Essays on the Statistics of Financial Markets

Dissertation
zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von

Johannes Bleher, MSc.

geboren in Münsingen

Tübingen, 2020

1. Betreuer:	Prof. Dr. Thomas Dimpfl
2. Betreuer:	Prof. Dr. Joachim Grammig
3. Betreuer:	Prof. Dr. Martin Biewen
Tag der mündlichen Prüfung:	08.02.2021
Dekan:	Prof. Dr. Josef Schmid
1. Gutachter:	Prof. Dr. Christian Koziol
2. Gutachter:	Prof. Dr. Thomas Dimpfl

Acknowledgments

Working on the present dissertation project never was a burden to me, in fact I enjoyed it. The real heavy lifting was done by others: The persons that do not appear as author of this dissertation, but without whom it never would have been possible for me to pursue this project in the first place. This is the place to engrave my everlasting gratefulness to these persons.

I want to express the immense gratitude I feel towards my wife, Malin, for supporting me throughout this project. Her professional skills as a midwife to accompany the birth process of this dissertation were, during the past four years, taken up to their limits. Her endurance, patience, backing and constant support made this dissertation possible. While the nightly joint work with my two sons on this dissertation was not always as productive as I hoped – and their vivid, energetic and early start into a dawning day after such joint night sessions seldom was contagious, the balancing joy and leveling energy I experienced in every day's interactions with the two, as well as the pride I feel of the two motivated me immensely to keep going.

I am immeasurably grateful to Thomas Dimpfl and deeply indebted for all the resources he provided and doors he opened. He believed in me, gave me the chance to work on subjects I liked and was always ready to give profound advice. I am thankful that he provided me with vast freedoms that enabled me to balance family, life and interesting work. Even though I was his first doctoral student, he truly was and is the greatest mentor and doctoral thesis supervisor I could have wished and anyone else could wish for. Without Thomas, this dissertation simply would not exist. Without Thomas, balancing family and work would not have been possible. Without Thomas, so much would have been different, and I am deeply grateful to him that things are as they are.

I am also deeply thankful for the support, advice, resources and possibilities that Joachim Grammig provided during the last four years. Like Thomas, he believed in my abilities and enabled this dissertation project from the start. Without the resources of his chair, the XETRA data set as well as the server infrastructure, this dissertation would not have been possible. I am grateful for the opportunity he gave me to work at his chair, among him and his team of scientific titans. It was a truly exhilarating experience. The permanent provisional office in the Sigwartstraße, also known as the MEGA flat, the shared flat that Makes Econometrics Great Again, is a truly wonderful place to work and ponder

econometric problems. The meetings around the coffee table and the joint lunches with invigorating academic and other discussions, as well as the kind atmosphere make this place stand out. For this wonderful work experience, I am thankful to my current and former colleagues at the chair and flatmates in the Sigwartsraße: Martin Biewen, Sylvia Bürger, Lea Eiting, Dalia Elshiaty, Constantin Hanenberg, Eva-Maria Küchlin, Marian Rümmele, Jakob Schwerter, Matthias Seckerl, Jantje Soenksen and Miriam Sturm.

Another shared flat was also an essential contribution to this dissertation: Susanne Wellmann and her husband Jean-Paul Sezawo kindly provided me a place to stay during my days in Tübingen. Thank your for letting your attic be my bivouac, a very comfortable one indeed.

I am also gratefully indebted to my brother and coauthor, Michael Bleher and his wife Dr Katharina Bleher who had to endure our long discussions about the paper at family gatherings. Even though Michael is a theoretical physicist, he is a true handyman – a mathematical one, though. He is a great explainer. Without his enduring and fortunately pedantic explanation of the theoretical concepts of operator algebra, I would not have been able to come up with the idea for the order book model in Chapter 4. An idea is not everything. Equipped with his mathematical toolbox and his rigor, he polished the rough-edged theoretical part of the limit order book model into its current shiny shape. I am also grateful to him, that he was willing to screen the entire dissertation for errors and am looking forward to now discussing other recreational math problems, aside from our joint paper in the future.

Also my sister Ruth Bleher, deserves a special mention and thanks. Only with her insightful comments on how to google for a new and cheap cellphone case, she incepted the index for prices searched online (IPSO) in Chapter 1.

This dissertation especially also profited from suggestions and remarks of Martin Biewen, Julie Schnaitman, Jantje Sönksen. I also thank Roxana Halbleib and Winfried Pohlmeier for their invitation to the Econometrics Colloquium in Konstanz and their insightful comments. The annual joint doctoral research conference of the econometrics departments of the Universities Konstanz, Hohenheim and Friedrichshafen was a wonderful occasion to pitch new ideas and receive feedback on work in progress. I appreciated these annual gatherings.

In this context, special thanks is due to my sister Hannah Bleher, Jonathan Ulrich Baumann and Susanne Wellmann who all reviewed this manuscript and kindly pointed out my orthographical weaknesses.

I also thank the dean of the Faculty of Economics and Social Sciences, Josef Schmid who created an uncomplicated work environment for me at the deanery. I am also much obliged to Sven Bauer, Gabriele Baumann, Daniela Hedrich as well as Dominik Papies and his team. I also want to thank Philipp Kurzendörfer and Alexander Reining for their studious

support in the preparation of the math prep classes, as well as the matrix game. I also acknowledge the support by the state of Baden-Württemberg through bwHPC.

Last but not least, I want to thank my entire family. My mother Andrea Bleher and father Helmut Bleher for their constant support as well as all my other siblings Lena, Lisbeth and Stefan Bleher who on several occasions took care of our sons. I also want to thank the parents of my wife Christiane Schnepf-Balle and Volker Balle for their support. Especially, the weekly Saturday gathering of all their grandchildren is an institution that helped finalizing this dissertation monumentally.

And now, my wife's grandmother, Hedwig Schnepf, can relax. I am not a student anymore, I really work now. My dissertation is finalized.

Contents

List of Figures

List of Tables

Introduction

In the year 1898 after a visit to the Liverpool Exchange Newsroom, Joseph Chamberlain, father of the then future British Prime Minister Neville Chamberlain and Colonial Secretary held a speech in Liverpool's Conservative Club. In his speech, he uttered the phrase which not only described the general newsroom feeling back then, at the dawn of the 19th century, but probably could also be considered today as a fair description of nowadays volatile environment:

> I think that you will all agree that we are living in most interesting times. I never remember myself a time in which our history was so full, in which day by day brought us new objects of interest, and, let me say also, new objects for anxiety.[1]

Since Chamberlain's time, history has filled further and is fuller than ever. At the time of writing this introduction, the world is faced with a pandemic of the Corona virus which has led governments to impose lock-downs earlier this year of 2020 and to pass all sorts of measures to stop the spreading of the virus. Yet, a second wave of the virus seems to emerge in Germany, while in France and other European countries daily infections are rising. Until now the pandemic is expected to have caused the greatest recession in history or at least since 1929. 12 years after the 2008 financial crisis whose aftermaths also caused the European Sovereign debt crisis in the early 2010s, governments around the globe have again poured unprecedented amounts of their tax-payers money into their economies. Central Banks have flooded the markets with liquidity. Amidst this pandemic, Britain, four years after the referendum in which it decided to leave the European Union, is actually about to leave the EU without a legal framework as negotiations about a free trade agreement are stuck. Also, elections are about to be held in the United States of America and their ramifications on world trade and international relations are uncertain. In these times, not only every day brings more 'objects of interest', but every hour, every minute and even every second does. Today, we often see more news, more information pouring in continuously. Information is constantly published and distributed around the globe, available to everybody, almost instantly. No wonder, the flood of new information

[1] The Western Daily Press, Mr. Chamberlain at Liverpool: A Series of Speeches, Patriotism Still a Live Force, Quote Page 8, Column 3, Bristol, England. (British Newspaper Archive). January 21, 1898.

and the change that spreads with it frightens us at times. The racist Victorian colonialist, that Joseph Chamberlain was, had a similar feeling at the turn of the century. Under the impression of a diverging Empire, he obviously felt that change was coming his and Britain's way. Even though we are skeptical and anxious about new information, and we hawk bad news more widely, more detailed and longer than good news (Hornik, Satchi, Cesareo and Pastore 2015), we often lack the skills to adequately process it. The ability to use present information to predict future events is at times deranged. History holds plenty of examples where the distinction between relevant information and irrelevant noise has gone awefully wrong. We are, at times even tragically, bad at adequately predicting certain aspects of the future. Four years ago, when this dissertation project began, nothing in the current global environment was foreseeable. Or was it?

Now and more than a century ago when Joseph Chamberlain visited the Newsroom nearby the Exchange Flags in Liverpool and held his speech on the interesting times he lived in, financial markets – as the venue where people trade expectations about the future – have been and still are especially keen on having the most current news available. Naturally, the result of market participants' interactions, transaction prices and volume, is highly sensitive to new information. When trading financial instruments, knowing in advance pays off. But what information is relevant? Information that drives prices may concern the macroeconomic scale, e.g., information about prevailing inflation expectations, or it may as well be rooted in microeconomic information, such as, whether a certain financial asset attracts the interest of many. Price driving information may also originate from the mechanics of markets' microstrucure, e.g. the information how incoming and canceled orders where distributed in the last five minutes may be predictive of future price movements. This dissertation will investigate all three examples and end with a new econometric method to determine the predictive power of almost any quantitative information. The question, which information really is relevant for price movements will be the recurrent theme of this dissertation.

In Chapter 1, we[2] develop an algorithm to sensibly concatenate Google's SVI so that it can be used for research purposes. The regression-based algorithm allows to construct arbitrarily many comparable, multi-annual, consistent time series on monthly, weekly, daily, hourly and minute-by-minute search volume indices based on the scattered data obtained from Google Trends. The accuracy of the algorithm is illustrated using old datasets from Google that have been used previously in the literature. The algorithm is used to construct an index of prices searched online (IPSO). Out-of-sample, the IPSO improves monthly inflation and consumption forecasts for the US and the Euro Area. In-sample it is contemporaneously correlated with US consumption, when controlling for

[2] Chapter 1 is based on Bleher and Dimpfl (2019) available at SSRN https://ssrn.com/abstract=3357424.

seasonality, and Granger causes US inflation on a monthly frequency. Chapter 1 serves as a basis for Chapter 2.

Chapter 2 starts with analyzing the question whether increased searches on Google have a predictive ability for the transaction prices and volatility of several cryptocurrencies. The analysis is based on a new algorithm which allows to construct multi-annual consistent time series of Google Search Volume Indices (SVIs) on various frequencies. As cryptocurrencies are actively traded on a continuous basis and react very fast to new information, the analysis is initially conducted on a daily basis, lifting the data imposed restriction faced by previous research. In line with the literature on financial markets, we[3] find that returns are not predictable while volatility is predictable to some extent. A number of reasons are discussed why the predictive power is poor. One aspect is the observational frequency which is therefore varied. The results of unpredictable cryptocurrency returns holds on higher (hourly) and lower (weekly) frequencies. Volatility, in contrast, is predictable on all frequencies and we document an increasing accuracy of the forecast when the sampling frequency is lowered.

In Chapter 3 I review and with concrete examples the mathematical tools used in Chapter 4. Then, in Chapter 4, the eagle-eyed perspective on financial markets is left for a microscopic one. A financial market microstructure model for the limit order book is subsequently presented in Chapter 4. In the model, the limit order book (LOB) is described as a continuous Markov process. We[4] develop an algebra to describe its dynamics based on the fundamental events of the book: order arrivals and cancellations. It is shown how all observables (prices, returns, and liquidity measures) are governed by the same variables which also drive arrival and cancellation rates. This is where the influx of news can be observed. It is where individual decisions of traders, based on the latest information, are directly related to the price formation process. 'Interesting times' where lots of news are generated, such as the ones Chamberlain referred to, directly affect the price mechanism as arrival and cancellation rates are shifted across price levels. The sensitivity of the model developed in Chapter 4 is evaluated in a simulation study and an empirical analysis. Several linearized model specifications based on the theoretical description of the LOB are estimated and in- and out-of-sample forecasts on several frequencies conducted. The in-sample results based on contemporaneous information suggest that the model describes up to 90% of the variation of close-to-close returns, the adjusted R^2 still ranges at around 80%. In the more realistic setting where only past information enters the model, we still observe an adjusted R^2 in the range of 15%. The direction of the next return can be predicted, out-of-sample, with an accuracy of over 75% for short time horizons below

[3] Chapter 2 is based on Bleher and Dimpfl (2019) published in the International Review of Financial Analysis.

[4] Chapter 4 is based on Bleher, Bleher and Dimpfl (2020) available at SSRN https://ssrn.com/abstract=3589763 and arXiv https://arxiv.org/abs/2004.11953.

10 minutes. Out-of-sample, on average, we obtain R^2 values for the Mincer-Zarnowitz regression of around 2-3% and an $RMSPE$ that is 10 times lower than values documented in the literature. These are remarkable results for high-frequency data which are usually considered stochastically independent.

Last but not least, Chapter 5 presents a new estimation technique for relative entropy measures. Especially, its application to transfer entropy is promising to answer whether information from one random variable X is helpful in predicting another random variable Y. In certain situations, transfer entropy may also be interpreted in the sense of information flow between the two variables. In information abundant times, it may provide a measure to distinguish the relevant from the irrelevant information. The estimation of relative entropy measures, such as mutual information or transfer entropy, requires the estimation of conditional and joint densities. When the data are continuous, a multi-variate kernel density estimation or a discretization scheme is usually applied. I propose to estimate the necessary joint and conditional frequencies by means of quantile regression. This allows me to avoid arbitrary binning and all associated problems. Moreover, due to the semi-parametric nature of this approach, the computational burden is decisively reduced compared to multi-variate kernel density estimation. Instead, I show that one can flexibly use quantile regressions to estimate joint and conditional densities in order to calculate relative entropy measures such as transfer entropy and mutual information. The estimation technique requires little restrictive assumptions and can help to analyze variables in situations where only few data points are available. Furthermore, by casting the estimation approach into a Generalized Method of Moments framework, I develop the basis for an asymptotic theory to conduct inference on relative entropy measures for multiple variables. In two short applications of the technique I analyze the temporal relationship between Credit Default Swap premia and credit spreads, as well as transatlantic information flows. I find that one minute returns on the German DAX contained predictive information for the S&P500 one-minute returns.

In essence, this dissertation presents several studies, albeit each with a different focus, all are connected by questions at the heart of financial econometrics: what affects prices, how can we make sense of abundantly available information, what information does matter and how do you separate the relevant from the irrelevant?

Chapter 1

Knitting Multi-Annual High-Frequency Google Trends to Predict Inflation and Consumption[1]

There is a well-established branch in the academic literature which relies on Google's search volume indices (SVIs) for prediction. The very first application goes back to Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009) who use SVIs to detect influenza epidemics prior to their official acknowledgment or diagnosis. The main assumption is that individuals rely on Google to gather subject related information at the time the information is needed. Google's SVI makes this information demand transparent and can therefore serve as a good predictor in many fields. In Finance, for example, Bank, Larch and Peter (2011), Da, Engelberg and Gao (2015), Dimpfl and Jank (2016), or Perlin, Caldeira, Santos and Pontuschka (2017), among others, rely on Google's SVI to improve predictions of stock returns and/or volatility. Again, the main assumption is that retail investor's use Google to collect stock and stock market specific information before they trade. Hence, Google's SVIs are said to proxy retail investor attention to the market (cp. Chen, De, Hu and Hwang 2014), in contrast to institutional investors who rely on other means (like Bloomberg) to collect (real-time) information. Google SVIs are also used in other fields of economics and business administration to now- and forecast key variables of interest. Choi and Varian (2012), for example, predict vehicle sales or claims for unemployment benefits, Qadan and Nama (2018) focus on the oil price, Rochdi and Dietzel (2015) consider real estate investments, and Li, Shang, Wang and Ma (2015) predict inflation.

The identifying assumption which is common to the above cited articles is that for each research question, increases or decreases in certain (patterns of) search-terms precede economically relevant, individual behavior. To be able to exploit this relationship, nowadays a careful construction of the included SVIs is paramount when using multi-annual SVIs on a frequency higher than monthly. Ever since the first studies emerged, Google has repeatedly changed the way it makes Google Trends time series available. Initially, they were provided on a daily frequency and a reference date could be specified. The latter served to standardize the time series so that the SVIs could be concatenated immediately.

[1] This chapter is based on Bleher and Dimpfl (2019) available at SSRN https://ssrn.com/abstract=3357424.

Currently, the length of the time series is limited and no flexible reference date is available which makes it impossible to download, for example, three years of daily search query index values directly. Daily data are only provided for a 270 day period, but a reference date cannot be fixed.

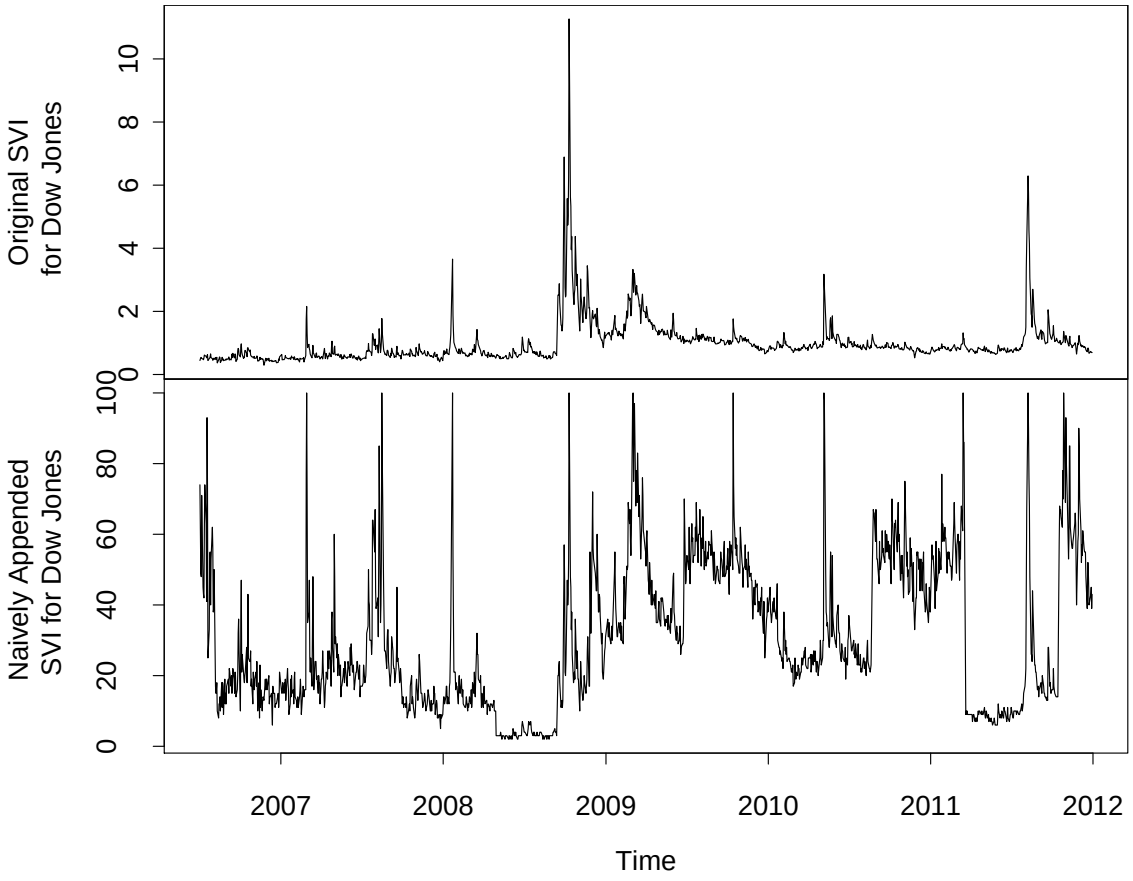
In this chapter, we propose and evaluate an algorithm which allows to knit multi-annual, consistent Google Trends time series. To circumvent the problem that long consistent time series are not directly available, recent research working with Google Trends SVIs is usually based on low frequency time series, i.e., weekly or monthly (Kristoufek 2013, Scott and Varian 2015, Dimpfl and Kleiman 2017) in order to cover longer time spans. To use daily SVI time series, the short samples of 270 days have to be concatenated somehow and different approaches have been proposed in the literature already. We contribute to the literature with a regression-based construction algorithm (RBC algorithm) which relies on a level and scale adjustment of Google SVIs based on sufficient non-zero overlapping search volume data. Based on a mathematical formulation of Google's adjustment procedure, we are not only able to construct consistent, multi-annual time series, but also to compare multiple series among each other.

The existing approaches range from a naive, direct concatenation to methods which are similar to our approach. For example, Panagiotidis, Stengos and Vravosinos (2018b) interpolate the available weekly data points. Garcia, Tessone, Mavrodiev and Perony (2014) re-scale the directly downloadable daily SVI time series by transforming them in such a way that the mean of the daily data series matches the weekly observations. In the online appendix to Garcia et al. (2014), the method is sparsely described and evaluated using a random walk simulation. As already mentioned, another approach is to naively concatenate the downloaded data (e.g. Dastgir, Demir, Downing, Gozgor and Lau 2019). When working with the SVI in levels, this is rather problematic as can be seen in Figure 1.1. Due to the different scaling of the concatenated time frames, the levels are not comparable over time anymore and the time series exhibits jumps at the break points. Some authors argue that the problem is solved by using logarithmic first differences. We will show that this argument does not always hold as the distributional properties are affected if the data are not concatenated carefully before taking first differences.

Kristoufek (2015) uses another methodology to construct SVI time series on a daily frequency. Unfortunately, the concatenation procedure is only described in two sentences: "To obtain daily series for Google searches, one needs to download Google Trends SVI in three months blocks. The series are then chained and rescaled using the last overlapping month." (Kristoufek 2015, p.5) From this brief explanation, we assume that he might have constructed the daily price series similar to our proposed algorithm. Zhang, Wang, Li and Shen (2018) used Google Trends SVIs by applying a similar approach, using an overlapping period of two months.

Figure 1.1: Original SVI and Naively Concatenated SVI

In the top graph the time series of the original SVI is depicted. It was directly downloadable from google.com/trends prior to January, 2011. The bottom graph shows the naively concatenated SVI time series that can be downloaded in time frames of 270 days, as of November, 2019.



Recently, Google added a functionality to compare time frames. Based on this new feature, Chronopoulos, Papadimitriou and Vlastakis (2018) describe an algorithm how to retrieve consistent time series for several years. Hence, we argue that there are two viable alternatives to obtain consistent Google Trends time series for longer time ranges: Either one constructs a consistent time series by reversing the standardization employed by Google as outlined below, or one may use the recently added comparison feature for different time ranges offered by Google itself as described in Chronopoulos et al. (2018). We will refer to this method as the time-frame comparison algorithm (TFC algorithm). We show that our RBC algorithm performs better in situations where Google search volume exhibits unprecedented peaks while average search volume is comparatively low. To this end, we review the two methods and test their accuracy (along with a naive concatenation scheme) using data sets provided by Dimpfl and Jank (2016).

The advantage of our methodology lies in its capability to make multiple SVIs comparable. To date, Google offers the comparison of only up to five different search-terms. Our methodology is suited to override this limitation which turns out to be important for

our application where we predict inflation and consumption. Searches for consumption products conducted via Google are often accompanied by a limiting price (e.g. 10 US dollar). Based on the distribution of multiple price levels, we create an index which reflects the willingness of online buyers to spend money on any product: the index of prices searched online (IPSO). Subsequently, we use the IPSO to forecast inflation and consumption measures for the US and the Euro Area. Using data limited to the US, we show that the index is strongly contemporaneously correlated with monthly US consumption and Granger causes monthly US-inflation, when controlling for seasonality (in-sample). In out-of-sample one-step ahead forecasts for the US, the index is also able to reduce the root mean squared prediction error ($RMSPE$) by around 30% compared to the $RMSPE$ of a benchmark autoregressive process. On a daily frequency, when forecasting the changes of US break-even inflation rates, the index reduces the out-of-sample forecasts by more than 50% compared to the autoregressive baseline model. Cross-checking our results for Europe, we find no evidence in-sample that the index is contemporaneously correlated with the European Harmonized Consumer Price Index (HCPI). Nonetheless, in out-of-sample one-step ahead forecasts for Euro Area inflation, the index reduces the $RMSPE$ by almost 30% compared to an autoregressive baseline model, when controlling for seasonality. As monthly consumption data for Europe are not available, we check the predictive ability of the index against monthly consumer credit growth. The out-of-sample performance of the index in predicting the European consumption measure turns out to be very good as it reduces the $RMSPE$ by 70% compared to the baseline model, while accounting for annual seasonality.

