibral et al. [2012], Lizier [2014]). The permutation testing is currently the most common method to assess statistical significance of information theoretic estimates (see Lindner et al. [2011], Lizier [2014], Wibral et al. [2013], Wollstadt et al. [2018]), hence it is also employed in this thesis.

When performing the statistical assessment with permutation technique, we are essentially assessing whether the estimated KL divergence between transitional probabilities $p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})$ and $p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})$ is indeed statistically significant. Consequently, we are assessing if the information provided by the source process does in fact contribute unique information about the future state of the target process. In practice, we are interested in testing for the null hypothesis that the state changes $\boldsymbol{y}_t^{(d_y)} \to y_{t+1}$ have no temporal dependence on the source process $\boldsymbol{x}_t^{(d_x)}$ (Lizier et al. [2011]), hence:

$$\mathbf{H_0}: p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}) = p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \implies \widehat{\mathrm{TE}}_{X_t \to Y_t}^{(d_y,d_x)} = 0 \tag{4.20}$$

$$\mathbf{H_1}: p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}) \neq p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \implies \widehat{\mathrm{TE}}_{X_t \to Y_t}^{(d_y,d_x)} > 0 \tag{4.21}$$

The one-sided alternative hypothesis is motivated by Lemma 3.39, which states that transfer entropy is a non-negative measure. To test the above-presented hypotheses, we need to generate an empirical distribution of transfer entropy estimates under null hypothesis. This can be done by generating a large number of source process surrogates (X_t^s), which preserve transitional probability $p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})$, but destroy the dependence in $p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})$ (Lizier [2014]). While the surrogates can be generated in many different ways, the most important aspect is to ensure that the vectors of past states of source process $\boldsymbol{x}_t^{(d_x)}$ are preserved (unless $d_x = 1$). This is crucial, because we need to retain the reconstructed state-spaces (Lizier [2014]). Thus, the easiest way to generate the surrogates would be shuffling the vectors of past states

$\boldsymbol{x}_t^{(d_x)}$ among the set of $\{y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}\}$ tuples - i.e. interchanging their time indices within the time-series (Lizier [2014]). This approach is also employed in this thesis.

Let us assume that we generated S of source process surrogates (X_t^s). Then, the p-value is determined by simply counting the number of cases when $\widehat{\mathrm{TE}}_{X_t^s \to Y_t}^{(d_y,d_x)} > \widehat{\mathrm{TE}}_{X_t \to Y_t}^{(d_y,d_x)}$. In other words, if the transfer entropy estimated for a surrogate source process is greater than the actual transfer entropy estimated for the original source process, then this suggests that our transfer entropy estimate is not significant. This is because, such scenarios would indicate that even if we destroy the temporal precedence structure our estimator still can yield higher value of transfer entropy (Wibral et al. [2014b]), which undermines the significance of the original estimate. Thus, the p-value is determined in the following manner:

$$\text{p-value} = \frac{1}{S} \sum_{i=1}^{S} \mathbb{1} \left(\widehat{\mathrm{TE}}_{i \; X_t^s \to Y_t}^{(d_y,d_x)} > \widehat{\mathrm{TE}}_{X_t \to Y_t}^{(d_y,d_x)} \right) \tag{4.22}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and S is the number of surrogate source processes. For a given significance level α, we reject $\mathbf{H_0}$ if p-value $< \alpha$ (Lizier et al. [2011]). Additionally, one should note that since we are going to conduct multiple hypothesis tests, we need to employ Bonferroni correction.

The decision on how many permutations are needed to provide a robust p-value is not straightforward. Of course, the more permutations performed, the more robust the computed p-value. However, in practice we are restricted by the computational limitations since estimating the transfer entropy for each surrogate of the source process is computationally equivalent to estimating a transfer entropy for the original source process. Hence, while we should investigate the robustness of the p-value to the number of permutations performed, in our information-theoretic network inference algorithm the number of permutations is set to $S = 500$. S is set to 500 because this is the largest number of permutations that could be performed within the time limits of this thesis project. Note that we are expecting to compute a total of 236,544 entropy estimates (only for the EUR/USD data set). Thus, if we perform 500 permutations for each estimate, we are looking at computing approximately up to 118 million entropy estimations, which is already quite a large number.

Finally, one should note that $S = 500$ is not a low number of permutations; for example, Lizier et al. [2011] use only $S = 300$ to assess the statistical significance of their estimates. As discussed later in the results section, only a very small portion of statistically significant apparent or conditional transfer entropies are found to have a p-value greater than zero.

4.4 Implementation remarks

Since the econometric model and network inference algorithm are employed on an extensive data set, a data pipeline had to be developed. The pipeline in Figure 4.4 is developed using three programming languages; `Python` and `C++` wrapped together with `Cython`. The transition between `Python` and `C++` via `Cython` takes place on the verges of purple box, where modules within the box were developed in `C++`. All of the modules are compiled together as one `Python` library. As presented in Figure 4.4, the developed pipeline is comprised of multiple modules that perform various tasks from data transformation to analysis with an econometric model and network inference algorithm.

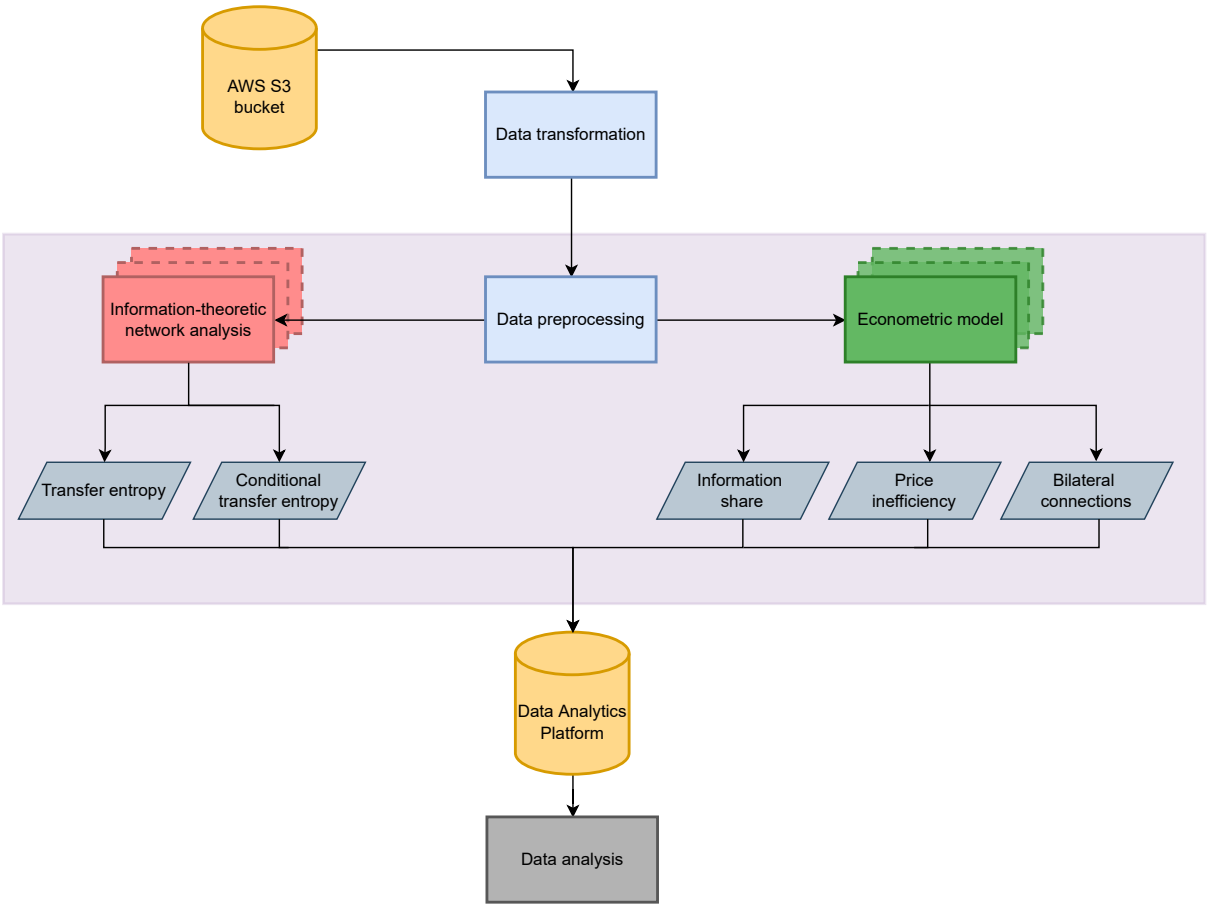

FIGURE 4.4: Flowchart illustrating the data pipeline developed for this thesis.

The pipeline starts with the data for a particular day of the week getting fetched from the Amazon Web Services (AWS) S3 bucket. In the data transformation module, raw data are first transformed from a `qtable` format to `Pandas DataFrame`. Next, from the `DataFrame` columns with relevant information and the best bid-and-ask prices are extracted. For example, the choice can be made to use Timestamp 1 instead of Timestamp 2. Next, the data is cleaned from observations that have missing entries in any remaining columns. The number of removed and

remaining data observations is recorded and later reviewed to ensure the integrity of the data, data cleaning process, and the subsequent analysis.

In the preprocessing module, the data is first resampled to a defined sampling frequency, followed by data extraction for chosen hours of the day, e.g., between 8:00 and 16:00. Afterwards, the time series is differenced to ensure stationarity and split into five-minute-long subsamples fed to the econometric model. Effectively, an 8-hours-long time window yields 96 five-minute-long subsamples. The analysis that is part of the econometric model is parallelized over all available CPU cores, where each CPU performs the analysis on one of the subsamples at a time.

For the network inference algorithm, in addition to being differenced the data is also standardized, and a low-amplitude noise is added. The network inference algorithm is also parallelized, where each CPU performs analysis on different subsamples from a particular day. The data produced from the econometric model and network inference algorithm are then merged and saved as multidimensional `NumPy` matrices into the Data Analytics Platform's S3-backed filesystem. Further data analysis-related activities are performed on this platform.

While the econometric analysis is swift, as it takes roughly a few seconds to complete the entire day's worth of data, the network inference algorithm is considerably more computationally demanding. On average, it takes 24 hours to perform the information-theoretic analysis on eight hours' worth of data[8]. Consequently, it takes the entire pipeline about a month to process all of the EUR/USD data in the scope of this thesis.

While a month-long run time may appear lengthy, it is expected, given that the KSG algorithm is very computationally demanding. As discussed in Section 4.3.1, for each transfer and conditional transfer entropy, we need to perform four multidimensional nearest-neighbors searches. To be exact, one nearest neighbor search involves determining distances to the K^{th} nearest neighbors of each data point in the joint space, and the other three need to determine the number of nearest neighbors that can be found within these distances in the other joint and marginal spaces. Moreover, the nearest neighbor search is also employed in the algorithm that determines the optimal dimension and delay embedding for each source-target pair. Finally, the whole process is repeated for each surrogate used for permutation testing. To put things into perspective, for EUR/USD data set up to 473 million multidimensional nearest neighbor searches need to be performed. In practice, however, the algorithm is optimized by aborting the permutation testing when the p-value of the estimate exceeds the predefined minimum significance level required.

A naive algorithm for the nearest neighbors search using Chebyshev distances scales with $O(DN^2)$, where D is the dimension of the data, and N is the number of data points (Pedregosa et al. [2011], Lizier [2014]). Note that for a joint space, the dimension D is equivalent to the

[8]For a network of 8 dealers with the time series of quotes resampled to 100 milliseconds, using 8 CPU cores and 500 permutations per entropy estimate. Moreover, the maximal embedding dimension is set to 4.

sum of embedding dimensions of all variables. Hence the dimension can get quite large. For fast nearest neighbor algorithm such as KD-tree, the nearest neighbors search complexity scales with $O(DN \log N)$ (Pedregosa et al. [2011], Lizier [2014]).

The performance of the KD-tree is the bottleneck of the network inference algorithm; therefore, much attention was put into developing a high-performance KD-tree algorithm. Not only was the algorithm first developed from scratch, but also many openly available `C++` implementations were tested. None of the algorithms were able to perform fast enough to complete the analysis of such a large data set within the time limits of this thesis. Finally, it was determined that `Scipy's` implementation of KD-tree in `Python` was outperforming any implementation in `C++`. The source code for this implementation turned out to be also written in `C++`. Hence, `Scipy's` source code was adjusted, compiled as a separate dynamic library, and used in the network inference algorithm (Virtanen et al. [2020]). As a result of this change, the KD-tree nearest neighbor search for 3000 data points in 1D takes as little as 70 milliseconds. Consequently, the performance KSG algorithm for transfer entropy was able to significantly outperform[9] the implementation from JIDT Libary (Lizier [2014]). Thus, it was the most significant breakthrough that made it possible to meet the thesis goals and perform network inference investigation on a large data set.

Finally, it should be noted that the same attention for high-performance was paid to each algorithm employed in the network inference module. Any extra one millisecond per one entropy estimate leads to approximately an additional hour of run time needed for the EUR/USD data set. On the final note, the implementation utilizing GPUs was also considered; however, given the time limitations of this thesis and no prior experience with `CUDA`, this idea appeared unfeasible. However, this is the next step that should be considered to improve the performance of the network inference algorithm.

[9] Not considering the usage of GPUs.

Chapter 5

Application

5.1 The foreign exchange market

As briefly discussed in Chapter 1, the foreign exchange market is a worldwide, decentralized marketplace that facilitates trading of currencies between various market participants. Over the past years, we have observed a significant change in the structure of the FX market with the emergence of new types of market players (King et al. [2011]). Since in this thesis we attempt to quantify the information flows between various market participants, it is necessary to outline the structure of the FX market, explain how it operates, and introduce the types of market participants involved in the market. This is necessary to provide the reader with a general overview of how information is created and flows through the market as well as establish the differences in customers that each dealer caters to.

The key market participants who naturally emerge in the FX market are the so-called market makers, broadly referred to as dealers. For a large part, banks perform the role of market makers; however, they are not the only market participants. In this thesis, one non-bank market maker is considered. Market makers' primary function is to provide liquidity to the market, which means they are ready to trade with anyone at any time. They facilitate the trading process by allowing clients to quickly make trades without the need to wait for a counter-party that would match the trade on the other side. Additionally, they contribute toward maintaining competitive bid-ask rates on the market. In the FX market, dealers not only trade with each other but also cater to their private clients. Such clients could be, for example, a multinational company or importer/exporter operating in Euro and occasionally making a deal in US dollars.

Now, let us consider how the FX market operates. Clients approach a dealer, for example through their bank, which quotes the bid and ask the prices based on their clients' need. Clients can complete the transaction or search for a better quote from a different dealer. When the client's order is executed, the FX dealer trades in the inter-dealer market to unwind his customer's trade (Vitale [2006]). Trading on the inter-dealer market can take place in two different

ways: the dealer can approach another dealer directly or indirectly via public limit order books operated by electronic brokerage platform such as EBS. When a trade takes place directly between two dealers, such transaction is termed a bilateral or private meeting. On the other hand, on electronic brokerage platforms, subscribers can add limit orders or match outstanding ones (Vitale [2006]). When the orders of both parties are matched, only the dealers involved in the transaction know each other's identity.

As mentioned before, in the microstructure approach, we recognize that private information in the decentralized markets is produced locally and then transferred between the dealers (Hagströmer and Menkveld [2019], Hasbrouck [1996]). Relevant private information is obtained from direct clients of each dealer. The clients observe the fundamentals and basing upon them update their views. Next, they trade with FX dealers, who observe the orders of all of their clients – the order flow. Dealers observe the order flow and accordingly set the price for the currency exchange. Finally, this information is further conveyed to other dealers, which is reflected in the exchange rates updates send by the dealer. In this manner, the information is produced directly based on the demand of private clients and conveyed via FX dealers to the dealer-network. On the other hand, when we consider public information, the microstructure approach proposes that all public information is directly impounded into the price by dealers; however, there may exist disparities between dealers in the interpretation of the news.

5.2 Model assumptions

In the following subsection, the underlying model assumptions are clearly stated.

- Dealers see the quotes of all other dealers in the dealer-network.
- The latency between the action of a dealer updating their quote and the moment other dealers observe that particular update is smaller than the time interval used in data resampling.
- The asymmetry in the information is attributed to the dealer's private information. Private information is attributed to the internal order flow observed by each dealer.
- Hasbrouck's information share is not a true measure of the dealer's contribution to the price discovery process; however, we use it as a point of reference.
- We assume that CTE quantifies the unique transfer of information between dealers.

5.3 Data

As mentioned in Section 4.1.1 our goal is to perform an analysis on the FX spot data for EUR/USD currency pair traded between February 27 2020 and March 27 2020. Additionally, for USD/JPY currency pair, we are going to consider the period starting from January 1 2019 until January 17 2019.

Foreign exchange market is open 24 hours per day from 23:00 (UTC+1)[1] on Sunday until 22:00 UTC+1 on Friday. Thus, FX is open all around the clock, as it needs to cater to customers located across different time zones. However, the activity on the FX market varies over time, with dealers providing more frequent updates on the quotes on particular days of the week and at particular times of the day. Since we are going to work with a lot of data, we would like to sample the time windows for the days on which the dealers are most active.

The changes in activity on the FX market can be explained by considering the timezones for each major region. For example, on Saturday we do not observe any activity on the FX market as all of the sessions are closed, and on Sunday there is very little activity given that only the Sydney session is open. For this reason, we exclude the data from the Saturdays and Sundays from the analysis as relatively very little to no activity is observed on these days. This decision is also in line with methodology of Hagströmer and Menkveld [2019].

Following further the methodology of Hagströmer and Menkveld [2019], we investigate the intraday quote frequency in the EUR/USD dataset. For each trading day in the data set, the days are split into 30-min time windows. In each time window we count the number quote updates from each dealer and then the average number of quote updates per second is determined. The intraday quote frequency aggregated per each 30-minute time windows in the day is presented in Figure 5.1. Note that the figure illustrates the average quote frequency observed cumulatively for all dealers in the data set. Thus, in order to obtain an average quote frequency per second per dealer, the values would need to be normalized by the number of the dealers.

From Figure 5.1 it is clear that changes in the activity closely align with different timezones overlaps. Let us consider the generally accepted timezones for four major sessions; New York 14:00-23:00, Tokyo 1:00-10:00, Sydney 23:00-8:00 and London 9:00-18:00 (UTC+1). In Figure 5.1 we observe a spike in activity at 8:00, an hour before the opening of the London session, and relatively high activity prevails until an hour before end of London session. Additionally, we can observe another spike in activity at 13:00, which is an hour before opening of the New York session. The highest average quote frequency is observed when both London and New York sessions are open. Based on the results from this investigation, we sample from time windows between 8:00 and 16:00 on consecutive days, since clearly in this time window the highest

[1] Note that we are going to operate in London time, hence all times are expressed in UTC+1 time zone.

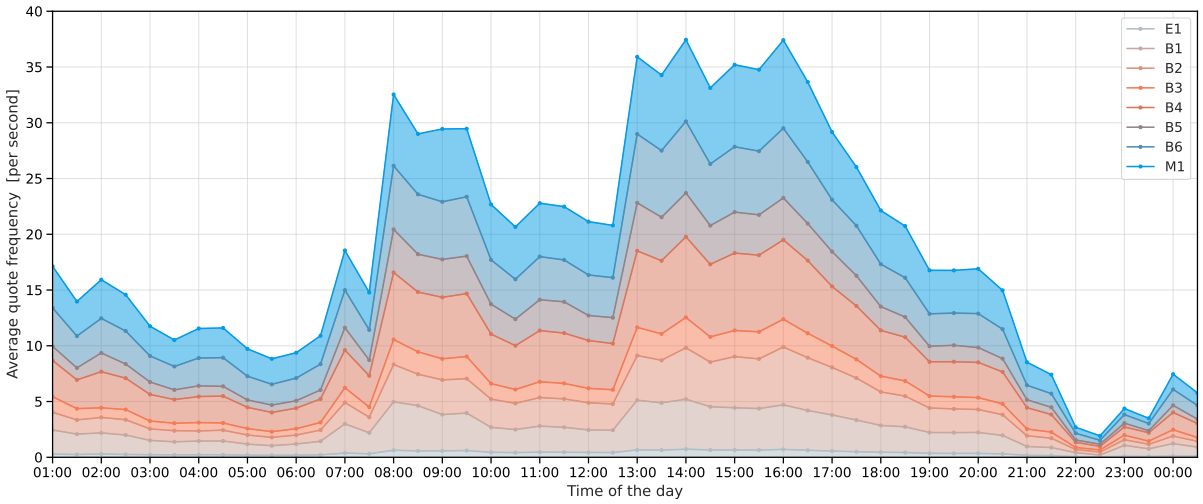

FIGURE 5.1: Average intraday EUR/USD quote frequency for trading days in the period February 27 2020 and March 27 2020. Figure illustrates how average quote frequency of all dealers in the data set evolves throughout the day. The average quote frequencies are aggregated per 30 minute time windows. For example value at 13:00 represents the quote frequency observed between 13:00 and 13:30.

activity is observed.

The same analysis is performed for USD/JPY data set. For USD/JPY the intraday quote frequency aggregated for all 30-minute time windows in the day is presented in Figure E.1. For this data set we also choose to sample from each day time windows between 8:00 and 16:00. Moreover, it should be noted that on January 1, 2019, significantly reduced activity on the FX market can be observed, hence this day is also excluded from the analysis.

5.4 Results

5.4.1 Econometric model

First, we present the econometric information-revelation map for the baseline sample. As explained in Section 5.3, the EUR/USD baseline[2] is computed from (22days · 96time windows) 2112 samples of five-minute-long subsamples obtained from trading days between 27th February and 27th March 2020. From 2112 subsamples, a total of 95 subsamples were excluded; in 67 subsamples the cumulative impulse response function did not converge, and in 28 subsamples all of the dealers had less than 10 midpoint updates within five-minute time window. Additionally, only in 12 subsamples not all dealers were included in the analysis due to some of them having insufficient number quote updates. Note that the requirement that dealers have at least 10 quote midpoints updates is also enforced by Hagströmer and Menkveld [2019]. This condition is

[2]From this point on, we will refer to the average information map obtained from the entire data set of particular currency pair as a baseline.

necessary to ensure that appropriate cointegration relations can be captured within the vector error correction model.

Dealers	Information Share [%]	Price inefficiency				Centrality $\tau = 0$		
		$\tau = 0$	$\tau = 1$	$\tau = 5$	$\tau = 10$	All	>0.1	>0.2
B1	2.454	0.472	0.229	0.159	0.116	1.085	4.314	2.121
B2	0.933	0.452	0.239	0.171	0.128	0.973	3.981	1.845
B3	2.699	0.385	0.229	0.167	0.127	0.850	3.977	1.199
B4	1.712	0.347	0.219	0.162	0.123	1.069	4.421	1.748
B5	0.508	0.543	0.286	0.204	0.151	0.888	3.480	1.746
B6	0.779	0.437	0.221	0.166	0.127	0.876	4.409	0.992
M1	15.489	0.131	0.094	0.079	0.073	0.902	3.465	1.683
E1	3.932	0.603	0.238	0.170	0.130	0.394	1.352	0.308

TABLE 5.1: Summary of econometric metrics characterizing the of EUR/USD baseline presented in Figure 5.2.

To meet out first thesis goal, we attempt to reproduce the relationships between econometric metrics that Hagströmer and Menkveld [2019] determined to provide statistically significant explanatory power for Hasbrouck's information share. In particular, Hagströmer and Menkveld [2019] show that there is a strong and statistically significant relationship between dealer centrality and Hasbrouck's information share, as well as dealer's centrality and price inefficiency.

The networks maps[3] generated with econometric approach are presented in Figure 5.2. All of the metrics presented on these maps are also summarized in Tables 5.1 and 5.2. Panels in Figure 5.2 illustrate what Hagströmer and Menkveld [2019] call the process of information revelation. Each panel is a snapshot τ time steps since the introduction of the multivariate information shock to the system. Since the data is resampled to 100ms long time intervals, each time step corresponds to 100ms since the shock. In Figure 5.2, we observe the time evolution of dealers responses, i.e. changes in price quotes, to multivariate shock introduced at $\tau = 0$. Panel A illustrates the information-revelation map the moment the shock is introduced to the system. Panel B, C, and D reveal how dealers' price inefficiencies and bilateral connections change after 100ms, 500ms and 1000ms ($\tau = 1$, 5 and 10) respectively.

On the Panel A of Figure 5.2, we observe that the market maker (M1) is positioned significantly closer to the center of the map than any other dealer is. In Table 5.1 it can be verified

[3] A viewer of electronic version of the thesis can zoom in onto each subfigure. Figures are vectorized hence each detail of the plot is clearly visualized even when the figure is zoomed on.

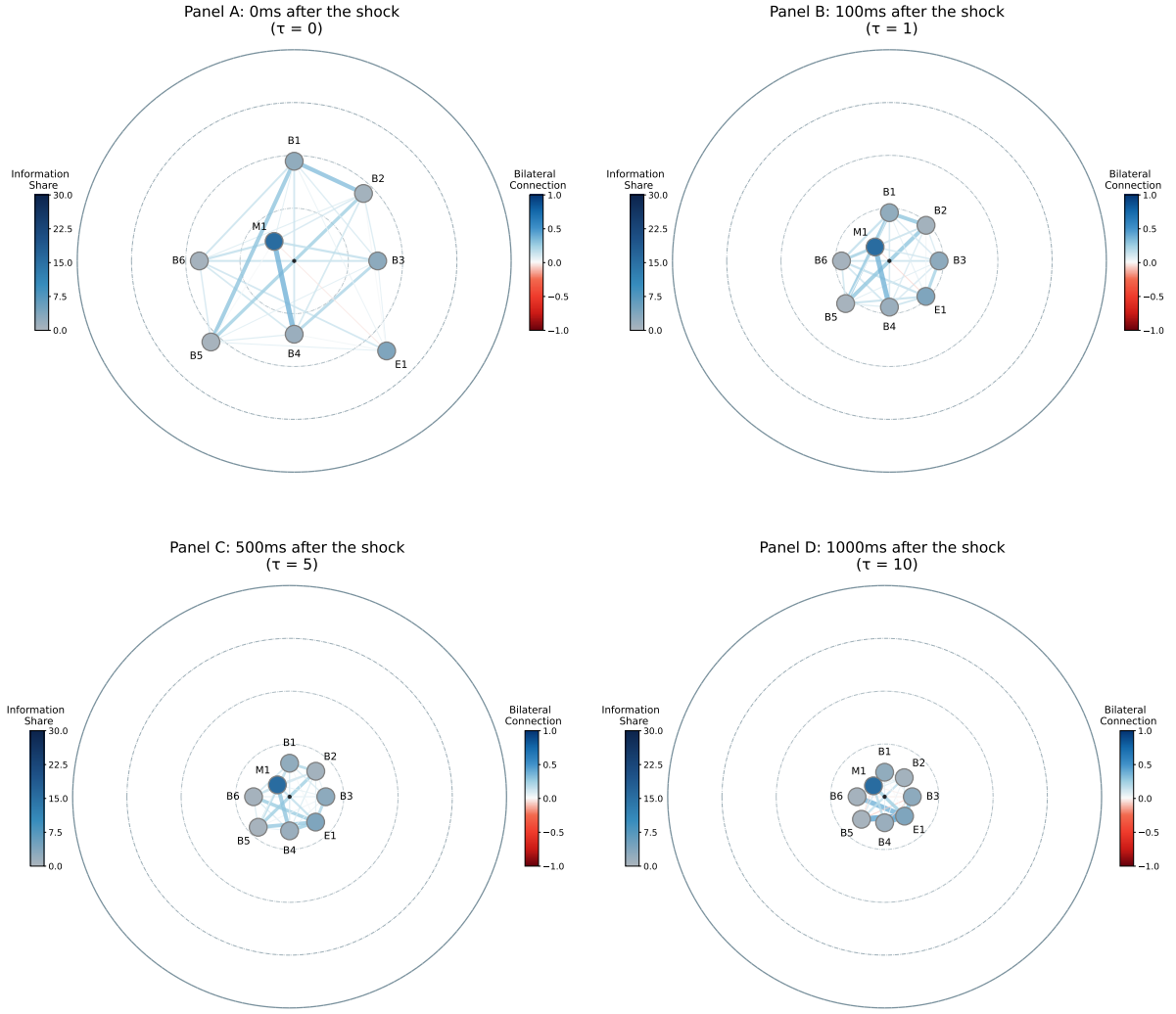

FIGURE 5.2: Baseline econometric map illustrating the process of information revelation in the EUR/USD dealer-network for period between 27th February and 27th March 2020. The network vertices correspond to the dealers, as noted in Table 4.1. Each panel visualizes averages of metrics computed from all five-minute-long subsamples of trading days τ time steps since the introduction of the multivariate shock. Edges between the vertices correspond to partial correlation matrix of cumulative responses of markets, and dealers distances to the center correspond to their price inefficiencies at particular τ. The color of vertices represent dealers' lower bound information share (metric independent of τ and presented in percentage units). The four grey circles provide the reference for the price inefficiency metric. The outer most circle is at a distance of 1 from the center. The subsequent circles are at 0.75, 0.5, and 0.25 distance from the center. More price efficient dealers are located closer to the center of the circle.

that in fact M1 is the most price efficient dealer, with price inefficiency of only 0.131. In addition to being the most price efficient dealer, M1 is determined to have on average highest information share of 15.5%, approximately four times larger than the second highest information share of any other dealer. In contrast, the electronic trading platform (E1) is not only the most price inefficient market with price inefficiency equal to 0.603, but also least central one as

very weak bilateral connections are observed for this market (see also Table 5.2). Interestingly, the information share of E1 is 3.9%, the second highest of all dealers, thus implicating against any relationship between price inefficiency and information share. Note that this observation will be further quantitatively investigated in this section. Finally, all banks, apart from B5, are located within the third inner circle – i.e. price inefficiency < 0.5. For banks the lowest price inefficiency of 0.347 is observed for B4 and highest one 0.543 determined for B5. The information share for the banks ranges from 0.5% for B5 and 2.7% for B3. Finally, it is also interesting to observe that the strongest bilateral connection is observed between M1 and B4 - the two most price inefficient dealers.

On Panel B, C and D we observe how the price inefficiency and the strength of bilateral connections changes with time steps since the introduction of the shock. Comparison of Panel A to Panel B reveals that price inefficiency disparities between dealers observed in Panel A are significantly reduced after 100ms. Clearly, all dealers are now located within the range of the inner-most circle, i.e. price inefficiency < 0.25. This observation suggests that it takes less than 100ms for dealers to impound a large portion of the informational shock to their price quotes. A striking pattern in all panels is that M1 is considerably more price inefficient than any other dealer. It also experiences the smallest changes in its price inefficiency between consecutive τ's, while maintaining it's price inefficiency dominance through all snapshots presented.

Comparison of all panels reveals that the largest change in price inefficiency for all dealers happens between $\tau = 0$ and $\tau = 1$. This observation is, however, very different from the one made by Hagströmer and Menkveld [2019]. The convergence of price inefficiencies is determined to be much faster in EUR/USD and USD/JPY data sets, than it was determined in EUR/CHF data sets by Hagströmer and Menkveld [2019]. For example, the snapshot presented at $\tau = 10$ resembles a EUR/CHF snapshot at $\tau = 50$. Note that we are working with exactly the same sampling frequency, hence $\tau = 100$ms is also used in their investigation. We propose that this dissimilarity can be attributed to the enormous reduction in latencies since 2015, and thus considerable increase in responsiveness of dealers to quote updates of other dealers. Since, the latencies were reduced, the dealers have been able to observe the changes in other dealers' quotes faster and to respond to them more quickly. Consequently, much faster responses to informational shocks can be seen. Following the approach of Hagströmer and Menkveld [2019] we choose to further focus on the snapshot at $\tau = 0$ as biggest disparities between the dealers are observed in this particular snapshot.

The adjacency matrix for the information-revelation map from Panel A is presented in Table 5.2. Thus, the table represents the partial correlation matrix for all dealers obtained at the moment when informational shock is introduced ($\tau = 0$). All of the bilateral connections are determined to be family-wise statistically significant at the 5% level. Additionally, the adjacency matrix reveals that all edges are positive apart from the bilateral link between M1

Dealers	Dealers						
	B2	B3	B4	B5	B6	M1	E1
B1	0.307**	0.096**	0.133**	0.285**	0.137**	0.072**	0.050**
B2		0.094**	0.120**	0.236**	0.079**	0.108**	0.013**
B3			0.207**	0.074**	0.141**	0.159**	0.064**
B4				0.087**	0.121**	0.364**	0.029**
B5					0.107**	0.008**	0.071**
B6						0.160**	0.137**
M1							-0.037**

TABLE 5.2: Partial correlation matrix (bilateral connections) of dealers based on EUR/USD baseline at $\tau = 0$. (**) indicates edges that are family-wise statistically significant at 5% level.

and E1, the most and least price inefficient dealers, respectively. As Hagströmer and Menkveld [2019] argue, negative partial correlations can be attributed to disagreements between the two dealers (Hagströmer and Menkveld [2019]). Since Hagströmer and Menkveld [2019] associate this partial correlation matrix with the strength of bilateral connections between markets they attempt to use it as an explanatory variables for the distribution of information share. We follow the steps of Hagströmer and Menkveld [2019] and also attempt to recover their findings, which leads us to our first hypothesis.

Hypothesis 1. **Central dealers are more informed**.
The hypothesis is inspired by work of Babus and Kondor [2018], who show that dealer's perspective on the fundamental value depends on opinion of all the other dealers that a given dealer interacts with. Babus and Kondor [2018] further show that under the assumption that the information is equally distributed between all dealers, since more central dealers source their information from larger group of dealers, they are able to come up with a better fundamental value, i.e. be more informed.

Hagströmer and Menkveld [2019] test this hypothesis in two ways. First, by investigating if there is a relationship between dealers' centralities and their lower bound information shares. Secondly, by determining if more central dealers become more price inefficient as we move in time since time shock τ. We attempt to reproduce this results on our data sets. To do so, we use the results obtained from the baseline sample. The dealers' informativeness metrics are regressed against their centralities. The two scatter plots uncovering the relationship between price inefficiency and dealer's centrality as well as lower bound information share and the dealer's centrality are presented in Figure 5.3.

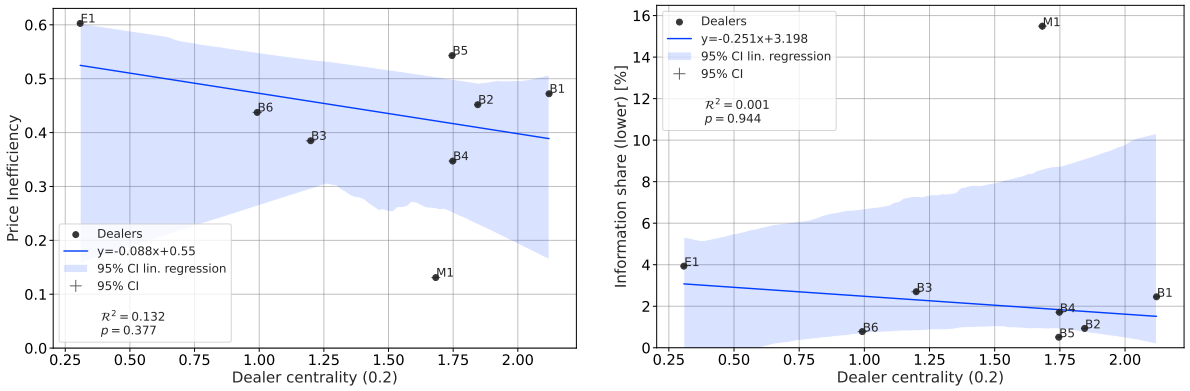

FIGURE 5.3: Scatter plot of dealer centrality (0.2) versus price inefficiency and lower bound information share for EUR/USD baseline. The solid line is generate with a linear regression, and the shaded area surrounding it represents the 95% confidence interval. Additionally, for each dealer's metric a 95% confidence interval is plotted, which may appear invisible given the number of samples were used to generate the means.

The figure presented on the left hand side illustrates the relationship between average dealers centralities and their average price inefficiencies as determined in the EUR/USD baseline. Following the steps of Hagströmer and Menkveld [2019] we use dealer centrality (0.2) – which is defined as the number of connections a dealer has, whose weights are larger than 0.2 (Hagströmer and Menkveld [2019]). The figure illustrates weak and statistically insignificant negative relationship between dealers' centralities and their price inefficiencies. Especially noteworthy is the fact that market maker M1 strongly deviates from the recovered trend, thus suggesting against existence of relationship between investigated variables. In conclusion, we observe that a very small proportion of price inefficiency variance is explained with dealer centrality as $\mathcal{R}^2 = 0.132$, and the relationship is statistically insignificant with p-value = 0.377. Hence, we are unable to recover the relationship observed by Hagströmer and Menkveld [2019].

The scatter plot presented on the right hand side of Figure 5.3 reveals the relationship obtained between dealer centrality (0.2) and Hasbrouck's lower bound information share. The scatter plot recovers negative trend, thus opposite to the one expected. As explained in Hypothesis 1, the expectation is that more central dealers are more informed, which we do not observe here. Moreover, this relationship is found to be statistically insignificant with $\mathcal{R}^2 = 0.001$, with p-value = 0.944.

The same investigation is performed for USD/JPY data set. The map illustrating the process of information revelation in USD/JPY dealer-network is presented in Figure E.2 and metrics are summarized in Table E.4. The linear regressions for the metrics under investigation are presented in Figure E.3. In fact, for the USD/JPY data set, we recover a relationship much closer to the expected ones. The relationship between dealer centrality (0.2) and price inefficiency was found to be significant at 10% level with $\mathcal{R}^2 = 0.549$ and p-value = 0.092; whereas a positive

relationship between information share and dealer centrality was found to be significant at 5% level with $\mathcal{R}^2 = 0.733$ and p-value = 0.03.

In conclusion, while we managed to find statistically significant relationship between the econometric metrics based on baseline from USD/JPY data set, the relationship was not significant for the baseline from EUR/USD data set. Therefore, we conclude that there is some evidence that such relationships exist, however our results do not provide evident, unequivocal support for this hypothesis.

Before moving to the next hypothesis, it would be also interesting to investigate if there is any statistically significant relationship between information share and price inefficiency. While this particular relationship is not investigated by Hagströmer and Menkveld [2019], this investigation is motivated by our preliminary observations. In Figure 5.2, it can be seen that market maker M1 is both the least price inefficient and holds the highest information share. Moreover, the plausibility of such relationship is much clearer in Figure E.2, which shows that all most price efficient banks have higher information share than less price efficient dealers. Therefore, for completeness purposes the relationship between price inefficiency and information share metrics based on baseline samples is also investigated and the results are presented in Figure 5.4.

The scatter plot suggests a negative relationship as expected, i.e. the higher the price inefficiency of a given dealer, the smaller their information share. While the relationship is statistically significant at 5% level, because of the dominance of M1 dealer and the trend that does not explain the variations for other dealers, the explanatory power is determined to be only $\mathcal{R}^2 = 0.583$.

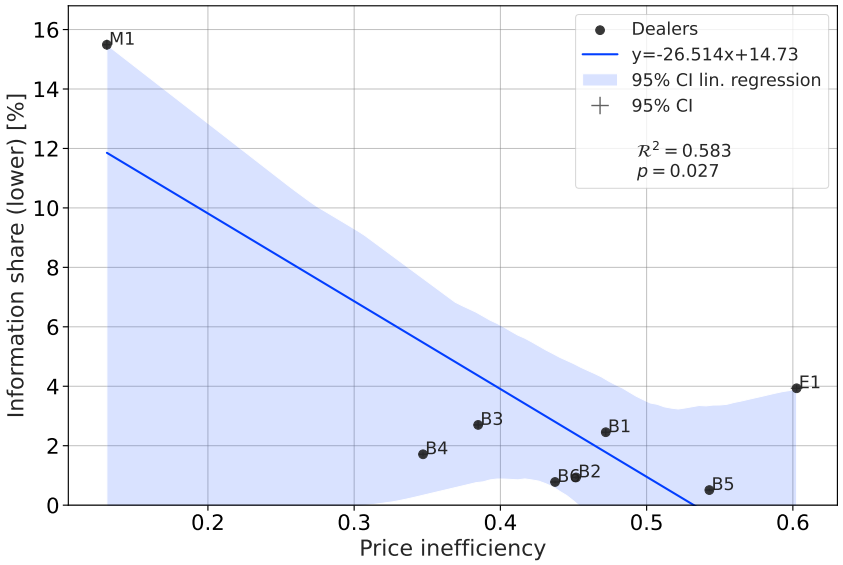

FIGURE 5.4: Scatter plot of dealer price inefficiency versus Hasbrouck's lower bound information share for EUR/USD baseline.

For the USD/JPY baseline we observe significantly stronger relationship. Figure E.4 reveals that almost all dealers perfectly line up with the recovered trend, with electronic trading platform E1 being the only dealer significantly deviating from the trend. Note that the same observation about E1 can be made on Figure 5.4. Consequently, for the USD/JPY baseline relationship is found to be statistically significant at 5% level with the explanatory power being $\mathcal{R}^2 = 0.86$. In conclusion, we find the relationship statistically significant for both EUR/USD and USD/JPY data sets.

5.4.2 Information-theoretic network inference

After presenting the results from the econometric model, we should reevaluate the use of partial correlation to capture the bilateral connections between dealers. First of all, centrality is just a sum of connections with some arbitrarily predefined minimum weight. Hence, this metric does not explicitly account for the strengths of all connections to particular dealer, but only counts them. While filtering out correlations below some threshold weight provides the means to disregard weaker and less relevant connections; centrality does not distinguish a partial correlation of e.g., 0.3 and 0.9. As a result, when centrality metric is employed, a lot of potentially relevant information about the nature of dealer connections is lost.

Second of all, partial correlation is not a directional measure, hence by utilizing the partial correlation and centrality metric, we accept the underlying assumption that both dealers benefit from the interaction in the exact same way. However, this assumption is at flaw when the microstructure framework is considered. Recall that the microstructure approach recognizes that some information is not publicly available, thus there exists an information asymmetry between dealers, which we attribute to the existence of private information (Lyons et al. [2001]). Thus, given the information asymmetry, there must be an asymmetric benefit from an interaction of two dealers. Thus, as Hagströmer and Menkveld [2016] state, it is reasonable to assume that there must exist some flow of information between dealers, in which less informed counterparties learn from more informed ones. Therefore, to better understand and explain the information share of each dealer, it may be necessary to employ a directional measure that can be used to quantify these flows of information. To do so, we propose to employ the transfer entropy and conditional transfer entropy measures, to quantify the information flows between pairs of dealers which brings us to the following two hypothesis.

Hypothesis 2. **The price discovery process is dominated by dealers who share the most private information to less informed dealers.**

First, recall that the market dominates the price discovery process if it has a dominant influence on the change of the long-term cointegration equilibrium price, i.e. efficient price. Additionally, the information share is a measure that quantifies relative contribution of each market to the

innovations in the efficient price. As previously stated, we hypothesize that the information flows from more informed dealers to less informed dealers. Thus, we propose that Hasbrouck's information share metric can be explained with directional informational flows captured by transfer entropy and conditional transfer entropy. The reason for this is that information flows computed especially with conditional transfer entropy quantify direct contributions of a dealer to changes in quotes of other dealers. Since the common efficient price is dependent on the prices of all dealers, as it is a reflection of all public and private information available in the dealer-network, we postulate that the dealer that has the most significant impact on the changes of the efficient price is the one that is the main source of information for other dealers. Consequently this hypothesis provides the bridge between the information flows computed with information-theoretic metrics and Hasbrouck's information share.

Hypothesis 3. **The largest disparities between information contributions are observed between different types of dealers; i.e., banks, non-bank market makers and electronic trading platforms.**
In the microstructure framework, as Lyons et al. [2001] state, the private information are attributed to the internal order flow observed by each dealer. For example, banks see demands of private clients, usually larger institutions or central banks, which are not publicly observable (Lyons et al. [2001]). Lyons et al. [2001] further show that such information forecasts the changes in the exchange rates. Additionally, Evans and Lyons [2002] find that order flow can explain up to 60% of changes in the FX rates. This notion is also supported by the European Central Bank (ECB), which also claims that *order flow is an important determinant of exchange rate dynamics in the short term and possibly even in the medium term* (Vitale [2006]). Moreover, it should be noted that pricing that takes place within the scope of each dealer is a reflection of internal liquidity. Thus the idea of the efficient price for each dealer is a reflection of the demand observed in the order flow. This idea is also supported by Hagströmer and Menkveld [2016]. Consequently, we hypothesize that given the differences in types of clients that banks and non-bank market makers cater to, and consequently the differences in the order flows observed by these types of dealers, there are significant differences in information gathered and shared by these dealers. In particular, the fact that non-bank market maker has freedom to cater to a larger and more diversified pool of clients suggests that it should be faster than banks in the price discovery process. When it comes to electronic trading platform, given the functional role played by ECNs, we hypothesize that these platforms provide very little to no unique information to the dealer-network. As previously discussed, ECNs only collect the quotes from different dealers, hence they act more as transmitters of information rather than sources of information.

Dealers (TE outflow)	Dealers (TE inflow)							
	B1	B2	B3	B4	B5	B6	M1	E1
B1	0.0000	0.0053	0.0016	0.0005	0.0088	0.0049	0.0006	0.0510
B2	0.0031	0.0000	0.0014	0.0004	0.0074	0.0051	0.0006	0.0462
B3	0.0136	0.0120	0.0000	0.0006	0.0180	0.0094	0.0005	0.0792
B4	0.0220	0.0183	0.0062	0.0000	0.0261	0.0173	0.0012	0.0809
B5	0.0018	0.0014	0.0008	0.0003	0.0000	0.0020	0.0004	0.0425
B6	0.0090	0.0087	0.0022	0.0009	0.0126	0.0000	0.0008	0.0532
M1	0.0265	0.0219	0.0116	0.0037	0.0288	0.0222	0.0000	0.0744
E1	0.0021	0.0028	0.0005	0.0003	0.0039	0.0008	0.0004	0.0000

TABLE 5.3: Table summarizing the mean of family-wise statistically significant at 10% level transfer entropies detected between pairs of dealers in EUR/USD data set for the period between between 27th February and 27th March 2020. The information flows from dealers (outflow) to dealers (inflow).

To investigate our hypothesis we generate maps of information flows in the dealer-network at $\tau = 0$. TE and CTE information maps are generated by retaining Hasbrouck's lower bound information share and price inefficiency metrics, and replacing the partial correlation of cumulative price changes with transfer and conditional transfer entropies. An important distinction must be made here, namely, while the flows in TE network are composed purely based on transfer entropies, the edges in CTE network may be computed based on averages of both transfer and conditional transfer entropies. This is because, in the case when in particular time window we detect only one statistically significant TE to a given dealer (information inflow), then there is no need it to involve conditional transfer entropy. In that scenario, we can treat transfer entropy as the flow of unique information to that dealer, i.e. TE=CTE. Moreover, it should be noted that the subsamples that were excluded for the econometric baseline were also excluded from the TE and CTE baselines. Consequently, the econometric, TE and CTE networks are generated based on the same set of subsamples. Finally, given our reflections on the potentially relevant information being lost when centrality metric is used, we introduce two new metrics: total inflow and total outflows. The total inflow (outflow) is simply the sum of all flows that are directed towards (away from) particular dealer. In this way we explicitly account in the strengths of all informational inflows and outflows detected for each dealer.

The TE and CTE information maps are presented in Figures 5.5 and 5.7, respectively, and the informational flows computed with transfer entropies are present in Table 5.3. Moreover, all the metrics visualized on the TE and CTE information maps are summarized in Table 5.4.

	Information Share [%]		TE [nats]		CTE [nats]		Price inefficiency
Dealers	Lower	Upper	(inflow)	(outflow)	(inflow)	(outflow)	$\tau = 0$
B1	2.454	47.812	0.076	0.070	0.009	0.002	0.472
B2	0.933	43.602	0.068	0.062	0.008	0.002	0.452
B3	2.699	50.183	0.023	0.129	0.009	0.009	0.385
B4	1.712	60.013	0.007	0.166	0.005	0.011	0.347
B5	0.508	34.494	0.102	0.048	0.007	0.001	0.543
B6	0.779	43.911	0.060	0.084	0.009	0.003	0.437
M1	15.489	78.717	0.004	0.183	0.003	0.033	0.131
E1	3.932	22.000	0.414	0.010	0.010	0.001	0.603

TABLE 5.4: Summary of key information-theoretic metrics characterizing the EUR/USD baseline presented in Figures 5.5 and 5.7.

5.4.2.1 TE information network

First, we consider the results obtained with TE network algorithm. The flows in the EUR/USD dealer-network presented in Figure 5.5 represent the averages of family-wise statistically significant at 10%[4] level transfer entropies between pairs of dealers. Like in Figure 5.2, the TE information map is generated for the period between 27th February and 27th March 2020.

In Figure 5.5, we observe many directional connections between dealers of various strengths. The first observation that can be made is that it appears as though the strongest information flows are observed flowing from more price efficient to less price efficient dealers. In fact the most price efficient dealers, such as M1, B4 and B3, are sources of the most total information outflows, with TE outflows of 0.183, 0.166 and 0.129 nats, respectively. Additionally, we observe relatively very strong information flows from all dealers to electronic trading platform E1 - the most price inefficient dealer. In Table 5.3, it can be further verified that indeed the strongest outflows for each dealers are the ones directed towards E1. The total TE inflow for E1 is determined to be 0.414 nats. On the other hand, we also observe that on average very little information is flowing out of E1. This can be further verified in Table 5.4, where TE total inflows and outflows to each dealer are summarized. This is in line with our expectations, given Hypothesis 3, and the fact that electronic trading platforms aggregates quotes submitted by different dealers.