The chapter proceeds as follows. In the following section, we describe Google's SVI and how it is adjusted by Google before publication, and propose an algorithm to construct multi-annual high-frequency Google SVI time series. The section also contains a comparison of our proposed approach to an alternative approach using the time frame comparison offered by Google. In the third section, we lay out the construction of the IPSO and apply it in the forecasting of inflation and consumption measures. The last section summarizes the main results and concludes.

1.1 Constructing Multi-Annual, Comparable SVIs

With Google Trends, Google offers a service that allows to compare the relative popularity of search-terms. There are two important issues to consider when using Google Trends data. The first is concerned with the interpretation of the measure, while the second with data preparation and the construction of multi-annual time series. Both issues are usually treated subordinately in the literature.

Google computes and publishes a SVI that compares the occurrence of searches to the entire volume of searches based on a data sample (Stephens-Davidowitz and Varian 2014). Hence, a falling SVI does not (necessarily) mean that there are less searches than in the past, but it means that a smaller share of searches in the drawn sample is dedicated to the respective search-term. According to *smartinsights.com*[2], Google's total search volume increased from a level of 1.2 billion searches per day in 2012 to about 4.5 billion in 2017. This has a number of implications. First, it shows the importance of Google in the overall internet search market today. Second, leaving the sample variance aside, the SVI numbers provide a useful estimate for the propensity of people searching for a certain query at a given time. Third, one has to bear in mind that the composition of Google users might have changed over time. Thus, in essence, falling relative popularity of a search-term over longer time frames and, therewith, a decreasing SVI, does not necessarily mean that less people were searching for it; it may just mean that a whole lot of (new) Google users were searching for something else at a certain time resulting in a lower SVI even if the exact same number of searches for the respective search-term has been submitted. Thus, the interpretation of the SVI as a proxy for search propensity within a given time frame is useful. As the market share of Google in the search engine market lies around 90% worldwide since 2009[3], relating the volume of search-terms queried on Google to the search-terms of all internet users seems justified.

Triggered by the market penetration of smartphones, a connection to the internet and, thus, to Google's search engine is omnipresent. These trends suggest that the overall number of searches does not fluctuate much from day to day, but rather grows gradually over the years as internet services become more widespread. Taking the sampling variance into consideration, however, the fluctuations for search-terms with a small search volume are large. Also, SVIs are only reported by Google as non-zero if the share indicates a sufficient popularity. Unfortunately, it remains unclear which threshold Google defines as 'sufficient'. For search-terms with a small search volume, the use of SVIs is, thus, limited.

[2] Source: https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/, last accessed: 2017-01-11

[3] Source: https://gs.statcounter.com/search-engine-market-share#monthly-200901-201910, last accessed: 2019-10-10.

Table 1.1: Downloadable Frequencies and Time Frames

The table lists the maximum length of a time frame and the corresponding frequency of Google's SVI downloadable from www.google.com/trends as of October 2019.

Length	Frequency	Earliest available Date
no limit	monthly	January 1, 2004
5 years	weekly	January 1, 2004
270 days	daily	January 1, 2004
7 days	hourly	January 1, 2015
72 hours	16 mins	January 1, 2015
36 hours	8 mins	January 1, 2015
5 hours	1 min	January 1, 2015

Keeping these perils in mind, we suggest to interpret Google's SVI as relative interest in a certain topic only for search-terms that exhibit very few missing or zero values.

Besides this interpretation issue, there are further limitations when using Google Trends. For one, only five search-terms can simultaneously be compared with each other, and, second, the maximum time frame for a download varies with the desired frequency. For example, monthly Google Trends time series are made available without any limitation for the entire history of searches, while daily data can only be obtained for time frames with a maximal length of 270 days. Intraday data are only available for time frames starting on or after January 1, 2015. Table 1.1 provides an overview of the available frequencies (as of October 2019) along with the respective maximum length of the time series and the earliest available date. To construct daily or intraday Google Trends time series for a longer period of time, say several years, multiple SVIs have to be downloaded for smaller time frames and concatenated.

1.1.1 The Rules of Google Trends

Before making the Google Trends time series available to the public, Google standardizes the values of the SVI to the time frame the user wants to download. The values of the SVI are also rounded to integers. Google does not reveal how exactly the standardization is conducted. The description provided on Google's FAQ website[4], and reproduced here for convenience, contains the following three rules (quote):

> *1) Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest.*

[4] Source: https://support.google.com/trends/answer/4365533?hl=en , last accessed: 2019-10-15.

2) *The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics.*

3) *Different regions that show the same search interest for a term don't always have the same total search volumes.*

As Google does not provide a formal mathematical description, we present our understanding of the above explanation in a mathematical framework. Note that rule 3 follows from rules 1 and 2, and we thus only have to look at the first two in detail. Denote by $\mathcal{T}$ the cover of the interval $[0, T]$, which is the overall time frame for which we want to download a SVI time series. Let $n_{j,i,t}$ be the number of searches for search-term $j \in \mathcal{S}$ in region $i \in \{1, \ldots, N_R\}$ in time interval t. $\mathcal{S}$ is the set of all distinct queries searched and N_R is the number of regions. $t \subset \mathcal{T}$ denotes the specific sub-interval for which the SVI is calculated. Depending on the frequency, it may represent a month, a week, a day, an hour etc.

In rule 1, Google formulates the procedure to arrive at some comparable share of total search volume $s_{j,i,t}$. This is the relative propensity of searches for term j in region i for time interval t:

$$s_{j,i,t} = \frac{n_{j,i,t}}{\sum_{j \in \mathcal{S}} n_{j,i,t}} = \frac{n_{j,i,t}}{n_{\cdot,i,t}}.$$

In rule 2, Google describes the standardization procedure applied if the user chooses a set of $\mathcal{M} = \{j_1, j_2, j_3\}$ topics to compare. If the user chooses only one topic j_1 (i.e., $|\mathcal{M}| = 1$) and selects region i_1, this results in an index

$$SVI_{j_1,i,t} = \text{round}\left(\frac{s_{j_1,i_1,t} - L}{\max_{t \in \mathcal{T}}(s_{j_1,i_1,t}) - L}\right).$$

where L denotes the (unknown) threshold which defines the propensity level for which Google deems that insufficient data are available.

Following the description in Google's FAQs, we can extend this definition to a bundle of search-terms, $\mathcal{M}$, in region i during the time frame $\mathcal{T}$. Google's SVI for one specific search-term $j \in \mathcal{S}$ is constructed as

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = \text{round}\left(\frac{s_{j,i,t} - L}{\max_{\substack{m \in \mathcal{M} \\ t \in \mathcal{T}}}(s_{m,i,t}) - L}\right) \tag{1.1}$$

where $s_{m,i,t}$ denotes the relative propensity of searches in region i for time interval t. In this way, the relative ratio between the subjects is preserved.

Ignoring the rounding of the index to an integer number, we can write Google's SVI as an affine-linear transformation of the scaled search propensity as

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = \alpha_{\mathcal{M},i,\mathcal{T}} + \beta_{\mathcal{M},i,\mathcal{T}} s_{j,i,t} + \nu_{j,i,t}, \tag{1.2}$$

where the parameters α and β are given as

$$\alpha_{\mathcal{M},i,\mathcal{T}} = -\frac{L}{\max_{m\in\mathcal{M},t\in\mathcal{T}}(s_{m,i,t}) - L}, \beta_{\mathcal{M},i,\mathcal{T}} = \frac{1}{\max_{m\in\mathcal{M},t\in\mathcal{T}}(s_{m,i,t}) - L}.$$

The rounding error $\nu_{j,i,t}$ can be assumed to be independently and identically distributed (i.i.d.). In particular, it is independent of the total search volume $s_{j,i,t}$.

Even though Google limits the length of the time frame which the user is allowed to choose, the structure of the SVI as outlined in Equations (1.1) and (1.2) allows to construct a consistent multi-annual SVI of arbitrary length based on downloading overlapping SVIs. To do so, one can exploit the linear relationship between the SVIs obtained for two time frames $\mathcal{T}$ and $\mathcal{T}'$ for the same point in time $t \in \{\mathcal{T} \cap \mathcal{T}'\}$, which is formally described as

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = \gamma + \delta SVI_{j,i,t|\mathcal{M},\mathcal{T}'} + \varepsilon_{j,i,t}. \tag{1.3}$$

δ and γ are the parameters of this linear relation and clearly depend on the region i, the time interval t as well as the time frames $\mathcal{T}$ and $\mathcal{T}'$ as well as sets of simultaneously downloaded search terms $\mathcal{M}$ and $\mathcal{M}'$. For simplicity, all these dependencies are suppressed in the notation of Equation (1.3).

Again, the rounding error $\varepsilon_{j,i,t}$ is assumed to be i.i.d. More details on the derivation of Equation (1.3) are provided in the appendix. We will use this linear relationship in the next section to construct consistent multi-annual Google Trends time series.

1.1.2 Linear Regression and Evaluation

As there is little explanation made available by Google on how the SVI is calculated exactly, and since the scientific literature that uses daily Google Trends SVIs is rather unconcerned with a detailed explanation of constructing coherent time series, we deem it necessary to clearly describe how we arrive at our algorithm. We assume, according to the description Google provides, that Google adjusts the search volume according to Equation (1.1) for a single search-term.

Another possibility, used by Google up to the end of 2011, is to standardize the time series of search volume index values. To distinguish this standardization approach, we denote the resulting index with $v_{j,t}$, for some search-term j for some interval $t \in [t_0, \mathcal{T}]$. Back

then, Google subtracted the mean $\mu_{t_0,\mathcal{T}}$, divided by the standard deviation $\sigma_{t_0,\mathcal{T}}$ of the number of searches within a certain time frame. Google then transformed the series to unit mean $\bar{\mu} = 1$ and unit standard deviation $\bar{\sigma} = 1$ to obtain the index

$$v_{j,t} = \frac{n_{j,t} - \mu_{t_0,T}}{\sigma_{t_0,T}} \bar{\sigma} + \bar{\mu}. \tag{1.4}$$

We know that Google made SVIs available according to Equation (1.4) in 2011[5]. Back then, the user could choose on which time frame the mean $\mu_{t_0,T}$ and standard deviation $\sigma_{t_0,T}$ would be calculated on. In 'relative mode', mean and standard deviation were calculated on the chosen time frame $[t_0, T]$, whereas in fixed mode the user could choose a reference time period $[\tau_0, \tau_1]$. The fixed mode allowed the construction of multi-annual, consistent time series. Unfortunately, this is not the case anymore and only (another form of the former) 'relative mode' is available which, in our understanding, can be formalized by Equation (1.1).

Due to Equation (1.3), however, we can knit separately scaled time series that are downloadable from Google together if there are overlapping points in the data sets. In theory, two overlapping points in time would suffice to identify the parameters γ and δ in Equation (1.3). Since the relationship only holds approximately, we suggest at least 30 overlapping days. We estimate the parameters via standard ordinary least-squares (OLS) regression. If the overlapping points contain a lot of zeros in both sets, an even longer overlapping period is advisable. In our algorithm, we require that there are at least 30 days in the overlapping window where at least one of the two data sets has a non-zero value. Furthermore, we require that within the overlapping time period each of the two data sets taken alone exhibits at least 20 non-zero values.

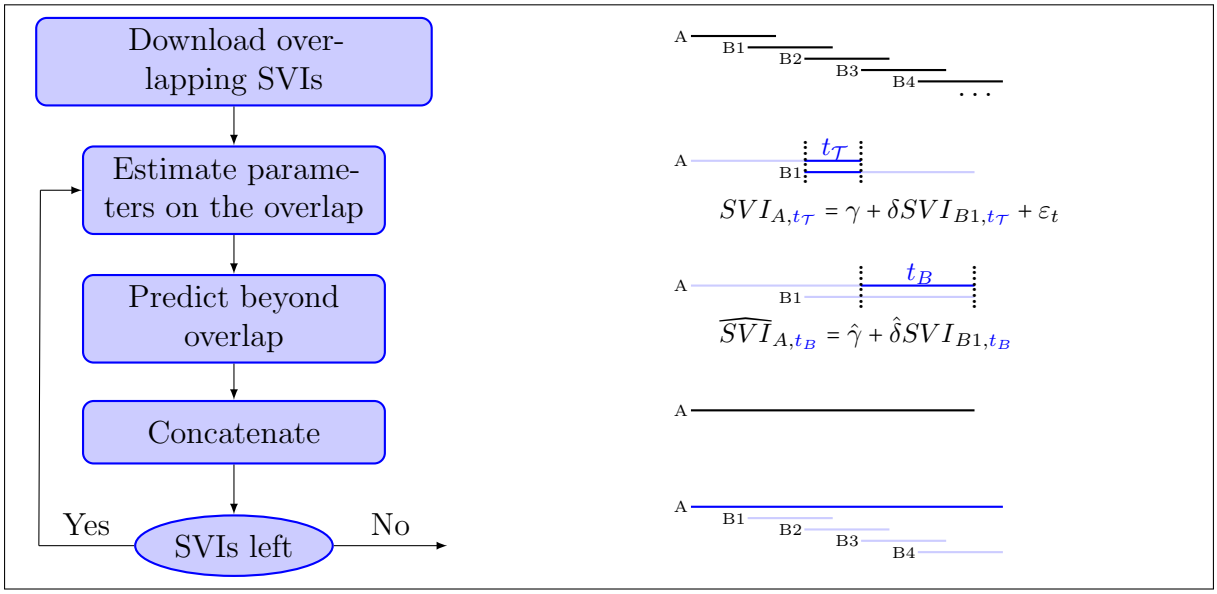
According to whether we start with the youngest or oldest time frame when knitting the time series together, we distinguish between the *backward* and *forward* method. Furthermore, for each concatenation step, i.e., each time Equation (1.3) is used, we can test whether our estimate for the constant parameter γ is statistically significant on a 5% significance level. To calculate the test statistic, we use robust standard errors. If the null hypothesis is not rejected on the 5% significance level in a two-sided test, we can choose to re-estimate the linear relationship based on the model

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = \delta SVI_{j,i,t|\mathcal{M},\mathcal{T}'} + \varepsilon_{j,i,t}. \tag{1.5}$$

[5] Source: Question 8 on https://web.archive.org/web/20101229150233/http://www.google.de:80/intl/en/trends/about.html (Last access: February 13, 2018.)

Figure 1.2: The Regression Based Construction Algorithm

The figure illustrates the forward method of the regression based construction algorithm.



The regression-based construction algorithm can be summarized in the following steps:

1. Download 270-day SVI data sets from Google for the time period of interest. Make sure that each two subsequent data sets overlap by at least 30 non-missing values.

2. Estimate Equation (1.3) on the overlapping data points (do not exclude zeros). Begin with the two data sets containing the youngest (backward method) or oldest (forward method) SVI observations for a search-term. We call the data set containing the starting point A and denote the values in it with $SVI_{j,i,t|\mathcal{M},\mathcal{T}_A}$. The subsequent 270-day data set is called B and the SVI values in it are denoted with $SVI_{j,i,t|\mathcal{M},\mathcal{T}_B}$.
   Test if the hypothesis for the intercept $H_0 : \hat{\gamma} = 0$ can be rejected. If so, keep estimates for Equation (1.3). If not, estimate Equation (1.5).

3. Predict the $SVI_{j,i,t|\mathcal{M},\mathcal{T}_A}$ out of sample (over the time range of $SVI_{j,i,t|\mathcal{M},\mathcal{T}_B}$ without the overlap) by using the estimates $\hat{\gamma}$ and $\hat{\delta}$ for the relation in Equation (1.3) or only $\hat{\delta}$ if Equation (1.5) is used.

4. Concatenate the original $SVI_{j,i,t|\mathcal{M},\mathcal{T}_A}$ and the predicted values $\widehat{SVI}_{j,i,t|\mathcal{M},\mathcal{T}_B}$ to one data set. This data set takes the place of data set A whereas B is replaced with the next data set to be attached.

5. Repeat steps 2 to 4 until there are no further data sets left.

Figure 1.2 summarizes the steps of the algorithm (left) and illustrates the implementation in an abstract way (right).

Table 1.2: Correlations of Constructed and Original SVI

The table reports the correlation coefficients of the RBC SVI using the respective method with the original search volume index as downloaded in 2012 by Dimpfl and Jank (2016).

	With Intercept		Optional Intercept	
Index	forward	backward	forward	backward
CAC	0.9786	0.9777	0.9813	0.9804
DAX	0.9578	0.9758	0.9704	0.9779
DJIA	0.9911	0.9854	0.9913	0.9886
FTSE	0.9471	0.9610	0.9642	0.9615

We have two options to evaluate the accuracy of our proposed algorithm. First, we compare a so-constructed data set to a data set which was obtained from Google when immediate concatenation was still possible. Second, we can aggregate the RBC SVI to a lower frequency and compare it to an SVI on this frequency obtained directly from Google.

The first option relies on the data sets used by Dimpfl and Jank (2016). In 2011, when the authors collected the data, it was possible to download Google Trends SVI scaled to a fixed reference date and simply string them together. Back then, the SVI was also not rounded. Dimpfl and Jank (2016) downloaded data sets for the search-terms *CAC* (related to the French stock index CAC40), *DAX* (related to the German stock market index), *Dow Jones* and *FTSE* (related to the British Financial Times Stock Exchange Index). The data cover Google's SVI from July 3, 2006 until January 30, 2011 for searches originating from the country in which the respective market is located.

For the construction of the SVI from currently accessible Google Trends time series, we downloaded 24 separate data sets ranging back until 2004. Each data set contains 270 days and overlaps with the previous data set in at least 30 non-zero observations. We use the data from Google Trends based on searches originating from the country in which the respective index is located. The timezone is fixed to UTC+1.[6]

As we can either use the forward or the backward method, and choose to always include an intercept or only if it is found to be statistically significant, we have 4 options to construct the time series. Table 1.2 reports the correlation coefficients of the 4 methods with the benchmark SVI times series. For all methods and search-terms, we find correlation coefficients larger than 0.94. It turns out that we can increase the accuracy of the RBC SVI time series by only optionally including the intercept parameter in the estimation.

Figure 1.3 compares the forward (upper panel) and backward (lower panel) RBC SVI for the search-term *Dow Jones* when we always include an intercept to the benchmark

[6] With the HTTP-request to Google Trends, a parameter `tz` is set to −60 if the request is made from Germany which corresponds to a time-zone offset of 1 hour. We extended the `gtrendsR`-package available for `R` to include the possibility to fix the time zone.

Table 1.3: Correlation Between Naively Concatenated and RBC SVI with the Original SVI

The table presents the correlation of the naively concatenated and the RBC SVI with the original SVI in levels and returns. The RBC SVI is calculated using the backward method including an intercept. The biased returns are dropped from the naively concatenated SVI. The backward method including an intercept consistently exhibits a higher correlation with the original SVI than using naively concatenated SVI returns. For the backward method with optional intercept, this is not always the case. When considering levels, the correlation of the RBC SVI with the backward method and optional intercept has a high correlation with the original SVI.

	In Levels		Returns	
Index	RBC	Naiv	RBC	Naiv
CAC	0.9777	0.2432	0.5078	0.4584
DAX	0.9758	0.2628	0.6358	0.5961
DJIA	0.9854	0.4036	0.7294	0.6496
FTSE	0.9610	0.2285	0.5837	0.5374

time series. Figure 1.4 compares the two methods when the intercept is only optionally included when it turns out statistically significant in step 4 of the algorithm. Comparing Figures 1.3 and 1.4, as well as Table 1.2, we can see that for the search-terms *CAC*, *DAX*, *Dow Jones* and *FTSE*, all the methods perform well, but it seems admissible to use the intercept only for concatenation if it is statistically significant.

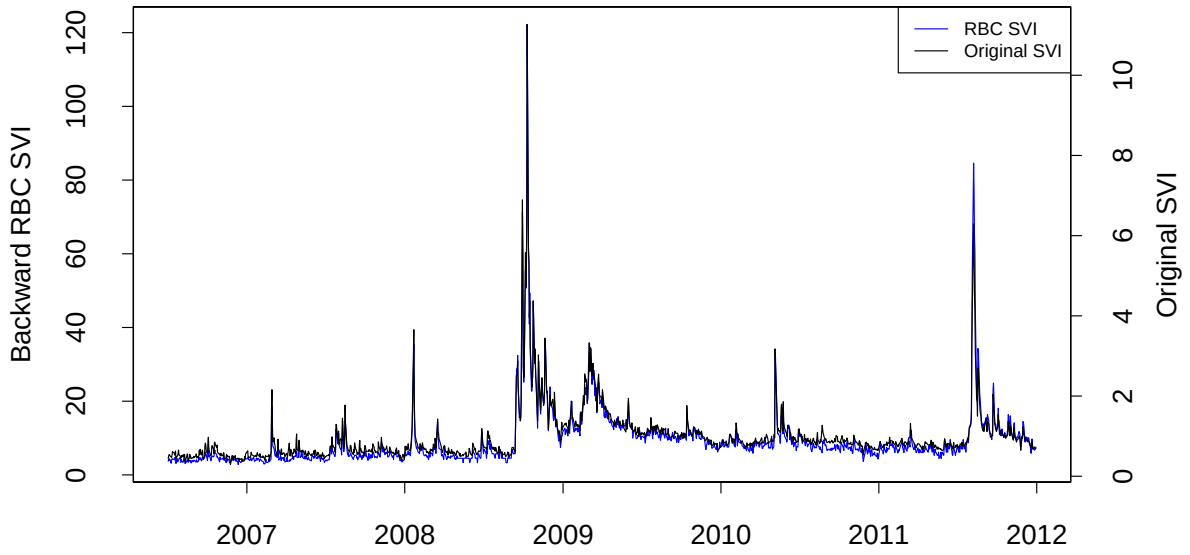
When using SVIs in empirical work, usually the logarithmic growth rates of the SVI or logarithmic first differences are used. To evaluate our method, we therefore report in Table 1.3 the correlation between levels and first differences of the original SVIs of Dimpfl and Jank (2016), the RBC SVI, and a naive concatenation where downloaded series are attached to each other without adjustment. We interpret a correlation coefficient smaller than 1 as a measure for the loss of information from the construction of the index.

As can be seen, the correlation between our RBC index in levels and the original one is very close to one. In contrast, the naive concatenation comes at the cost of a huge loss of information. This is in line with Figure 1.1 which shows that the naive concatenation method results in an SVI time series which does not correspond to the original SVI series at all. When using returns, the backward RBC SVI (with intercept) consistently exhibits a higher correlation with the original time series than the naive SVI log-returns.

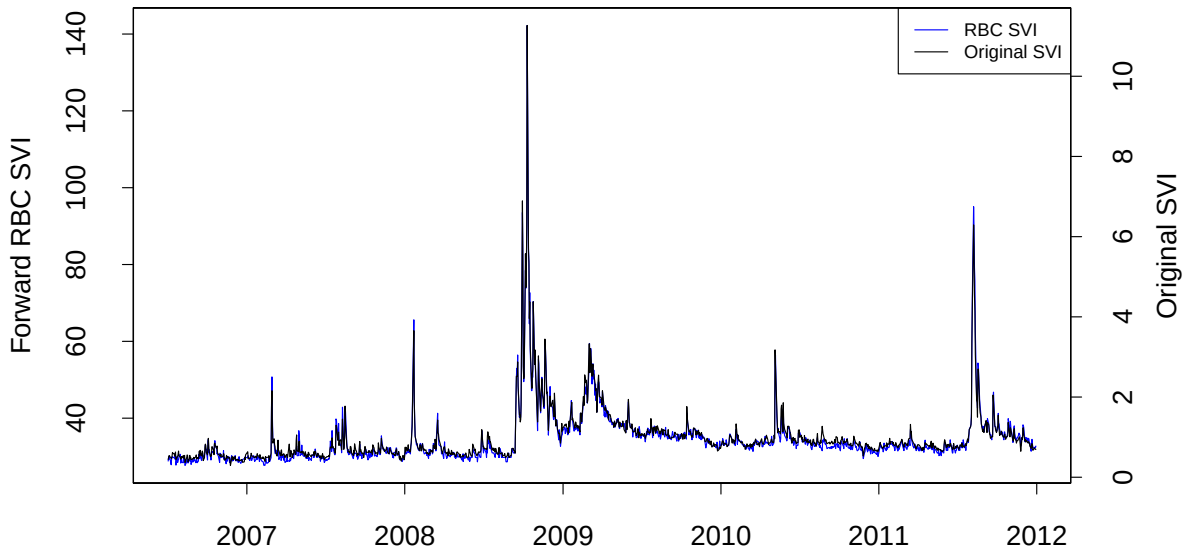
In order to evaluate whether our proposed regression-based construction method preserves the statistical properties of the SVI, kernel densities and moments based on log-returns of the original SVI, log-returns of the RBC SVI as well as log-returns from the naively concatenated SVI are calculated. The kernel densities are displayed in Figure 1.5. For the return series, it turns out that constructing the SVI backwards and always including an intercept is the best choice for all series as this kernel density is closest to the one of the original data. The naive concatenation always results in the worst approximation of

Figure 1.3: Comparison: RBC SVI and Original Google SVI – Search-Term *Dow Jones*

Google's original SVI index as downloaded on 30-1-2011 (right scale, black line) compared to the RBC SVI based on currently available data (left scale, blue line). For the construction, a linear transformation is used that always contains a constant.