[4]Note that 10% is the lowest significance level that we were able to set given 500 permutations per statistical test and Bonferroni correction. In practice we allow at most one surrogate out of 500 permutations to exceed the true transfer entropy. Moreover, further analysis revealed that for TE and CTE entropies less than 5% and 10%, respectively, of the statistically significant values were determined to have their p-values greater than zero.

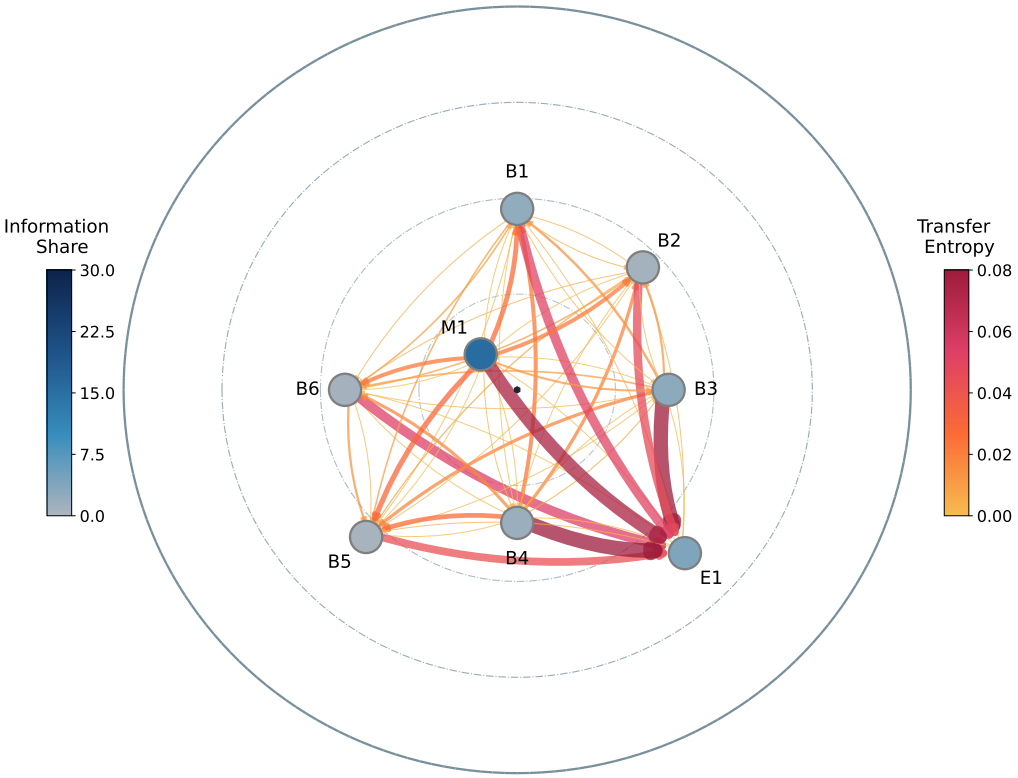

FIGURE 5.5: TE information map illustrating mean information flows in the EUR/USD dealer-network between 27th February and 27th March 2020. Information share is presented in percentage units, whereas transfer entropy is in nats. The width of information flows scales with the strength of the flow. Additionally, colors of both Hasbrouck's lower bound information share and transfer entropy reflect their true values as indicated in the supporting color scales. The remarks about price inefficiency and Hasbrouck's information share metrics from Figure 5.2 apply to all figures with TE and CTE information maps.

Now let us consider the inflows and outflows of non-bank market maker M1. As already observed in the econometric map, the dealer M1 is the most price efficient dealer. In Figure 5.5 we further observe many strong information flows stemming from the non-bank market maker M1. Further inspection of Tables 5.3 and 5.4 reveals that indeed M1 is the dominant source information with the highest TE total outflow of 0.183 nats. Additionally, it can be seen that M1 is in fact a source of the largest information flows for all dealers except for E1.

To test Hypothesis 2 we generate scatter plots that illustrate the relationship between Hasbrouck's information share and the total TE outflow. For the analysis, we consider both lower and upper bounds of Hasbrouck's information share. Consideration of both upper and lower bounds information share is motivated by the nature of these metrics, which is in a way analogous

to transfer entropy and conditional transfer entropy. Recall that the upper bound information share accounts not only for market's own unique contribution to the changes in the efficient price, but also its correlation with other markets. On the other hand, the lower bound information share is considered to only account for the direct and unique contribution to changes in the efficient price that is uncorrelated with any other market (Baillie et al. [2002]). In an analogous manner, transfer entropy does not quantify just unique information, but it also potentially accounts for other flows to the same dealer. On the other hand, conditional transfer entropy, while not perfect as discussed in Chapter 3, it is here used in an attempt to capture the unique contribution of each dealer.

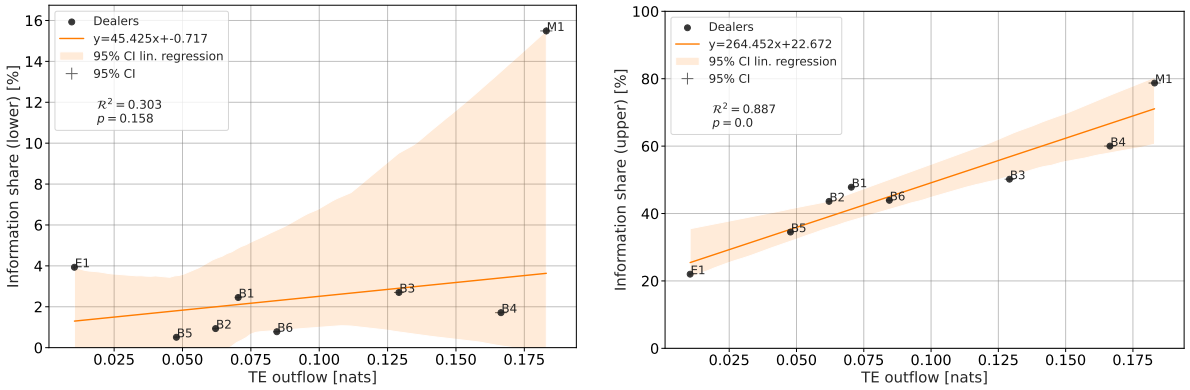

FIGURE 5.6: Scatter plot of TE outflows versus Hasbrouck's lower and upper bound information shares for EUR/USD baseline. Note that all orange scatter plots are related to TE information networks, whereas the violet ones present relationships for metrics from CTE information network.

In Figure 5.6, we present the scatter plots uncovering the relationship between lower and upper bound of the information share, and the total TE outflow of each dealer. The figure presented on the left-hand side illustrates the relationship between lower bound information share and TE outflow. While a positive trend is recovered, as expected, it does not clearly explain the variance of the information share, and especially the information share of market maker M1. Consequently, the relationship is not found to be statistically significant with p-value $= 0.158$ and $\mathcal{R}^2 = 0.303$. On the other hand, the regression of TE outflows on the upper bound information shares reveals a clear trend. The relationship between these two variables is determined to be statistically significant at 1% level with p-value$= 0$ and $\mathcal{R}^2 = 0.887$. As can be observed, nearly all dealers are captured within 95% confidence level of the linear regression.

The same investigation is performed for the metrics obtained in the USD/JPY baseline. The TE information map is presented on the left-hand-side of Figure 5.11. Scatter plots for this data set are presented in Figure E.5. In contrast to EUR/USD baseline, in the USD/JPY baseline the relationship between lower bound information share and TE outflows is found to

be statistically significant at 1% level, with p-value=0.001 and $\mathcal{R}^2 = 0.95$. Additionally, the relationship between upper bound information share and TE outflows was also found to be statistically significant at 1% level, with p-value=0.002 and $\mathcal{R}^2 = 0.937$.

In conclusion, while the relationship between Hasbrouck's lower bound information share and TE outflows is not exposed in the EUR/USD dataset, it is determined to be statistically significant at 1% level for the USD/JPY baseline. On the other hand, the relationship between Hasbrouck's upper bound information share and TE outflows was determined to be statistically significant for both baselines. Thus, to some extend our expectations for the stronger relationships between TE and upper bound information share are met. However, we have yet to explore CTE information maps, and determine whether there is any relationship between CTE outflows and lower bound information shares.

5.4.2.2 CTE information network

In this subsection, we will consider the results obtained with the CTE information map, which is presented in Figure 5.7. The CTE information map is generated for the same time period as the TE information map. Note that the edges in the CTE map are much weaker, hence, to make the connections visible on the plot, the scale for the weights is reduced four times as compared to the scale used in the TE information map.

The reduction in the average weights of the information flows is a consequence of conditioning, and suggests that conditioning successfully filtered out influences from other information contributors. As a result, we are left with unique contributions from each dealer. Of course, it cannot be determined whether any synergistic interactions were captured; however, since all of the CTE edges are significantly smaller than the TE edges, at least we can infer that overall more redundant information was filtered out than synergistic information created.

One observation that can be made is that CTE does not expose previously observed large information flows from all dealers to electronic trading platform E1. Out of all inflows to E1 observed in the TE information map, the only remaining and relatively large information flow is observed from B3 to E1. This suggests that on average B3 is the contributor of the most of the unique information to the electronic trading platform. Another observation that can be made, is that market maker M1 now clearly appears to be the source of the largest flows of information. Interestingly, an identical observation can be made on the CTE information map for USD/JPY data set presented in Figure 5.11. On the USD/JPY TE information map, we observe four relatively strong flows towards bank B1, however, on the CTE information map, the only remaining strong flow is one directed towards B1.

To test out the hypothesis, we generate scatter plots to expose the relationship between Hasbrouck's information share and CTE outflows for each dealer. The scatter plots are presented in Figure 5.8. The relationship between CTE outflow and lower bound information share is

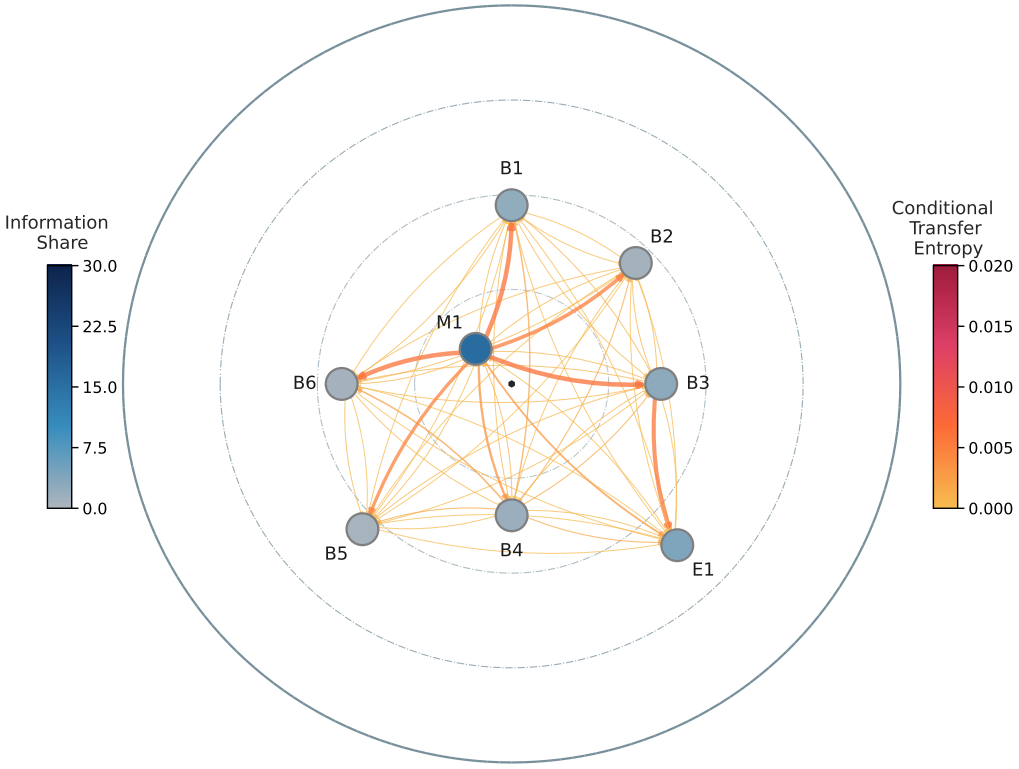

FIGURE 5.7: CTE information map illustrating mean information flows in the EUR/USD dealer-network for period between 27th February and 27th March 2020. Information share is presented in percentage unit whereas conditional transfer entropy in nats

determined to be significant at 1% level, with p-value= 0.001 and $\mathcal{R}^2 = 0.849$. Consequently, the CTE outflows appear to explain a considerably larger portion of the variance of the Hasbrouck's lower bound information share than TE outflows or dealer centrality do. Furthermore, the scatter plot reveals that market participants E1, B1, B2, B5, and B6, contribute very little unique information to the dealer-network, whereas banks B3 and B4 contribute significantly more information to the dealer-network. And finally, as also observed on the TE information map, market maker M1 is the dominant source of the information for the dealer-network. The relationship between the upper bound information share and CTE outflows is determined to be significant at 1% level, with p-value= 0.005 and slightly weaker explanatory power of $\mathcal{R}^2 = 0.764$.

The same investigation is also conducted for the USD/JPY baseline. The scatter plots generated for USD/JPY information map are presented in Figure E.5. Scatter plots expose statistically significant (at 5% level) relationships between the upper and lower bound information

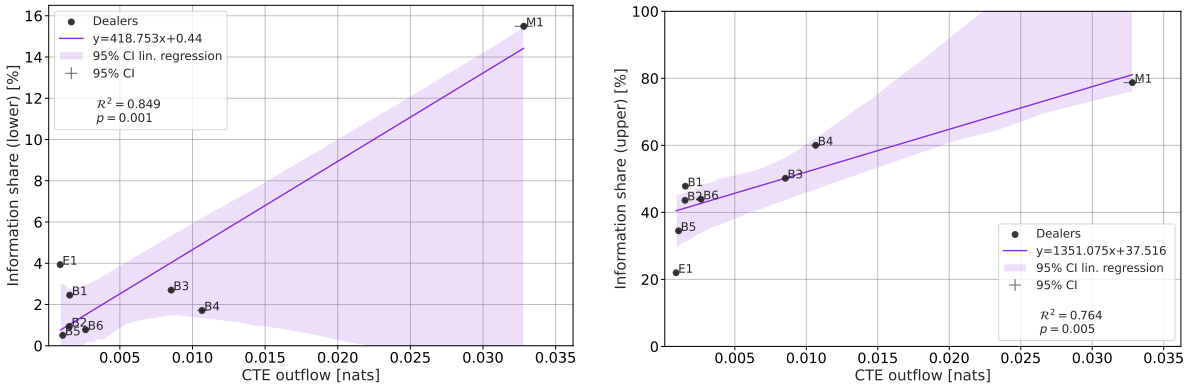

FIGURE 5.8: Scatter plot of CTE outflows versus Hasbrouck's lower and upper bound information shares for EUR/USD baseline.

shares and CTE outflows. For the lower bound information share, the p-value is determined to be 0.013 with $\mathcal{R}^2 = 0.817$, and for the upper bound information share, the p-value = 0.029 with $\mathcal{R}^2 = 0.736$.

In conclusion, we find that CTE outflows can explain much greater portion of the lower bound information share's variance than dealer centrality does. Additionally, in both EUR/USD and USD/JPY baselines, CTE outflows provide larger explanatory power for the lower bound than for the upper bound of the information share.

Before we go on to analyze the reaction of the dealer-network to the a flash crash or the announcement of quantitative easing by ECB, let us first investigate if there is any relationship between the responsiveness of dealers to information shocks and information flows. Recall that in Section 5.4.1 we made an observation that after 100ms large portions of disparities in price inefficiencies are reduced for all dealers. We propose, that the reduction in price inefficiency can be attributed to amount of information learned by these dealers. Hence the following hypotheses are made.

Hypothesis 4. **The more a given dealer learns from other dealers, the faster is his response to the informational shocks, i.e. the faster is his price discovery process.** This hypothesis is an extension to Hypothesis 2. Recall that informational flows are reflected by quote updates of other dealers. Thus, gaining more information from other dealers should lead to faster price discovery, which should be reflected by more efficient quotes. The speed of the response of the dealer to the information shock is captured by his changes in price inefficiency. Thus we postulate that there is a positive relationship between the informational inflows and the responsiveness of dealers to the informational shocks. The responsiveness of the dealers is captured by the absolute change in the price inefficiency. Note that, as we progress in τ all dealers become less price inefficient, hence the absolute value of price inefficiency change is considered.

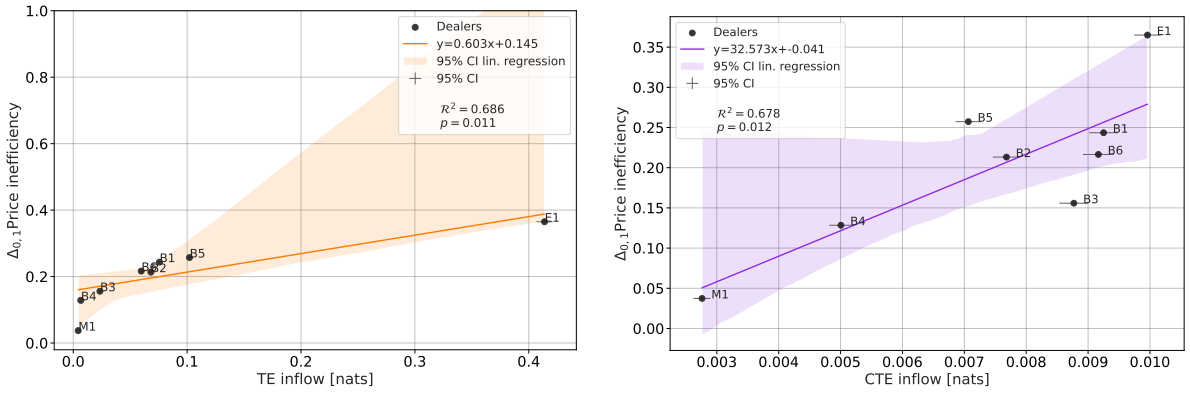

FIGURE 5.9: Scatter plot of TE and CTE inflows versus absolute change in price inefficiency from $\tau = 0$ to $\tau = 1$ for EUR/USD baseline.

In order to test this hypothesis, we consider the change in price inefficiency between $\tau = 0$ and $\tau = 1$, simply because the largest changes in price inefficiency are observed between these two snapshots. The scatter plots revealing the relationship between the change in price inefficiency and TE and CTE inflows are presented in Figure 5.9. The regression plot reveals a positive relationship between TE inflows and change in price inefficiency The relationship is determined to be statistically significant at 5% level with p-value= 0.011 and $\mathcal{R}^2 = 0.686$. While TE inflows relatively well explain the changes in price inefficiencies for banks and electronic trading platforms, the change in price inefficiency for non-bank market maker is not captured very well. For CTE inflows, the relationship is also determined to be positive and statistically significant at 5% level with p-value= 0.012 and $\mathcal{R}^2 = 0.678$. On this scatter plot, we observe much more deviations from the linear regression line.

The results of the same investigation for USD/JPY baseline are presented in Figure E.6. The regression plot reveals a positive relationship between change in price inefficiency and both TE and CTE inflows. For TE and CTE inflows, the relationship is determined to be statistically significant at 1% level with p-value= 0.002 and $\mathcal{R}^2 = 0.934$, and p-value= 0.001 and $\mathcal{R}^2 = 0.947$, respectively. Consequently, we have determined that the relationship is statistically significant for both TE and CTE networks in EUR/USD and USD/JPY baseline. Clearly, much larger proportion of price inefficiency change is explained by information-theoretic metrics in the USD/JPY baseline.

Up to this point of the analysis we have observed that in the network, there are dealers that are predominant sources of information for others, like M1, and dealers who are predominately recipients of the information. As we determined, M1 is characterized by very large outflows but has very low inflows, whereas the opposite is observed for electronic trading platform E1. Given these observations, it would be interesting to investigate whether there is any relationship between dealers' inflows and outflows in the dealer-network. The investigation reveals that in fact there exists a log-linear relationship between outflows and inflows in both TE and CTE

networks. For EUR/USD and USD/JPY networks the scatter plots visualizing this relationships are presented in Figure 5.10 and Figure E.7, respectively.

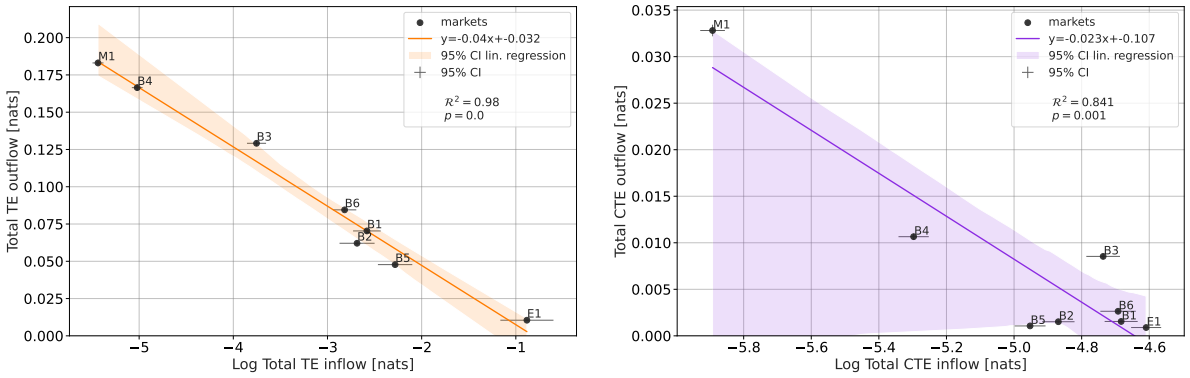

FIGURE 5.10: Scatter plot of log TE inflows versus TE outflows (LHS) and log CTE inflows versus CTE outflows (RHS) for EUR/USD baseline.

On the left-hand side of Figure 5.10 the relationship between log of TE inflows and TE outflows is presented. A negative relationship can be seen, i.e., the more information flows in, the less information flows out. Additionally, we observe a very good fit for our data, hence the relationship is determined to be statistically significant at 1% level with p-value= 0 and $\mathcal{R}^2 = 0.98$. On the right-hand-side of Figure 5.10 the relationship between log of CTE inflows and CTE outflows is presented. The log-linear relationship is significantly weaker than the one observed for TE metric. The relationship is determined to be statistically significant at 1% level with p-value = 0.001 and $\mathcal{R}^2 = 0.841$. However,many more deviations from the linear regression line can be observed than for the scatter plot of TE inflows vs outflows.

In USD/JPY baseline, we also observe a stronger relationship between log TE inflows and TE outflows, than for the CTE network. For the TE network, the relationship is determined to be statistically significant at 1% level with p-value = 0.003 and $\mathcal{R}^2 = 0.911$. On the other hand, the relationship between log CTE inflows and CTE outflows is determined to be statistically significant at 10% level with p-value = 0.05 and $\mathcal{R}^2 = 0.659$.

5.4.3 Flash crash

On January 2nd 2019, the USD/JPY currency pair has experienced the so called flash crash. Generally, there is no official definition of flash crash simply because each crash is unique, and is usually caused by a different factor. However, broadly a flash crash can be defined as a large-scale decline in price of financial instrument in a short period of time (Tsai [2018]). In case of USD/JPY, on the day of flash crash at around 22:35 London time, the US Dollar depreciated by approximately 3% against the Japanese Yen in just about 30 seconds (Wehrli and Sornette [2022]). The currency pair movement is presented in Figure 4.2, where we observe a sudden large drop in the exchange price of USD/JPY currency pair, followed by turbulent price movements and gradual stabilization when an broader agreement on a new efficient price is being reached. Events like the flash crash are of considerable interest from scientific perspective for two reasons. First, because they expose the dynamics of the dealer-network and price discovery process on a much large time scale. Second, because while they are very rare, they have an influence on everyone with an exposure to the affected risk factor.

In an attempt to expose the dynamics observed in the dealer-network before, during and after the flash crash, we will investigate the informational maps generated for each hour of the day of the flash crash. In particular, we will zoom into the period between 16:00 on the 2nd of January and 8:00 on 3rd of January. Additionally, we will also use the TE and CTE baseline information maps obtained for this currency pair presented in Figure 5.11 as a point of reference for our information maps from the flash rash period.

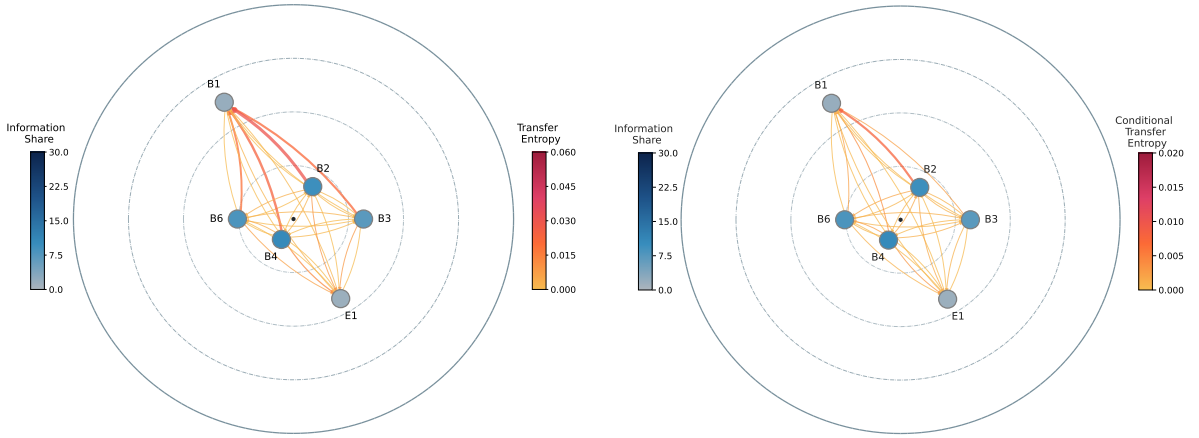

FIGURE 5.11: TE (LHS) and CTE (RHS) information maps illustrating mean information flows in the USD/JPY dealer-network for period between 2nd and 17th January 2019.

Our investigation will be led by two hypotheses. The first one pertaining to the changes in the information flows during and after the extreme event like the flash crash. The second

hypothesis is motivated by the findings presented by Hagströmer and Menkveld [2019].

Hypothesis 5. **There is more exchange of information between dealers when there is more to be learned.**
Duffie et al. [2009] show that the incentive to seek information is reduced at the time of public announcements. This hypothesis is tested by Hagströmer and Menkveld [2019], who find out that during the times of public announcements dealers connections are weaker. However, we are going to take a slightly different approach. We postulate that the opposite happens during the times of extreme events in the market. Namely, during for example a flash crash, dealers increase their price discovery efforts to establish a new price and protect themselves from being arbitraged. Therefore, during and right after the flash crash, we are expecting to observe more information flows in the dealer-network.

Hypothesis 6. **During a flash crash, the dealer-network deviates from a steady-state structure and eventually converges back to it.**
Hagströmer and Menkveld [2019] show that after the EUR/CHF peg was dropped, the dealer-network gradually converged to a steady-state structure, which they defined as the structure observed in the baseline generated from the two-weeks-worth of data following[5] the flash crash. In an analogous manner, we investigate this hypothesis based on the USD/JPY flash crash. We postulate that before the flash crash we will observe the dealer-network in it's steady-state structure. Additionally, we hypothesize that after the flash crash, the dealer-network will gradually converge back to the steady-state structure.

In order to test these hypotheses, we generate CTE information maps for each one-hour-long time window between 16:00 on the 2nd of January and 8:00 on 3nd of January. The CTE information maps are presented in Figures 5.12 and 5.13. In this section we will focus our analysis on the CTE information map, however the same observations can be made in the TE information maps.

In Figure 5.12, we observe that between 16:00-21:00, there are very little and also relatively weak information flows between dealers. We also observe that the strongest informational flow is observed from bank B2 to bank B1, as in the baseline (right-hand side of Figure 5.11). Furthermore, the price inefficiency of all dealers besides E1 is approximately the same as in our baseline. E1 is determined to be significantly more price inefficient than it is found to be on average. In time between 16:00-21:00 very little variability in dealers' informational flow, price inefficiency or information share is observed.

Panel F of Figure 5.12 presents the average informational map observed between 21:00 and 22:00, i.e. approximately up to 30min before the flash crash. On this panel, some considerable

[5] Note that Hagströmer and Menkveld [2019] use two-weeks-worth of data following the flash crash, just because before the flash crash Euro was pegged to Swiss franc.

changes in the dealer-network ca be seen. For instance, bank B6 and the electronic trading platform become substantially more price inefficient, as they move away from the center of the informational map. The same is observed for banks B2, B3 and B4 but to a much lesser extent. Thus, we observe that all dealers become more price inefficient. Additionally, it can be noticed that the information share of the most informed banks slightly decreases as well. Interestingly, these observations are made even before the flash crash takes place.

Panel G illustrates the dealer-network structure between 22:00 and 23:00. As previously mentioned the flash crash takes place at around 22:35 London local time, hence this figure captures the changes taking place during the flash crash and roughly in the first 25 minutes following the event. On the informational map, we can see that the price inefficiencies of all dealers are extended even more since the last snapshot. The only bank that does not experience an extreme increase in price inefficiency is bank B2. Additionally, we observe more informational flows emerging as compared to Panel F; for example, the flow from B6 to B3, or from B1 to B2 is not observed in any previous snapshot. Curiously, we also observe an increase in information share for electronic trading platforms. It is thought-provoking since this is the only snapshot in which we can observe any relatively large information share for E1.

As Panel H reveals, dealers become even more connected. Additionally, we observe that B2 and B1 take over a much larger portion of information share, with B1 dominating temporarily in this time window. However, further inspection of the following panels unveils that in fact B2 has the largest information share until the last investigated time window.

In Panel I, B2's information share becomes very large. Additionally, it is interesting to notice that banks B2, B3, B4 and B6 become the most price inefficient in the whole dealer-network. It is notable given that in fact these four banks were the most price inefficient before the flash crash (in Panels A-F) and are also found to be on average the most price inefficient in our baseline (Figure 5.11)). Evidently, the dealer-network structure slowly converges towards the steady-state structure.

In the following panels, we observe that all dealers gradually become more price inefficient. In Panel P, we observe that banks B1, B2, B3, B4 and B6 re-establish their steady state price inefficiencies. The electronic trading platform is still slightly away from its steady state price inefficiency. Additionally, the distribution of the Hasbrouck's information share also very closely aligns with the one observed in the baseline.

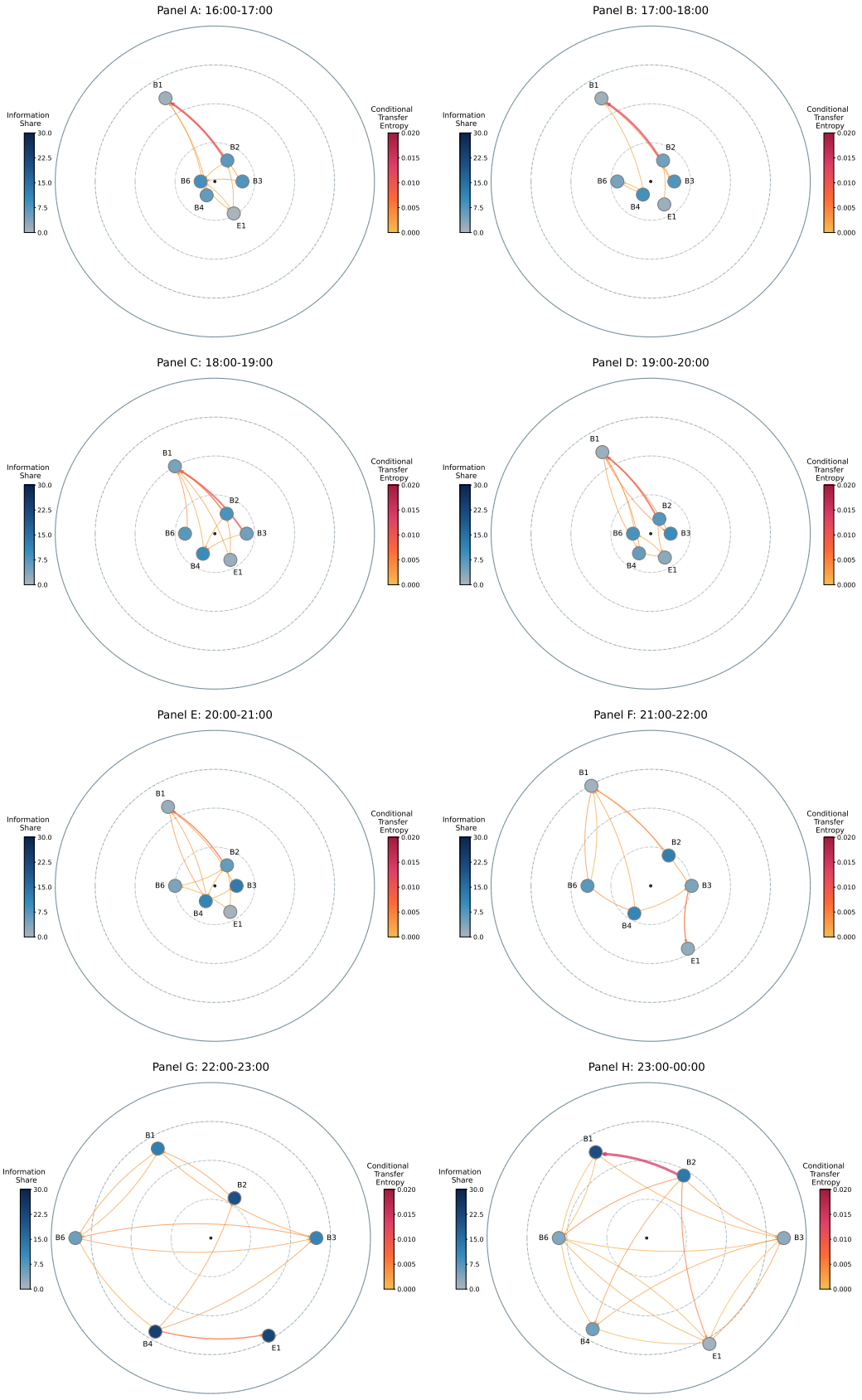

FIGURE 5.12: CTE information maps illustrating changes in the USD/JPY dealer-network between 16:00 and 24:00 on the 2nd January 2019 - the day of flash crash. Each panel is generated from 12 of five-minute-long subsamples.

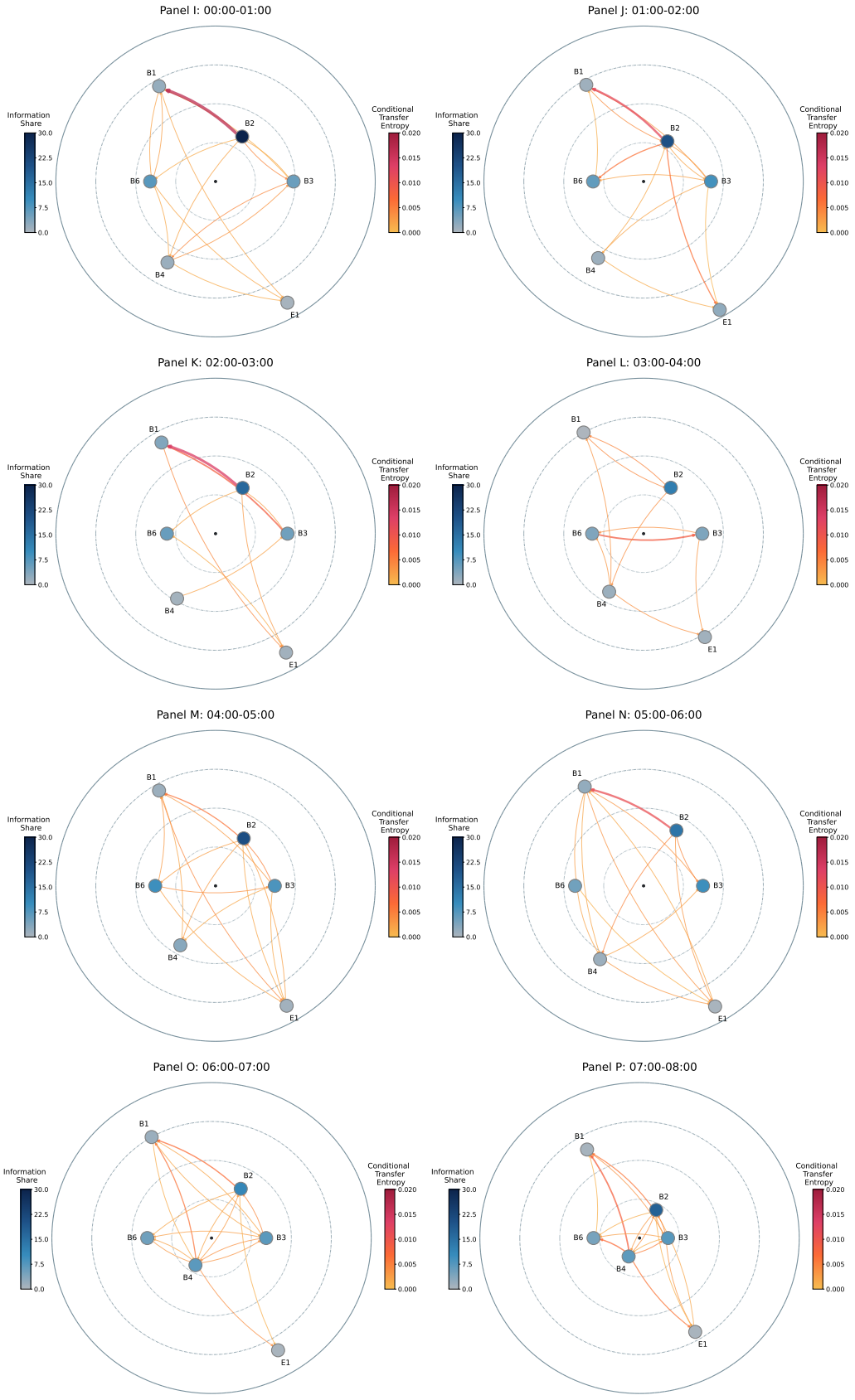

FIGURE 5.13: CTE information maps illustrating changes in the USD/JPY dealer-network between 24:00 and 8:00 on the 3rd January 2019 - the next day after the flash crash. Each panel is generated from 12 of 5-minute-long subsamples.

5.4.4 Quantitative easing

In December 2019, a novel coronavirus infectious disease (COVID-19) was first identified in Wuhan, China (Huang et al. [2020]). On the 21st of January 2020, CDC confirmed the first case of the 2019 novel coronavirus in the United states (Staff). Soon after, the disease rapidly spread throughout the globe, paralyzing the global economy. The spread of the virus led to a shutdown of financial markets, corporate offices, business and schools (Ozili and Arun [2020]). As a result of the uncertainty associated with the pandemic, we observed decreased consumption among consumers, investors and international trade partners (Ozili and Arun [2020]). Furthermore, the pandemic led to an increase in global inflation and unemployment rates (Ozili and Arun [2020]).

In stressful times like this, central banks, such as Federal Reserve or European Central Bank (ECB), may employ monetary strategies to stimulate economic activity and prevent financial crisis. An example of such policy is the so-called quantitative easing (QE), during which a central bank purchases a large number of financial assets, such as corporate or government bonds (Jackson and Curry [2022]). Introduction of the quantitative easing measures increases the amount of money circulating in the economy, which, broadly speaking, results in the reduction of the long-term interest rates (Jackson and Curry [2022]). This further stimulates the economic growth by encouraging lending.

An extension of the ongoing quantitative easing policy was announced by ECB on the 12th March 2020, in an attempt to to further cushion the impact of the COVID-19 pandemic. At 13:30 UTC+1, ECB announced additional measures to support bank liquidity conditions and money market activity (European Central Bank [2020]). From the perspective of the FX market, this announcement has important implications, namely, it sends a powerful message to the market that ECB is acting to stimulate economic growth (Jackson and Curry [2022]).

The ECB announcement is an example of a public news that may have a strong impact on the dynamics of FX market. Therefore, from the perspective of the price discovery framework, it is interesting to investigate whether we can observe any structural changes in the FX dealer-network during and following the QE announcement. To do so, we generate CTE informational maps for each hour of the announcement day. In particular, we will zoom into the period between 8:00 and 20:00 on the 12th of March 2020. Again, as a point of reference for our hourly informational maps we will use the EUR/USD CTE baseline, previously presented in Figure 5.7. For this investigation, we make the following hypothesis.

Hypothesis 7. **Public announcements lead to a decrease in the information exchange between dealers.**

As previously mentioned, Duffie et al. [2009] show that the incentive to seek information reduces after public announcements. This phenomenon is also observed empirically by Hagströmer and Menkveld [2019], who determine that the bilateral connections between dealers are significantly weaker following public announcements. Therefore, we postulate that the ECB announcement resulted in a reduction in the information flows in the dealer-network. Furthermore, we hypothesize that since all dealers are reacting to the same event, we should observe a temporary synchronization of the changes in their quotes. Consequently, the synchronization should reveal itself in a uniform reduction in price inefficiency of all dealers.

For the purpose of this analysis, we focus on the CTE information maps generated for each hour between 8:00 and 20:00 presented in Figures 5.14 and 5.15. In Panels A and B of Figure 5.14, it can be seen that between 8:00 and 10:00 hours the CTE information network closely resembles the baseline structure. For example, the non-bank market maker is also located close to the center of the map, and is a source of many strong information outflows to other dealers. Furthermore, the price inefficiency of all dealers is roughly the same as in the baseline, with the only exception being E1. Electronic trading platform is determined to be significantly more price inefficient than it is found to be on average.

In Panels C and D, the information flows between dealers are much weaker than in the baseline and in the previous hourly snapshots. Additionally, in Panel C, dealers B3, B4 and E1, experience slight decrease in their price inefficiencies. On the other hand, in Panel D non-bank market maker M1 experiences a slight increase in price inefficiency.

Panel E presents the CTE information map about 30 minutes before the ECB announcement. At first glance, we notice the emergence of many strong information flows stemming from M1. Additionally, the information share of M1 appears to be significantly lower than it is in the baseline. And finally, all dealers' price inefficiencies are approximately the same as they appear in the baseline (again apart from the one of E1).

Panel F shows the dealer-network structure between 13:00 and 14:00 UTC+1. The ECB announcement of the monetary policy took place at around 13:30 London local time, hence this map captures the changes that happened during and right after the announcement. In the panel, it can be seen that all dealers except for B5 become much more price efficient. They are now located much closer to the center of the map, as compared to the previous snapshots from the same day and the baseline map. We also observe relatively weak information flows, with strongest ones flowing from M1. However, the information flows are not weaker than in Panel C or D, for instance. Finally, the information share of M1 remains relatively low as compared to the average found in the baseline.

In Panel G, the snapshot approximately one hour following the announcement is presented. In this information map, all dealers appear to be even more price efficient. In particular, the largest increase in price efficiency is observed for dealer B1, B3, B5 and M1. Moreover, the characteristic strong information flows stemming from M1 are present as well. The same is observation can be made in Panel H, where the dealer-network reaches overall the highest price efficiency. In the following two panels (I and J), dealers gradually move away from the center of the map, as they become more price inefficient. In Panel K, the information map very closely aligns with steady state structure observed in the baseline. The only difference being the fact that E1 is more price inefficient than it is determined to be in the baseline. Additionally, the information flows between dealers are much weaker than in the baseline. Panel L further verifies that the network reached its steady-state structure, as very little changes in the dealer-network are observed.

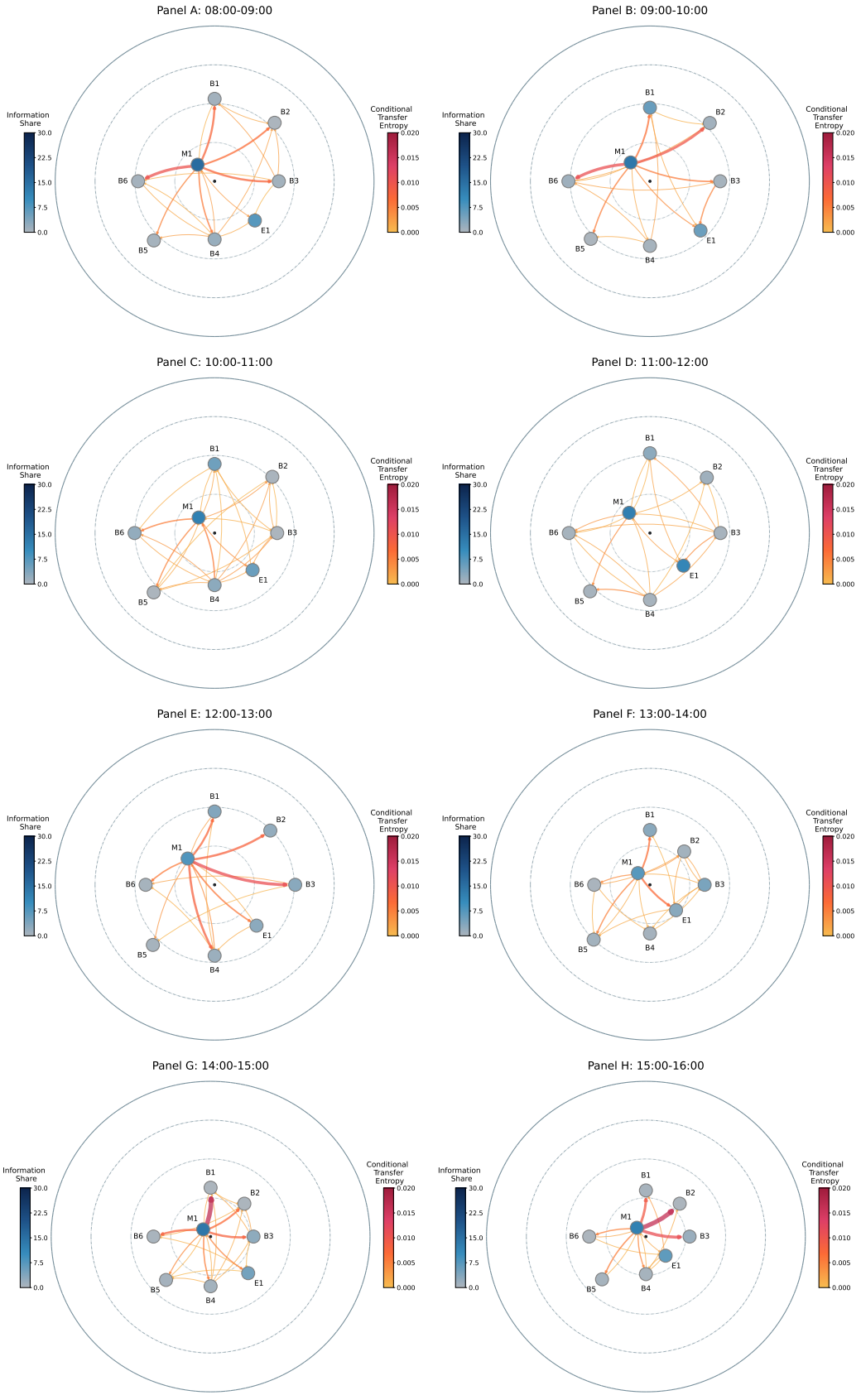

FIGURE 5.14: CTE information maps illustrating changes in the EUR/USD dealer-network between 8:00 and 16:00 on the 12th March 2020 - the day of ECB's quantitative easing announcement. Each panel is generated from 12 of five-minute-long subsamples.

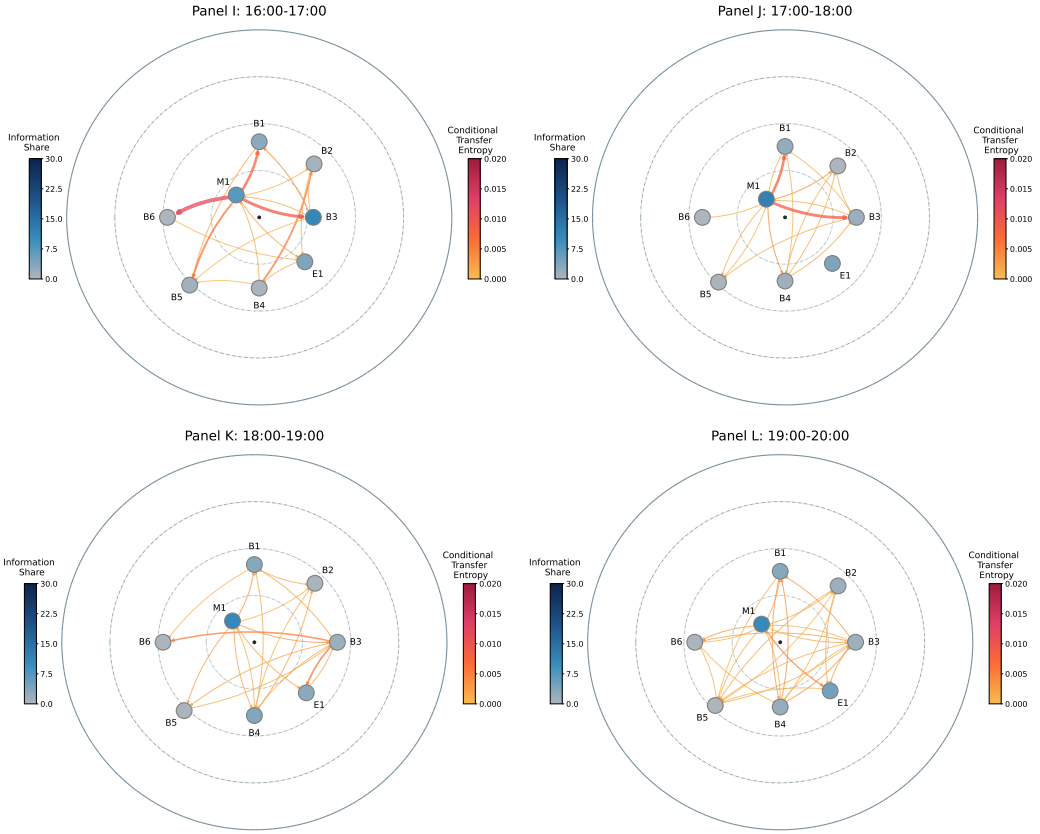

FIGURE 5.15: CTE information maps illustrating changes in the EUR/USD dealer-network between 16:00 and 20:00 on the 12th March 2020 - the day of ECB's quantitative easing announcement. Each panel is generated from 12 of five-minute-long subsamples.