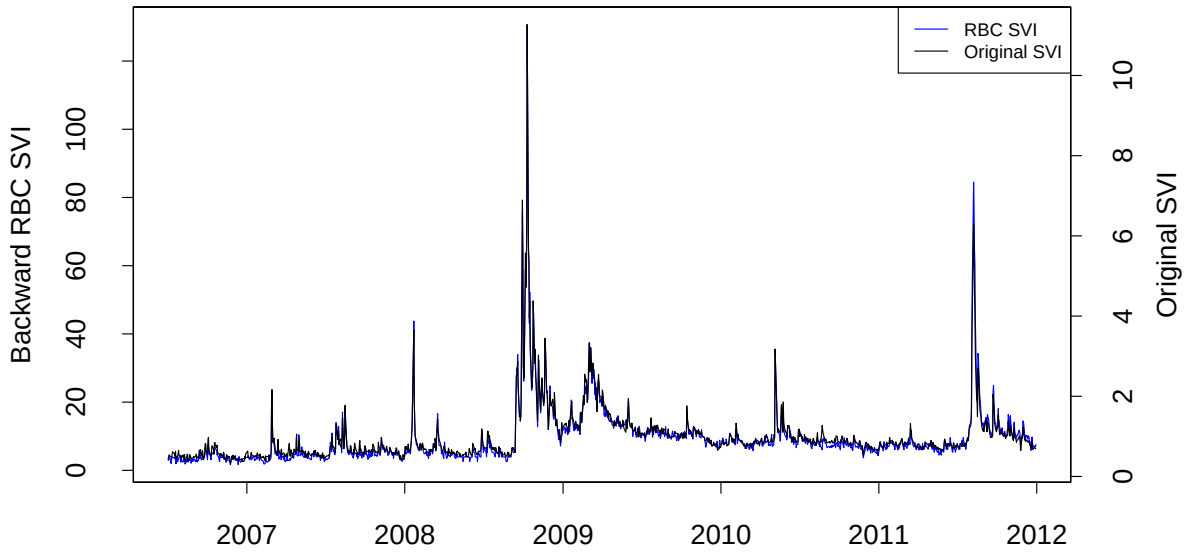
(a) SVI for Search-Term 'Dow Jones' Backwards Constructed



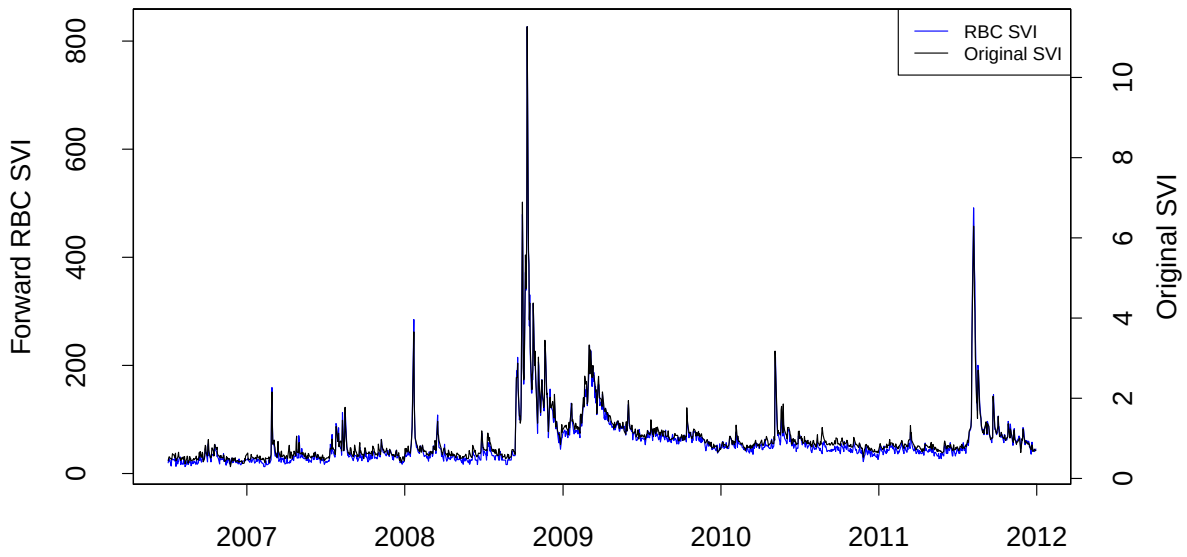
(b) Search-Term 'Dow Jones' Forward Constructed

Figure 1.4: Comparison
of RBC SVI (Optional Intercept) and Original Google SVI – Search-Term *Dow Jones*

Google's original SVI index as downloaded on 30-1-2011 (right scale, black line) compared to the RBC SVI based on currently available time series (left scale, blue line). When constructing the SVI, in this case we excluded the constant from the linear transformation, when we were not able to reject the hypothesis $\gamma = 0$ based on a t-test with robust standard errors.



(a) Backward RBC SVI Compared to Original SVI



(b) Forward RBC SVI Compared to Original SVI

Table 1.4: Moments of the Original, Naive and RBC SVI

The table displays the mean μ, standard deviation σ as well as the skewness and kurtosis of the returns of the original SVI (Original) and of the backward regression-based constructed (RBC) for various search-terms. When constructing the SVI returns backwards, an intercept is always included. The third line (Naive) presents the moments, if returns are calculated on a naively concatenated SVI time series. As the naively, concatenated SVI simply chains data time frames of 270 days together, the fourth line (Naive Ex.) tables the moments, if the biased inter-time-frame-returns are excluded from the naively concatenated time series.

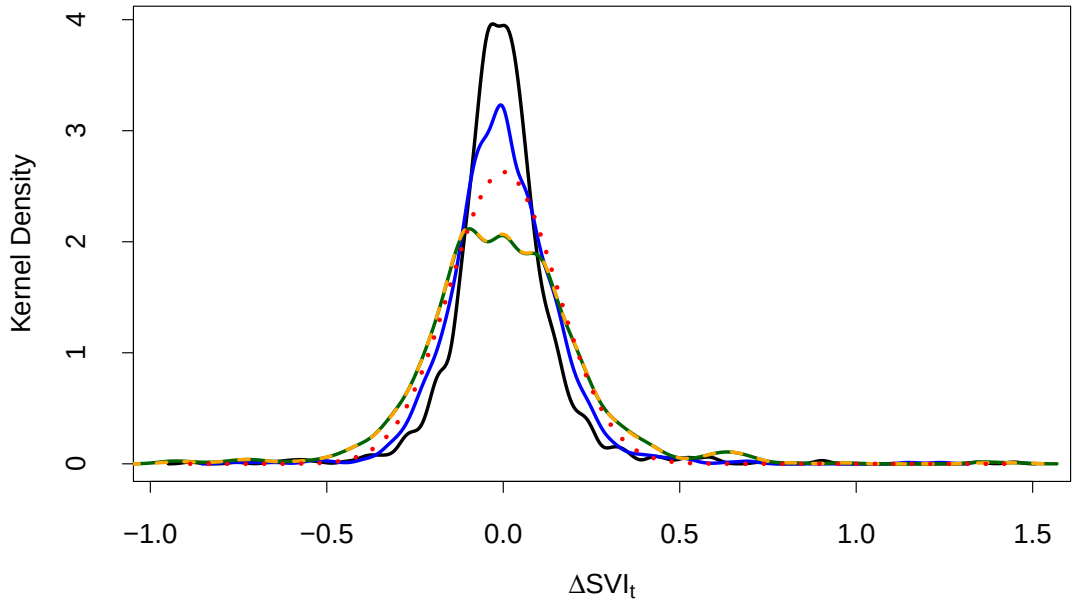
Query	Series	μ	σ	Skewness	Kurtosis
CAC	Original	0.00	0.15	0.84	9.60
	RBC	0.00	0.15	0.60	7.27
	Naive	-0.00	0.26	0.18	4.40
	Naive Ex.	-0.00	0.25	0.15	4.01
DAX	Original	0.00	0.15	1.51	19.12
	RBC	-0.00	0.15	0.81	10.38
	Naive	-0.00	0.23	0.28	9.28
	Naive Ex.	0.00	0.22	0.53	7.61
DJIA	Original	0.00	0.17	1.67	15.57
	RBC	0.00	0.20	0.95	9.53
	Naive	-0.00	0.27	0.43	10.60
	Naive Ex.	0.00	0.26	0.75	8.95
FTSE	Original	-0.00	0.16	1.52	14.73
	RBC	-0.00	0.14	0.60	7.72
	Naive	-0.00	0.25	0.41	5.90
	Naive Ex.	-0.00	0.24	0.40	5.43

the original data, even if returns across the border points at which adjacent time frames are concatenated are excluded. The comparison of moments is presented in Table 1.4. The means of the logarithmic growth rates of the original as well as all RBC/naive SVIs are centered around zero. However, the log-returns of the naive SVI are (in some cases decisively) more volatile. Also, naive concatenation reduces skewness and kurtosis by much more than our proposed algorithm, alienating the distributional properties further from the original data. Considering volatility, skewness and kurtosis together, the returns from the backward RBC SVI (with intercept) reflect the moments of the original SVI best and in particular much better than the returns from the naively concatenated SVI.

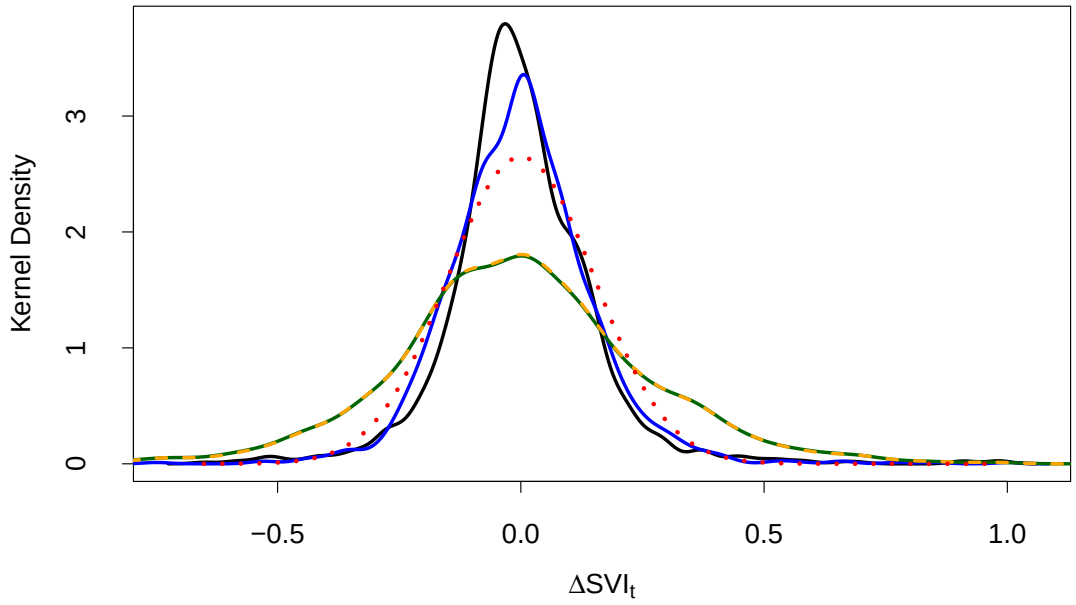
Based on all the criteria above, we conclude that the regression-based construction of the SVI according to our algorithm is sensible and useful. It is able to mimic the statistical properties which a hypothetical time series that Google could provide might have. This is most important if the data are to be used in levels (which is often the case in forecasting applications). If first differences are used, our methodology still performs better than a naive concatenation, but the differences are not as pronounced any more as for the levels.

Figure 1.5: Density Comparison of the Logarithmic Growth Rates of SVIs

This figure compares the kernel density of the logarithmic growth rates of Google's original SVI as downloaded on 30-1-2011 (black line) to the kernel density of the logarithmic growth rates of the RBC SVI based on currently available data (blue line). For the construction, the backwards method is used. The density of a normal distribution with the same mean and standard deviation as the original SVI is displayed with a dotted red line. In green, the kernel density estimation for the naively concatenated SVI returns are displayed, which is almost identically with the naively kernel density estimate of the concatenated SVI returns without the biased inter-time-frame-returns. The latter is depicted by the orange dashed line.



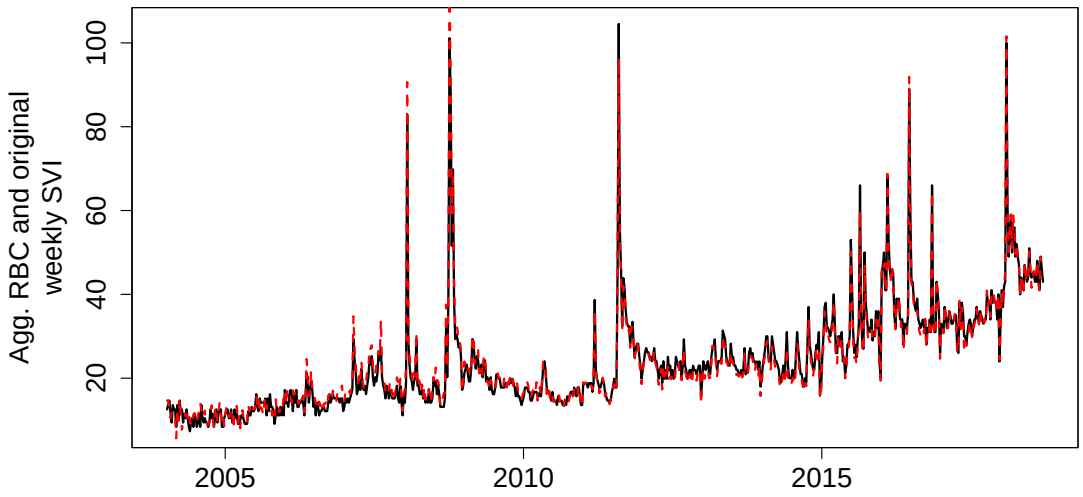
(a) Kernel Density of the SVI for the Search-Term *DAX*



(b) Kernel Density of the SVI for the Search-Term *CAC*

Figure 1.6: Comparison of Original and RBC Weekly SVI – Search-Term "DAX"

The graph compares Google's original weekly SVI (black line) and our transformed, aggregate weekly RBC SVI (red line) for the term "DAX".



As Google makes SVI time series available for longer time horizons with weekly resolution and in order to evaluate the RBC algorithm with another data set directly obtained from Google, we aggregate our constructed time series to a weekly frequency. For this comparison, we limit ourselves to the SVI which turned out most accurate in the evaluation above, i.e., the SVI based on the backwards construction with optional estimation of the intercept. We aggregate it by taking the weekly sum of the daily observations.

After this aggregation step we still need to account for the scaling of the time series. Therefore, we regress the downloaded weekly time series on the aggregated RBC SVI and calculate the fitted values. The success of the method is illustrated in Figure 1.6 for the DAX in which fitted values and the downloaded SVI series are shown. The two time series can almost not be distinguished by the naked eye. The high fit is also supported by the high R^2s that result in the auxiliary regressions (not reported). These are above 98% for all considered search-terms.

1.1.3 Time Frame Comparison

With the recently added functionality of comparing SVI values over different time frames, Chronopoulos et al. (2018) suggest an algorithm to concatenate Google's SVI over different time frames to a consistent time series. Google Trends allows to download (up to five) different time ranges for comparison. All values are then scaled to the maximum search intensity within the (up to) five selected time ranges. Hence, once the time frame with the

highest intensity is found, one can download other time frames and compare them to this time frame. This ensures that the reference point for Google's standardization is fixed and a consistent time series can be produced.

Following Chronopoulos et al. (2018), we employ the time-frame comparison algorithm (TFC algorithm) to download a multi-annual time frame T as follows:

1. Download monthly data since 2004 and find the maximum SVI within T.
2. Construct a time frame, A, that contains weekly data around the month with the maximum value within T.
3. Download up to 4 other (adjacent) different time frames (B, C, D, E) on a weekly frequency at the beginning of T together and in comparison to A. Concatenate them.
4. Check if the maximum value in the downloaded data is still in A. If the maximum value is in A, substitute the four time ranges, B, C, D, E, with new time ranges and concatenate the result. If the maximum value is not in A, substitute A with the time range that includes the maximum value and start anew at step 3. Repeat this step until the entire time period T is covered.
5. Find the maximum SVI in the concatenated data set that spans over T.
6. Start with step 2 again. Use the next lower frequency (daily, hourly, 16min, 8min, 1min).

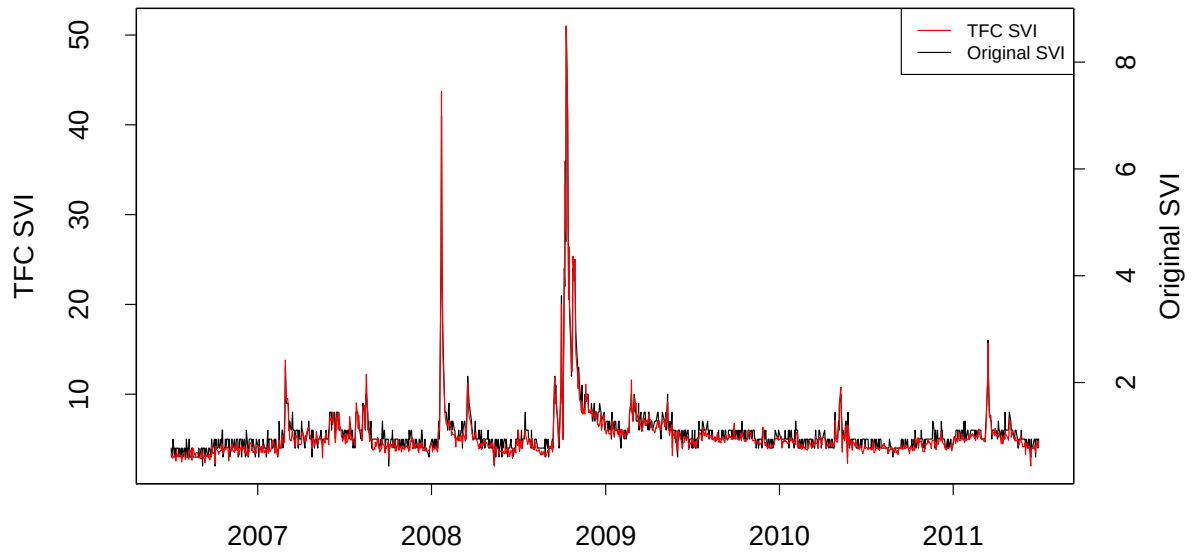
The method works for most search-terms which is illustrated in Figure 1.7 for the search-term "DAX". Even though the concatenated time series only contains integers between 0 and 100, it exhibits the same dynamic as the original SVI downloaded in 2011 that allowed for non-integer values.

However, an issue with the TFC method arises when there was a (short) hype for a certain query such as for the search-term *Bitcoin*, for example. Since all values are scaled, in the case of a hyped query, dynamics in periods with low search volume are set to zero by the concatenation method employed by Chronopoulos et al. (2018). In essence, we find that if the attention for a subject spikes at some point over several years, the dynamics at lower values of the SVI is diluted by Google's scaling between 0 to 100 and the unknown threshold value L (cp. Equation (1.1)). This pathological case is illustrated in Figures 1.8a and 1.8b.

When plotted for multiple years (Figure 1.8a), the TFC SVI very often takes a value of 0. In Figure 1.8b, where the TFC SVI time series (black line) is compared to the original SVI time series (blue line) that can be directly downloaded from Google for the time frame from April 5, 2012 until October 31, 2012, the problem becomes even clearer. The black line of the concatenated time series only spikes at the end of August and at the mid of

Figure 1.7: Comparison of Original and RBC SVI – Search-Term "DAX"

The figure compares Google's original weekly SVI for the search-term "DAX" (black line) and a concatenated SVI using Google's time frame comparison (red line), based on the methodology suggested by Chronopoulos et al. (2018).



November 2012. At all other dates it remains zero. Hence, it either decisively over- or underestimates the relative search propensity. It also does not provide information about its dynamics which is the key to many forecasting applications. For comparison, we include in Figure 1.8b also the RBC SVI time series using our methodology presented above (red line). As can be seen, the RBC SVI captures the dynamics of the original SVI very closely. This gives us reason to claim that our algorithm is a robust method which can deal with these kind of pitfalls that may come up when constructing consistent, multi-annual SVIs.

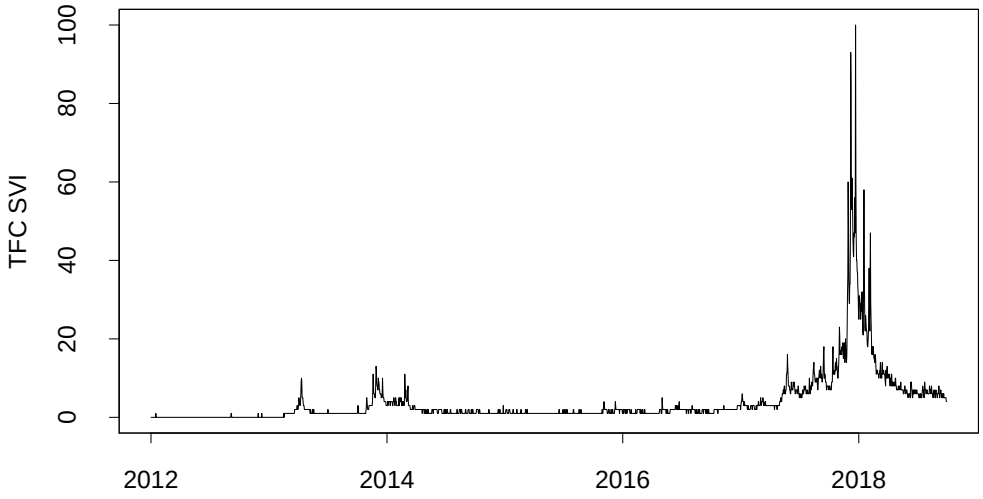
1.1.4 Coherent Scale

The algorithms presented and evaluated above work for up to five search-terms since Google limits the comparison to five search-terms. In order to assess the relative popularity of one search-term with more than five, we can again use the linear relation of the search-terms of Equation (1.3). The structure of the SVI allows to circumvent Google's limitation and to compare an arbitrary number of search-terms.