Chapter 6

Discussion

6.1 Results

In Chapter 5, a hypothesis based on the microstructure and information percolation literature was formulated. A number of absorbing and some unexpected observations were made. We will now discuss our findings further, and evaluate the evidence found in support of each hypothesis. Additionally, in this chapter remarks on the econometric model and information-theoretic network inference algorithms are made.

6.1.1 Hypothesis 1 - Central dealers are more informed.

Our first hypothesis was motivated by the work of Babus and Kondor [2018] and the empirical findings of Hagströmer and Menkveld [2019]. Hagströmer and Menkveld [2019] found out that the dealers who are more central, i.e., have more positive partial correlations with other dealers, are better informed. It is further verified whether the evidence supporting this hypothesis can be found in EUR/USD and USD/JPY dealer-networks. Following the steps of Hagströmer and Menkveld [2019] the hypothesis was tested for econometric informativeness metrics: i.e., price inefficiency and Hasbrouck's lower bound information share. Some evidence supporting this hypothesis was found, as we determined the relationship between dealer's centrality and informativeness metrics statistically significant at 5% and 10% levels for USD/JPY baseline. However, no statistically significant relationships between these metrics was uncovered for the EUR/USD baseline. Considering the fact that we validated our model implementation against the original model provided by Hagströmer and Menkveld [2019][1] (see Appendix C) , we are confident that the discrepancies in the results are not due to flawed implementation of the model. Therefore, we cannot convincingly conclude that evidence supporting this hypothesis was found.

[1]The model is originally written in `R` and is available at (Mekveld).

Furthermore, the investigation related to this hypothesis made us realize that from the perspective of microstructure framework, dealers' centrality metric is not fit for measuring the exchange of information between dealers. As previously mentioned, there are two main reasons for that. First, dealer centrality metric is derived from partial correlation matrix, which is a non-directional measure. Second, centrality metric counts the connections to a dealer without accounting for the strength of each connection.

6.1.2 Hypothesis 2 - The price discovery process is dominated by dealers who share their private information the most.

The goal of the second hypothesis was to determine whether TE and CTE information flows inferred by network inference algorithms can explain the variance in Hasbrouck's information share metric. Furthermore, this proposition was extended with a hypothesis that TE and CTE information flows are better explanatory variables for the Hasbrouck's upper and lower bound information share, respectively. Findings presented in Chapter 5 revealed that the TE outflows explain a larger portion of the variance of the upper bound than of the lower bound information share in both EUR/USD and USD/JPY dealer-networks. The relationship between TE outflows and the lower bound information share was determined to be statistically significant for the USD/JPY baseline, but not for the EUR/USD baseline. CTE outflows turned out to provide large explanatory power for lower bound information share in both baselines. And finally, it should be noted that the explanatory power of the CTE outflows was substantially larger than that of the dealer centrality metric proposed by Hagströmer and Menkveld [2019].

An observation that we have not discussed yet is that TE information outflows were also determined to provide a lot of explanatory power for the lower bound of Hasbrouck's information share. While the linear regression fit between these metrics was not particularly good in EUR/USD dealer-network, a significant relationship was found for USD/JPY dealer-network with $\mathcal{R}^2 = 0.95$. If we recall that TE does not filter out potential information contributions from other processes, then the fact that we uncover a significant relationship between these two metrics undermines the notion that lower bound information share is a measure of market's unique information contributions. Since there is no true measure for this phenomenon, we cannot further investigate the plausibility of this statement.

In conclusion, a lot of evidence was found in support of Hypothesis 2. The most information outflows were observed for the dealers with highest contributions to the innovations in the efficient price. Additionally, we demonstrated that to a large extent information flows computed with information theoretic metrics align with the variance-decomposition-based metric proposed by Hasbrouck. Finally, we determined that TE information flows can explain a considerable portion of variance of Hasbrouck's upper bound information share, whereas CTE information flows provide better explanatory power for the lower bound of information share.

6.1.3 Hypothesis 3 - The largest disparities in information contributions are observed between different types of dealers; i.e., banks, non-bank market makers and electronic trading platforms.

Evidence that supports the third hypothesis was also found. For instance, in both EUR/USD and USD/JPY dealer-networks, we found that the fewest information flows out of electronic trading platforms. Additionally, the platform was determined to be the recipient of the most information in EUR/USD dealer-network, and also had the second largest information inflow in the USD/JPY dealer-network. On the other hand, the analysis revealed that the non-bank market is the source of the largest number of information outflows, while receiving the least information inflows. No clear characteristics for banks were determined; some banks provided more information to the dealer-network, while others were predominantly recipients of the information.

In order to draw valid conclusions about non-bank market makers, it would be necessary to explore other dealer-networks with more dealers of this type. Only one non-bank market maker was included in both the dealer-networks investigated, which is not sufficient to draw reliable conclusions. However, in both dealer-networks, the electronic trading platform was predominantly the recipient of the information. Hence, we conclude that only some evidence supporting Hypothesis 3 was found. Further investigation is required.

6.1.4 Hypothesis 4 - The more a given dealer learns from other dealers, the faster is his response to the informational shocks, i.e. the faster is his price discovery process.

In regards to Hypothesis 4, it was determined that dealers with larger total information inflows experience greater corrections in their price inefficiencies between tau = 0 and tau = 1. For both EUR/USD and USD/JPY dealer-networks, we discovered a statistically significant, positive relationship between the information-theoretic inflows and changes in the dealers' price inefficiencies. For EUR/USD dealer-network, a large proportion of the variance of the dependent variable was explained by the inflows. However, a much stronger relationship was observed in the USD/JPY dealer-network. In conclusion, we find strong and statistically significant evidence in support of Hypothesis 4.

6.1.5 Hypothesis 5 - There is more exchange of information between dealers when there is more to be learned.

Evidence supporting this hypothesis was found in the CTE information maps for each one-hour-long time window before, during, and after the USD/JPY flash crash. Interestingly, even before the flash crash, we observed considerable changes in the dealer-network structure, as

some dealers became substantially more price inefficient as compared to the previous one-hour snapshots and the baseline information map. In the information map capturing the time when flash crash took place, we observed a sudden, substantial increase in price inefficiency of all dealers. Additionally, during the flash crash more information flows were present. Therefore, our investigation presents strong evidence supporting the hypothesis that dealers become more inclined to interact during extreme events like flash crashes. On the maps, we observed formation of more information flows between dealers as they all became more price inefficient as a result of an abrupt change in the market.

6.1.6 Hypothesis 6 - During a flash crash, the dealer-network deviates from a steady-state structure and eventually converges back to it.

The analysis of CTE information maps from the day of the USD/JPY flash crash also provided evidence supporting Hypothesis 6. As we observed in Figures 5.12 and 5.13, during and right after the flash crash, the structure of the dealer-network drastically changed. As discussed in Section 6.1.5, all dealers become significantly more price inefficient, and we also observe emergence of new information flows as well as changes in the distribution of information share. Furthermore, we noticed that the distribution of the information share and price inefficiencies, gradually converged to a steady-state structure after about 9 hours after the flash crash.

Based on the presented results we cannot conclude whether the information flows reached their steady-state. First, because, in the last panel investigated we still observe slightly more information flows than before the flash crash. Second, because the baseline presents all possible connections, we cannot determine whether the steady-state information flows are reached. More in-depth investigation of the hourly information flows distributions is needed to draw reliable conclusions. Hence it needs to be concluded that only some evidence in support of the hypothesis was discovered.

6.1.7 Hypothesis 7 - Public announcements lead to a decrease in the information exchange between dealers.

This last hypothesis was explored by observing the structural changes in the dealer-network on the day of ECB's quantitative easing announcement. The investigation did not reveal any evidence supporting the idea that the exchange of private information between dealers is reduced at the times of public announcement. However, an exciting observation was made instead. Namely, after the ECB's announcement, all dealers became more price efficient. Additionally, we observed that dealer-network converged back to its steady-state structure about five hours after the announcement.

6.2 Methodology

In the following subsection, we make a few additional remarks on the econometric model and information-theoretic inference algorithm. While many observations were made during the development we have chosen to further discuss only the most important ones.

6.2.1 Econometric model

- The econometric model is susceptible to numerical errors. This is because the method of computing partial correlations of returns as proposed by Hagströmer and Menkveld [2019] involves the inversion of the matrix. In the case a covariance matrix of cumulative responses of dealers has a large condition number, we are exposed to instability which may lead to major numerical errors. This particular problem is observed primarily for large τ, at which the cumulative responses of all dealers become nearly identical. The problem is also presented in Figure C.3, in which larger mean relative errors for higher τ's can be seen.

- The assumption that markets' quotes are cointegrated and integrated of order one is valid to a large extent. However, markets tend not to be cointegrated during market stress times. This observation was made during cointegration testing for USD/JPY data set. Therefore, the econometric model may not be particularly fit for the investigation of the dealer-network dynamics during market stress times.

- The method of computing the partial correlation matrix and price inefficiencies heavily depends on the outcome of the VECM fitting. Consequently, the VECM lag selection is critical. Incorrect selection of the VECM lags results in the impulse response function not converging. The convergence of the impulse response function is crucial because it is used to determine the vector of the long-term response of variables to a shock. This vector is further used to compute Hasbrouck's information share and price inefficiency. Thus, much attention should be paid to proper VECM lag selection.

6.2.2 Information-theoretic model

- The information-theoretic network inference algorithm is computationally very demanding, which is the method's main disadvantage. However, as mentioned before, this limitation can be resolved by extending the algorithm implementation with non-uniform embedding and employing GPUs to speed up the permutation testing.

- Kraskov algorithm estimation is not fit for sparse data. This observation was made during the development of the algorithm and initial testing on the EUR/CHF data set. The

EUR/CHF data set was very sparse during the flash crash. Since it is recommended to add low-level noise to the time series when the Kraskov algorithm is employed, the transfer entropies calculated between two sparse processes turned out to be strongly impacted by the stochastic factor, and hence not reproducible. While this issue was not observed in EUR/USD or USD/JPY data sets, it constitutes a significant limitation of the method, which ought to be kept in mind.

- Even though we use conditional transfer entropy to measure the transfer of unique information, it should not be treated as such. As previously established, while CTE reduces the transfer of redundant information considering information from other dealers in the system, it also captures synergistic contributions from multisource interactions. Our analysis determined that overall, more redundant information was filtered out than synergistic information was created. Nevertheless, the results obtained with conditional transfer entropy should be treated with caution. This limitation of the CTE metric is widely recognized by the research community (James et al. [2016]), and many efforts have gone into developing a metric that would be able to quantify unique information contributions. For example, a recent alternative measure for information flows is proposed by Oizumi et al. [2016].

6.2.3 Other remarks

- The last remark pertains to the use of best bid and ask prices for the price discovery process (suggested by Hagströmer and Menkveld [2019]). As previously explained, dealers usually provide multiple ask and bid quotes for different volumes at the same point in time. The price points are different for different volumes of particular financial instruments. The best bid and ask prices are the ones that represent the narrowest spread, or in other words, the most competitive ask and bid prices. However, by using the best bid and ask price, we ignore the disparities between the prices associated with different volumes being priced by dealers. Take the situation where one dealer provides a quote for a standard lot (100,000 units of currency), whereas the other one prices a mini lot (10,000 units of currency). In this case, the difference in their quotes may be purely due to the difference in the quoted volumes, as usually a larger volume is offered at a better price. In the current analysis, this factor is not considered. Therefore, the weighted average price (WAP) should be used instead to put all quotes on the same foot.

Chapter 7

Conclusion and future work

7.1 Summary

In this thesis, we have successfully met all our goals. In Chapters 1 and 2 the theoretical framework for the microstructure approach to exchange rates as well as the notion of price discovery were thoroughly discussed. We also reviewed the fundamental econometric techniques for quantifying dealers' contribution to the price discovery process. In particular, we have addressed in detail the information share metric proposed by Hasbrouck [1995], and the notion of the common trend component established by Stock and Watson [1988]. Finally, in Chapter 2 we also discussed recent developments in the information theory. In particular, we elaborated on the growing interest in applying transfer entropy metric to different scientific settings and introduced recent developments in information-theoretic network inference algorithms.

Chapter 3 provided a mathematical formulation behind the econometric model proposed by Hagströmer and Menkveld [2019]. All the econometric metrics and methods were extensively treated and interpreted. Additionally, an intuitive example of the econometric model analysis was presented. Chapter 3 also lays the foundation for the information-theoretic network inference algorithm. In this chapter, we introduce and explain all fundamental information-theoretic metrics. The chapter is concluded with a detailed treatment of the transfer entropy and conditional transfer entropy metrics. Potential pitfalls, particularly the notion of synergy associated with information-theoretic metrics, are also addressed. The main goal of this chapter was achieved as we presented similarities and differences between both methods and demonstrated the adequacy of both methods for the FX market setting.

In Chapter 4, we presented all the data processing and parameter selection methods employed in the econometric model and the information-theoretic network inference algorithm. Most importantly, we discussed the concept of time stamp synchronization and addressed this issue in our data. Additionally, the Kraskov algorithm used to estimate the information-theoretic

metrics was extensively discussed. Lastly, the notion of state-space reconstruction, delay reconstruction, and the permutation testing scheme were explained.

The analysis results for both econometric and information-theoretic models were presented in Chapter 5, and the their implementations were validated in Appendix C. Our investigation revealed that the dealer centrality metric is inconsistent with Hasbrouck's information share or price inefficiency metric. The analysis of the results obtained with the TE and CTE information maps provided evidence supporting many of the investigated hypotheses. A particularly important finding is that the information flows inferred with the information-theoretic network inference are, to a large extent, consistent with Hasbrouck's information share, as well as the prince inefficiency metric proposed by Hagströmer and Menkveld [2019]. While some relationships between information flows and econometric metrics were weaker for the EUR/USD data set, all of the relationships discovered in the USD/JPY dealer-network were significant. Consequently, these findings provide evidence that TE and CTE information flows offer new insight into the information flows within the dealer-network, consistent with the fundamental and novel econometric approaches. Interestingly, we also observed that both methods consistently suggest that non-bank market makers dominate the price discovery process. The information-theoretic maps also captured the changes in the dealer-network structure during the USD/JPY flash crash and the ECB's quantitative easing announcement.

7.2 Contributions

In this thesis, the following contributions were made:

- This thesis constitutes, to the best of our knowledge, the first analysis of the FX dealer-network with an information-theoretic network inference algorithm. Moreover, we also present the results obtained with the Hagströmer and Menkveld [2019]'s econometric model for the same data sets. Thus, not only do we offer the results obtained with each model separately, but we also provide an extensive comparison of both methods. Finally, both econometric and information-theoretic analyses are performed on unique, very large data sets, with over one-month worth of high frequency data.

- Complete implementation of the econometric model by Hagströmer and Menkveld [2019] in `C++`, made available for ING as both `C++` and `Python` libraries.

- The first implementation of an information-theoretic network inference algorithm in `C++` also available as `Python` library. The algorithm is very versatile, as it has many parameters that can be adjusted per user needs. Among many functionalities, the user can, for example, define the maximal dimensions for automatic optimal embedding dimensions

detection or use pre-defined constant embedding dimensions instead. Moreover, the algorithm is faster than other openly available implementations. Hence, other users may find this implementation useful, as it is a powerful and efficient tool to infer networks from extensive data sets.

7.3 Future work

While many directions for future work were mentioned throughout this thesis, here we compile the list of the most important ones.

- The network inference algorithm could be extended with an option to use the non-uniform embedding. This would not only bridge the gap between the algorithm developed for this thesis and novel state-of-art network inference algorithm, but also significantly improve the algorithm's performance.
- Next, the network inference algorithm could be extended with an option to utilize GPUs for permutation testing. This change would further improve the performance of the algorithm.
- Additionally, a 'smarter' method for permutation testing could be considered. Random permutations is not the most efficient method. 'Slight' permutations could be considered instead, where only a tiny portion of the time series is permuted. Completely random permutations are more likely to produce a time series for which no information flow can be detected. Whereas, slight permutations should provide more robust p-values of the estimates as a function of the number of permutations performed.
- Given that the above extension is implemented, the analysis could be performed on the data resampled to a much lower sampling interval, ideally at the clock's precision. Then we could fully extract all the information ingrained in the data and map out dealer interaction on the most granular level. This would undoubtedly provide a different perspective and uncover more information flows that cannot be observed when data is resampled.
- The information-theoretic network inference algorithm could be further validated with synthetic data used by Novelli et al. [2019].
- As already discussed in Chapter 6, weighted average prices could be used for the analysis instead of the best bid and ask prices.
- After all the steps above are completed, the next step would be a comparison of the information-theoretic network inference approach with recently developed alternative methods for the inference of information flows. For example, it would be interesting to explore

the partial information decomposition approach proposed by Williams and Beer [2010] or a unified framework for information integration based on information geometry recently introduced by Oizumi et al. [2016].

We believe that these steps would add further value to this area of research.

Appendix A

Example for information revelation method

In Section 3.1.2 (Vector-Error Correction Model), we have already introduced an intuition behind the dynamics of price quotes of two markets that price the same currency pair. We also established that the changes in the price quotes of markets would depend on the deviations from each other quotes in the previous time step $t-1$, and we have shown that their price changes can be modelled using Equations (3.12) and (3.13). Now, let us assume that fitting the data into our VECM model yielded the following parameters

$$\boldsymbol{\alpha} = \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} 1 & -1 \end{bmatrix}, \boldsymbol{\gamma}_i = \mathbf{0} \quad \forall\, i \tag{A.1}$$

Using Equation (3.16) we can present the relationship between prices of the two markets with the following model

$$\begin{bmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \tag{A.2}$$

where $\epsilon_{1,t}$, $\epsilon_{2,t}$ are residuals of VECM model associated with market 1 and market 2, respectively. In a more explicit form the changes in the prices can be modelled by the following system of equations,

$$\Delta y_{1,t} = -0.5\,(y_{1,t-1} - y_{2,t-1}) + \epsilon_{1,t} \tag{A.3}$$

$$\Delta y_{2,t} = \ \ 0.25\,(y_{1,t-1} - y_{2,t-1}) + \epsilon_{2,t} \tag{A.4}$$

In order to keep this example as simple as possible, let us further assume that the covariance matrix of the residuals identified by the VECM model is and identity matrix,

$$\boldsymbol{\Omega} = \boldsymbol{I} \tag{A.5}$$

The VECM fitting allowed us to recover the inherent cointegration relation between both markets Equations (A.3) and (A.4) characteristic for the particular time window. Having the aforementioned equations we are now able to perform an impulse response analysis.

The first step of the IRF analysis is to investigate and quantify both markets short-term and long-term responses to unit shocks. Let us begin with investigating the influence of a unit shock introduced to the first market (M1). Since we are interested in the cumulative price

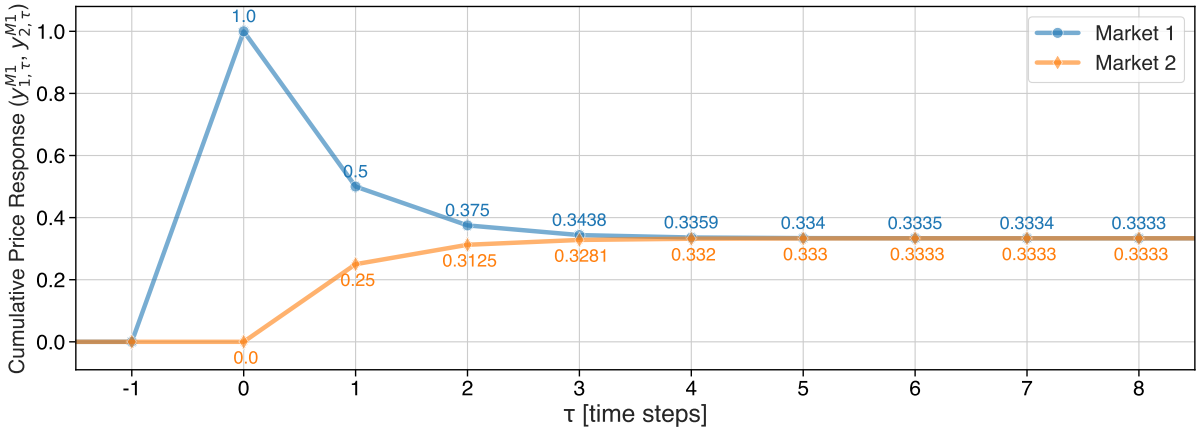

FIGURE A.1: Impulse response of markets to a unit shock in market 1 (M1). Figure illustrates how the cumulative price changes diverge for a short time as the shock is introduced and converge in a long run.

response, in other words the cumulative change in the price, the initial efficient price can be set to any number. Therefore, let us assume that both markets prices are initially at equilibrium $y^{M1}_{1,\tau} = y^{M1}_{2,\tau} = 0$ for $\tau < 0$. The superscript M1 signifies that we are investigating response of a system to a unit shock to market 1 (M1), i.e. $\boldsymbol{\epsilon}_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Let us iterate the system dynamics using Equation (A.2). We start with $\tau = 0$, at which the shock is impounded into price of market 1,

$$\begin{bmatrix} \Delta y^{M1}_{1,0} \\ \Delta y^{M1}_{2,0} \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix} \begin{bmatrix} 0 - 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{A.6}$$

As a result of this price change, the new prices on both markets are

$$\begin{bmatrix} y_{1,0}^{M1} \\ y_{2,0}^{M1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{A.7}$$

To determine the price change at $\tau = 1$, we need to account for the cumulative price changes on both markets up to $\tau = 0$ - as determined in Equation (A.7). Thus, the next price change in both markets is,

$$\begin{bmatrix} \Delta y_{1,1}^{M1} \\ \Delta y_{2,1}^{M1} \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix} \begin{bmatrix} 1 - 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix} \tag{A.8}$$

And the new cumulative price changes in both markets are

$$\begin{bmatrix} y_{1,1}^{M1} \\ y_{2,1}^{M1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix} \tag{A.9}$$

In this iterative manner, one will find that at $\tau = 15$,

$$\begin{bmatrix} \Delta y_{1,15}^{M1} \\ \Delta y_{2,15}^{M1} \end{bmatrix} \approx \begin{bmatrix} -1.863 \cdot 10^{-9} \\ 9.313 \cdot 10^{-10} \end{bmatrix} \tag{A.10}$$

And the cumulative price changes in both markets are

$$\begin{bmatrix} y_{1,15}^{M1} \\ y_{2,15}^{M1} \end{bmatrix} \approx \begin{bmatrix} 0.3333 \\ 0.3333 \end{bmatrix} \tag{A.11}$$

Hence, clearly the system reached a new equilibrium as $y_{1,15}^{M1} = y_{2,15}^{M1}$. The results of the IRF iteration are recorded in Table A.1.

Next, it is necessary to perform the same impulse response analysis for when a unit shock is introduced to market 2 (M2), i.e. $\boldsymbol{\epsilon}_t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. This is not going to be derived here step by step, instead the impulse response of the system is visualized on Figure A.2 and the results of the IRF analysis are summarized in Table A.2.

From the impulse response analysis and Tables A.1 and A.2 it is clear that the long-term cumulative responses of markets to unit shocks are

$$\lim_{\tau \to \infty} \tilde{\boldsymbol{\Psi}}_\tau = \begin{bmatrix} 0.333 & 0.667 \\ 0.333 & 0.667 \end{bmatrix} \Rightarrow \boldsymbol{\psi} = \begin{bmatrix} 0.333 & 0.667 \end{bmatrix} \tag{A.12}$$

Impulse Response Analysis - Shock M1					
Step	**Price Changes**		**Total Price Changes**		**Abs. Difference**
τ	$\Delta y_{1,\tau}^{M1}$	$\Delta y_{2,\tau}^{M1}$	$y_{1,\tau}^{M1}$	$y_{2,\tau}^{M1}$	$\mid y_{1,\tau}^{M1} - y_{2,\tau}^{M1} \mid$
-1	0	0	0	0	0
0	1	0	1	0	1
1	-0.5	0.25	0.5	0.25	0.25
2	-0.125	0.0625	0.375	0.3125	0.0625
3	-0.03125	0.01562	0.3438	0.3281	0.01562
4	-0.007812	0.003906	0.3359	0.332	0.003906
5	-0.001953	0.0009766	0.334	0.333	0.0009766
6	-0.0004883	0.0002441	0.3335	0.3333	0.0002441
7	-0.0001221	6.104e-05	0.3334	0.3333	6.104e-05
8	-3.052e-05	1.526e-05	0.3333	0.3333	1.526e-05

TABLE A.1: Table summarizing the results of the impulse response of the system to a unit shock to market 1 (M1). The table reveals how the absolute difference between cumulative price changes of both markets decrease as time passes. The results of the impulse response analysis align with the impulse response presented on Figure A.1.

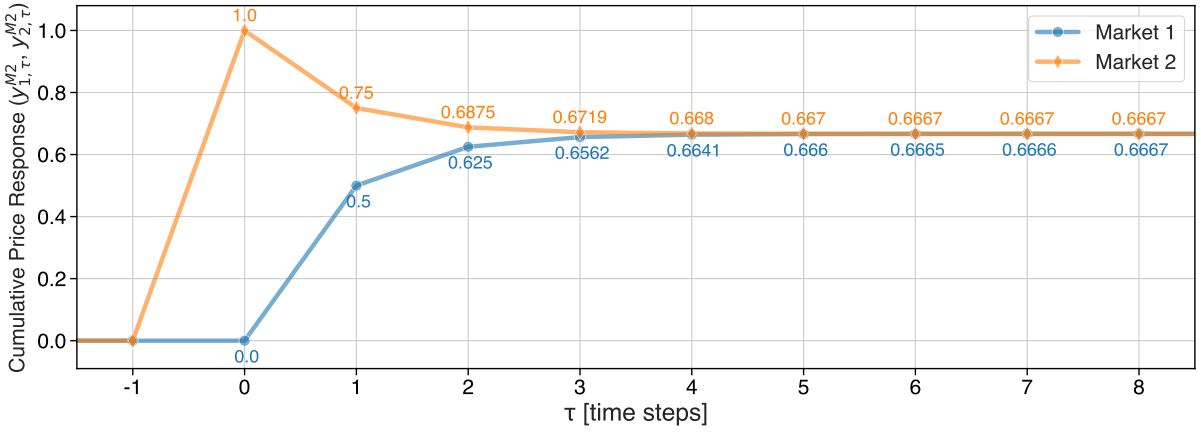

FIGURE A.2: Impulse response of markets to a unit shock in market 2 (M2). Figure illustrates how the cumulative price changes diverge for a short time as the shock is introduces and converge in a long run.

Furthermore, we can now use the data collected in the IRF analysis to construct $\tilde{\mathbf{\Psi}}_\tau$ for any τ by simply choosing the right terms, as follows

$$\tilde{\mathbf{\Psi}}_\tau = \begin{bmatrix} y_{1,\tau}^{M1} & y_{1,\tau}^{M2} \\ y_{2,\tau}^{M1} & y_{2,\tau}^{M2} \end{bmatrix} \tag{A.13}$$

For example,

$$\tilde{\boldsymbol{\Psi}}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{\boldsymbol{\Psi}}_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}, \quad \tilde{\boldsymbol{\Psi}}_2 = \begin{bmatrix} 0.375 & 0.625 \\ 0.312 & 0.688 \end{bmatrix} \tag{A.14}$$

Having a recipe to construct matrices of short-term and long-term cumulative responses of markets, we can now calculate price inefficiencies and the strength of bilateral connections between markets for each τ. To illustrate, first we calculate the beta coefficient using Equation (3.75) for both markets at e.g. $\tau = 1$:

$$\beta_{1,1} = \frac{\tilde{\boldsymbol{\Psi}}_{1,0}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal}{\boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal} = \frac{\begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.667 \end{bmatrix}}{\begin{bmatrix} 0.333 & 0.667 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.667 \end{bmatrix}} = 0.90 \tag{A.15}$$

$$\beta_{2,1} = \frac{\tilde{\boldsymbol{\Psi}}_{2,0}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal}{\boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal} = \frac{\begin{bmatrix} 0.25 & 0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.667 \end{bmatrix}}{\begin{bmatrix} 0.333 & 0.667 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.667 \end{bmatrix}} = 1.05 \tag{A.16}$$

Impulse Response Analysis - Shock M2					
Step	**Price Changes**		**Total Price Changes**		**Abs. Difference**
τ	$\Delta y_{1,\tau}^{M1}$	$\Delta y_{2,\tau}^{M1}$	$y_{1,\tau}^{M1}$	$y_{2,\tau}^{M1}$	$\mid y_{1,\tau}^{M1} - y_{2,\tau}^{M1} \mid$
-1	0	0	0	0	0
0	0	1	0	1	1
1	0.5	-0.25	0.5	0.75	0.25
2	0.125	-0.0625	0.625	0.6875	0.0625
3	0.03125	-0.01562	0.6562	0.6719	0.01562
4	0.007812	-0.003906	0.6641	0.668	0.003906
5	0.001953	-0.0009766	0.666	0.667	0.0009766
6	0.0004883	-0.0002441	0.6665	0.6667	0.0002441
7	0.0001221	-6.104e-05	0.6666	0.6667	6.104e-05
8	3.052e-05	-1.526e-05	0.6667	0.6667	1.526e-05

TABLE A.2: Table summarizing the results of the impulse response of the system to a unit shock to market 2 (M2). The table reveals how the absolute difference between cumulative price changes of both markets decrease as time passes. The results of the impulse response analysis align with the impulse response presented on Figure A.2.

Having the beta coefficient, the price inefficiency of each market at $\tau = 1$ are calculated using Equation (3.80), hence

$$PI_{1,1} = \mid 1 - \beta_{1,1} \mid = 0.10 \tag{A.17}$$

$$PI_{2,1} = \mid 1 - \beta_{2,1} \mid = 0.05 \tag{A.18}$$

Next, the covariance matrix of cumulative responses until time $\tau = 1$ is determined using Equation (3.87)

$$\boldsymbol{\Sigma}_1 = \tilde{\boldsymbol{\Psi}}_1 \boldsymbol{\Omega} \tilde{\boldsymbol{\Psi}}_1^{\intercal} = \begin{bmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.5 & 0.25 \\ 0.5 & 0.75 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.625 \end{bmatrix} \tag{A.19}$$

The inverse of the covariance matrix of cumulative responses at $\tau = 1$ is

$$\boldsymbol{\Sigma}_1^{-1} = \begin{bmatrix} 10 & -8 \\ -8 & 8 \end{bmatrix} \tag{A.20}$$

And finally, the partial correlation matrix associated with the covariance matrix of cumulative responses of the markets is determined using Equation (3.88),

$$\boldsymbol{R}_1 = \boldsymbol{D}_{\boldsymbol{\Sigma}_1^{-1}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{D}_{\boldsymbol{\Sigma}_1^{-1}}^{-\frac{1}{2}} \circ \boldsymbol{K} \tag{A.21}$$

$$= \begin{bmatrix} 10 & 0 \\ 0 & 8 \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} 10 & -8 \\ -8 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 10 & 0 \\ 0 & 8 \end{bmatrix}^{-\frac{1}{2}} \circ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.8944 \\ 0.8944 & 1 \end{bmatrix} \tag{A.22}$$

Note that since $\boldsymbol{R}_\tau$ is a partial correlation matrix, $\rho_{i,j,\tau} = \rho_{j,i,\tau}$ as expected. Finally, it is

time to determine the information share of each market. Recall, that in this section we assumed that the covariance matrix of the residuals is an identity matrix. Furthermore, as discussed in Section 3.1.6.1, since the innovations of our markets are uncorrelated, the lower and the upper bound of information share are equivalent. Also, the lower triangular of Cholesky's decomposition of identity matrix is an identity matrix. Therefore, we can simply compute the

Step	**Total Price Changes**				**Beta Coeff.**		**Price Ineff.**		**PCorr**
τ	$y_{1,\tau}^{M1}$	$y_{2,\tau}^{M1}$	$y_{1,\tau}^{M2}$	$y_{2,\tau}^{M2}$	$\beta_{1,\tau}$	$\beta_{2,\tau}$	$PI_{1,\tau}$	$PI_{2,\tau}$	$\rho_{1,2,\tau}$
-1	0	0	0	0	-	-	-	-	-
0	1	0	0	1	0.6	1.2	0.4	0.2	0
1	0.5	0.25	0.5	0.75	0.9	1.05	0.1	0.05	0.8944
2	0.375	0.3125	0.625	0.6875	0.975	1.012	0.025	0.0125	0.9935
3	0.3438	0.3281	0.6562	0.6719	0.9937	1.003	0.00625	0.003125	0.9996
4	0.3359	0.332	0.6641	0.668	0.9984	1.001	0.001563	0.0007812	1
5	0.334	0.333	0.666	0.667	0.9996	1	0.0003906	0.0001953	1
6	0.3335	0.3333	0.6665	0.6667	0.9999	1	9.766e-05	4.883e-05	1
7	0.3334	0.3333	0.6666	0.6667	1	1	2.441e-05	1.221e-05	1
8	0.3333	0.3333	0.6667	0.6667	1	1	6.104e-06	3.052e-06	1

TABLE A.3: Table summarizing the impulse response analysis of both markets. The row highlighted in grey represents the value calculated explicitly in this section.

information share for both markets using Equation (3.65), as follows

$$\text{InfoShare}_i = \frac{([\boldsymbol{\psi} \boldsymbol{F}_x])^2}{\boldsymbol{\psi}\boldsymbol{\Omega}\boldsymbol{\psi}^\intercal} = \frac{\left(\left[\begin{bmatrix}0.333 & 0.667\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right]\right)^2}{\begin{bmatrix}0.333 & 0.667\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}0.333\\ 0.667\end{bmatrix}} = \tag{A.23}$$

$$= \frac{\begin{bmatrix}0.111 & 0.444\end{bmatrix}}{0.556} = \begin{bmatrix}0.2 & 0.8\end{bmatrix} \tag{A.24}$$

Thus, the information share of the first market is $\text{InfoShare}_{M1} = 0.2$, while the information share of the second market is $\text{InfoShare}_{M2} = 0.8$. Considering the fact that $\psi_{M1} = 0.333$ and $\psi_{M2} = 0.667$, and the fact that we assumed that correlation matrix of innovations in the markets Ω is an identity matrix, it becomes clear that market 2 has a much larger contribution to the innovations in the efficient price.

Appendix B

Proofs of information-theoretic theorems and lemmas

B.1 Proof of Lemma 3.13 - Maximum of Shannon entropy

Proof. In order to prove this lemma, let's utilize the method of Lagrange multipliers in order to find the maximum.

$$\mathcal{L} = \left\{ -\sum_{x \in X} p(x) \log\left(p(x)\right) - \lambda \left(\sum_{x \in X} p(x) - 1 \right) \right\} \tag{B.1}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\sum_{x \in X} p(x) + 1 = 0 \tag{B.2}$$

$$\implies \sum_{x \in X} p(x) = 1 \tag{B.3}$$

$$\frac{\partial \mathcal{L}}{\partial p(x)} = -\log\left(p(x)\right) - 1 - \lambda = 0 \tag{B.4}$$

$$\implies p(x) = e^{-(\lambda+1)} \tag{B.5}$$

From eq. (B.5) it is clear that $p(x)$ that maximizes the Shannon entropy measure is independent of x. If we further assume that cardinality of X is $n = |\mathcal{X}|$:

$$\sum_{x=1}^{n} p(x) = 1 \tag{B.6}$$

$$\implies np(x) = 1 \tag{B.7}$$

$$\implies p(x) = \frac{1}{n} \tag{B.8}$$

Now, given that $p(x) = \frac{1}{n}$ maximizes the Shannon entropy, let's determine the maximum Shannon entropy. Using eq. (3.90):

$$H(X)_{\max} = -\sum_{x=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) = \log(n) \tag{B.9}$$

□

B.2 Proof of Lemma 3.23 - Additivity

Proof. If $X \perp\!\!\!\perp Y$:

$$p(x,y) = p(x)p(y) \implies \log\left(p(x,y)\right) = \log\left(p(x)p(y)\right) \tag{B.10}$$

$$\implies \log\left(p(x,y)\right) = \log\left(p(x)\right) + \log\left(p(y)\right) \tag{B.11}$$

$$\implies \log\left(\frac{1}{p(x,y)}\right) = \log\left(\frac{1}{p(x)}\right) + \log\left(\frac{1}{p(y)}\right) \tag{B.12}$$

$$\implies \mathbb{E}\left[\log\left(\frac{1}{p(x,y)}\right)\right] = \mathbb{E}\left[\log\left(\frac{1}{p(x)}\right)\right] + \mathbb{E}\left[\log\left(\frac{1}{p(y)}\right)\right] \tag{B.13}$$

$$\implies H(X,Y) = H(X) + H(Y) \tag{B.14}$$

□

B.3 Proof of Lemma 3.16 - Sub-additivity

Proof. From theorem 3.18, we know that $H(X,Y) = H(X) + H(Y|X)$, hence :

$$H(X,Y) \leq H(X) + H(Y) \tag{B.15}$$

$$\implies H(X) + H(Y|X) \leq H(X) + H(Y) \tag{B.16}$$

$$\implies H(Y|X) \leq H(Y) \tag{B.17}$$

which is always true given definitions 3.7 and 3.17. □

B.4 Proof of Theorem 3.18 - Chain rule

Proof.

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x,y)) \tag{B.18}$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(p(x)p(y|x)\right) \tag{B.19}$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(p(x)\right) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(p(y|x)\right) \tag{B.20}$$

$$= -\sum_{x\in X} p(x)\log\left(p(x)\right) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(p(y|x)\right) \tag{B.21}$$

$$= H(X) + H(Y|X) \tag{B.22}$$

□

B.5 Proof of Corollary 3.19 - Chain rule for N random variables.

Proof. Let's rewrite $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)$

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n)\log\left(p(x_1, x_2, \ldots, x_n)\right) \tag{B.23}$$

$$= -\sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n)\log\left(\prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)\right) \tag{B.24}$$

$$= -\sum_{x_1,x_2,\ldots,x_n}\sum_{i=1}^{n} p(x_1, x_2, \ldots, x_n)\log\left(p(x_i|x_{i-1}, \ldots, x_1)\right) \tag{B.25}$$

$$= -\sum_{i=1}^{n}\sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n)\log\left(p(x_i|x_{i-1}, \ldots, x_1)\right) \tag{B.26}$$

$$= -\sum_{i=1}^{n}\sum_{x_1,x_2,\ldots,x_i} p(x_1, x_2, \ldots, x_i)\log\left(p(x_i|x_{i-1}, \ldots, x_1)\right) \tag{B.27}$$

$$= -\sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{B.28}$$

The proof follows the steps presented in Thomas and Joy [2006]. □

B.6 Proof of Theorem 3.22 - Information inequality

Proof. Let $X_{\mathcal{S}} = \{x \in X : p(x) > 0\}$ be the support set of $p(x)$.

$$-D(p||q) = -\sum_{x \in X_{\mathcal{S}}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{B.29}$$

$$= \sum_{x \in X_{\mathcal{S}}} p(x) \log\left(\frac{q(x)}{p(x)}\right) \tag{B.30}$$

$$\leq \log\left(\sum_{x \in X_{\mathcal{S}}} p(x)\frac{q(x)}{p(x)}\right) \tag{B.31}$$

$$= \log\left(\sum_{x \in X_{\mathcal{S}}} q(x)\right) \tag{B.32}$$

$$\leq \log\left(\sum_{x \in X} q(x)\right) \tag{B.33}$$

$$= \log(1) \tag{B.34}$$

$$= 0 \tag{B.35}$$

The transition from eq. (B.30) to eq. (B.31) is supported by Jensen's inequality, which establishes that for a convex function f and random variable X the following relation holds $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. Analogously, it states that for a concave function the opposite relation holds, i.e. $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$. The proof follows the steps presented in Thomas and Joy [2006]. □

B.7 Proof of Lemma 3.25 - Mutual information to entropy

Proof. Using definition 3.20,

$$I(X;Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{B.36}$$

$$= \sum_{x,y} p(x,y) \log\left(\frac{p(x|y)}{p(x)}\right) \tag{B.37}$$

$$= -\sum_{x,y} p(x,y) \log(p(x)) + \sum_{x,y} p(x,y) \log(p(x|y)) \tag{B.38}$$

$$= -\sum_{x,y} p(x,y) \log(p(x)) - \left(-\sum_{x,y} p(x,y) \log(p(x|y))\right) \tag{B.39}$$

$$= H(X) - H(X|Y) \tag{B.40}$$

The proof follows the steps presented in Thomas and Joy [2006]. □

B.8 Proof of Theorem 3.30 - Mutual information to entropy

Proof. Using definition 3.20,

$$I(X_1, X_2, \ldots, X_n; Y) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) \tag{B.41}$$

$$= \sum_{i=1}^{n} H\left(X_i|X_{i-1}, X_{i-2}, \ldots, X_1\right) - \sum_{i=1}^{n} H\left(X_i|X_{i-1}, X_{i-2}, \ldots, X_1, Y\right) \tag{B.42}$$

$$= \sum_{i=1}^{n} \left[H\left(X_i|X_{i-1}, X_{i-2}, \ldots, X_1\right) - H\left(X_i|X_{i-1}, X_{i-2}, \ldots, X_1, Y\right)\right] \tag{B.43}$$

$$= \sum_{i=1}^{n} I\left(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1\right) \tag{B.44}$$

□

B.9 Proof of Theorem 3.46 - Differential entropy vs discrete entropy

Proof. Let's assume that we divide the space of X into equisized bins of length Δ, thus X^{Δ} is simply a discretized analog of continous random variable X. If we further assume that the density is continous throughout the bins, then by mean value theorem, there must exist x_i such that (Thomas and Joy [2006], Michalowicz et al. [2013])

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx \tag{B.45}$$

Thus, X^{Δ} is simply defined as

$$X^{\Delta} = x_i \quad \text{if} \quad i\Delta \leq X < (i+1)\Delta \tag{B.46}$$

Consequently, the probability mass function $p(x_i)$ associated with x_i is

$$p(x_i) = f(x_i)\Delta \tag{B.47}$$

The entropy of the discretized continous random variable X can be calculated in the followin manner

$$H(X^{\Delta}) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{B.48}$$

$$= -\sum_{i=1}^{n} f(x_i)\Delta \log(f(x_i)\Delta) \tag{B.49}$$

$$= -\sum_{i=1}^{n} f(x_i)\Delta \left[\log(f(x_i)) + \log(\Delta)\right] \tag{B.50}$$

since $\sum_{i=1}^{n} f(x_i)\Delta = \int f(x) = 1$, the equation simplifies to

$$= -\sum_{i=1}^{n} f(x_i)\Delta \log(f(x_i)) - \log(\Delta) \tag{B.51}$$

Clearly, if we compare the above formula to eq. (3.90), it is apparent that $-\log(\Delta)$ is an extra term. Moreover, as $\Delta \to 0 \implies \log(\Delta) \to \infty$, hence

$$\lim_{\Delta \to 0} h(X) - h(X^{\Delta}) = \log(\Delta) \tag{B.52}$$

□

B.10 Proof of Theorem 3.40 - Transfer entropy is conditional time-delayed mutual information

Proof. From theorem 3.37 we know that transfer entropy can be expressed as Kullback-Leibler divergence, as follows:

$$\mathrm{TE}_{X_t \to Y_t}^{(d_y,d_x)} = \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log \left(\frac{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})}{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})} \right) \tag{B.53}$$

$$= \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log \left(p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \right) \tag{B.54}$$

$$- \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log \left(p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}) \right)$$

$$= - \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log \left(\frac{1}{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)})} \right) \tag{B.55}$$

$$+ \sum_{\boldsymbol{x}_t^{(d_x)} \in \mathcal{X}^{d_x}} \sum_{\boldsymbol{y}_t^{(d_y)} \in \mathcal{Y}^{d_y}} \sum_{y_{t+1} \in \mathcal{Y}} p(y_{t+1}, \boldsymbol{y}_t^{(d_y)}, \boldsymbol{x}_t^{(d_x)}) \log \left(\frac{1}{p(y_{t+1}|\boldsymbol{y}_t^{(d_y)})} \right)$$

Now, using the conditional entropy definition 3.17

$$= -H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) + \left(H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}) \right) \tag{B.56}$$

$$= H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}) - H(Y_{t+1}|\boldsymbol{Y}_t^{(d_y)}, \boldsymbol{X}_t^{(d_x)}) \tag{B.57}$$

From the definition of the conditional entropy (definition 3.29)

$$= I(Y_{t+1}; \boldsymbol{X}_t^{(d_x)}|\boldsymbol{Y}_t^{(d_y)}) \tag{B.58}$$

□

Appendix C

Metrics validation

In this appendix, we present the validation results for all the metrics used in the econometric model and the transfer entropy. Econometric model validation is performed using 96 of the 5-minute-long subsamples from the EUR/CHF data set. Using all the subsamples, we generate baseline maps and compare them to the baselines generated with Hagströmer and Menkveld [2019]'s model[1]. Each econometric metric from the baseline is validated independently. Next, the transfer entropy metric is validated with the analytical solution for transfer entropy between two Gaussian coupled auto-regressive processes. We use the analytical solution derived by Kaiser and Schreiber [2002].

[1] Hägrstromer and Menkveld's model is available at Mekveld.

C.1 Information Share

The first metric that we validate is Hasbrouck's information share. The mean relative errors for the lower and upper bound of information share are presented in the top row of Figure C.1. The mean relative error for lower bound information share ranges between approximately 0.1% and 1.4%, whereas for the upper bound, the mean relative error runs between 0.1% and 0.8%.

The mean relative errors for each dealer's long run efficient price are presented in the bottom row of Figure C.1. The mean relative error of this metric ranges between approximately 0.1% and 0.5%.

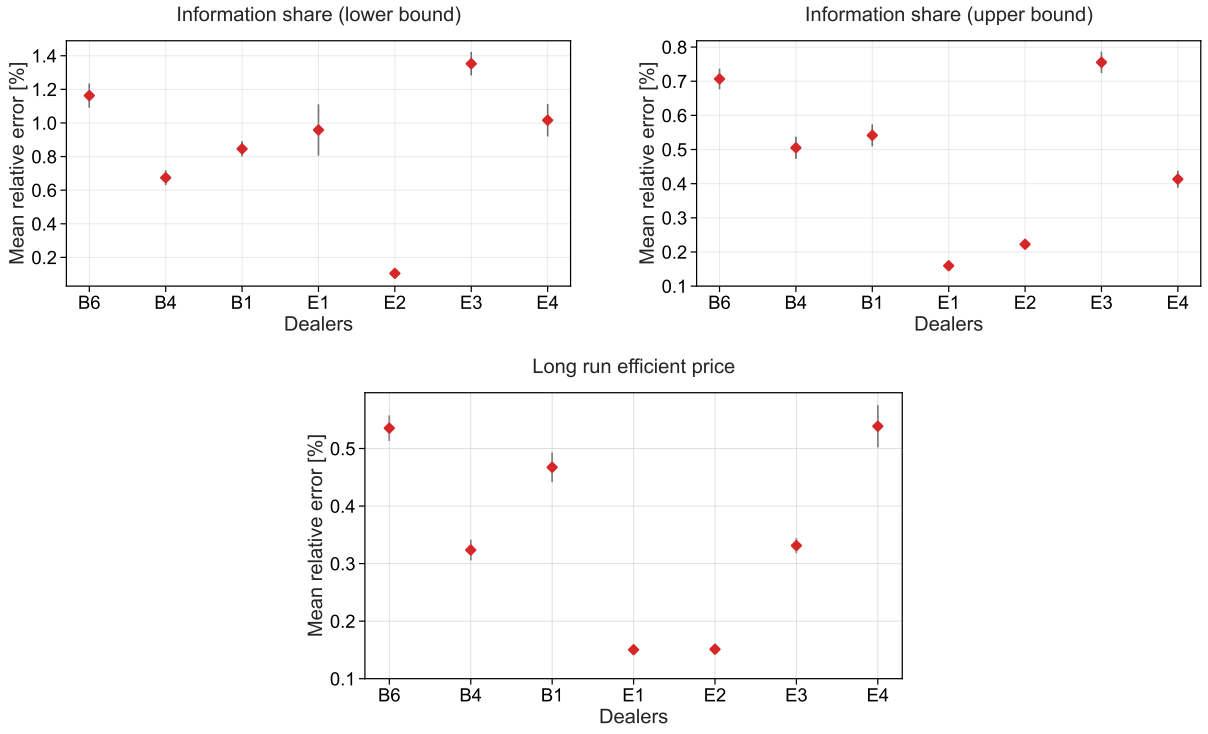

FIGURE C.1: Validation of Hasbrouck's information share metric and the long run efficient price computed with MIRF. The top plots present the mean relative error of Hasbrouck's lower bound (left) and upper bound (right) information share metric calculated for each dealer in the baseline. The bottom plot presents the mean relative error of dealers' long run efficient prices. Additionally, for each mean relative error, a 95% confidence interval is plotted, which may sometimes be too small to be visible on the plot.

C.2 Price inefficiency

The next metric we validate is Hägstromer and Menkveld's price inefficiency metric. The mean relative errors price inefficiency metric and the corresponding standard error of the estimates are presented in the top row of Figure C.2. The mean relative error price inefficiency ranges between 0.2% and 2.9%. However, for most estimates, the mean relative error is smaller than 1%.