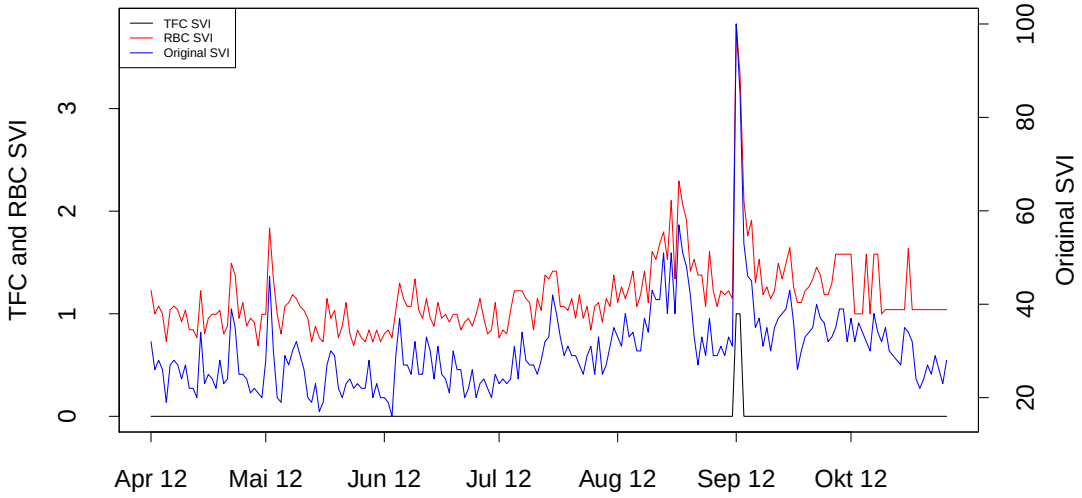
For this purpose, suppose we have two sets of search-terms $\mathcal{M} = \{j_1, j_2, j_3, j_4, j_5\}$ and $\mathcal{M}' = \{j_1, j_6, j_7, j_8, j_9\}$. In order to convert all SVIs for all of these search-terms into a coherent scale, we proceed as follows:

Figure 1.8: Comparison Original SVI and TFC SVI

The top graph presents the concatenated SVI using Google's time frame comparison (TFC) for the search-term *Bitcoin*. The TFC SVI is plotted over several years. As can be seen, for periods prior to 2013 the series exhibits no dynamics and is often zero or close to zero. However, the original time series within 2012 is far from being constant, as can be seen in the bottom graph, where the original time series from 05-04-2012 until 31-10-2012 is compared on a daily basis to the TFC SVI and the RBC SVI. In the bottom graph, the TFC as well as the RBC SVI use the same scale giving the impression that they fall apart. The important thing is, however, that they exhibit a rather strong co-movement.



(a) Multi-Annual TFC SVI



(b) Original, RBC and TFC SVIs During 2012

Table 1.5: Comparison of Concatenation Procedures

The table compares our regression-based algorithm (RBC) to construct multi-annual time series with a naive concatenation method (naive) and the time-frame comparison approach used by Chronopoulos et al. (2018) (TFC). ✓ means that the respective feature is fully available, with a (✓) limited availability is indicated, and × signifies that the feature is not supported.

	RBC	Naive	TFC
Produces consistent time series	✓	×	(✓)[1]
Handles outliers	✓	✓	×
Construct useful level data	✓	×	(✓)[1]
Construct useful return data	✓	(✓)[2]	(✓)[1]
Comparable Searchterms	✓	×	(✓)[3]

[1] Feature is restricted if search-term is subject to a short hype.

[2] Distributional properties are partially matched.

[3] Only up to 2 search-terms can be compared.

1. Estimate the following regression model

$$SVI_{j_1,i,t|\mathcal{M}',\mathcal{T}} = \gamma + \delta SVI_{j_1,i,t|\mathcal{M},\mathcal{T}} + \varepsilon_{j,i,t},$$

   and calculate the fitted values $\widehat{SVI}_{j_1,i,t|\mathcal{M}',\mathcal{T}}$.

2. Calculate the ratios

$$R_{j,i,t} = \frac{SVI_{j,i,t|\mathcal{M},\mathcal{T}}}{SVI_{j_1,i,t|\mathcal{M},\mathcal{T}}} \quad \forall t \in \mathcal{T}, \forall j \in \mathcal{M}.$$

3. Multiply the ratios with the fitted values from step 1 to obtain the adjusted SVIs converted to the scale used in $\mathcal{M}$.

Note that if $s_{j,i,t} \gg L$ and $s_{j_1,i,t} \gg L$, $R_{j,i,t}$ is approximately equal to the ratio of searches.

$$R_{j,i,t} = \frac{s_{j,i,t} - L}{s_{j_1,i,t} - L} \approx \frac{s_{j,i,t}}{s_{j_1,i,t}} = \frac{n_{j,i,t}}{n_{j_1,i,t}}.$$

In any case, $R_{j,i,t}$ can be interpreted as the ratio of distances of the search propensity to the threshold value L.

To compare the different concatenation methods presented in this chapter, Table 1.5 summarizes the supported features of each of these methods.

1.2 Empirical Application: Predicting Inflation and Consumption

1.2.1 The Index of Prices Searched Online (IPSO)

As argued in the introduction, search queries should reflect information demand of individuals, i.e., of consumers. When it comes to consumption decisions, they may be interested in a product, but may not be willing to pay more than a certain amount. Hence, they might gather product related information and also look for stores (online or offline) where the price threshold is respected. If they are successful, they may subsequently engage in buying a product. Hence, search queries with a limit price (irrespective of the product) bear information on consumption behavior of online users and, in addition, about their expectation of future prices, i.e., inflation.

Recently, D'Acunto, Malmendier, Ospina and Weber (2019) report that based on their daily errands, consumers infer individual inflation expectations. If at least some consumers use Google and search for prices online, these individuals reveal information about either prevailing prices and/or their willingness to pay. The share of consumers that search for prices online and have searched for \$10 compared to those that have searched for \$50 should, thus, contain information on the current price level (online) and/or the aggregate willingness to pay (online).

The methodology to construct comparable multi-annual time series presented in the preceding section bears the possibility to construct an *Index of Prices Searched Online* (IPSO). In this context, it is also interesting that Google offers the possibility to explore the most common queries submitted from users who searched for a given search-term. This functionality is called *related queries*. For our purpose, we can deduce from this functionality that when people search specific prices online, e.g. *\$1*, they usually want to find a product that is available for less or exactly \$1. Table 1.6 lists the top 25 search queries related to the search-term *\$1*. Instead of the official symbol for US-dollars, people may also use the word *dollar* or the currency acronym *USD*. With these other dollar references, the related queries function of Google shows that users also search for specific price levels when they would like to know the exchange rate of the specified amount into another currency. Anecdotal evidence also suggests that such searches are often connected to a product, found online, of which the price is only given in dollar, but the user is more acquainted with another currency. We do not expect that currency traders or other professionals that regularly exchange large amounts of money use Google to gather information about exchange rates.

Table 1.6: Related Queries to $1 in the US

The table lists the top 25 related queries on Google to the search-term $1 originating in the US.

$1 million	$1 coin value
$1 movies	$1 drinks
$1 coin	stocks under $1
$1 bill	summer movies 2019
applebees $1	$1 books
applebee	$1 summer movies 2019
$1 movie	$1 taco bell
$1 store	$1 pizza
regal $1 movies	$1 theater
applebees $1 drink	krispy kreme
$1 tree	krispy kreme $1 dozen
$1 in pesos	regal $1 summer movies
$1 silver certificate	

As the search behavior of Google users when searching for prices is related to consumption decisions (especially for price levels below 10,000 Dollar or Euro), we can take Google's SVI and construct an expected-value-like index of the price levels searched online and relate it to consumption and inflation. The construction of this expected value is possible, since for each point in time, Google's SVIs preserve the relative popularity of a search-term or at least it's distance to the threshold value. In other words, the SVI is proportional to the search propensity (minus the threshold value) of a certain search-term and only differs from the propensity by a normalizing factor, i.e.,

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} \propto (s_{m,i,t} - L).$$

For the price levels $\mathcal{P} = \{$ 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100; 200; 300; 400; 500; 600; 1,000$\}$, having at hand a set of identically scaled SVIs, denoted $\mathcal{M}_P$, for each point in time t, we can construct an approximation of the distribution function for the probability that the price level a Google user searches is smaller than p. Careful examination of the SVIs related to price levels above 1,000 has led us to the conclusion that the data quality of the SVIs used is paramount. For that matter SVIs for price levels above 1,000, which are often below the threshold or were not part of Google's sample within a specific day or month, exhibit a lot of missing values and had to be excluded from the analysis. These SVIs for sparsely searched price levels would cause large jumps in the IPSO since they have positive values only on a few days or months, when searches reach the threshold. Also, price levels of above 1,000 denoted in Euro or Dollar are supposedly less likely connected to consumption products and might have less informational content

with respect to inflation or consumption prediction. There are also practical reasons as for very large six-digit price levels the search-terms become to long for the Google Trends API.

In order to approximate the cdf from the SVIs, we first calculate for all $p \in \mathcal{P}$ a discrete (empirical) distribution function over our set of prices $\mathcal{P}$

$$F(p) = \text{Prob}(P \leq p) \approx \frac{\sum_{s=1}^{p} SVI_{j_s,i,t|\mathcal{P},\mathcal{T}}}{\sum_{s \in \mathcal{P}} SVI_{j_s,i,t|\mathcal{M}_P,\mathcal{T}}}. \tag{1.6}$$

In a second step, we approximate the probability density function by a discrete probability mass function for our set of prices $\mathcal{P}$. Therefore, we calculate the differential quotient from the points obtained by Equation (1.6) as

$$f(p) \approx \frac{\Delta F(p)}{\Delta p}.$$

For these calculations, we insert an extra point and assume that zero prices are searched with probability zero.

In a third step, we then calculate the expected value of prices searched online through the discrete probability mass function which we call the index of prices searched online

$$\text{IPSO} = \mathbb{E}[P] \approx \sum_{p \in \mathcal{P}} p f(p).$$

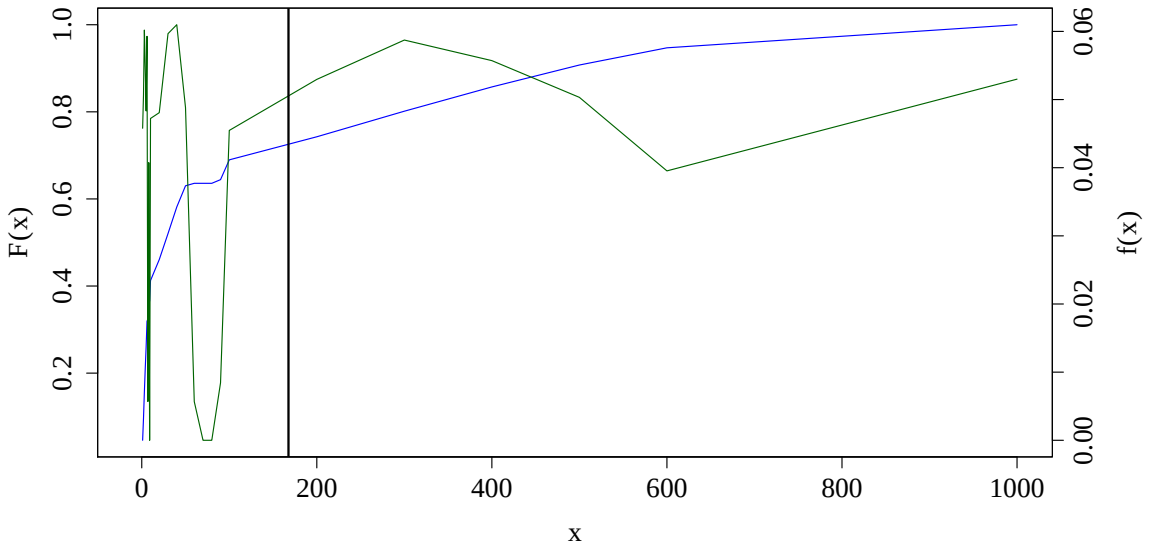
Figure 1.9 shows the estimated cdf in blue as well as the corresponding linearly approximated density function in green. The estimated expected value is shown as a vertical black line. Such an approximated density function would assume that searches for prices are equally likely between the discrete price levels. However, online users do not search with equal probability for price levels between the levels included in our analysis. To make this more clear, online users are less likely to search for \$1.54 than \$1 or \$2. As rounded values are usually searched by online users, we assume that our discrete version provides a sufficient approximation to the true expected value of online searched prices. The fact that the SVIs for higher price levels exhibit a lot of missing values speaks for a neglectable probability mass related to the excluded upper tail of the price distribution.

The set of coherent SVIs, denoted $\mathcal{M}_p$, which we use to construct the IPSO, is based on a bundle of search-terms. Since US users may reference the currency searched for with different symbols, names or abbreviations, we construct for each price level $p \in \mathcal{P}$ a search-term as follows:

$$\$p + p\$ + p\ \$ + p\text{usd} + p\ \text{usd} + p\ \text{dollar} - \text{million} - \text{billion}$$

Figure 1.9: Empirical Distribution Function

The graph shows the empirically approximated distribution function for searched Dollar amounts on April 30th, 2018, in blue and the related density function in green. The black vertical line is the estimated expected value.



Note that + is the OR-operator when using Google Trends while − is the exclusion operator.[7] Also note that Google Trends interprets a blank space as an AND operator. Furthermore, Google is not case sensitive. Since in the English price notation the currency is specified first, we include all permutations in the search. The terms million(s) and billion(s) are (in the US) frequently searched together with Dollar signs and numbers. Hence, we exclude them from our search-term as they clearly belong to a much higher price level.

For users that are searching for Euro prices, there are even more possibilities to reference the Euro. This is of particular importance when worldwide searches are considered in our application. Google does not offer the possibility to restrict the SVI to include only searches originating from countries in the Euro Area. Since the HTTP-request to Google Trends cannot incorporate all possibilities, we stick to the Euro names in the largest economies as well as to the symbol and the abbreviation EUR. The words millions and billions, or any other quantification of Euros, is not as frequently searched together with Euro prices compared to Dollar prices. Therefore, the search-term for amounts denoted in Euro is constructed for each $p \in \mathcal{P}$ as

$$p \text{ euro} + p \text{ eur} + \text{eur}p + p€ + €p + p\ € + p\ \text{ευρώ} + p \text{ euros} + p \text{ euroa}$$

[7] Source: https://support.google.com/trends/answer/4359582?hl=en (last visited: 09-09-2019)

In addition, we omit the Slovenish, Estonian, Latvian, Maltese, and Lithuanian names for the Euro.

As a check for robustness, we use the possibility Google offers to limit the geographic origin of searches for Euro prices to Germany or France (as the two largest economies in the Euro Area), separately. When we consider searches from Germany or France, we construct the search-term for each $p \in \mathcal{P}$ as

$$p \text{ euro} + p \text{ eur} + \text{eur}p + p€ + €p + p\ € + p \text{ euros}$$

To download the data, we use a slightly modified version of the *gtrendsR*-package maintained by Massicotte and Eddelbuettel (2018) in R.[8]

The IPSO can be constructed on a number of frequencies. Monthly IPSOs, for the US and the Euro Area, are shown in Figures 1.10a and 1.10b, while daily IPSOs are shown in Figure 1.10c. For the daily IPSO, the method presented in the previous section is used to construct multi-annual coherent time series, while for the monthly IPSOs coherent time series can be downloaded directly. For the multi-annual daily and monthly time series, the method laid out above has to be used to make them comparable. Descriptive statistics are available in Table 1.7. For the US the expectation of the prices searched online, the daily IPSO, is on average at around \$165 and varies with a daily standard deviation of around \$3. The average monthly IPSO for the US is around \$111. The lower scale of the monthly average may be due to Google's scaling of the data. First, on a monthly frequency, Google's threshold as referred to in Equation (1.1) might be higher and since a larger time frame is considered the maximum search intensity is higher as well. Since for price levels above \$1,000, search volume seems to be lower, higher price levels are weighted less. Additionally, recall that the (average of the) daily expected values does not have to be equal to the (average of the) monthly expected value. This means that monthly and daily levels are not comparable. Interestingly, however, the level of the IPSO for prices in Euro based on worldwide searches, as well as for the Dollar based on American searches are comparable. However, the dynamics of the IPSO may contain valuable information. This is because, in theory, expected price inflation should lead to a shift in search propensities of different price levels. As inflation expectations rise, Google users should exhibit a tendency to more often search for higher price levels. This is then reflected in a higher IPSO. However, since Google estimates the SVIs based on samples of searches, the sample variation may introduce noise into the measurement of the IPSO. Furthermore, other factors that influence search behavior, such as the publication or availability of collector coins, may also play a role for certain price levels and introduce noise unrelated to inflation.

[8] Our pull request to incorporate the operator functionality into the development version of the package has been accepted by the maintainers on September 9$^{\text{th}}$, 2019, available on github.com/PMassicotte/gtrendsR.

Table 1.7: Descriptive Statistics

In the table below, descriptive statistics of the times series used in the analysis of this chapter are displayed. The table presents the mean μ, the standard deviation σ, the median $x[0.5]$ as well as the minimum and maximum values of each series. Note that daily and monthly ΔIPSO, inflation π_t, the monthly change in private consumption for the US Δc_t as well the change in consumption loans for the Euro Area Δl_t are logarithmic growth rates measured in percent.

Panel A - US Data

Frequency	Series	Origin	Start to End	μ	σ	$x[0.5]$	min	max
Daily	IPSO	US	03/06/16 to 10/06/19	164.837	2.734	164.965	124.456	169.141
	ΔIPSO		04/06/16 to 10/06/19	0.002	1.390	0.031	-26.544	17.657
	ΔBEIR		06/06/16 to 10/06/19	0.000	0.031	0.000	-0.120	0.159
Monthly	IPSO	US	30/06/06 to 30/06/19	111.121	14.170	107.339	87.898	167.296
	ΔIPSO		31/07/06 to 30/06/19	-0.164	6.633	-0.744	-18.906	17.949
	π_t		31/08/06 to 31/05/19	0.149	0.389	0.168	-1.934	1.003
	Δc_t		31/01/07 to 30/04/19	0.149	0.277	0.160	-0.949	0.916

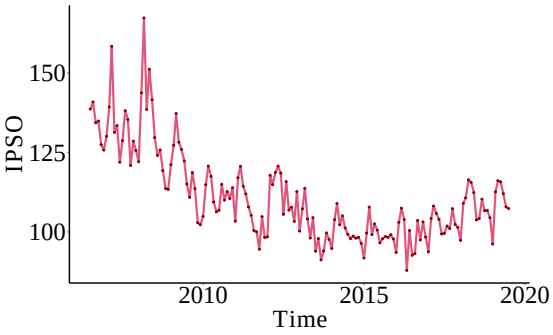
Panel B - Euro Area Data

Frequency	Series	Origin	Start to End	μ	σ	$x[0.5]$	min	max
Monthly	IPSO	world	30/06/06 to 30/06/19	111.543	15.160	114.130	78.076	143.291
		DE	30/06/06 to 30/06/19	140.098	14.940	139.248	94.824	168.498
		FR	30/06/06 to 30/06/19	109.117	18.699	102.115	67.697	162.694
	ΔIPSO	world	31/07/06 to 30/06/19	-0.071	6.925	0.157	-33.341	28.052
		DE	31/07/06 to 30/06/19	-0.133	8.228	-0.060	-38.217	40.283
		FR	31/07/06 to 30/06/19	-0.255	11.200	-0.099	-59.400	60.475
	π_t	EA	31/08/06 to 30/04/19	0.125	0.510	0.179	-1.555	1.341
	Δl_t	EA	31/07/06 to 31/12/18	14.863	3.338	13.931	9.278	22.498

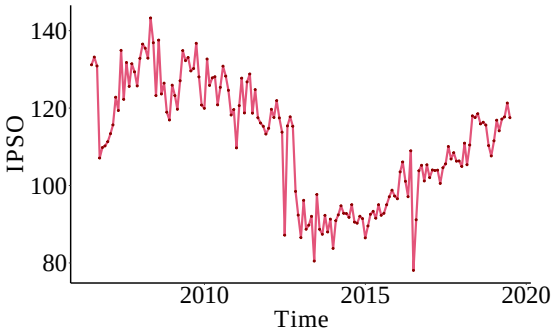
Figure 1.10: The Index of Prices Searched Online

The figure shows the constructed time series for the IPSO as set out in the section above for the US in Dollar and the Euro Area, naturally, in Euro. The absolute value of the IPSO constructed on a daily basis is not entirely comparable to that constructed on a monthly. This is because Google's monthly threshold may differ from the daily threshold, and the monthly IPSO is not simply the average of the daily IPSOs. The daily IPSOs range back till January 1, 2010, while the monthly start in January, 2006. However, the levels within one time series are comparable. The IPSOs depicted are constructed on search-terms for the price levels $\mathcal{P}$ = { 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100; 200; 300; 400; 500; 600; 1,000}

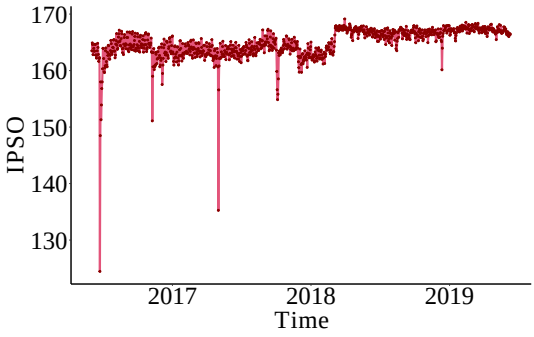
(a) Monthly US



(b) Monthly EU



(c) Daily US



Thus, we consider in our analysis the logarithmic growth rates of the IPSO in percent, referred to in Table 1.7 as ΔIPSO. Also, inflation and the consumption measures are reported as logarithmic growth rates in percent. The average 0.149% for monthly US-inflation can be converted to an average annual inflation rate of 1.788%. Similar conversions can be calculated for all other monthly and daily logarithmic growth rates. There are some puzzling results. The average monthly ΔIPSO in the US implies an annual reduction of around −2%, while the median ΔIPSO implies an annual reduction almost −9%. For the EU the monthly ΔIPSO seems to be more centered.

1.2.2 Macroeconomic Data

As official inflation and consumption data are available on a monthly frequency or lower, we test the forecasting ability of our measure on monthly macroeconomic data. For the US, we consider inflation measured as the logarithmic growth rate of the Consumer Price Index as published by the U.S. Bureau of Labor Statistics (2019) and the logarithmic growth rate of real personal consumption expenditure from the U.S. Bureau of Economic Analysis (2019). For the Euro Area, we consider the growth rate of the harmonized consumer price index (HCPI) published by the ECB.

With respect to personal consumption expenditure in the Euro Area, we are faced with the situation that the ECB only publishes quarterly data. Thus, to keep our analysis simple, we consider instead the growth of loans granted to households for consumption goods (excluding transportation and equipment) in the Euro Area.

To support our analysis for the US inflation, we also use inflation linked T-Bills to infer a daily measure for inflation, the so called break-even inflation rate. We calculate the break even inflation rate from

$$\text{BEIR}_t = \frac{1 + i_{t,\text{nom}}}{1 + i_{t,\text{real}}} - 1,$$

where we use the market yield on U.S. Treasury securities at 5-year constant maturity, quoted on investment basis and inflation-indexed[9], denoted as i_{real} and the market yield on U.S. Treasury securities at 5-year constant maturity, quoted on investment basis[10], denoted as i_{nom}.

1.3 Econometric Approach

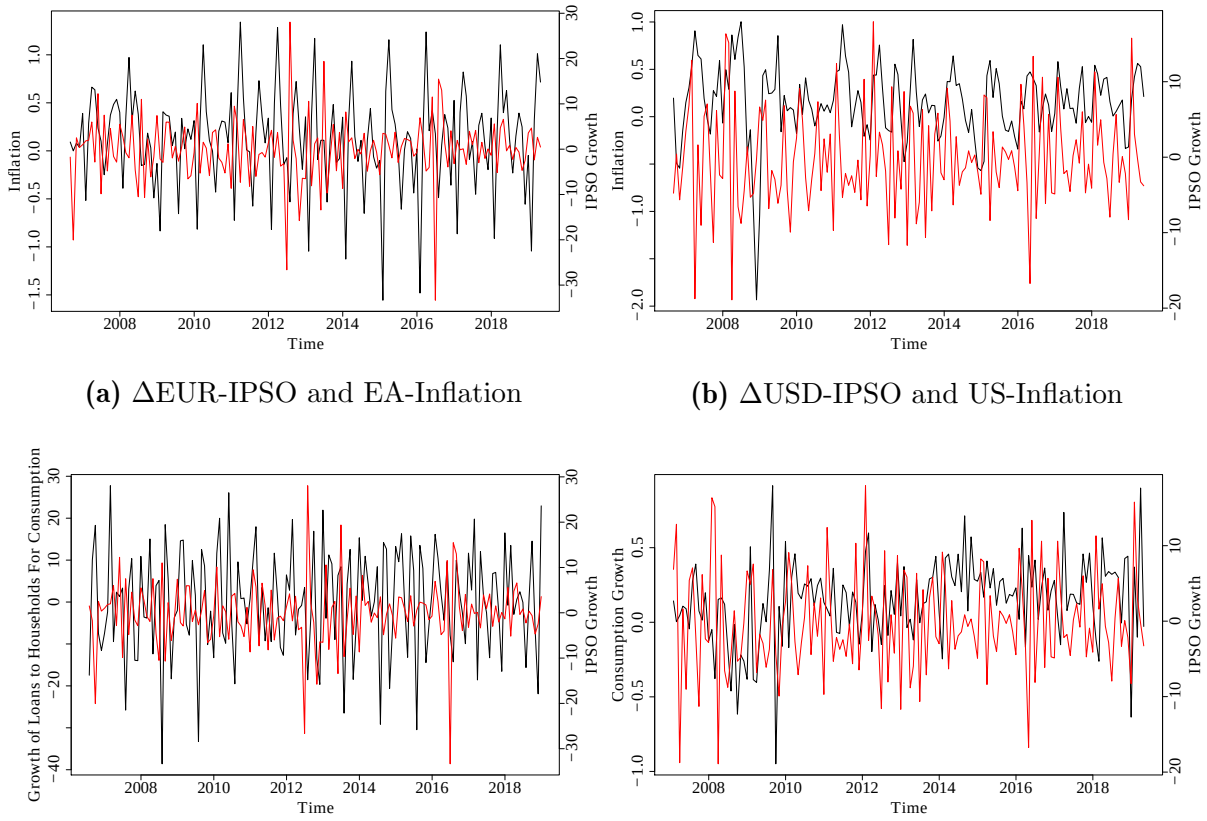
To evaluate the forecasting ability of the IPSO, we use pairwise vector autoregressive models (VARs). In each VAR, we use the IPSO and one other macroeconomic time series. For each pair, we conduct a Granger causality test, i.e., we test the hypothesis whether past values of the IPSO can help to predict inflation or consumption expenditure. The analysis is conducted in R using the *vars*-package (Pfaff 2008). In a first analysis, we only consider a VAR(1). We then extend the analysis on the basis of the Schwarz-Bayes Information Criterion (BIC) to select the lag length of the model p.