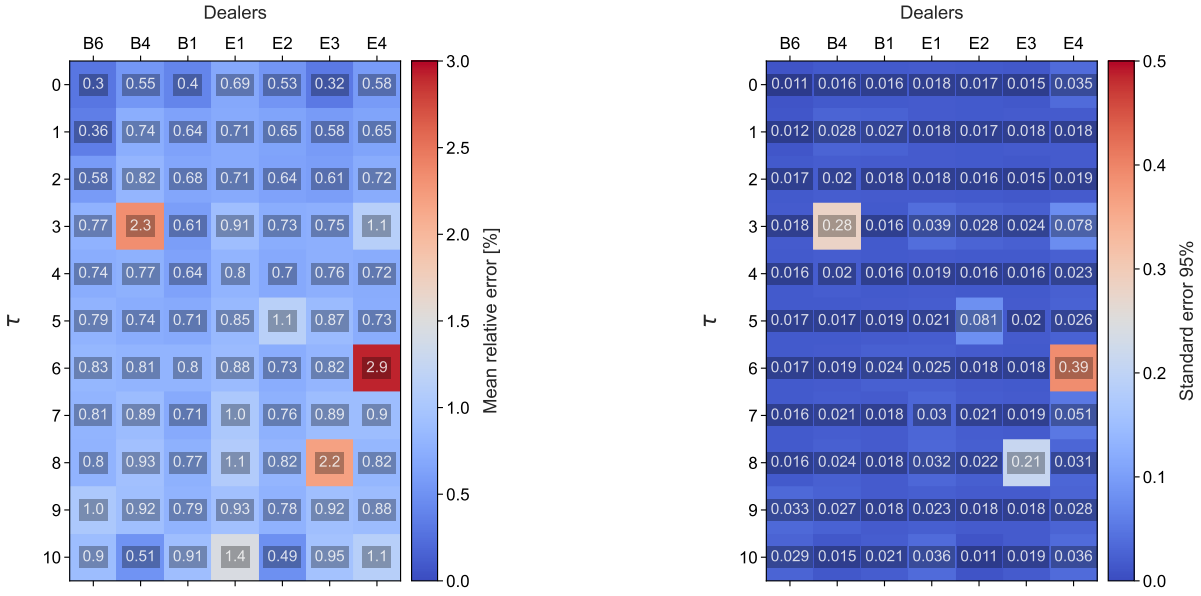

FIGURE C.2: Validation of Hägstromer and Menkveld's price inefficiency metric. The matrix on the left-hand side shows the mean relative errors of dealer's price inefficiencies from $\tau = 0$ until $\tau = 10$. The matrix on the right-hand side presents the corresponding 95% standard errors.

C.3 Bilateral connection

The last econometric metric required to validate is the partial correlation matrix of the dealer's cumulative price changes (bilateral connections). The mean relative errors of the partial correlation matrices for $\tau = 0, 1, 5$ and 10 are presented in Figure C.3. The figure shows that the mean relative error tends to increase for larger τ's. The largest mean relative error of 3.0% is observed for $\tau = 10$. Further investigation revealed that the problem lies in the methodology of computing partial correlations of cumulative price changes, particularly the fact that matrix inversion is used. At large τ's, the cumulative responses of all dealers become very similar, yielding a covariance matrix of dealers' cumulative responses with high condition numbers. Even though various numerically stable algorithms were used to perform the matrix inversion, it is not surprising that we observe relatively large numerical errors. Nevertheless, as discussed in Chapter 5, in the analysis of our results, we only make use of matrices for $\tau = 0$ and $\tau = 1$.

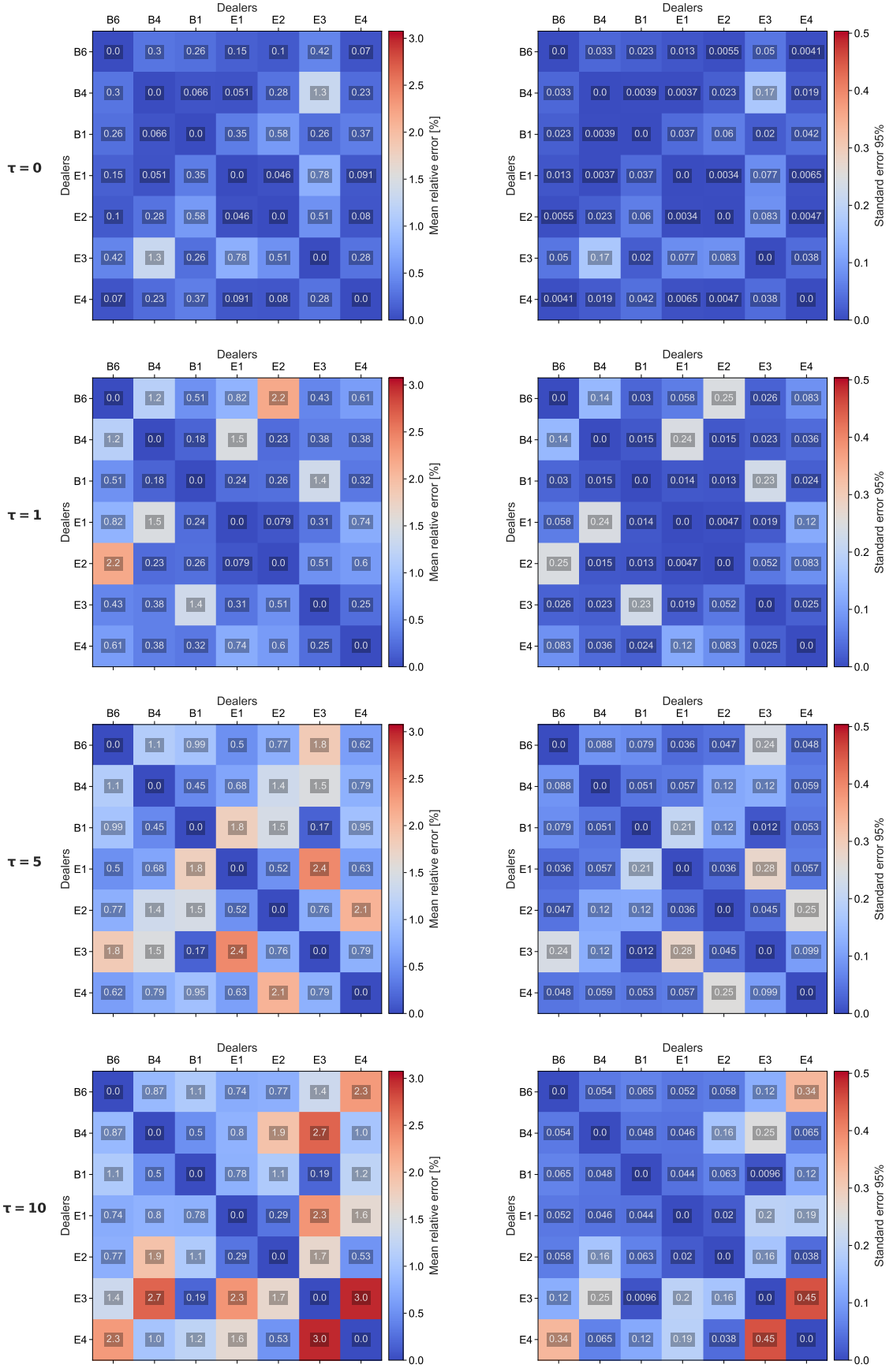

FIGURE C.3: Validation of Hägstromer and Menkveld's bilateral connection metric. Matrices on the left-hand side show the mean relative errors for each dealer pair in the partial correlation matrix of the dealer's cumulative price changes. The matrix on the right-hand side presents the corresponding 95% standard errors.

C.4 Transfer entropy

To validate the KSG algorithm's implementation for the transfer entropy estimation, we compare our transfer entropy estimates with the analytical solution provided by Kaiser and Schreiber [2002]. The analytical solution is provided for Gaussian coupled auto-regressive processes, which need to be simulated. The following system of equations defines the coupled processes X_t and Y_t.

$$X_{t+1} = \alpha X_t + \epsilon_t \tag{C.1}$$

$$Y_{t+1} = \beta Y_t + \gamma X_t + \omega_t \tag{C.2}$$

where ϵ_t and ω_t are independent standard normal random variables.

Following the steps of Kaiser and Schreiber [2002] we use $\alpha = 0.5$ and $\beta = 0.6$. Next, we simulate processes X_t and Y_t for two different sample sizes; 3000 and 1 million. Additionally, the data is generated for 20 different coupling parameter values γ. And finally, the transfer entropy between the processes is estimated with our KSG algorithm and computed using the analytical solution. The comparison of the KSG estimate to the analytical solution is presented in Figure C.4. From Figure C.4, it is clear that the analytical solution nearly perfectly aligns

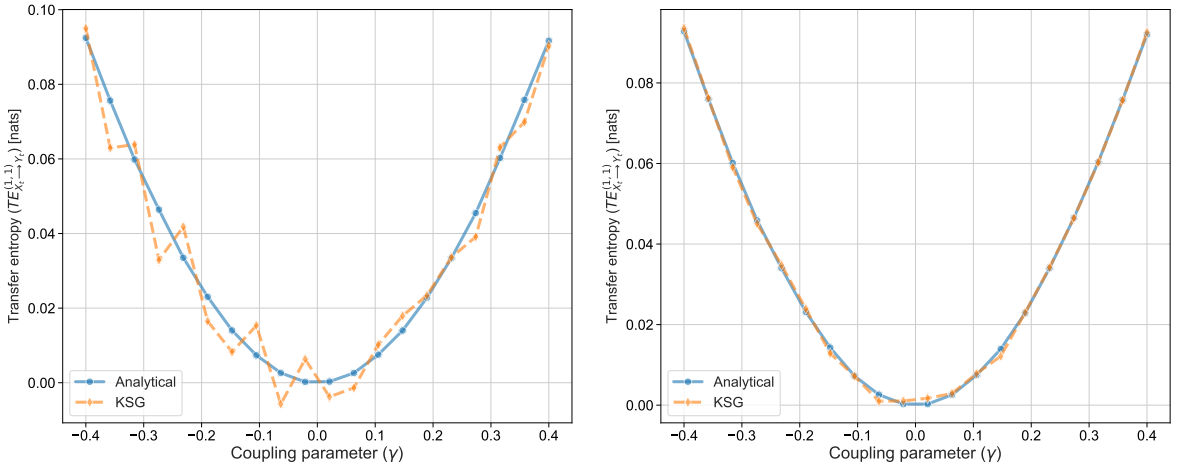

FIGURE C.4: Validation of the KSG algorithm implementation against the analytical solution for transfer entropy. The plot on the left-hand side presents the transfer entropies estimated for a sample size of 3000. At the same time, the plot on the right-hand side compares both estimates for a sample size of 1 million data points.

with our KSG transfer entropy estimates for a large sample size. Therefore, small disparities observed for the smaller sample size can be disregarded because they result from finite-size effects.

Appendix D

Supplementary results for EUR/USD data set.

D.1 Informativeness metrics vs other types of dealer centrality.

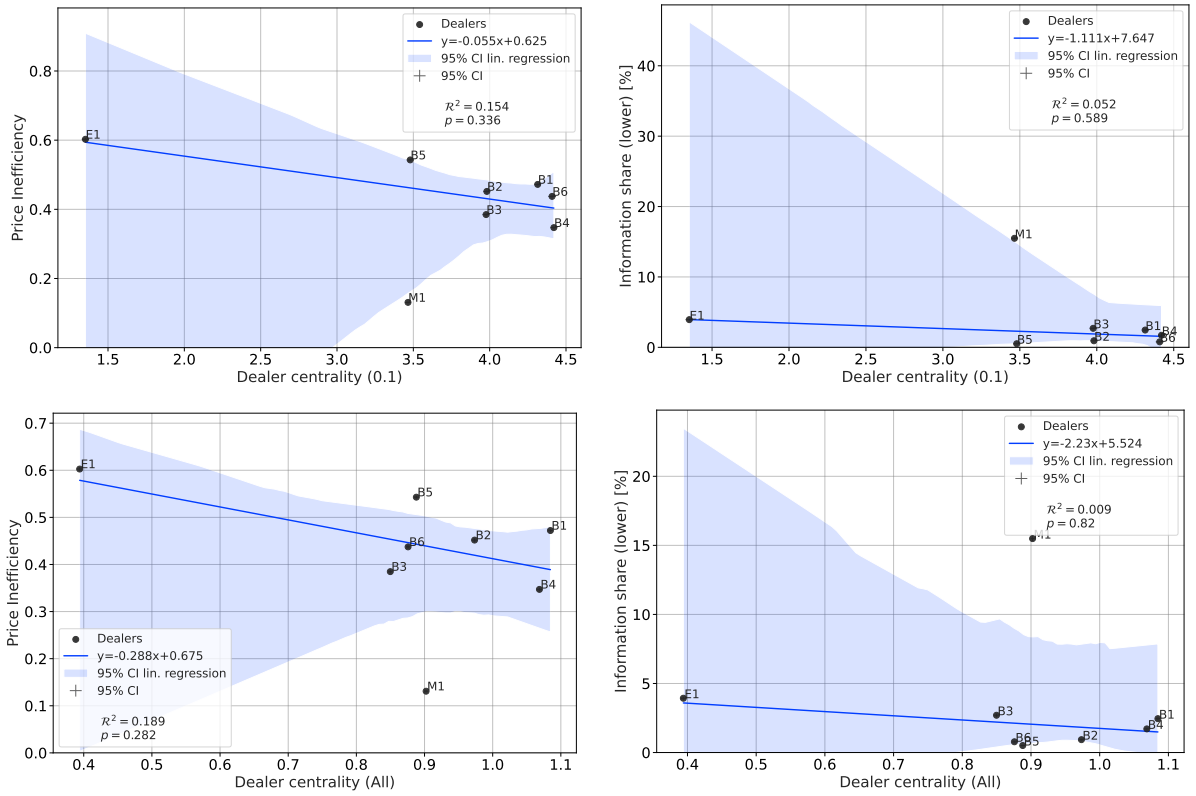

FIGURE D.1: Scatter plot of dealer centrality (0.1) (top row) and (All) (bottom row) versus price inefficiency and lower bound information share for EUR/USD baseline. The linear regressions for dealer centrality (0.1) and dealer centrality (All) are presented on top and bottom, respectively.

Appendix E

Results for USD/JPY dealer-network

E.1 Stationarity & Cointegration Testing

Dates	Dealers					
	B1	B2	B3	B4	B6	E1
2019-01-02	1	2	0	0	0	2
2019-01-03	1	2	0	1	3	0
2019-01-04	1	4	1	0	2	0
2019-01-07	0	2	1	0	0	2
2019-01-08	0	2	0	0	1	1
2019-01-09	3	2	3	3	2	6
2019-01-10	1	0	1	3	2	5
2019-01-11	1	0	1	1	3	1
2019-01-14	6	0	0	0	1	1
2019-01-15	1	3	3	0	0	0
2019-01-16	0	8	2	0	1	4
2019-01-17	3	5	1	2	2	5

TABLE E.1: Result of augmented Dickey-Fuller tests for USD/JPY data set. The tables reveals the number of subsampled that were determined to be stationary for a significance level $\alpha = 0.05/96$. For each dealer-date combination there are 96 subsamples tested, hence the Bonferroni correction is applied. The subsamples used for testing are 5-minute-long time-series resampled to 100ms time intervals. Total of 6912 subsampled were investigated.

Dealers	Dealers				
	B2	B3	B4	B6	E1
B1	1	96	2	6	94
B2		73	0	3	67
B3			71	106	8
B4				4	75
B6					118

TABLE E.2: Result of augmented Engle-Granger two-step cointegration test for USD/JPY data set. The test is performed for each dealer-dealer pair. The table presents the number of subsamples that were determined to be not cointegrated at significance level $\alpha = 0.05/96$. For each dealer-dealer combination there are $12 \cdot 96 = 1152$ subsamples tested. The subsamples used for testing are 5-minute-long time-series resampled to 100ms time intervals.

E.2 Average intraday quote frequency

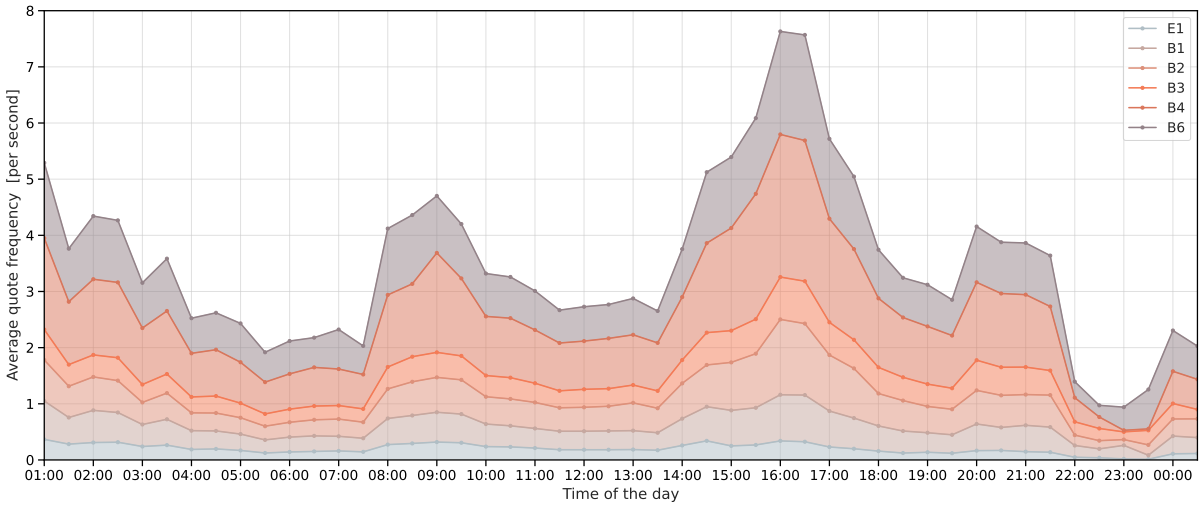

FIGURE E.1: Average intraday EUR/USD quote frequency for trading days in the period February 27 2020 and March 27 2020. Figure illustrates how average quote frequency of all dealers in the data set evolved throughout the day. The For example value at 13:00 represents the quote frequency observed between 13:00 and 13:30.

E.3 Econometric baseline map

FIGURE E.2: Econometric map illustrating the process of information revelation in the USD/JPY dealer-network for period between 2nd January and 17th January 2019. The network vertices correspond to the dealers, as noted in Table 4.1. Each panel visualizes averages of metrics computed from all five-minute-long subsamples of trading days τ time steps since introduction of the multivariate shock.

E.4 Scatter plots for econometric baseline

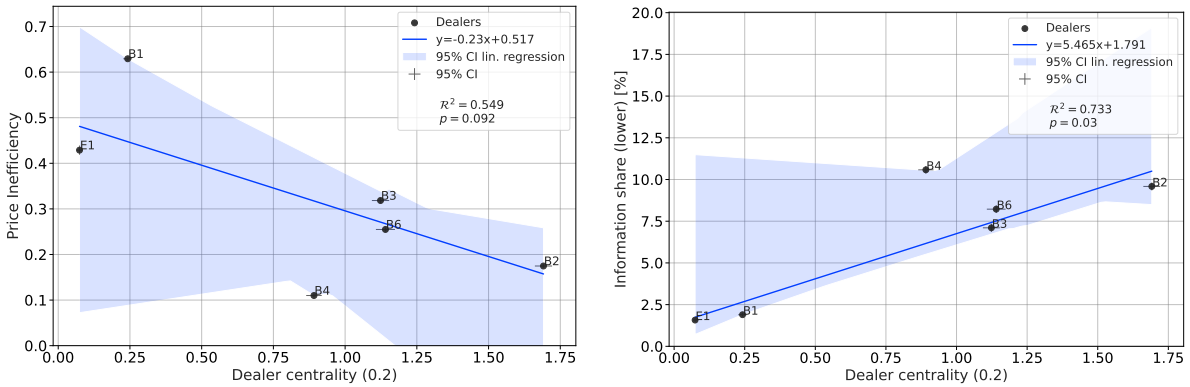

FIGURE E.3: Scatter plot of dealer centrality (0.2) versus price inefficiency and lower bound information share for USD/JPY baseline.

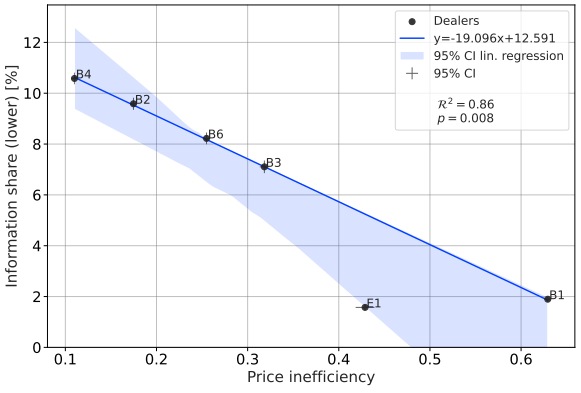

FIGURE E.4: Scatter plot of dealer price inefficiency versus Hasbrouck's lower bound information share for USD/JPY baseline.

E.5 Summary of key information theoretic metrics

	Information Share [%]		TE [nats]		CTE [nats]		Price inefficiency
Dealers	Upper	Lower	(inflow)	(outflow)	(inflow)	(outflow)	$\tau = 0$
B1	17.274	1.895	0.072	0.003	0.012	0.001	0.629
B2	54.526	9.585	0.003	0.034	0.003	0.009	0.175
B3	43.451	7.104	0.003	0.026	0.003	0.004	0.318
B4	48.874	10.581	0.002	0.033	0.002	0.009	0.110
B6	46.669	8.222	0.008	0.021	0.004	0.004	0.255
E1	10.353	1.573	0.030	0.002	0.005	0.001	0.429

TABLE E.4: Summary of key information theoretic metrics characterizing the USD/JPY TE and CTE information maps presented in Figure 5.11.

E.6 Scatter plots for information-theoretic baseline

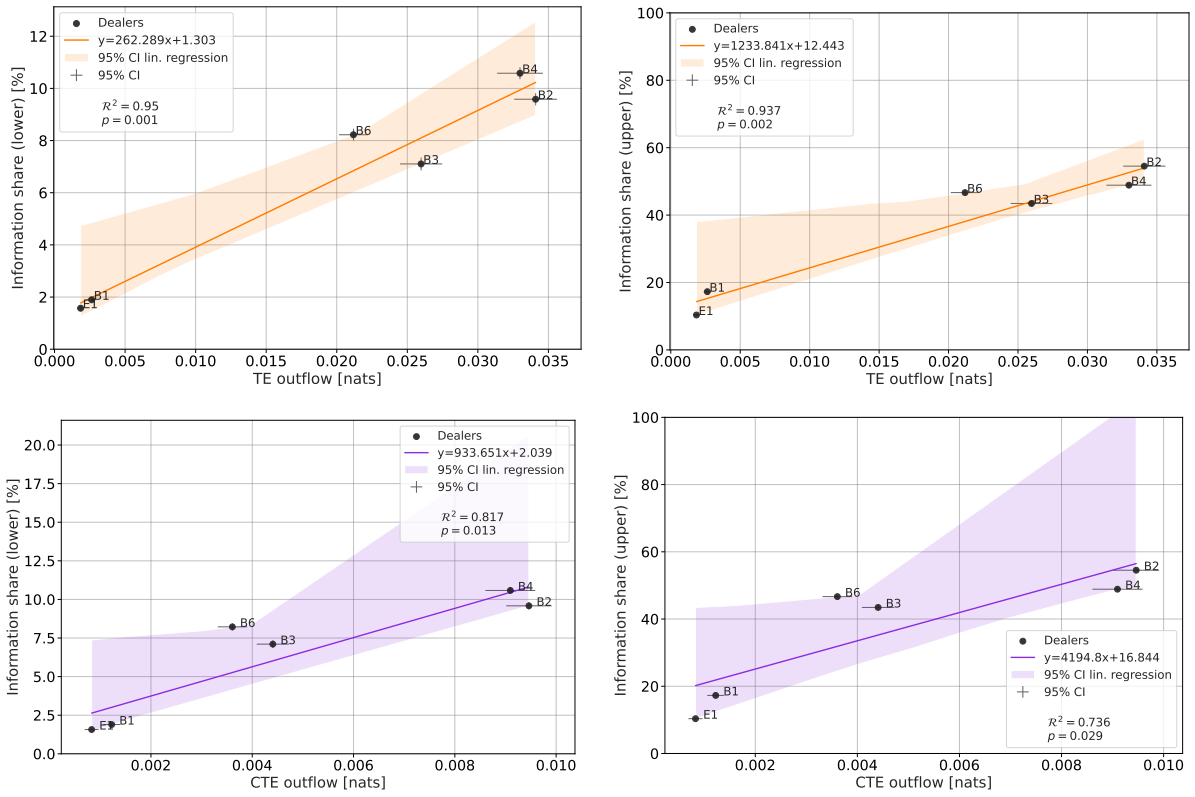

FIGURE E.5: Scatter plot of TE (top row) and CTE (bottom row) outflows versus Hasbrouck's lower (LHS) and upper (RHS) bound information shares for USD/JPY baseline.

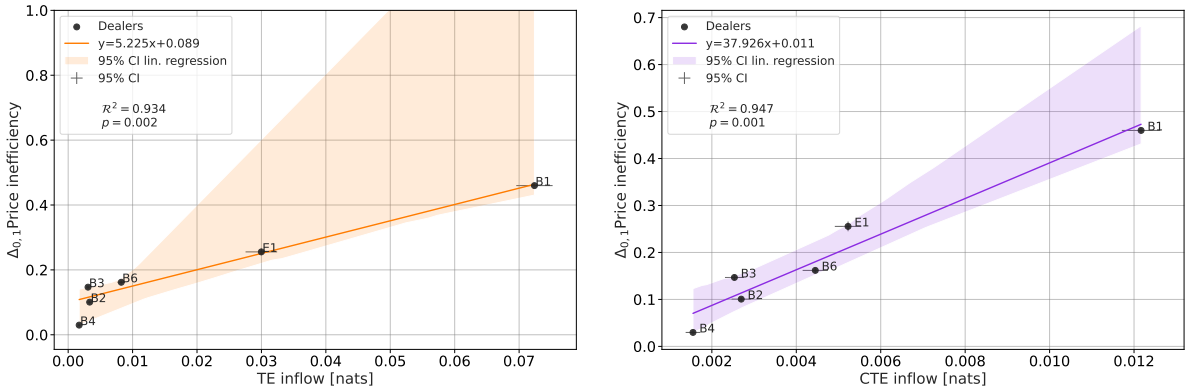

FIGURE E.6: Scatter plot of TE (LHS) and CTE (RHS) inflows versus absolute change in price inefficiency from $\tau = 0$ to $\tau = 1$ for USD/JPY baseline.

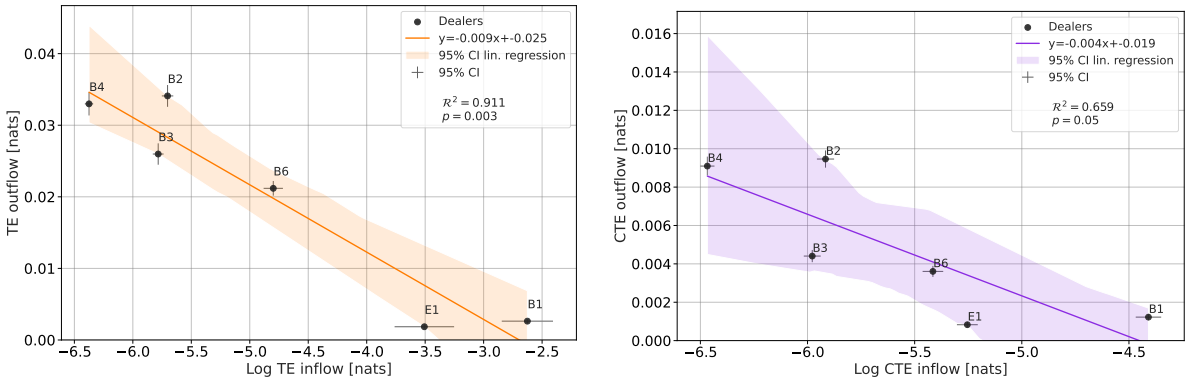

FIGURE E.7: Scatter plot of log TE inflows versus TE outflows (LHS) and log CTE inflows versus CTE outflows (RHS) for USD/JPY baseline.

Bibliography

Triennial Central Bank Survey - Foreign exchange turnover in April 2019, Sep 2019a.

Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets in 2019, Sep 2019b. URL https://www.bis.org/statistics/rpfx19.htm.

Andrew Addison, Charles Andrews, Newas Azad, Daniel Bardsley, John Bauman, Jeffrey Diaz, Tatiana Didik, Komoliddin Fazliddin, Maria Gromoa, Ari Krish, et al. Low-latency trading in the cloud environment. In *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 272–282. IEEE, 2019.

David J Albers and George Hripcsak. Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. *Chaos, Solitons & Fractals*, 45(6):853–860, 2012a.

David J Albers and George Hripcsak. Using time-delayed mutual information to discover and interpret temporal correlation structure in complex populations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(1):013111, 2012b.

Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5):056221, 2004.

Daniel Andrei and Julien Cujean. Information percolation, momentum and reversal. *Journal of Financial Economics*, 123(3):617–645, 2017.

Yuri Antonacci, Laura Astolfi, Giandomenico Nollo, and Luca Faes. Information transfer in linear multivariate processes assessed through penalized regression techniques: validation and application to physiological networks. *Entropy*, 22(7):732, 2020.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

Ana Babus and Péter Kondor. Trading and information diffusion in over-the-counter markets. *Econometrica*, 86(5):1727–1769, 2018.

Ehung Baek and William Brock. A general test for nonlinear Granger causality: Bivariate model - Working Paper. *Iowa State University and University of Wisconsin at Madison Working Paper*, 1992.

Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52. IEEE, 1986.

Richard T Baillie, G Geoffrey Booth, Yiuman Tse, and Tatyana Zabotina. Price discovery and common factor models. *Journal of financial markets*, 5(3):309–321, 2002.

Anat Baniel. *Move into life: the nine essentials for lifelong vitality.* Harmony, 2009.

Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical review letters*, 103(23):238701, 2009.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559, 2012.

Søren Bisgaard and Murat Kulahci. *Time series analysis and forecasting by example.* John Wiley & Sons, 2011.

Max R Blouin and Roberto Serrano. A decentralized market with common values uncertainty: Non-steady states. *The Review of Economic Studies*, 68(2):323–346, 2001.

Ludwig Boltzmann. On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium. *Entropy*, 17(4):1971–2009, 2015.

G Geoffrey Booth, Raymond W So, and Yiuman Tse. Price discovery in the German equity index derivatives markets. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 19(6):619–643, 1999.

Alexander Borst and Frédéric E Theunissen. Information theory and neural coding. *Nature neuroscience*, 2(11):947–957, 1999.

T Bossomaier, L Barnett, M Harré, and JT Lizier. An Introduction to Transfer Entropy: Information Flow in Complex Systems. Cham, Switzerland: Springer International Publishing; 2016, 2016.

Morten Brandvold, Peter Molnár, Kristian Vagstad, and Ole Christian Andreas Valstad. Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money*, 36:18–35, 2015.

Andres Buehlmann and Gustavo Deco. Optimal information transfer in the cortex through synchronization. *PLoS computational biology*, 6(9):e1000934, 2010.

Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

Liangyue Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110(1-2):43–50, 1997.

Kalok Chan. A further analysis of the lead–lag relationship between the cash market and stock index futures market. *The Review of Financial Studies*, 5(1):123–152, 1992.

Chris Chatfield and Haipeng Xing. *The analysis of time series: an introduction with R*. Chapman and hall/CRC, 2019.

Shu-Heng Chen, Mak Kaboudan, and Ye-Rong Du. *The Oxford handbook of computational economics and finance*. Oxford University Press, 2018.

Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. Analyzing multiple nonlinear time series with extended Granger causality. *Physics letters A*, 324(1):26–35, 2004.

Daniel Chicharro and Anders Ledberg. When two become one: the limits of causality analysis of brain dynamics. *PloS one*, 7(3):e32466, 2012.

Frank De Jong. Measures of contributions to price discovery: A comparison. *Journal of Financial markets*, 5(3):323–327, 2002.

Luc Devroye and László Györfi. On the consistency of the Kozachenko-Leonenko entropy estimate. *IEEE Transactions on Information Theory*, 2021.

Francis X Diebold and Kamil Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics*, 182(1):119–134, 2014.

Thomas Dimpfl and Franziska Julia Peter. Using transfer entropy to measure information flows between financial markets. *Studies in Nonlinear Dynamics and Econometrics*, 17(1):85–102, 2013.

Roland L'vovich Dobrushin. A simplified method of experimentally evaluating the entropy of a stationary sequence. *Theory of Probability & Its Applications*, 3(4):428–430, 1958.

Brent Donnelly. *The Art of Currency Trading: A Professional's Guide to the Foreign Exchange Market*. John Wiley & Sons, 2019.

D Duffie. Dark Markets: Asset Pricing and Information Percolation in Over-the-Counter Markets. *Princeton Lecture Series*, 2012.

Darrell Duffie and Gustavo Manso. Information percolation in large markets. *American Economic Review*, 97(2):203–209, 2007.

Darrell Duffie, Semyon Malamud, and Gustavo Manso. Information percolation with equilibrium search dynamics. *Econometrica*, 77(5):1513–1574, 2009.

Darrell Duffie, Semyon Malamud, and Gustavo Manso. Information percolation in segmented markets. *Journal of Economic Theory*, 153:1–32, 2014.

Yfke Dulek and Christian Schaffner. Information Theory, University of Amseterdam, Master of Logic, 2017.

Steven Durlauf and Lawrence Blume. *Macroeconometrics and time series analysis*. Springer, 2016.

Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

Michael Eichler. *Causal inference in time series analysis*. na, 2012.

Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

European Central Bank European Central Bank. ECB announces measures to support bank liquidity conditions and money market activity, Mar 2020. URL https://www.ecb.europa.eu/press/pr/date/2020/html/ecb.pr200312_2~06c32dabd1.en.html.

Martin DD Evans and Richard K Lyons. Order flow and exchange rate dynamics. *Journal of political economy*, 110(1):170–180, 2002.

Paweł Fiedor. Networks in financial markets based on the mutual information rate. *Physical Review E*, 89(5):052801, 2014.

Isabel Figuerola-Ferretti and Jesús Gonzalo. Modelling and measuring price discovery in commodity markets. *Journal of Econometrics*, 158(1):95–107, 2010.

Christodoulos A Floudas and Panos M Pardalos. *Encyclopedia of optimization*. Springer Science & Business Media, 2008.

Stefan Frenzel and Bernd Pompe. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters*, 99(20):204101, 2007.

Carley Garner. *The Basics of Forex Trading*. Pearson Education, 2011.

Deniz Gençağa. Transfer entropy, 2018.

Deniz Gencaga, Kevin H Knuth, and William B Rossow. A recipe for the estimation of information flow in a dynamical system. *Entropy*, 17(1):438–470, 2015.

John Geweke. Inference and causality in economic time series models. *Handbook of econometrics*, 2:1101–1144, 1984.

Mikhail Golosov, Guido Lorenzoni, and Aleh Tsyvinski. Decentralized trading with private information. *Econometrica*, 82(3):1055–1091, 2014.

Germán Gómez-Herrero, Wei Wu, Kalle Rutanen, Miguel C Soriano, Gordon Pipa, and Raul Vicente. Assessing coupling dynamics from an ensemble of time series. *Entropy*, 17(4):1958–1970, 2015.

Jesus Gonzalo and Clive Granger. Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics*, 13(1):27–35, 1995.

Boris Gourévitch, Régine Le Bouquin-Jeannès, and Gérard Faucon. Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biological cybernetics*, 95(4):349–369, 2006.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Clive WJ Granger. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, 16(1):121–130, 1981.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided self-organization: inception*, pages 159–190. Springer, 2014.

Damodar N Gujarati and Dawn C Porter. Basic econometrics (ed.). *Singapore: McGrew Hill Book Co*, 2003.

Zhenghao Guo, Verity M McClelland, Osvaldo Simeone, Kerry R Mills, and Zoran Cvetkovic. Multiscale Wavelet Transfer Entropy with Application to Corticomuscular Coupling Analysis. *IEEE Transactions on Biomedical Engineering*, 69(2):771–782, 2021.

Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, 477(2251):20210110, 2021.

Björn Hagströmer and Albert J Menkveld. A network map of information percolation. *Working Paper*, 2016.

Björn Hagströmer and Albert J Menkveld. Information revelation in decentralized markets. *The Journal of Finance*, 74(6):2751–2787, 2019.

James Douglas Hamilton. *Time series analysis*. Princeton university press, 1994.

Frederick H deB Harris, Thomas H McInish, Gary L Shoesmith, and Robert A Wood. Cointegration, error correction, and price discovery on informationally linked security markets. *Journal of financial and quantitative analysis*, 30(4):563–579, 1995.

Frederick H deB Harris, Thomas H McInish, and Robert A Wood. Common factor components versus information shares: a reply. *Journal of Financial Markets*, 5(3):341–348, 2002a.

Frederick H deB Harris, Thomas H McInish, and Robert A Wood. Security price adjustment across exchanges: an investigation of common factor components for Dow stocks. *Journal of financial markets*, 5(3):277–308, 2002b.

Joel Hasbrouck. One security, many markets: Determining the contributions to price discovery. *The journal of Finance*, 50(4):1175–1199, 1995.

Joel Hasbrouck. 22 Modeling market microstructure time series. *Handbook of statistics*, 14: 647–692, 1996.

Joel Hasbrouck. Stalking the "efficient price" in market microstructure specifications: an overview. *Journal of Financial Markets*, 5(3):329–339, 2002.

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

Joel Hasbrouck and Gideon Saar. Low-latency trading. *Journal of Financial Markets*, 16(4): 646–679, 2013.

Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, 2009.

Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

Alfred Hero and Bala Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078, 2012.

Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

Ali Hirsa and Salih N Neftci. *An introduction to the mathematics of financial derivatives*. Academic press, 2013.

Katerina Hlavácková-Schindler. Equivalence of granger causality and transfer entropy: A generalization. *Applied Mathematical Sciences*, 5(73):3637–3648, 2011.

Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, 395(10223):497–506, 2020.

Anna-Louise Jackson and Benjamin Curry. Quantitative Easing Explained, Jan 2022. URL https://www.forbes.com/advisor/investing/quantitative-easing-qe/#:~:text=Quantitative%20easing%E2%80%94QE%20for%20short,lending%20to%20consumers%20and%20businesses.

Ryan G James, Nix Barnett, and James P Crutchfield. Information flows? A critique of transfer entropies. *Physical review letters*, 116(23):238701, 2016.

Seung-Hyun Jin, Peter Lin, and Mark Hallett. Linear and nonlinear information flow based on time-delayed mutual information method and its application to corticomuscular interaction. *Clinical Neurophysiology*, 121(3):392–401, 2010.

Søren Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, pages 1551–1580, 1991.

Søren Johansen et al. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand, 1995.

Guy Jumarie. Relative Information—What For? In *Relative Information*, pages 1–11. Springer, 1990.

Alexandra M Jurgens and James P Crutchfield. Shannon entropy rate of hidden Markov processes. *Journal of Statistical Physics*, 183(2):1–18, 2021.

Andreas Kaiser and Thomas Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1-2):43–62, 2002.

Michael R King, Carol L Osler, and Dagfinn Rime. Foreign exchange market structure, players and evolution. 2011.

Robert L Kissell. *The science of algorithmic trading and portfolio management*. Academic Press, 2013.

AN Kolmogorov. Information Theory and the Theory of Algorithms, volume III of Selected Works of AN Kolmogorov, 1993.

Lyudmyla F Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Okyu Kwon and J-S Yang. Information flow between stock indices. *EPL (Europhysics Letters)*, 82(6):68003, 2008.

David C Lay. Linear Algebra and its applications 5th edition. *Pearson*, 2016.

Hodong Lee, Changsoo Kim, Sanha Lim, and Jong Min Lee. Data-driven fault diagnosis for chemical processes using transfer entropy and graphical lasso. *Computers & Chemical Engineering*, 142:107064, 2020.

Bruce N Lehmann. Some desiderata for the measurement of price discovery across markets. *Journal of Financial Markets*, 5(3):259–276, 2002.

Ruben M Leisink. On The Quantification Of Information Transfer Between Discrete-Time Stochastic Processes Using Information Theory. Master's thesis, University of Amsterdam, 2019.

Jianping Li, Changzhi Liang, Xiaoqian Zhu, Xiaolei Sun, and Dengsheng Wu. Risk contagion in Chinese banking industry: A Transfer Entropy-based analysis. *Entropy*, 15(12):5549–5564, 2013.

Songting Li, Yanyang Xiao, Douglas Zhou, and David Cai. Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information. *Physical Review E*, 97(5):052216, 2018.

Wentian Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5):823–837, 1990.

Zheng Li, Ping Li, Arun Krishnan, and Jingdong Liu. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics*, 27(19):2686–2691, 2011.

Michael Lindner, Raul Vicente, Viola Priesemann, and Michael Wibral. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC neuroscience*, 12(1):1–22, 2011.

Joseph Lizier. Is KSG estimator deterministic? - [Java Information Dynamics Toolkit (JIDT) discussion], 2015. URL https://groups.google.com/g/jidt-discuss/c/EmLEXo9BGcA/m/ReOAnwc6DgAJ.

Joseph Lizier. Ragwitz auto-embedding in conditional transfer entropy - [Java Information Dynamics Toolkit (JIDT) discussion], 2022. URL https://groups.google.com/g/jidt-discuss/c/TEcGwPQ__7U.

Joseph Lizier and Mikail Rubinov. Multivariate construction of effective computational networks from observational data. 2012.

Joseph T Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.

Joseph T Lizier and Mikhail Prokopenko. Differentiating information transfer and causal effect. *The European Physical Journal B*, 73(4):605–615, 2010.

Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. Information modification and particle collisions in distributed computation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):037109, 2010.

Joseph T Lizier, Jakob Heinzle, Annette Horstmann, John-Dylan Haynes, and Mikhail Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *Journal of computational neuroscience*, 30(1):85–107, 2011.

Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. Local measures of information storage in complex distributed computation. *Information Sciences*, 208:39–54, 2012.

Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. A framework for the local information dynamics of distributed computation in complex systems. In *Guided self-organization: inception*, pages 115–158. Springer, 2014.

Warren M Lord, Jie Sun, and Erik M Bollt. Geometric k-nearest neighbor estimation of entropy and mutual information. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3): 033114, 2018.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Richard K Lyons et al. *The microstructure approach to exchange rates*, volume 333. Citeseer, 2001.

David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Ananth Madhavan. Market microstructure: A survey. *Journal of financial markets*, 3(3):205–258, 2000.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

María Rodríguez Martínez, Alberto Corradin, Ulf Klein, Mariano Javier Álvarez, Gianna M Toffolo, Barbara di Camillo, Andrea Califano, and Gustavo A Stolovitzky. Quantitative modeling of the terminal differentiation of B cells and mechanisms of lymphomagenesis. *Proceedings of the National Academy of Sciences*, 109(7):2672–2677, 2012.

Albert J Mekveld. Albert J. Menkveld - My two cents. URL https://albertjmenkveld.com/published-papers/.

Albert J Menkveld. High frequency trading and the new market makers. *Journal of financial Markets*, 16(4):712–740, 2013.

Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007:1–9, 2007.

Joseph Victor Michalowicz, Jonathan M Nichols, and Frank Bucholtz. *Handbook of differential entropy*. Crc Press, 2013.

Alessandro Montalto, Luca Faes, and Daniele Marinazzo. MuTE: a MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy. *PloS one*, 9(10): e109462, 2014.

Leonardo Novelli, Patricia Wollstadt, Pedro Mediano, Michael Wibral, and Joseph T Lizier. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, 3(3):827–847, 2019.

Masafumi Oizumi, Naotsugu Tsuchiya, and Shun-ichi Amari. Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, 113(51):14817–14822, 2016.

Peterson K Ozili and Thankom Arun. Spillover of COVID-19: impact on the Global Economy. *Available at SSRN 3562570*, 2020.

Shaowu Pan and Karthik Duraisamy. On the structure of time-delay embedding in linear models of non-linear dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(7):073135, 2020.

A Papana, D Kugiumtzis, and PG Larsson. Reducing the bias of causality measures. *Physical Review E*, 83(3):036207, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

H Hashem Pesaran and Yongcheol Shin. Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29, 1998.

Mikhail Prokopenko, Joseph T Lizier, and Don C Price. On thermodynamic interpretation of transfer entropy. *Entropy*, 15(2):524–543, 2013.

Tālis J Putniņš. What do price discovery metrics really measure? *Journal of Empirical Finance*, 23:68–83, 2013.

Rick Quax, Omri Har-Shemesh, and Peter Sloot. Quantifying synergistic information using intermediate stochastic variables. *Entropy*, 19(2):85, 2017.

Mario Ragwitz and Holger Kantz. Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Physical Review E*, 65(5):056201, 2002.

Neil Record. *Currency overlay*. John Wiley & Sons, 2004.

YV Reddy and A Sebastin. Interaction between forex and stock markets in India: An entropy approach. *Vikalpa*, 33(4):27–46, 2008.

Jakob Runge. Detecting and quantifying causality from time series of complex systems. 2014.

Michael J Sager and Mark P Taylor. Under the microscope: the structure of the foreign exchange market. *International Journal of Finance & Economics*, 11(1):81–95, 2006.

Leonidas Sandoval. Structure of a global network of financial companies based on transfer entropy. *Entropy*, 16(8):4443–4482, 2014.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Roberto Serrano and Oved Yosha. Information revelation in a market with pairwise meetings: the one sided information case. *Economic theory*, 3(3):481–499, 1993.

Roberto Serrano and Oved Yosha. Welfare analysis of a market with pairwise meetings and asymmetric information. *Economic Theory*, 8(1):167–175, 1996.

Payam Shahsavari Baboukani, Carina Graversen, Emina Alickovic, and Jan Østergaard. Estimating Conditional Transfer Entropy in Time Series Using Mutual Information and Nonlinear Prediction. *Entropy*, 22(10):1124, 2020.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Yidan Shu and Jinsong Zhao. Data-driven causal inference based on a modified transfer entropy. *Computers & Chemical Engineering*, 57:173–180, 2013.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

AJMC Staff. A timeline of covid-19 developments in 2020. URL https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. 2012.

James H Stock and Mark W Watson. Testing for common trends. *Journal of the American statistical Association*, 83(404):1097–1107, 1988.

Ewa M Syczewska and Zbigniew R Struzik. Granger causality and transfer entropy for financial returns. *Acta Physica Polonica A*, 127(3A), 2015.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

Frédéric Theunissen, J Cooper Roddey, Steven Stufflebeam, Heather Clague, and John P Miller. Information theoretic analysis of dynamical encoding by four identified primary sensory interneurons in the cricket cercal system. *Journal of Neurophysiology*, 75(4):1345–1364, 1996.

MTCAJ Thomas and A Thomas Joy. Elements of information theory, 2006.

I-Chun Tsai. Flash crash and policy uncertainty. *Journal of International Financial Markets, Institutions and Money*, 57:248–260, 2018.

Vasily A Vakorin, Bratislav Misic, Olga Krakovska, and Anthony Randal McIntosh. Empirical and theoretical aspects of generation and transfer of information in a neuromagnetic source network. *Frontiers in systems neuroscience*, 5:96, 2011.

Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1):54–59, 1976.

Martin Vejmelka and Milan Paluš. Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2):026214, 2008.

Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

Alejandro F Villaverde, John Ross, Federico Morán, and Julio R Banga. MIDER: network inference with mutual information distance and entropy reduction. *PloS one*, 9(5):e96732, 2014.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Paolo Vitale. A market microstructure analysis of foreign exchange intervention. 2006.

Anders Warne. *A common trends model: identification, estimation and inference*. IIES, 1993.

Alexander Wehrli and Didier Sornette. Classification of flash crashes using the Hawkes (p, q) framework. *Quantitative Finance*, 22(2):213–240, 2022.

Michael Wibral, Patricia Wollstadt, Ulrich Meyer, Nicolae Pampu, Viola Priesemann, and Raul Vicente. Revisiting Wiener's principle of causality—interaction-delay reconstruction using transfer entropy and multivariate analysis on delay-weighted graphs. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3676–3679. IEEE, 2012.

Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T Lizier, and Raul Vicente. Measuring information-transfer delays. *PloS one*, 8(2):e55809, 2013.

Michael Wibral, Raul Vicente, and Michael Lindner. Transfer entropy in neuroscience. In *Directed information measures in neuroscience*, pages 3–36. Springer, 2014a.

Michael Wibral, Raul Vicente, and Joseph T Lizier. *Directed information measures in neuroscience*. Springer, 2014b.

Michael Wibral, Conor Finn, Patricia Wollstadt, Joseph T Lizier, and Viola Priesemann. Quantifying information modification in developing neural networks via partial information decomposition. *Entropy*, 19(9):494, 2017.

Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Paul L Williams and Randall D Beer. Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*, 2011.

Asher Wolinsky. Information revelation in a market with pairwise meetings. *Econometrica: Journal of the Econometric Society*, pages 1–23, 1990.

Patricia Wollstadt, Joseph T Lizier, Raul Vicente, Conor Finn, Mario Martinez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. IDTxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. *arXiv preprint arXiv:1807.10459*, 2018.

Aaron D Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1):51–59, 1978.

Bingcheng Yan and Eric Zivot. A structural analysis of price discovery measures. *Journal of Financial Markets*, 13(1):1–19, 2010.

Yue-Jun Zhang and Yi-Ming Wei. The crude oil market and the gold market: Evidence for cointegration, causality and price discovery. *Resources Policy*, 35(3):168–177, 2010.