[9] Board of Governors of the Federal Reserve System (US), 5-Year Treasury Constant Maturity Rate, retrieved from https://www.federalreserve.gov/datadownload/Choose.aspx?rel=H15, Unique Identifier: H15/H15/RIFLGFCY05_XII_N.B

[10] Board of Governors of the Federal Reserve System (US), 5-Year Treasury Constant Maturity Rate, retrieved from https://www.federalreserve.gov/datadownload/Choose.aspx?rel=H15, Unique Identifier: H15/H15/RIFLGFCY05_N.B

Figure 1.11: IPSO Growth and Macroeconomic Time Series

The figure shows in each panel the growth of the IPSO for the Euro Area or the US constructed on search-terms for the price levels $\mathcal{P} = \{$ 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100; 200; 300; 400; 500; 600; 1,000$\}$ in red. The black lines are either US consumption growth, US inflation, EA inflation or EA consumption loan growth.



(a) ΔEUR-IPSO and EA-Inflation

(b) ΔUSD-IPSO and US-Inflation

(c) ΔEUR-IPSO and EA Consumption Loan Growth

(d) ΔUSD-IPSO and US-Consumption Growth

The model set up for the VAR(p) is

$$\mathbf{x}_{i,t} = \boldsymbol{\mu}_i + \sum_{j=1}^{p_i} \mathbf{A}_{i,j}\mathbf{x}_{i,t-j} + \sum_{r=1}^{S-1} \mathbf{b}_i \, s_{t-r} + \boldsymbol{\varepsilon}_{i,t}, \tag{1.7}$$

where $\mathbf{x}_{i,t}$ is a 2×1 vector that contains the logarithmic growth rate (in percent) of the IPSO_t as well as the logarithmic growth rate (in percent) of the variable of interest in the various specifications. For monthly data, we only consider logarithmic growth rates as the level prices are non-stationary and we are interested whether or not changes in the IPSO are helpful to predict consumption growth or inflation. Also in the case of the daily BEIR, we consider logarithmic growth rates.

The subscript i refers to the different combinations of time series. $\mathbf{A}_{i,j}$ are the 2×2 parameter matrices while $\boldsymbol{\mu}_i$ is a vector of constants. s_t is a centered seasonal control variable. When controlling for annual seasonality, i.e., $S = 12$, we include $S - 1 = 11$ lags of

the control variable s_t. It is constructed as $s_t = D_t - \frac{1}{S}\sum_{r=1}^{12} D_t$ where D_t is a month specific dummy variable which is 1 for a certain month and otherwise 0. In all of our VAR-models for monthly time series, we control for annual seasonality. In the case of the BEIR time series, on a daily frequency, we do not include any seasonal control variables. We assume the innovations $\boldsymbol{\varepsilon}_{i,t}$ to be i.i.d. p_i is the lag-length, either fixed to 1 or selected by the BIC.

To test for Granger causality (Granger 1969), we use the heteroskedasticity consistent jackknife estimator of Efron (1982). Furthermore, we test for contemporaneous correlation, termed instantaneous Granger causality, between the measures based on the test-statistic developed by Granger (1969). However, Granger causality tests the in-sample relevance of the measure. In order to assess the out-of-sample forecasting ability, we analyze the out-of-sample forecasting performance with a rolling one-step ahead forecast. For each prediction $x_{i,t+1}$, we re-estimate a model based on the preceding $L = 84$ monthly observations equivalent to the last 7 years of monthly data. For the break-even inflation rate, which can be measured on a daily frequency, we use the last $L = 250$ observations equivalent to the observations for the business days of the last year. In each window the lag-length is either fixed to 1 or newly selected by the BIC.

To evaluate the forecasts, we report the out-of-sample root mean squared prediction error $RMSPE_i$ of the predicted macroeconomic time series,

$$RMSPE_i = \sqrt{\frac{1}{T-L}\sum_{t=L}^{T-1}(x_{i,t+1} - \hat{x}_{i,t+1})^2},$$

where $x_{i,t+1}$ is the observed macroeconomic variable of interest, i.e., inflation or consumption. For simplicity, we mostly report the $RMSPE_i$ only for out-of-sample forecasts in which the lag-length in each rolling window has been selected by the BIC. However, in the case of the test developed by Clark and West (2006, 2007), we have to use a one-size-fits-all estimation strategy in which we fix the lag-length to $p_i = 1$ across all forecasting windows. Only then, we can test whether the $RMSPE$ is significantly reduced when including the IPSO by using the methodology of Clark and West (2006, 2007) for nested models. For that purpose, we have to estimate a baseline model without the IPSO on the macroeconomic time series of interest. The base model needs to be nested in the extended model including the IPSO. In our case the base model is an auto-regressive model of order one, AR(1) and the extended model is a VAR(1), including the IPSO. The null-hypothesis of the test is that including the IPSO in the model setup does not affect the out-of-sample $RMSPE_i$ and will result in the same forecast error as for the base model.

The test statistic is calculated as

$$z = RMSPE_0 - (RMSPE_i - \kappa_i),$$

where $RMSPE_0$ is the root-mean-squared error of the base model, $\kappa_i = \frac{1}{T-i} \sum_{t=p}^{T} \hat{x}_{i,t+1} - \hat{x}_{0,i,t+1}$ and $\hat{x}_{0,i,t+1}$ denotes the forecast of the base model for the macroeconomic time series i. Critical values are available from Clark and West (2006, 2007).

We also calculate the R^2 of the Mincer and Zarnowitz (1969) (R^2_{MZ}) regression of the realizations $x_{i,t+1}$ on the out-of-sample forecasts $\hat{x}_{i,t+1}$ from the respective model. The regression equation reads

$$x_{i,t+1} = a_0 + a_1 \hat{x}_{i,t+1} + e_{m,t+1}.$$

The higher R^2_{MZ} the better the forecast.

1.4 Empirical Results

Our analysis focuses on two aspects. As set out above, we test for Granger and contemporaneous causality in-sample. With these tests, we check whether it does help to know the current and the past values of the IPSO to predict consumption or inflation on a monthly basis or not. For the US, we also analyze the relevance of the IPSO on daily basis for the break-even inflation rate backed-out of bond yields. We check both the in-sample fit and the out-of-sample prediction to examine whether they improve with the inclusion of the IPSO or not.

1.4.1 US Results

We find that the IPSO as well as US-inflation exhibit strong seasonal components. The results for the US are displayed in Table 1.8. In the case of consumption the in-sample $RMSE$ is not affected by the inclusion of the seasonal control variables. For US inflation the $RMSE$, however, is slightly reduced. We conclude that controlling for seasonal dummies is more important for consumption than inflation.

When controlling for the annual seasonal component ($S = 12$), the IPSO Granger causes US-inflation at least on the 10% significance level. Not controlling for annual seasonality, renders a bi-directional Granger causality on the 5% significance level as well as a contemporaneous correlation on the 10% significance level.

Interestingly, when controlling for the annual seasonality, the hypothesis that US-consumption is not contemporaneously correlated with the IPSO can be rejected on a 5% significance

Table 1.8: USD-IPSO: Causality Tests

The table shows the lag-length p selected by the BIC as well as the p-values for the Granger causality and contemporaneous causality test. Furthermore, the in-sample $RMSE$ is reported. The column denoted IPSO→ shows the p-value for the test on whether IPSO Granger causes the respective other variable; the column →IPSO shows the result on whether the IPSO is Granger caused by the respective variable.

Series	S	p	Granger		Contemp.	$RMSE$
			IPSO→	→IPSO		
US Inflation	0	2	0.030	0.000	0.086	0.309
	12	1	0.080	0.338	0.644	0.272
US Consumption	0	1	0.886	0.815	0.322	0.280
	12	1	0.875	0.324	0.039	0.281
US BEIR	0	2	0.306	0.427	0.036	0.030

Table 1.9: Out-of-Sample Fit: US Inflation and Consumption

The table displays the $RMSPE$ and R^2_{MZ} for the models with the IPSO (i.e., VAR(p_i) models) and without the IPSO (AR(p_i) models) for the listed macroeconomic time series. For each time series, different seasonality components are controlled for ($S \in \{0, 4, 12\}$ for inflation and consumption and $S \in \{0, 5, 30\}$ for the US-BEIR) and the respective $RMSPE$ and R^2_{MZ} are reported. If the difference between the $RMSPE$s of the models with and without IPSO is negative, then including the IPSO helps in predicting the respective time series. For the difference of the R^2_{MZ} it is the other way around: If the difference here is positive then the IPSO increases the quality of the out-of sample prediction.

		S	with IPSO	without IPSO	Difference
US Inflation	$RMSPE$	0	0.172	0.277	-0.104
		12	0.137	0.196	-0.059
	R^2_{MZ}	0	26.95	25.83	1.12
		12	54.89	54.05	0.84
US Consumption	$RMSPE$	0	0.180	0.364	-0.184
		12	0.189	0.266	-0.078
	R^2_{MZ}	0	1.12	5.96	-4.84
		12	1.76	1.45	0.31
US BEIR	$RMSPE$	0	0.0205	0.0459	-0.0253
	R^2_{MZ}	0	2.7696	4.7620	-1.9924

level. On a daily frequency, an instantaneous correlation between the IPSO and the BEIR can be established on a 5% significance level.

Out-of-sample the results for the $RMSPE$ and the R^2_{MZ}, reported in Table 1.9, give also clues that including the IPSO helps in forecasting inflation and surprisingly also consumption. Introducing the IPSO, reduces the $RMSPE$ of forecasted monthly inflation by 5.9 basis points when also controlling for seasonality i.e., a decrease of slightly less than a third. Converted to yearly inflation rates, this would amount to a sizable reduction in the $RMSPE$ of around 70.8 basis points. When controlling for annual seasonality, including the IPSO in forecasting US-inflation also increases the R^2_{MZ} slightly. For the daily BEIR measure, including the IPSO, effectively, more than halves the $RMSPE$. However, the R^2_{MZ} is decreased for the daily BEIR. In predicting consumption, the IPSO reduces the $RMSPE$ whether one controls for annual seasonality or not. However, the R^2_{MZ} is only increased slightly in the case when one includes seasonal dummies.

The results for the Clark-West test, presented in Table 1.10, calculated from the fixed lag-length VAR(1) and AR(1) models on the various time series, indicate that including the IPSO in the US inflation forecast reduces the $RMSPE$ significantly on a 5% significance level when controlling for monthly seasonal figures ($S = 12$). Not controlling for seasonality, the Clark-West test, also, is in favor of the IPSO inclusion for inflation ($S = 0$). For US-consumption, the results of the Clark-West test are entirely insignificant, as well as for the daily BEIR.

We can see that the seasonal figure included in inflation is, to a large extent, mimicked by the IPSO. However, beyond the seasonal figure, the IPSO still helps in predicting inflation and exhibits a precursory and contemporaneous Granger causality. Knowing today's IPSO, apparently, helps to forecast tomorrow and today's inflation in the US.

1.4.2 Euro Area Results

In order to check the robustness of the above findings, we repeat our analysis for the Euro Area. However, the situation for the Euro Area (EA) is rather complicated. As Google does not provide the possibility to limit the basis of searches on which the SVIs are calculated to the EA, we have to find proxy SVIs for the Euro price SVIs by setting other geographical limitations. Therefore, we simply take the basis of searches for Euro price levels worldwide. This might include searches for Euro price levels from other countries, and, thus, might add noise to the EA IPSO. This might hamper the correlation with inflation and consumption within the EA. In addition, we take the largest economies in the EA (France and Germany), and examine whether searches from within these two economies have any relation with EA consumption or inflation. The Google user basis in Germany or France, however, only partially matches the Euro Area's user basis in total. Possible

Table 1.10: Clark-West Test Results

The table displays the test statistics of the test developed by Clark and West (2006, 2007). In the test, an AR(1) for inflation or consumption is compared with an equivalent VAR(1) which also includes the IPSO as defined in Equation (1.7). For worldwide searches as well as searches from France and Germany, the forecasts for inflation and growth of consumption loans in the Euro Area underlie the test. For searches from the US, the US time series are used. S defines the frequency for which centered seasonal dummies are included. The null hypothesis of the test is that the AR(1) yields the same $RMSPE$ as the VAR(1). It can be rejected on a 10% significance level when the test statistic exceeds 1.280 or on a 5% significance level when the test statistics is larger than 1.645. The column 'Origin' displays the two letter country code for the origin of the searches with which the IPSO is constructed. If worldwide searches are used for the construction 'all' is displayed.

		S	
Series	Origin	0	12
Monthly Inflation	all	1.52	-0.18
	FR	-0.16	-0.03
	DE	-0.49	-0.79
	US	1.59	1.83
Monthly Consumption	all	-0.04	0.52
	FR	-0.32	1.30
	DE	-1.39	0.87
	US	-1.63	-0.69
Daily BEIR	US	0.07	–

relations with EA-inflation or consumption might therefore also be veiled. Furthermore, when it comes to consumption, the European Central Bank (ECB) and other European institution do not provide monthly estimates for private consumption expenditure. This is why we have to resort to the available monthly measure of consumption loans granted to private households. We are, thus, strictly speaking not analyzing the relation of the IPSO with consumption growth in the Euro Area, but its relation with the growth in consumption loans granted to private households. The results for the EA are displayed in Table 1.11.

They suggest that if no seasonal component is included, a bidirectional Granger-causality between the IPSO, based on worldwide and French searches, and inflation on at least a 10% significance level can be detected. For the IPSO based on worldwide searches, a contemporaneous correlation can also be detected on a 10% significance level. When German searches are used, the IPSO Granger causes inflation on a 1% significance level. However, these results are not robust to including centered seasonal dummies.

In-sample, in the case of consumption loan growth, when no seasonal dummies are included, on at least a 10% significance level, Granger causality can be found for the IPSO constructed on worldwide searches. For the index constructed on French searches, on every conventional significance level, the IPSO Granger causes consumption loan growth when controlling for

Table 1.11: EUR-IPSO: Causality Tests

The table shows the lag-length selected by the BIC, p, the seasonal component controlled for S, as well as the p-values for the Granger causality and contemporaneous causality test. Furthermore, the in-sample $RMSE$ is reported. The column denoted IPSO→ shows the p-value for the test on whether IPSO Granger causes inflation or consumption, respectively; the column →IPSO shows the result on whether the IPSO is Granger caused by the respective variables.

Series	S	p	Granger		Contemp.	$RMSE$
			IPSO→	→IPSO		
Euro Area Inflation	**Worldwide Searches**					
	0	2	0.000	0.062	0.092	0.481
	12	1	0.115	0.521	0.886	0.220
	German Searches					
	0	6	0.330	0.006	0.266	0.405
	12	1	0.593	0.997	0.479	0.221
	French Searches					
	0	2	0.027	0.100	0.971	0.496
	12	1	0.510	0.764	0.364	0.221
Euro Area Consumption	**Worldwide Searches**					
	0	2	0.009	0.236	0.855	10.753
	12	2	0.359	0.175	0.546	6.351
	German Searches					
	0	2	0.793	0.783	0.126	11.107
	12	2	0.168	0.557	0.121	6.322
	French Searches					
	0	2	0.618	0.077	0.852	11.082
	12	2	0.001	0.608	0.759	6.093

seasonality. For the IPSO constructed with worldwide searches, Granger causality of the IPSO on consumption loan growth can only be found at every conventional significance level when one is not controlling for seasonality.

Comparable to the US data, the out-of-sample results are a little more promising. When looking at the $RMSPE$s of the out-of-sample forecasts reported in the Tables 1.12 1.13 and 1.14, when controlling for annual seasonality, we find that the $RMSPE$ for inflation is always reduced by the inclusion of the IPSO by around 30%. When controlling for monthly seasonal dummies, in the case of inflation, the R^2_{MZ} is rather unaffected and increased or decreased marginally.

Table 1.12: Out-of-Sample Fit: Euro Area Inflation and Consumption (Worldwide)

The table displays the $RMSPE$ and R^2_{MZ} for the models with the IPSO (i.e., VAR(p_i) models) and without the IPSO (AR(p_i) models) for the listed macroeconomic time series. For each time series, different seasonality components are controlled for ($S \in \{0, 4, 12\}$ for inflation and consumption and $S \in \{0, 5, 30\}$ for the US-BEIR) and the respective $RMSPE$ and R^2_{MZ} are reported. If the difference between the $RMSPE$s of the models with and without IPSO is negative, then including the IPSO helps in predicting the respective time series. For the difference of the R^2_{MZ} it is the other way around: If the difference here is positive then the IPSO increases the quality of the out-of sample prediction.

		S	with IPSO	without IPSO	Difference
Euro Area Inflation	$RMSPE$	0	0.377	0.774	-0.397
		12	0.168	0.231	-0.064
	R^2_{MZ}	0	6.27	0.03	6.24
		12	82.64	83.65	-1.01
Euro Area Consumption	$RMSPE$	0	7.642	18.979	-11.337
		12	3.999	15.855	-11.856
	R^2_{MZ}	0	19.15	7.30	11.85
		12	80.98	19.96	61.02

For consumption loan growth the reduction in the $RMSPE$ is drastic for indices based on worldwide, French and German searches. For all indices, by including the IPSO when forecasting EA consumption loan growth, the $RMSPE$ is reduced to less than a third of the $RMSPE$ of the benchmark model. We also find that R^2_{MZ} is always increased strongly to levels over 70% for consumption loan growth by including the IPSO as well as centered seasonal dummies for a monthly frequency.

For the EA forecasts of inflation, when controlling for annual seasonality, the null hypothesis of the Clark-West test that a simple AR(1) model has the same $RMSPE$ as a VAR(1) model cannot be rejected on any significance level. When not controlling for seasonality, the Clark-West test turns up significant on a 10% significance level for the IPSO based on worldwide searches. For EA consumption, only when controlling for seasonality, the Clark-West test indicates a significant reduction (on the 10% level) of the $RMSPE$ for the IPSO based on French searches.

Again we find that the IPSO mimics the seasonality in the macroeconomic time series. When including seasonal dummies, for the Euro Area in-sample all results vanish, except for the Granger causality of the IPSO based on French searches and consumption loan growth. However, focusing only at the out-of-sample results, the improvements of including the IPSO in forecasting inflation are sizable. The gains when predicting consumer loan growth are very large.

Table 1.13: Out-of-Sample Fit: Euro Area Inflation and Consumption (German Searches)

The table displays the $RMSPE$ and R^2_{MZ} for the models with the IPSO (i.e., VAR(p_i) models) and without the IPSO (AR(p_i) models) for the listed macroeconomic time series. For each time series, different seasonality components are controlled for ($S \in \{0, 4, 12\}$ for inflation and consumption and $S \in \{0, 5, 30\}$ for the US-BEIR) and the respective $RMSPE$ and R^2_{MZ} are reported. If the difference between the $RMSPE$s of the models with and without IPSO is negative, then including the IPSO helps in predicting the respective time series. For the difference of the R^2_{MZ} it is the other way around: If the difference here is positive then the IPSO increases the quality of the out-of sample prediction.

		S	with IPSO	without IPSO	Difference
Euro Area Inflation	$RMSPE$	0	0.415	0.774	-0.359
		12	0.165	0.231	-0.067
	R^2_{MZ}	0	5.59	0.03	5.56
		12	83.31	83.65	-0.34
Euro Area Consumption	$RMSPE$	0	7.860	18.979	-11.119
		12	4.368	15.855	-11.487
	R^2_{MZ}	0	15.89	7.30	8.59
		12	76.96	19.96	57.00

Table 1.14: Out-of-Sample Fit: Euro Area Inflation and Consumption (French Searches)

The table displays the $RMSPE$ and R^2_{MZ} for the models with the IPSO (i.e., VAR(p_i) models) and without the IPSO (AR(p_i) models) for the listed macroeconomic time series. For each time series, different seasonality components are controlled for ($S \in \{0, 4, 12\}$ for inflation and consumption and $S \in \{0, 5, 30\}$ for the US-BEIR) and the respective $RMSPE$ and R^2_{MZ} are reported. If the difference between the $RMSPE$s of the models with and without IPSO is negative, then including the IPSO helps in predicting the respective time series. For the difference of the R^2_{MZ} it is the other way around: If the difference here is positive then the IPSO increases the quality of the out-of sample prediction.

		S	with IPSO	without IPSO	Difference
Euro Area Inflation	$RMSPE$	0	0.377	0.774	-0.397
		12	0.164	0.231	-0.068
	R^2_{MZ}	0	5.29	0.03	5.26
		12	83.67	83.65	0.02
Euro Area Consumption	$RMSPE$	0	7.531	18.979	-11.448
		12	3.973	15.855	-11.882
	R^2_{MZ}	0	21.77	7.30	14.47
		12	81.77	19.96	61.82

1.5 Summary

Google search queries are a popular addendum to autoregressive models used for prediction. Their use is justified if one is willing to accept the assumption that people gather information before taking action. From an econometric point of view, including Google's search queries or any derivatives based on search queries like our IPSO adds an additional source of information to the autoregressive model and allows a faster adjustment of the dynamics compared to a pure autoregressive model.

While Google data are therefore an attractive variable to improve predictions, they are not directly available on all desired frequencies over long time horizons. We have therefore proposed an algorithm which allows to construct multi-annual search volume indices based on overlapping periods of subsequently downloaded subsamples for the same search query where these subsamples contain a sufficient overlap. The method also paves the way to make more than five SVIs comparable where five is the maximum that Google allows to be compared directly on its website. During a detailed evaluation of our algorithm and a comparison with other approaches to concatenate SVIs (naive concatenation scheme and a method based on time frame comparison), it turns out that our algorithm is capable to circumvent all potential pitfalls (zeros in the index or sudden spikes) while preserving the statistical properties of the benchmark SVIs.

We illustrated the use of our algorithm in an application to forecast US and European inflation and consumption measures, thereby discussing again potential pitfalls in gathering adequate datasets. Multiple Google SVIs were made comparable and were aggregated to an Index for the Prices Searched Online (IPSO) which constitutes an expected average price level of individuals who engage in buying. The index based on US searches precedes the monthly US inflation rate and is contemporaneously correlated with monthly US inflation and consumption growth. When forecasting monthly US or Euro Area inflation out-of-sample, the $RMSPE$ can be reduced by around 30% when the IPSO is included. Similarly, the prediction of Euro Area consumption loan growth is decisively improved when the IPSO is included in the prediction model.

Appendix

Suppose, we download SVIs for a search-term $j \in \mathcal{M}$ and region i for two overlapping time frames $\mathcal{T}_A$ and $\mathcal{T}_B$. According to Equation (1.2), we can describe the SVIs at the point in time of the overlap, i.e., $t \in \mathcal{T}_A \cap \mathcal{T}_B$, as

$$SVI_{A,t} = \alpha_A + \beta_A s_t + \nu_{A,t}, \tag{1.8}$$

$$SVI_{B,t} = \alpha_B + \beta_B s_t + \nu_{B,t}. \tag{1.9}$$

Since the region i and the search-term j are fixed, we drop the respective subscripts in Equations (1.8) and (1.9). Furthermore, we use A and B to clearly relate the objects to the reference time frames $\mathcal{T}_A$ and $\mathcal{T}_B$. Solving both equations for s_t and equating the results yields

$$\frac{SVI_{A,t} - \alpha_A - \nu_{A,t}}{\beta_A} = \frac{SVI_{B,t} - \alpha_B - \nu_{B,t}}{\beta_B}. \tag{1.10}$$

Solving expression (1.10) yields Equation (1.3)

$$\begin{aligned} SVI_{A,t} &= \underbrace{\alpha_A - \tfrac{\beta_A}{\beta_B}\alpha_B}_{} + \underbrace{\tfrac{\beta_A}{\beta_B}}_{} SVI_{B,t} + \underbrace{\nu_{A,t} - \tfrac{\beta_A}{\beta_B}\nu_{B,t}}_{} \\ &= \quad \gamma \quad + \quad \delta \quad SVI_{B,t} + \quad \varepsilon_t \end{aligned}$$

Chapter 2

Today I Got a Million, Tomorrow, I Don't Know: On the Predictability of Cryptocurrencies by Means of Google Search Volumes[11]

In November 2017, buying Bitcoin and other cryptocurrencies seemed like the perfect investment. The media attention and news coverage was abundant and it seemed that prices could only continue to skyrocket as Bitcoin prices had already surged by approximately 1700% from January to mid December 2017. However, this period has also been accompanied by skepticism about the continued prosperity of cryptocurrencies as well as the introduction of Bitcoin futures on the CME and CBOE. Ultimately, the Bitcoin bubble (as it has been referred to by Corbet, Lucey and Yarovya (2018), Geuder, Kinateder and Wagner (2018) and others) started to deflate and the price of Bitcoin is on a steady decline since December 2017, having lost to date roughly 75% of its peak value.

A question that remains is whether the reversal was predictable which ultimately leads to the question whether cryptocurrency markets are predictable in general. The numerous events that surrounded in particular Bitcoin in 2017 (e.g. the fork into Bitcoin and Bitcoin Cash on August 1, the close down of cryptocurrency platforms in China by September 30, or the introduction of Bitcoin futures) created a need for information. Retail investors satisfied their demand by means of online searches, in particular through Google's search engine. If these investors subsequently act on their findings, they might ultimately trigger a price movement. In this chapter we reconsider the question whether Google's search volume indices (SVIs) can serve as a predictor for returns and volatility of cryptocurrency markets. As both trading of cryptocurrencies as well as the search for and provision of information is continuous and fast, we consider different high frequencies (hourly and daily) as compared to weekly frequencies which are dominant in the literature up to now. The reason for the latter is that Google has often changed the way data are provided since the introduction of Google Trends. Initially, daily data could be downloaded with a fixed reference date so that the SVI data could be concatenated without any problems. Today,

[11] This chapter is based on Bleher and Dimpfl (2019) published in the International Review of Financial Analysis.

daily data are provided for a 270 day period, but a reference day cannot be fixed. Hence, a long time series of daily data needs to be reconstructed by reversing the standardization employed by Google or using a feature recently added by Google to compare the SVI for different time frames. To this end, we propose in a separate accompanying article (Bleher and Dimpfl 2019), reprinted in Chapter 1, a concatenation method to prepare Google Trends data for high-frequency analysis.

While prior research found a more clear-cut relationship between Google search volume and cryptocurrency returns (in particular Kristoufek (2013) for Bitcoin), we only find a relationship for volatility. Returns are in general unpredictable. For volatility, we find that on average large search volumes precede higher volatility and price uncertainty. The forecast on lower frequencies turns out to be more accurate than on high frequencies. We discuss in detail a number of reasons that can explain our findings, in particular the time-varying relationship between Google Trends and Bitcoin returns, the impact of data frequency as well as the data sources.

The chapter proceeds as follows. Section 2.1 provides an overview of the existing literature and highlights our contribution. Section 2.2 describes the data used in this chapter and Section 2.3 presents the models and evaluation criteria. Section 2.4 contains the in-sample fit and out-of-sample forecast evaluation based on daily data. Section 2.5 discusses the sensitivity of these results with respect to sample period, sampling frequency, and model framework. Section 2.6 concludes.

2.1 Related Literature

Google search volume has been shown to be a useful predictor in various contexts. The first application is by Ginsberg et al. (2009) who predict influenza epidemics well ahead of the official registration. In economics, Choi and Varian (2012) predict vehicle sales or claims for unemployment benefits. Returns and/or volatility prediction using Google's SVI has been conducted by Bank et al. (2011), Da et al. (2015), Dimpfl and Jank (2016), Afkhami, Cormack and Ghoddusi (2017), or Perlin et al. (2017) to name but a few. These authors assume that retail investors first satisfy their need for information by means of an internet search which subsequently leads to trading activity. Hence, Google search volume is used as a proxy for retail investor interest in the respective asset (cp. Chen et al. 2014). While the early literature still made use of Google's possibility to concatenate daily data, recent research limits itself to weekly or monthly applications.

In our empirical analysis, we use daily time series constructed according to the algorithm presented in the accompanying article (Bleher and Dimpfl 2019), reprinted in Chapter 1, to see whether return and/or volatility prediction is possible in the cryptocurrency market on

a daily basis. Only for the cryptocurrency Bitcoin (when denominated in US dollars) the link to Google search queries is seemingly well-established. We contribute to the literature by extending the analysis to multiple cryptocurrencies traded in Euro. We also improve the data basis of former research as we use accurately constructed time series of Google Trends.[12]

Most closely related to our work is Kristoufek (2013) who analyses the connection between Bitcoin prices, Google's SVI for Bitcoin and the number of visits to the Wikipedia article of "Bitcoins" on a weekly basis. We overcome the limitation to weekly data imposed by the direct availability of Google's SVI and also construct daily data using 24 hours trading of the cryptocurrencies instead of a hypothetical 8-hours return which is aggregated to weekly returns. Based on impulse response analysis, Kristoufek (2013) finds that increased interest in Bitcoin leads to higher prices, which again causes higher search volume. He concludes that this forms the potential for a bubble development which might have been observed in December 2017. Recently, Urquhart (2018) investigates the relationship between Bitcoin returns, traded volume, and Google search queries and finds that search queries do not serve as a predictor for volatility. However, he documents that trading activity and volatility draw attention to Bitcoin which manifests in higher search activity.

Similarly, Garcia and Schweitzer (2015) use Google's SVI for the search-term Bitcoin among other variables (number of tweets or exchanged volume) to device a trading strategy. Their results suggest that the SVI variable carries no information which is useful as a trading signal, while variables measuring the sentiment of social activity provide robust trading signals. This contradicts the findings of Kristoufek (2013). We therefore revisit the question whether Google search volume indices, if constructed correctly, help to predict returns or volatility of cryptocurrencies.

In principle, every pricing relevant factor qualifies as a potential predictor for returns. The literature on Bitcoin and cryptocurrencies has identified a large number of such factors. Kristoufek (2013) states that Bitcoin is not comparable to standard currencies, and thus, has its own pricing relevant factors. In general, he classifies Bitcoin as a market without a "fair" value, driven by the sentiment of investors which suggests that prediction based on variables that are able to capture such sentiment is fruitful. Similarly, Garcia et al. (2014) identify two feedback loops that lead public interest towards Bitcoin pacing from booms to busts. Both loops suggest that individual investors satisfy their information demand using Google or Wikipedia which then leads to trading activity in Bitcoin. Furthermore, they find that search activity responds quicker to negative events (such as hacked Bitcoin exchanges) than prices. Hence, one of these pricing factors may be public attention which can be proxied by Google Trends data.

[12] Our results also hold when the entire analysis is conducted using cryptocurrencies traded in US dollars.

Using a LASSO approach, Panagiotidis et al. (2018b) find that gold and search intensity are the most important drivers of Bitcoin returns. However, the authors interpolate weekly data and model the relation to be contemporaneous which leaves the question whether Google Trends data are helpful in forecasting future Bitcoin returns out of sample. Zhang et al. (2018) also investigate contemporaneous cross-correlations between Google searches and Bitcoin based on daily data and document that cross-correlation between Google Trends and the Bitcoin prices is decreasing over time. Related to this finding is the work of Dastgir et al. (2019) who analyze the connection between Bitcoin returns and Google Trends based on a vector autoregressive model coupled with a specified volatility process. They employ a copula-based test for contemporaneous correlation and find a bi-directional effect in the left and right tail of the distribution.

Still, there might be additional fundamental factors. Hayes (2016) ascribes the determinants of the Bitcoin price to the cost of production, i.e., essentially electricity cost, but leaves demand aside. Ciaian, Rajcaniova and Kancs (2016) (who explicitly omit Google search data since daily time series are not available) find that macroeconomic factors do not influence the Bitcoin price while investor attractiveness does which suggests that the price is mainly determined by the demand side. Kristoufek (2015) identifies technical drivers like money supply, price level and usage in trade which are correlated with the price dynamic of Bitcoin. However, he ultimately concludes that the major driver of the Bitcoin price is only the public interest in the cryptocoin.

Furthermore, Alabi (2017) attributes the value of Bitcoin to network effects. He shows that the price is described well by Metcalfe's law (see e.g. Shapiro and Varian 1998) which conjectures that the value of a network is proportional to the squared number of people using it. However, whether the network effect is a dominant factor in pricing Bitcoin remains doubtful (see Poon and Dryja 2015).

The possibility to transact value without any middlemen and oversight by a bank or centralized authority constitutes a pricing relevant feature of Bitcoin or other cryptocurrencies. These features come at the cost that the high volatility of the currencies makes these transactions an unreliable and risky undertaking (cp. Baur and Dimpfl 2021, Baur, Hong and Lee 2018). A reliable forecast of volatility would allow to conduct transaction in low volatility phases and, thus, reduce transaction costs. It should be noted, however, that Bitcoin transactions are not fully anonymous (cp. Reid and Harrigan 2011).

Considering the above literature and the fact that the market is dominated by short-term investors, trend chasers and speculators (cp. Kristoufek 2013, Yelowitz and Wilson 2015), we would expect that the public interest measured with Google's SVI for a particular cryptocurrency should drive the price. Hence, Google Trends data should, in particular on a high frequency, be a good predictor for Bitcoin returns and volatility. Taking into account the characteristics of the various coins, we expect that some cryptocoins are

driven by the general interest in the cryptocurrencies and others may have coin-specific features that generate interest on their own. Thus, we contribute to the literature by an investigation whether Google's search volume indices (SVI) can be exploited systematically for the prediction of key characteristics, namely returns and volatility, of Bitcoin and other cryptocurrencies. This is accompanied by a concatenation algorithm which overcomes the limitations and shortcomings in previous studies imposed by the (un-)availability of Google Trends data on high frequencies.

2.2 Data Description

2.2.1 Cryptocurrency Price History

Focusing on the market for Euro denominated cryptocurrencies, we use prices of 12 cryptocurrencies traded on Kraken.com as of October 2018.[13] Historical prices for our cryptocurrency sample are obtained from CryptoCompare.com via its public access API. Kraken is currently the most important market for trading Bitcoin against the Euro. Figure 2.1 illustrates its market share in terms of transaction volume over time. By the end of 2017, for example, it had a market share of roughly 65% of total trading of Bitcoin in Euro. Up to 2013, MtGox was the dominant market, but it was closed down February 25, 2014. After MtGox' bankruptcy, Kraken's market share grew and it claims now to be the most liquid market for trading Euro denominated cryptocurrencies.

CryptoCompare offers open, high, low, and close data on an hourly and daily frequency which are retrieved for the maximum period for which the respective currency is available on Kraken. Hence, the longest history is obtained for Bitcoin while the shortest time series results for Bitcoin Cash as can be seen in Table 2.1 which lists in column 1 the cryptocurrencies used and in column 2 the maximum available time span. Weekly data are constructed from the daily data.

From the price history for each cryptocurrency, we calculate compound returns $r_t = p_t - p_{t-1}$ based on the logarithmic close prices p_t. As a measure for volatility, we consider the root of the non-parametric variance measure of Garman and Klass (1980) which is estimated as follows:

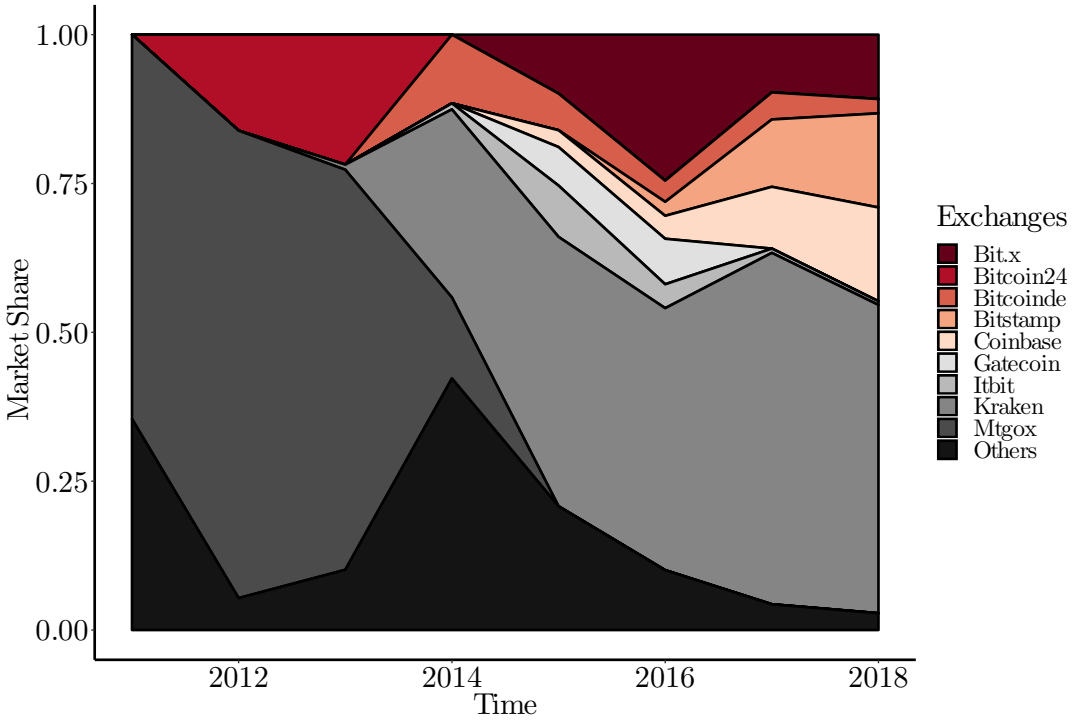
$$\sigma_t^2 = \frac{1}{2}(h_t - l_t)^2 - (2\log(2) - 1)(p_t - o_t)^2$$

where o_t, h_t, l_t, p_t denote logarithmic open, high, low, and close prices on day t. The measure developed by Garman and Klass (1980) does neglect large jumps between the

[13] In principle, there are 13 cryptocurrencies traded on Kraken in Euro. StellarLumens is excluded from the analysis due to an extended period of missing values.

Figure 2.1: Development of Market Shares of Exchanges

The graph depicts the market share of total annual traded volume of Euro against Bitcoin over time of the various exchanges. With the downfall of MtGox, Kraken takes up the market share of MtGox and currently is the most important exchange for trading Bitcoin against Euro. Source: `data.bitcoinity.org`, last accessed: 2018-10-04.



close price and the open price of the previous day (cp. Yang and Zhang 2000). However, as Kraken allows trading around the clock, the problem of overnight jumps does not exist.[14]

Table 2.1 reports time series mean and standard deviation of the returns and the Garman and Klass (1980) volatility measure σ on a daily basis. As can be seen, the average returns range from a daily 0.54% for Ethereum to -0.56% for zCash. The associated daily standard deviation of returns can be considered huge as compared to, for example, stock markets. The latter usually exhibit a daily volatility of roughly 0.1% whereas the volatility of the cryptocurrencies is 50 to 100 times higher.

2.2.2 Google Trends Data

With Google Trends, Google offers a service that allows to compare the relative popularity of search-terms. Google computes and publishes a Search Volume Index that compares the occurrence of searches to the entire volume of searches (Stephens-Davidowitz and Varian

[14] We have also calculated the measures of Rogers and Satchell (1991) and Parkinson (1980) and find that both are highly correlated with the measure of Garman and Klass (1980). The subsequent results are robust to the choice of the variance measure.

Table 2.1: Descriptive Statistics

The table summarizes the return and volatility series for the various cryptocurrencies. Dates are formatted in `dd/mm/yy`. Reported means and standard deviations in percent.

		Returns		**Volatility**	
	Range	Mean	S.D.	Mean	S.D.
Bitcoin Cash	03/08/17 to 30/09/18	-0.01	9.15	7.69	6.36
Bitcoin	14/09/13 to 30/09/18	0.23	4.29	3.55	3.63
Dashcoin	14/04/17 to 21/09/18	0.19	6.78	6.53	5.13
EOS Token	03/07/17 to 30/09/18	0.01	15.26	7.84	5.09
Ethereum Classic	28/07/16 to 30/09/18	0.25	7.30	6.64	5.21
Ethereum	08/08/15 to 30/09/18	0.54	6.94	6.16	5.72
Gnosis	05/05/17 to 30/09/18	-0.46	6.60	8.41	6.35
Litecoin	17/09/13 to 30/09/18	0.17	7.17	5.44	6.70
Augur Coin	05/10/16 to 29/09/18	0.06	8.78	8.81	6.35
Monero	04/01/17 to 30/09/18	0.27	7.07	6.44	4.50
Ripple	22/06/17 to 30/09/18	0.16	8.22	7.07	6.05
zCash	29/10/16 to 30/09/18	-0.56	10.59	8.36	9.74

2014). Hence, a falling SVI does not (necessarily) mean that there are less searches than in the past, but it means that a smaller share of searches is dedicated to the search-term. The measure therefore has to be interpreted carefully as it cannot per se be equated with a proxy for information demand. According to *smartinsights.com*[15], Google's total search volume increased from a level of 1.2 billion searches per day in 2012 to about 4.5 billion per day in 2017. Hence, if the exact same number of searches for one search-term would have been conducted in 2012 or 2017, the SVI would report a lower share in 2017 as opposed to 2012. To still allow for an interpretation as a valid measure of interest, we assume that the Google user base is a random sample of total internet users. As a measure that covaries with interest, we draw on the SVI as a predictor in accordance with the economic literature.

To select the relevant search-terms, we follow the guidelines provided by Stephens-Davidowitz and Varian (2014). We wish to identify the impact of the interest in a specific coin. There are several possible search-terms, namely the respective coin-name, the Kraken ticker symbol and the alternative ticker symbol[16]. We choose the most popular one as identified by Google Trends. The resulting search-terms and the corresponding ticker symbols are listed in Table 2.2.

[15] Source: https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/, last accessed: 2017-01-11.

[16] For some ticker symbols, such as XXBT Kraken also lists alternative ticker symbols XBT. A comprehensive list can be downloaded from https://api.kraken.com/0/public/Assets in JSON format, last accessed: 2017-02-11

Table 2.2: Coins and Corresponding Search-Terms

The table reports the ticker symbols used on *Kraken.com* and the associated search-terms. A minus in the search-term means that the following word is taken out of the searches that constitute Google's search volume index. For example, *Ethereum -Classic* considers all searches for Ethereum without the addendum of the word *Classic*.

Ticker	Searchterm	Cointicker	Ticker	Searchterm	Cointicker
BCH	Bitcoin Cash	BCH	GNO	Gnosis	GNO
XXBT	Bitcoin -Cash -Future	BTC	XLTC	Litecoin	LTC
DASH	Dashcoin	DASH	XREP	Augur Coin	REP
EOS	EOS Token	EOS	XXMR	Monero	XMR
XETC	Ethereum Classic	ETC	XXRP	Ripple	XRP
XETH	Ethereum -Classic	ETH	XZEC	zCash	ZEC

In general, when searching for e.g. *Bitcoin*, Google Trends subsumes all searches that contain this string. Therefore, when we are interested in Bitcoin only, we have to clean the searches which are related to other Bitcoin related subjects (but not directly to trading activity in Bitcoin itself) such as Bitcoin Cash or Bitcoin Futures which also contain the string *Bitcoin* in the search query. This can be done using the minus operator when downloading the SVI from Google Trends. The cryptocurrencies Ethereum (ticker symbol: XETH) and Ethereum Classic (ticker symbol: XETC) are two distinct crypto-tokens. Therefore, we choose to include the search-terms *Etherum -Classic* for XETH, i.e., all searches for Ethereum that do not include the word "Classic", and *Etherum Classic* for XETC. Similar to the case of Bitcoin and Bitcoin Cash, when we refer to the search-term *Etherum* what we actually mean is *Etherum -Classic*.

We deviate from the rule to take the most popular search-term in the following cases. The search-term *DASH* is more popular than the search-term *Dashcoin*. Still, we choose to include *Dashcoin* as it is more salient with respect to the currency. The word *dash* itself has several meanings and there are several brands, computer games and other products that include it. For the cryptocoin Augur, we do not use the search-term *Augur* or *REP* which are the most popular search-terms, but bear other meanings as well. Instead we use the search-term *Augur Coin* as it is more salient with respect to the cryptocurrency, and is according to Google Trends more popular than *Augur Reputation*, *Augur Reputation Token* and *Augur Token*. In order to construct a consistent SVI time series and due to the low search volume for the search-term *Augur Coin*, only for this search-term we relax the requirement of 30 non-zero elements in the overlapping time frame of two datasets containing SVIs for subsequent time periods to 15. We also require only 10 non-zero elements on each side instead of 20.

Aside from the coin-names, we also include the search-term *cryptocurrency* in our analysis. With the inclusion of this search-term, we evaluate if the overall interest in cryptocurrencies helps to forecast price developments. It can also be argued that the overall interest in cryptocurrencies is better reflected in the search-term *Bitcoin* which is the oldest and most actively traded cryptocoin. A comparison of the popularity on Google Trends supports this view. For the entire sample period, the search-term *Bitcoin* ranks highest in the search popularity, followed by *Ripple*, *Ethereum* and *cryptocurrency* (not necessarily in this order).

The popularity of search-terms is important for the quality of the SVI which raises several problems (see e.g. Stephens-Davidowitz and Varian 2014). As Google estimates the SVI for a search-term on a sample, if search volume is too low, the uncertainty about the SVI estimate becomes a problem. Furthermore, if searches do not surpass a threshold, the SVI value is set to 0 by Google. Google does not publish the used threshold. All SVIs are downloaded using package *gtrendsR* (Massicotte and Eddelbuettel 2018) in R (2018). Further details on the concatenation of the data are available in the accompanying article (Bleher and Dimpfl 2019), reprinted in Chapter 1.

Figure 2.2 illustrates our dataset for Bitcoin and Ripple. We observe a co-movement of the price or the volatility time series of Bitcoin with the respective Google SVI (depicted in the upper graphs in Figure 2.2). For the cryptocoin Ripple a similar co-movement can be observed (see bottom graphs in Figure 2.2). In the subsequent analysis we use log returns of the SVI. We checked that all data are stationary using an ADF test with the lag length suggested by the Schwartz Bayes information criterion (SIC).

2.3 Models and Forecast Evaluation Criteria

For each cryptocurrency, we can relate five time series with each other: the respective returns of the exchange rate with the Euro, the volatility of these returns and Google's SVIs for the search-terms *cryptocurrency*, *Bitcoin*, as well as a search-term related to the name of the respective cryptocurrency.

2.3.1 VAR Model for Returns and Volatility

From the basic asset pricing equation using the stochastic discount factor m_{t+1} (cp. Cochrane 2008), the conditional moment of future returns R_{t+1} is given by

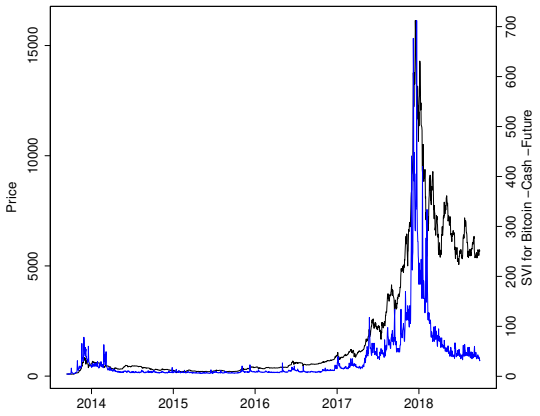
$$\mathbb{E}_t[R_{t+1}] = R_t^f + R_t^f \mathrm{cov}(m_{t+1}, R_{t+1}),$$

where R_t^f is the currently prevailing risk free rate at time t.

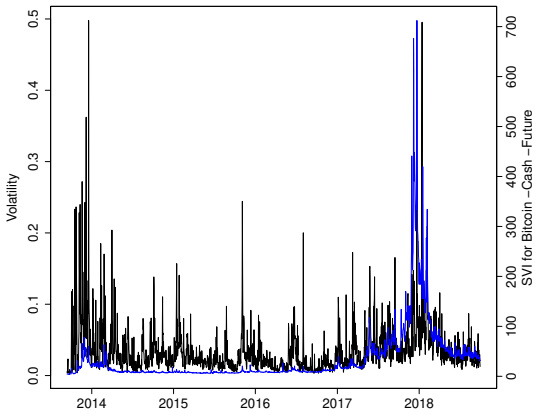
Figure 2.2: Closing Prices, Volatility and Search Volume Indices

The graphs compare the closing price and volatility (left scale, black line) with Google's search volume index for the coin-name (right scale, blue line). The two upper graphs refer to Bitcoin while the two bottom graphs refer to Ripple.

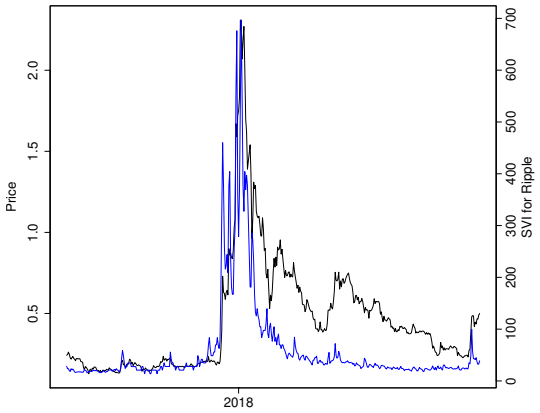
(a) Bitcoin Price and SVI



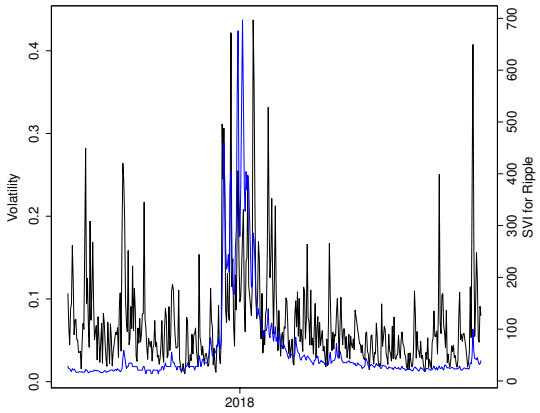
(b) Bitcoin Volatility and SVI



(c) Ripple Price and SVI



(d) Ripple Volatility and SVI



Assuming that the law-of-one price and the no arbitrage condition hold, the stochastic discount factor can be related to fundamental pricing factors. In our case, Google search volume may proxy or co-vary with one or several of these pricing factors. Hence, we assume that the future stochastic discount factor m_{t+1} can be proxied by a function of the present SVIs. The specific functional form may be non-linear. However, as our focus lies on predicting returns or volatility, we may approximate the conditional moment of returns by a linear function. As the conditional second centered moment is a function of the first moment and therewith a function of the factors determining it, a linear approximation for the conditional variance is suitable as well.

Table 2.3: Model Specification Overview

	Included SVIs		
Model	Coin Name	*Cryptocurrency*	*Bitcoin*
0	–	–	–
1	✓	–	–
2	–	✓	–
3	–	–	✓
4	✓	✓	–
5	✓	✓	✓

Thus, we estimate VAR-models for either the returns or the volatility of cryptocurrencies of the following form

$$\mathbf{x}_{i,t} = \boldsymbol{\mu}_i + \sum_{j=1}^{p_i} \mathbf{A}_{i,j}\mathbf{x}_{i,t-j} + \boldsymbol{\varepsilon}_{i,t} \tag{2.1}$$

where $\mathbf{x}_{i,t}$ is an $R \times 1$ vector that contains one or several SVIs and either the return or the volatility of the i^{th} cryptocurrency. $\mathbf{A}_{i,j}$ are the $R \times R$ parameter matrices while $\boldsymbol{\mu}_i$ is a vector of constants. The innovations $\boldsymbol{\varepsilon}_{i,t}$ are i.i.d. white noise. p_i is the lag-length selected by the SIC.[17]

We consider six models separately for either returns or volatility. Table 2.3 provides an overview of the SVIs that are included in each model specification in addition to autoregressive terms. Model 0, which reduces Equation (2.1) to a univariate AR(p)-model, serves as a benchmark. Model 1 is the specification which relates the search volume of a certain coin to the coin's price or volatility. Model 2 considers the relevance of the general interest in cryptocurrencies for forecasting returns and volatility as it includes the SVI for the search-term *cryptocurrency*. Model 3 assesses whether the interest in Bitcoin as the most pronounced cryptocoin helps to predict returns or volatility. With Models 4 and 5 we test if we can improve the forecasts by combining the general interest of Google users in cryptocurrencies and their interest in the respective cryptocoin. In the case of Bitcoin, Model 3 reduces to Model 1 and Model 5 reduces to Model 4 as the SVI for *Bitcoin* is both the coin name as well as the proxy for general interest.

The models are estimated using OLS. Estimation is conducted in R (2018) using packages *forecast* (Hyndman and Khandakar 2008) and *vars* (Pfaff 2008). Data and code are available at https://tinyurl.com/y7chh5r6.

[17] While an autoregressive model for the prediction of one variable suffices to predict one day ahead, forecasting returns or volatility over several days with the help of Google's SVI requires a VAR in order to also predict the SVI development.

2.3.2 Evaluation Measures

In order to assess whether Google's SVIs help to predict returns or volatility, Model 0 has to be outperformed by other model specifications according to the following measures. To evaluate the in- and out-of-sample fit of the models, we calculate the root mean squared error ($RMSE$) as

$$RMSE_m = \sqrt{\frac{1}{T-p-1}\sum_{t=p}^{T}(x_{t+1} - \hat{x}_{m,t+1})^2}$$

where x_{t+1} is the observed variable of interest and can either be the return series or the volatility series. m denotes any of the models 0 to 5. $\hat{x}_{t+1}$ denotes the forecasted value.

We then use the test developed by Clark and West (2006, 2007) for nested models, including their critical values, to assess whether the $RMSE$ is significantly reduced by the inclusion of Google's SVI in comparison to Model 0. The null-hypothesis of the test is that the models have the same forecast error whereas the alternative is that Model m has a smaller forecast error than the benchmark Model 0. The test statistic is calculated as

$$z = RMSE_0 - (RMSE_m - \kappa_m),$$

where $\kappa_m = \frac{1}{T-p}\sum_{t=p}^{T} \hat{x}_{m,t+1} - \hat{x}_{0,t+1}$. In our case, the models are only partially nested. Hence, it is not clear upfront which model is the more parsimonious one. In consequence, the adjustment κ_m can be positive or negative. We therefore require upfront that the $RMSE$ of the model including the SVIs is strictly lower than the $RMSE$ of the benchmark model.

We also run a Mincer-Zarnowitz regression (**?**) of the realizations x_{t+1} on the fitted values $\hat{x}_{m,t+1}$ from the respective model to evaluate the in-sample fit. Out-of-sample the fitted values are replaced by the forecasted values. The regression equation, thus, reads as follows:

$$x_{t+1} = a_0 + a_1\hat{x}_{m,t+1} + e_{m,t+1}.$$

The R^2 of this regression (denoted by R^2_{MZ} in the following) serves as a measure for the quality of the in-sample fit or the out-of-sample forecast performance.

For the volatility models, we also calculate the quasi-likelihood loss function (QL) as in Patton (2011) who shows that the QL is robust with regard to noise in the proxy measure (the Garman and Klass (1980) volatility measures in our case). The QL is calculated as follows:

$$QL_m = \frac{1}{T-p}\sum_{t=p}^{T}\left(\frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{m,t+1}} - \log\left(\frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{m,t+1}}\right) - 1\right).$$

The better the forecast, the smaller is the QL measure.

As volatility enters into the model in logarithmic form, before evaluation, we transform it back to the standard, non-logarithmic measure of Garman and Klass (1980). Although forecasting the logarithmic transform of a variable and then transforming it back, bears its problems (cp. Granger and Newbold 1976), Lütkepohl and Xu (2012) show that forecasting the logs can result in dramatic gains in forecast precision, when the resulting variance is more homogeneous. This is the case in our application.[18]

Furthermore, we conduct Wald-tests to i) check the model fit and ii) to see whether the SVIs Granger cause returns or volatility. The respective test scores w are constructed as Wald-statistic of a univariate model which corresponds to the return or volatility equation in the equation system (2.1). Hence, w is

$$w = (\mathbf{R}\hat{\mathbf{a}} - \mathbf{r})' \left(\mathbf{R}\,\hat{\boldsymbol{\Sigma}}\,\mathbf{R}'\right)(\mathbf{R}\hat{\mathbf{a}} - \mathbf{r}),$$

where $\mathbf{R}$ is the matrix that linearly combines the vector of parameter estimates $\hat{\mathbf{a}}$, and $\mathbf{r}$ is a vector of real numbers containing the numeric restrictions imposed on the so formed linear combinations of parameter estimates. $\hat{\boldsymbol{\Sigma}}$ is the estimated asymptotic variance-covariance matrix of the parameter estimates. We use the heteroskedasticity consistent jackknife estimator of Efron (1982) as recommended by Long and Ervin (2000) to estimate the variance/covariance matrix $\hat{\boldsymbol{\Sigma}}$

$$\hat{\boldsymbol{\Sigma}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,\mathrm{diag}\left(\frac{e_t^2}{(1-h_t)^2}\right)\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1},$$

where the residuals are denoted as $e_t = x_t - \mu - \sum_{j=1}^{p} \mathbf{x}_{t-j}\mathbf{a}_j$ (with x_t representing either the returns or the volatility). The matrix which collects all regressors in this equation is $\mathbf{X} = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-p})$, and $h_t = \mathbf{x}_t\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_t$. The division of e_t^2 by $(1-h_t)^2$ increases the variance estimate for the high contribution of outliers. Asymptotically, the test statistic w converges to a χ^2-distribution with the number of hypotheses, q, as degrees of freedom.

[18] Granger and Newbold (1976) also suggest multiplication with a corrective term to mimic the calculation of the expectation of a log-normal distributed variable with the first two moments of the underlying normally distributed variable in logarithms to get the optimal forecast in levels as $y_{t+h|t}^{\text{opt}} = \exp\left\{x_{t+h|t} + \frac{1}{2}\sigma_x^2\right\}$. Lütkepohl and Xu (2012) find that the naïve transform of the logarithmic forecast, $x_{t+h|t}^{\text{naïve}} = e^{x_{t+h|t}}$, performs just as well as the optimal transformation suggested by Granger and Newbold (1976). We tested both, and come to the same conclusion. Thus, we only use the naïve transformation.

2.4 Forecast Evaluation

2.4.1 In-Sample Fit

For the return models, evidence for model significance is mixed. The results are summarized in the upper panel of Table 2.4. For all cryptocoins but Bitcoin, EOS-Token, and Litecoin, the null hypothesis that the parameter estimates are jointly zero cannot be rejected. Hence, the specified models are not able to explain the variation in the return series of these cryptocurrencies significantly better than a simple time series average. Considering that we have conducted 60 tests, we would consider the statistical significance of the five models for Bitcoin, EOS-Token, and Litecoin to be pure coincidence.

For the volatility models, the picture is exactly the opposite as documented by the lower panel of Table 2.4. We can reject the null hypothesis that the respective model specification has no explanatory power on a 1% significance level for all cryptocurrencies for all models.

Concerning Granger causality, Table 2.5 provides an overview of the results. For the return models, the results of the test for no Granger causality are in line with the model specification tests. The predictive ability of Google SVIs is in general not given. The only two cryptocurrencies where the SVIs help to predict returns are EOSToken and Ripple. However, significance is still rather weak and might also be attributed to chance given the number of tests conducted.

The finding of no Granger causality for the vast majority of the coins comes at no surprise. In an efficient market, we would expect that asset price movements are not predictable in the short run. For the majority of cryptocoins, it seems, the market is aware of the demand driven nature of the coins. Available information is already largely incorporated in prices.

Concerning the volatility models, the picture is also mixed. For Gnosis and EOS-Token, we cannot reject the hypothesis of no Granger-causality of Google search volume on volatility on any conventional significance level. However, for all other coins at least for one of the models including Google's SVI we find that search volume Granger causes volatility.

From the comparison of the model specification test in Table 2.4 and the Granger-causality test in Table 2.5 we conclude that volatility is rather persistent and can be explained well exploiting its autoregressive dynamics. We also find that for the majority of the cryptocurrencies considered, Google SVIs have non-negligible predictive power for development of future volatility.

Whether public or coin specific interest is more important in predicting volatility can also be inferred from the two tables. For Bitcoin, we can reject the null hypothesis of no

Table 2.4: Model Significance

The table summarizes the results of Wald tests for the joint significance of all variables for the various model specifications and cryptocurrencies. * (**,***) denotes statistical significance on a 10% (5% and 1%, respectively) significance level.

Returns					
	Model 1	Model 2	Model 3	Model 4	Model 5
BitcoinCash					
Bitcoin		*	_		_
Dashcoin					
EOSToken				**	*
EthereumClassic					
Ethereum					
Gnosis					
Litecoin	*	**			
AugurCoin					
Monero					
Ripple					
zCash					
Volatility					
	Model 1	Model 2	Model 3	Model 4	Model 5
BitcoinCash	***	***	***	***	***
Bitcoin	***	***	_	***	_
Dashcoin	***	***	***	***	***
EOSToken	***	***	***	***	***
EthereumClassic	***	***	***	***	***
Ethereum	***	***	***	***	***
Gnosis	***	***	***	***	***
Litecoin	***	***	***	***	***
AugurCoin	***	***	***	***	***
Monero	***	***	***	***	***
Ripple	***	***	***	***	***
zCash	***	***	***	***	***

Table 2.5: Granger-Causality

The table summarizes the results from the results of the Granger-causality test for the various model specifications. The null hypothesis for each model is that the Google SVI variables do not Granger cause returns or volatility, respectively. * (**,***) denotes statistical significance on a 10% (5% and 1%, respectively) significance level.

Returns					
	Model 1	Model 2	Model 3	Model 4	Model 5
BitcoinCash					
Bitcoin			–		–
Dashcoin					
EOSToken		*		**	*
EthereumClassic					
Ethereum	*				
Gnosis					
Litecoin					
AugurCoin					
Monero					
Ripple	*	**	**	*	
zCash					
Volatility					
	Model 1	Model 2	Model 3	Model 4	Model 5
BitcoinCash		*		*	
Bitcoin	***		–	***	–
Dashcoin		***		**	*
EOSToken					
EthereumClassic	**	**	***	**	***
Ethereum	***	**	**	***	***
Gnosis					
Litecoin	***	***	***	***	***
AugurCoin	*			**	**
Monero		***	*		
Ripple	***	**	**	***	***
zCash			*		

Granger causality on a 1% significance level. Hence, the volatility of Bitcoin is clearly driven by the coin specific interest in Bitcoin. The volatility of BitcoinCash, Dashcoin, and Monero feed off the general attention in cryptocurrencies (Model 2). The volatility of zCash is driven partially by the attention to the cryptocurrency flagship Bitcoin (Model 3). For all other coins, a mix of coin-specific and general interest in cryptocurrencies precedes volatility (Model 4 and 5).

For Bitcoin, Ethereum Classic, Ethereum, Litecoin, AugurCoin, and Ripple, the searches for these cryptocurrencies Granger cause their own volatility. For the remaining coins, only the SVIs which capture general interest, Granger cause their volatility. As the literature associates Google queries with the trading of individual investors (see for example Dimpfl and Jank 2016) who add to overall volatility (e.g. Foucault, Sraer and Thesmar 2011), we may conclude that trading of Bitcoin, Ethereum Classic, Ethereum, Litecoin, AugurCoin and Ripple is attractive due to reasons rooted in the nature of these coins themselves as opposed to a general interest in cryptocurrencies documented for the other cryptocurrencies. Interestingly, those coins are also the most liquid ones (with the exception of AugurCoin).

Turning to the fit of the Mincer-Zarnowitz regression and the $RMSE$ for the fitted values of returns, we find that the in-sample fit is very low throughout the panel. Table 2.6 presents the detailed results. For the highly volatile return series of cryptocoins this is an expected result. When forecasting returns within the framework of traditional factor models, the R^2_{MZ} is usually low (see for example Cochrane 2008). Even though we find that Model 5 produces a significantly lower $RMSE$ as well as a higher R^2_{MZ} (compared to the benchmark Model 0), we conclude that the gains in forecasting are economically insignificant. Only for zCash the reduction in $RMSE$ is sizeable; when SVIs are included in the model, the $RMSE$ is almost divided in half.

For Ethereum, the $RMSE$ is reduced by roughly 20% when including all three SVIs for the coin-name, the search-term *cryptocurrency* and the search-term *Bitcoin* which is a non-negligible reduction from an economic point of view. For all other coins the reduction of the $RMSE$ is often limited to a few basis points, for example in the case of Bitcoin. While the $RMSE$ is always reduced when using Models 1 to 5 instead of Model 0, the forecast error is still huge. It ranges from 0.0426 to 0.15. By Chebyshev's inequality, in the case of Bitcoin which has the lowest $RMSE$ of 4.26%, this means that in up to 50% of all forecasts, the absolute value of the forecast error is larger than 6%.

For the volatility models, the evaluation measures are presented in Table 2.7. First, for Gnosis the $RMSE$ cannot be reduced at all when any of the SVI variables are added to model 0. For EOS-Token and zCash, the $RMSE$ is reduced, albeit the reduction not being statistically significant. The addition of SVIs for the coin-names improves the $RMSE$ significantly for Bitcoin, Dashcoin, Ethereum Classic, Ethereum, Litecoin, AugurCoin, and Ripple. With the exception of Dashcoin, the latter subsample is the one for which

Table 2.6: In-Sample Fit VAR Model for Returns

The table lists the root mean-squared error ($RMSE$) multiplied by 100 of the in-sample predictions of the models as well as the R^2 of a Mincer-Zarnowitz regression Mincer and Zarnowitz (1969) in percentages. Using the forecast evaluation test of **?** for nested models, the $RMSE$ of Models 1-5 can be tested whether on whether they result in smaller $RMSE$ than the one of Model 0, our benchmark model. One star indicates that the null hypothesis that the $RMSE$ of our benchmark model is smaller can be rejected on a 10% significance level, two stars signify rejection on the 5% significance level. For the $RMSE$ and the QL, the smallest value, for each cryptocoin, across the models is typeset in bold. For the R^2_{MZ}, the highest value is reported boldfaced.

	Model 0		Model 1		Model 2		Model 3		Model 4		Model 5	
	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}
BitcoinCash	8.96	1.66	8.89	2.14	8.77	1.41	**8.77**	1.47	8.89	2.19	8.89	**2.24**
Bitcoin	4.30	0.00	4.27**	0.90	**4.26****	**1.56**	–	–	4.26**	1.44	–	–
Dashcoin	6.77	0.00	6.74**	1.49	6.74**	1.21	6.73**	1.29	6.73**	1.44	**6.72****	**1.62**
EOSToken	15.24	0.00	15.11**	0.67	15.03**	1.78	**15.00***	1.26	15.03**	1.78	15.03**	**1.78**
EthereumClassic	7.30	0.00	7.27**	0.73	7.27**	1.00	7.26**	1.23	7.26**	1.00	**7.25****	**1.36**
Ethereum	8.05	0.00	6.90**	1.02	**6.79****	**1.98**	6.93**	0.36	6.90**	1.28	6.90**	1.09
Gnosis	6.60	0.00	**6.54****	**1.40**	6.54**	1.39	6.54**	1.28	6.58*	0.27	6.57*	0.40
Litecoin	7.17	0.00	6.99**	1.67	**6.99****	**1.71**	7.01**	0.94	7.09**	1.08	7.07**	1.66
AugurCoin	8.60	0.55	8.56	0.41	8.57	0.23	8.53	1.19	8.56	0.52	**8.51***	**1.66**
Monero	7.06	0.00	6.98**	1.20	6.98**	1.22	6.96**	1.79	6.97**	1.27	**6.95****	**1.88**
Ripple	8.21	0.00	8.08**	3.52	8.09**	3.28	8.14**	2.03	8.05**	4.02	**8.03****	**4.48**
zCash	14.45	**19.44**	8.36**	9.77	8.37**	9.56	8.39**	9.10	8.34**	10.20	**8.33****	10.53

we also document Granger causality (cp. Table 2.5). With the exception of Litecoin, the addition of general interest SVIs can further reduce the $RMSE$ and improve the model fit. For Monero, the best fitting model is Model 2 which contains the SVI for the search-term *cryptocurrency*, but not the SVI for its name. Lastly, for Bitcoin cash, Model 4 (which includes search-terms for the coin's name and *cryptocurrency*) is selected unanimously.

In general, the comparison of the in-sample fit across models is only clear for a few coins. In summary, according to the in-sample results, the search volume of yesterday (for one term or another) helps to predict today's volatility for the majority of coins, but is not a reliable predictor of today's return.

2.4.2 Out-of-Sample Forecast Evaluation

To evaluate whether the focus of the search interest on Google really helps to predict returns or volatility, we perform a one day, one week and 2 weeks ahead out-of-sample forecast of returns and volatility by estimating model specifications 0 to 5 on a growing sample size. We start with 180 observations and add one observation at a time to predict the next day, next week (7 days) or the next two weeks (14 days). In order to arrive at a 2 week forecast for volatility, after transforming the forecasted, daily, log-volatility values to volatility, we sum them up. Log-returns are also summed up. During each estimation step the optimal lag-length is determined anew via the SIC based on the observations within the growing sample window. As the one-step-ahead forecasts are exemplary for the one-week (7-steps-ahead) and two-week (14-steps-ahead) forecasts, we only report these. All other results are available from the authors upon request.

The one-day-ahead forecasts bear some remarkable results. In the case of the return models, for four coins (BitcoinCash, Ethereum, Monero, and zCash) the $RMSE$ is significantly improved when Google search volumes are added to model 0 as can be seen from Table 2.8. In all other cases, albeit we might observe a reduction of the $RMSE$, it is not statistically significant. The R^2_{MZ} is very low: In general, the variation in the forecast can explain less than 1% of the variation in the observed returns. Hence, the prediction can be considered random which is illustrated in Figures 2.3a and 2.4a.

Figure 2.3a presents a scatter plot of the forecasted and the observed returns for Bitcoin. The black line is the 45° line which marks the location of a perfect fit. As can be seen, the forecasted values do not vary a lot, the prediction is always close to zero. Hence, the location of the points is limited to a wide, ellipse-like area around the origin. For all other cryptocurrencies, the shape of such a scatter plot is similar.

Figure 2.4a provides further details. We zoom into the prediction based on the best fitting model for Bitcoin returns. The autoregressive process specified in Equation (2.1) cannot

Table 2.7: In-Sample Fit VAR Model for Volatility

The table lists the root mean-squared error ($RMSE$) multiplied by 100 of the in-sample predictions of the models as well as the R^2 of a Mincer-Zarnowitz regression **?** in percentages. Furthermore, the Quasi-maximum likelihood measure (QL) is reported as an outlier robust measure (cp. Patton 2011). Using the forecast evaluation test of **?** for nested models, the $RMSE$ of Models 1-5 can be tested whether they result in a smaller $RMSE$ than Model 0, our benchmark model. One star indicates that the null hypothesis (that the $RMSE$ of the benchmark Model is smaller) can be rejected on a 10% significance level, two stars signify rejection on the 5% significance level. For the $RMSE$ and the QL, the smallest value, for each cryptocoin, across the models is typeset in bold. For the R^2_{MZ}, the highest value is reported boldfaced.

	Model 0			Model 1			Model 2			Model 3			Model 4			Model 5		
	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}
BitcoinCash	0.92	6.05	30.84	0.83	5.37	35.46	0.81	5.36	33.47	0.81	5.39	32.62	**0.79**	**5.22****	**37.55**	0.82	5.35	35.95
Bitcoin	1.11	3.01	33.75	1.05	2.97**	35.52	1.11	3.01**	34.01	–	–	–	**1.05**	**2.97****	**35.53**	–	–	–
Dashcoin	0.97	4.73	18.78	**0.91**	4.66**	21.44	0.91	4.65**	21.73	0.94	4.69**	20.00	0.91	4.65**	21.72	0.91	**4.64****	**21.94**
EOSToken	0.82	6.66	32.03	0.68	5.57	37.12	0.61	4.10	30.28	**0.61**	**4.09**	30.70	0.68	5.57	37.15	0.67	5.52	**38.93**
EthereumClassic	0.91	4.85	30.70	0.76	4.28**	**35.79**	0.76	4.29**	35.50	**0.74**	**4.26****	35.77	0.80	4.34**	33.52	0.80	4.34**	33.40
Ethereum	1.17	7.95	22.07	0.91	4.65**	34.38	0.92	**4.62****	33.98	0.90	4.63**	33.70	0.90	4.64**	**34.54**	**0.89**	4.64**	34.48
Gnosis	**0.99**	**5.87**	**19.44**	1.01	5.88	19.23	1.00	5.88	19.38	1.00	5.89	19.11	1.01	5.88	19.33	1.03	5.97	16.70
Litecoin	1.39	5.83	27.45	1.38	5.77**	28.91	1.35	5.77**	28.39	1.32	5.74**	28.98	1.37	5.75**	29.27	**1.30**	**5.68****	**30.21**
AugurCoin	0.96	6.20	21.87	0.85	**5.65****	25.20	0.89	5.78	23.65	0.89	5.79	23.48	**0.84**	5.69**	**26.62**	0.84	5.69**	26.51
Monero	0.68	3.85	30.13	0.68	3.85	29.92	0.68	**3.79****	**31.88**	0.66	3.86	29.72	0.66	3.83**	30.93	**0.66**	3.83**	30.58
Ripple	0.92	5.07	34.43	0.87	4.86**	**39.23**	0.92	5.05	34.48	0.91	5.10	32.31	0.87	**4.86****	39.10	**0.85**	4.89**	37.87
zCash	1.24	10.82	42.00	**0.96**	**6.60**	41.27	0.98	6.91	**50.57**	1.00	7.00	47.82	0.98	6.94	50.15	0.98	6.96	49.03

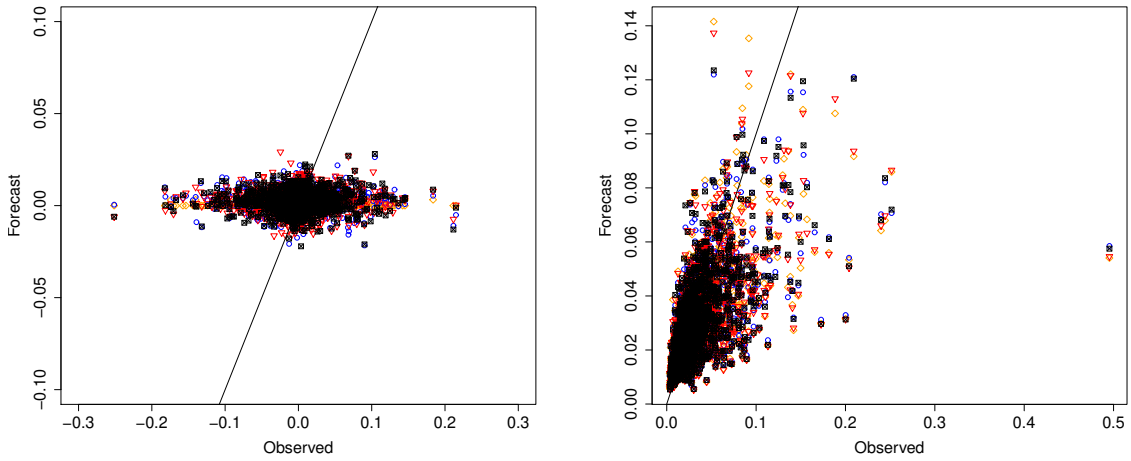
Table 2.8: Out-of-Sample Fit One-Day-Ahead Forecast – Returns

The table lists the root mean-squared error ($RMSE$) multiplied by 100 for the one-day-ahead out-of-sample predictions of returns for each model as well as the fit of a Mincer-Zarnowitz regression R^2_{MZ} **?** in percentages. Furthermore, the Quasi-maximum likelihood measure (QL) is reported as a outlier robust measure (cp. Patton 2011). Using the forecast evaluation test of **?** for nested models, the $RMSE$ of Models 1-5 can be tested whether they result in a smaller $RMSE$ than Model 0, our benchmark model. One star indicates that the null hypothesis (that the $RMSE$ of the benchmark Model is smaller) can be rejected on a 10% significance level, two stars signify rejection on the 5% significance level. For the $RMSE$ and the QL, the smallest value, for each cryptocoin, across the models is typeset in bold. For the R^2_{MZ}, the highest value is reported boldfaced.

	Model 0		Model 1		Model 2		Model 3		Model 4		Model 5	
	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}	$RMSE$	R^2_{MZ}
BitcoinCash	6.97	0.54	6.93*	0.13	7.02	**1.49**	6.98	0.73	**6.29**	0.19	6.32	0.42
Bitcoin	**3.87**	0.00	3.90	**0.11**	3.89	0.01	–	–	3.89	0.01	–	–
Dashcoin	**6.64**	0.11	6.69	0.16	6.69	0.16	6.67	**0.39**	6.72	0.08	6.71	0.22
EOSToken	**5.69**	0.00	5.79	0.44	5.96	**0.73**	6.00	0.00	5.93	0.18	5.98	0.55
EthereumClassic	**7.67**	0.00	7.75	**0.24**	7.74	0.08	7.75	0.02	7.81	0.17	7.81	0.07
Ethereum	6.68	0.08	**6.62****	**0.17**	6.67**	0.11	6.64**	0.06	6.68	0.02	6.68	0.00
Gnosis	**5.29**	0.00	5.35	0.49	5.37	2.02	5.31	0.24	5.43	**3.04**	5.44	2.81
Litecoin	**5.93**	0.00	6.04	0.00	6.06	0.00	6.01	**0.01**	6.08	0.00	6.10	0.00
AugurCoin	**9.40**	0.12	9.46	1.08	9.45	**1.63**	9.49	0.04	9.70	1.29	9.54	0.03
Monero	7.14	**0.79**	7.19	0.03	**7.14**	0.26	7.14*	0.60	7.18	0.06	7.19	0.23
Ripple	8.08	0.00	8.19	0.25	8.27	0.06	8.11	0.29	**7.49**	0.10	7.96	**0.31**
zCash	10.27	0.07	7.52**	0.00	**7.41****	0.02	7.49**	0.03	7.51**	0.06	7.57**	**0.19**

Figure 2.3: Fit of One-Day-Ahead Forecasts

The graphs depict the observed returns (horizontal axis) against the one-day-ahead forecasts based on Model 0 (orange $\diamond$), Model 1 (blue $\circ$), Model 2 (red $\triangledown$) and Model 5 (black $\boxtimes$) for Dashcoin (left) and Monero (right). A perfect fit would mean that the values are aligned on the 45°-degree line (black; note the scaling of the axes).



(a) Bitcoin Returns

(b) Bitcoin Volatility

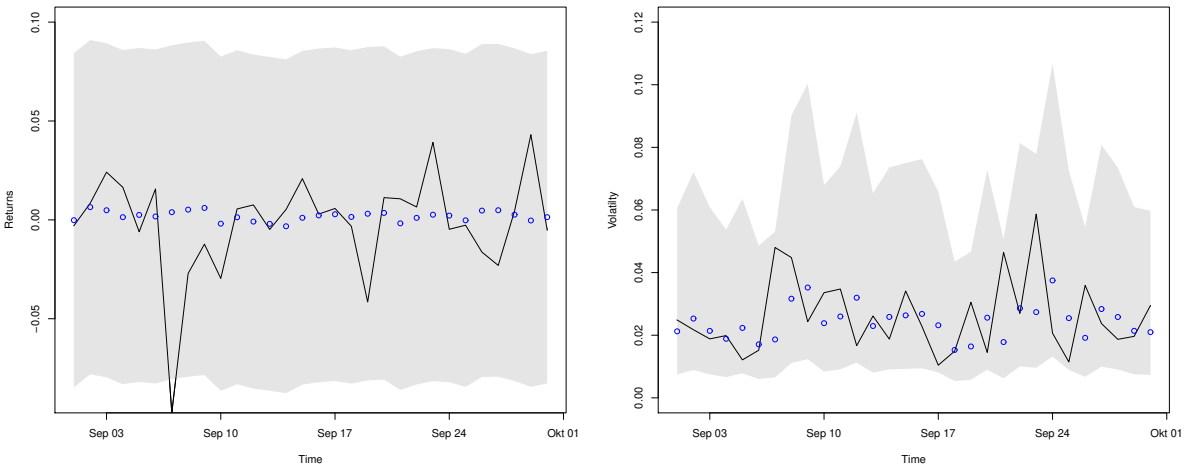
mimic the high volatility of the return series and therefore results in smooth forecasts centered around zero. Furthermore, individual large spikes result in a forecast that is further away from zero, albeit one period too late. In general, we conclude that for the out-of-sample prediction of returns on a daily basis, Google searches are not helpful.

Table 2.9 holds the forecast evaluation results for the volatility models. In general, the one-day-ahead forecast of volatility is improved when one or more of Google's SVIs are added to Model 0. Only for Ethereum Classic and AugurCoin none of the SVI models reduces the $RMSE$ significantly. For Dashcoin the QL selects the benchmark Model 0, while $RMSE$ and R^2_{MZ} favor Model 4. The R^2_{MZ} of the BitcoinCash-models favors the benchmark Model 0, while QL and $RMSE$ choose Model 4. With the exception of these four coins, the evaluation criteria of all other coins select a model in which Google's SVI is added.

Figure 2.3b shows the scatter plot of the actual and forecasted volatility for Bitcoin. There are two important takeaways. First, the model fit is much better than for returns (depicted in Figure 2.3a) as the values are much more clustered around the 45° line. Second, the points that belong to Model 0 (orange) are further away from the 45° line than the points associated with any other model. In particular, the points which belong to Model 5 (black) are the ones that are the closest to the line of the perfect fit, especially for high values of volatility. Hence, we conclude that in general, the addition of Google-search volume helps

Figure 2.4: Time Series of One-Day-Ahead Forecasts

The graph presents the observed time series of Bitcoin returns (a) and volatility (b) as a black solid line and the one day ahead forecast (blue dots) based on Model 1, only including the SVI of the search-term *Bitcoin*. Confidence intervals on the 0.95% level are shaded in gray.



(a) Bitcoin Returns

(b) Bitcoin Volatility

to predict volatility, which is further illustrated in Figure 2.4b. As in the case of returns, when large changes occur in the volatility series, the forecast does not react as quickly and picks up the movement with one period lag. It turns out that, in periods of extreme volatility, the inclusion of SVIs is more helpful than in periods with low volatility which is in line with the results of Dimpfl and Jank (2016).

2.5 Robustness and Sensitivity Analysis

Our results presented in Section 2.4 are to some extent in contrast to results reported in previous literature. In particular, Kristoufek (2013) documents predictability of Bitcoin returns while Urquhart (2018) finds no predictability of Bitcoin volatility. In the following, we therefore discuss potential reasons which can explain this discrepancy such as the considered time frame, the sampling frequency, or the restriction to linear models.

2.5.1 The Time Frame Matters

To analyze whether the time frame matters, we test Granger causality in rolling windows of 180 days over the entire sample period. For returns, the results are similar to the ones reported above in the sense that we do not find any extended period of time during which SVIs would Granger cause returns. For volatility, Figure 2.5 illustrates the resulting

Table 2.9: Out-of-Sample Fit One-Day-Ahead Forecast – Volatility

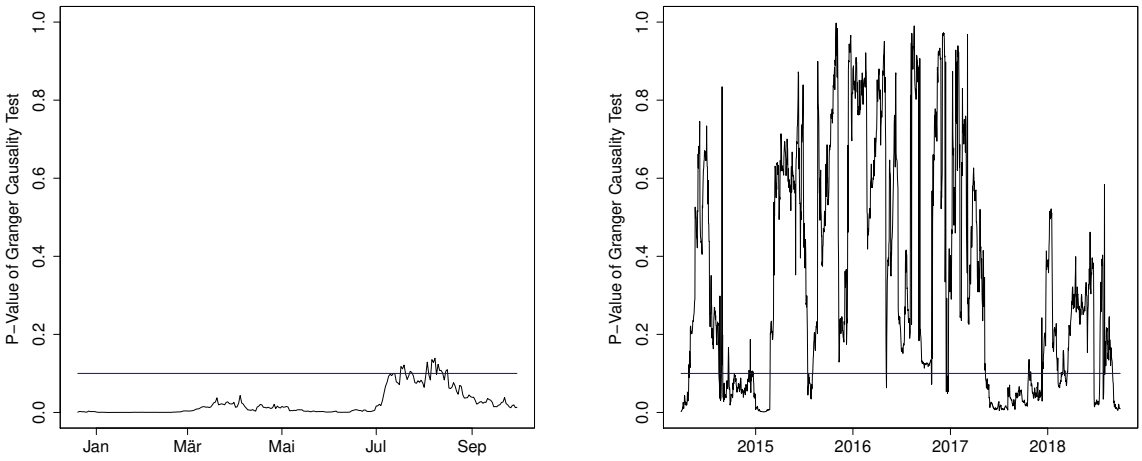
The table lists the root mean-squared error ($RMSE$) multiplied by 100 for the one-day-ahead out-of-sample predictions of volatility for each model as well as the fit of a Mincer-Zarnowitz regression R^2_{MZ} **?** in percentages. Furthermore, the Quasi-maximum likelihood measure (QL) is reported as a outlier robust measure (cp. Patton 2011). Using the forecast evaluation test of **?** for nested models, the $RMSE$ of Models 1-5 can be tested whether they result in a smaller $RMSE$ than Model 0, our benchmark model. One star indicates that the null hypothesis (that the $RMSE$ of the benchmark Model is smaller) can be rejected on a 10% significance level, two stars signify rejection on the 5% significance level. For the $RMSE$ and the QL, the smallest value, for each cryptocoin, across the models is typeset in bold. For the R^2_{MZ}, the highest value is reported boldfaced.

	Model 0			Model 1			Model 2			Model 3			Model 4			Model 5		
	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}	QL	$RMSE$	R^2_{MZ}
BitcoinCash	0.41	2.80	**25.73**	0.41	2.79	25.32	0.43	2.79*	25.67	0.41	2.78**	25.71	**0.39**	**2.50****	11.78	0.41	2.58	9.26
Bitcoin	0.84	2.42	37.83	**0.80**	**2.39****	**39.27**	0.84	2.42**	38.14	–	–	–	0.82	2.41**	39.13	–	–	–
Dashcoin	**0.83**	4.82	21.18	0.84	4.80*	22.23	0.84	4.77**	23.69	0.88	4.83	21.27	0.84	**4.76**	**24.43**	0.87	4.81	22.86
EOSToken	0.36	2.32	0.35	0.37	2.37	0.09	**0.33**	**2.25****	**1.00**	0.34	2.27**	0.48	0.36	2.35	0.55	0.35	2.41	0.17
EthereumClassic	0.74	**4.14**	34.00	0.76	4.16	33.93	0.74	4.16	**34.04**	0.78	4.24	30.77	**0.74**	4.18	33.74	0.76	4.20	31.77
Ethereum	0.94	4.63	28.92	0.90	**4.61****	30.17	0.91	4.62*	29.90	0.94	4.66	28.16	**0.90**	4.63	**30.58**	0.92	4.66	28.82
Gnosis	0.85	4.06	0.20	0.86	4.04	0.27	0.89	4.10	0.19	0.86	**4.03***	**0.30**	0.89	4.11	0.23	**0.81**	4.11	0.20
Litecoin	1.19	3.68	34.18	1.21	3.66**	34.53	1.21	3.67**	34.47	**1.16**	3.66**	34.58	1.20	**3.59****	**35.49**	1.20	3.65**	34.93
AugurCoin	1.07	6.58	19.90	**1.01**	**6.55**	**22.20**	1.10	6.63	19.59	1.13	6.72	18.54	1.05	6.80	21.48	1.04	6.71	21.74
Monero	0.69	3.98	23.94	0.71	4.00	23.35	0.70	**3.96***	**24.84**	**0.69**	4.00	23.46	0.71	3.99	24.30	0.71	3.97	24.71
Ripple	0.79	5.25	35.61	0.84	5.10**	**38.91**	0.82	5.24	35.33	0.83	5.31	32.75	**0.74**	**4.82**	35.25	0.77	4.90	36.05
zCash	0.97	5.50	19.58	1.00	5.49*	18.40	0.97	**5.45****	**19.86**	**0.96**	5.47*	19.34	0.99	5.53	19.11	0.98	5.46*	19.58

p-values for each window for Ripple and Bitcoin. Here we find that there exist multiple periods for which Granger causality is indeed a stable phenomenon. This holds particularly true for Ripple (Figure 2.5a). Hence, the time frame of the considered sample for such an analysis does matter. In sum, the propensity to find favorable results (i.e., significant Granger causality) is higher for volatility than for returns. For the latter, this is unlikely, but not impossible.

Figure 2.5: Granger Causality Test over Time: Daily Data

The graphs depict the p-values of a Granger causality test conducted in the context of Model 1 (which includes the SVI for the coin name) over time. The respective null hypothesis is that the SVI does not Granger cause volatility. The point in time associated with a p-value is the last date of a rolling window of 180 days. For values below the blue horizontal line, the null hypothesis of no Granger-causality can be rejected.



(a) Ripple Volatility

(b) Bitcoin Volatility

This result makes sense in the light of the findings of Garcia and Schweitzer (2015) who describe herding and bubble building behavior in the cryptocurrency market. Together with the time series characteristics of the herding model described by Alfarano, Lux and Wagner (2008), the results for the time varying nature of the relationship between Google Trends and volatility of cryptocurrencies becomes more plausible. The model for which Alfarano et al. (2008) describe time varying parameters was first introduced into the economic literature by Kirman (1993). The stochastic model of Kirman (1993) transfers the finding of Pasteels, Deneubourg and Goss (1987) into the economic realm. Pasteels et al. (1987) found that a colony of ants, faced with two equi-distant food sources, behave collectively asymmetric: They herd towards one food source. Kirman (1993) observed the behavior in recruitment processes and suggested a stochastic model to describe it. Alfarano et al. (2008) transferred the model then to financial markets. With its time varying moments, the model suggests a GARCH-like structure for the second moments

of returns. We find it interesting that in our context the time series characteristics of the model described by Alfarano et al. (2008) exhibit similar dynamics to the p-values of the rolling Granger Causality tests we conduct. While there are phases in which Google Trends seem to help predict cryptocoin volatility for a longer period of time (almost up to a year in 2017), the relationship breaks down in others. In the context of cryptocurrencies and Google Trends these dynamics hints towards herding behavior in the attention for and the investments made in the cryptocurrency markets. Aside from the fear of missing out (FOMO) interpretation connected to the herding behavior, active pump-and-dump operators in the market, which are more likely for smaller coins, might drive this result. The orchestration of online coin advertising might have generated interest and investment in a certain coin (pump-phase) reflected in rising Google SVIs and increased volatility in the market. When prices soar, the pumpers cash in (dump-phase), first adding volatility and thereby destroying the connection between volatility and Google's SVI. The findings of Baur and Dimpfl (2018b) also point in the direction of these results. In essence, we conclude that the relationship between Google Trends and cryptocurrency volatility is time varying. Attributing the time varying relationship to herding behavior, FOMO or pump-and-dump schemes described in the literature (e.g. by Garcia and Schweitzer (2015) or Baur and Dimpfl (2018b)) seems plausible.

2.5.2 Sampling Frequency Matters

As the literature uses different sampling frequencies, usually weekly data most recently, we investigate how changing it impacts on the predictability of returns and volatility. In the following, we consider weekly and hourly data.

Weekly Data

We conduct the weekly analysis for those cryptocurrencies for which we have more than 100 weekly observations. This leaves us with three cryptocurrencies: Bitcoin, Litecoin and Ethereum. The restriction is rooted in the Google SVIs which, when downloaded on a weekly basis, contain numerous missing values as the number of searches was below the reporting threshold set by Google.

For the volatility models, we find again overall significance of the models and Granger causality of Google's SVIs. Table 2.10 summarizes the results. Even if we switch to prices denoted in US Dollar and SVIs calculated from US searches only, the relationship between volatility and Google searches remains robust. We conclude that Google search volume helps to predict cryptocurrency volatility on a weekly basis.

In contrast, for returns the results are not as clear-cut. In the case of weekly Litecoin returns, the SVI based on the coin-name bears significant explanatory power and Granger causes returns. For weekly Ethereum returns, only for the model that includes all SVIs, search volume is overall significant in explaining variation in returns and Granger cause them. While for volatility, adding SVIs to the model reduces the $RMSE$ decisively, for the return models, the reduction in the $RMSE$ is limited. Only in the case of Litecoin the $RMSE$ of the return forecast is reduced by more than 50% when the SVI for the coin-name and the SVI of the search-term *cryptocurrency* is added. The results are also summarized in Table 2.10.

Table 2.10: Compact Results: Weekly Data

The table gives an overview of the weekly in-sample fit (ins) and out-of-sample forecast (oos) evaluation results. The reported pairs for returns (triples for volatility) contain the numbers m of the model selected by (QL,) $RMSE$, and $R^2_M Z$. Percentages are the p-values associated with the respective test. ✓ means that the $RMSE$ reduction is significant on a 10% significance level.

	Bitcoin		**Ethereum**		**Litecoin**	
	ret.	vola.	ret.	vola.	ret.	vola.
Model Significance:	–	all (1%)	5 (1%)	all (1%)	1,4,5 (1%)	all (1%)
Granger Causality:	–	all (1%)	–	all (1%)	–	1 (5%), 2-5 (1%)
Model with best in-sample fit:	(4,4)	(1,4,4)	(5,5)	(5,5,5)	(5,5)	(5,5,5)
Best out-of-sample fit:	(4,2)	(4,4,1)	(4,2)	(4,4,4)	(4,1)	(5,5,1)
Reduction in $RMSE$ significant (ins)?	✓	✓	✓	✓	✓	✓
Reduction in $RMSE$ significant (oos)?	–	✓	✓	✓	✓	✓

While we document that Google search volume helps to predict returns in- and out-of-sample on a weekly basis, one has to bear in mind that the $RMSE$ is huge. Albeit the significant reduction (which is in line with findings of Kristoufek (2013)), the prediction is not useful. Drawing again on Chebyshev's inequality, we may state that in the case of Bitcoin, the forecast will be more than 1400 basis points off the true value in up to 50% of all forecasts. Hence, investing a million today is a risky undertaking, as tomorrow still remains unknown.

Hourly Data

As discussed in Section 2.1, Bitcoin trading is fast and therefore the observational frequency of one hour might be too low already. On an hourly basis, we use data restricted by the availability on Google Trends which date back at most to January 2016.

The in-sample results of our analysis are rather promising. For volatility, we find again that on an hourly frequency all models including Google's SVI have significant explanatory power which is in line with the results documented in Section 2.4. The test results for Granger-causality are very pronounced as well.

For all coins the SVI for the search-term *Bitcoin* Granger-causes hourly volatility. We therefore conclude that hourly volatility in the cryptomarket is Granger-caused by the variation in the general search interest for the cryptomarket and Bitcoin as its flagship.

But also for returns, with the exception of Bitcoin Cash, the SVI models bear significant explanatory power at least on a 5% significance level. Due to the large sample size, of course the standard errors are smaller and, thus, smaller effects turn out statistically significant. However, when testing for Granger causality, the results are much weaker, similar to what is documented in Table 2.5 for daily data. Overall, the evaluation criteria select either Model 4 or 5.

However, when conducting the out-of-sample forecast, the relationship breaks down for both returns and volatility. Thus, we conclude that on an hourly frequency, Google search volume does not help to predict cryptocurrency returns or volatility.

2.5.3 Discussion of the Model Assumptions

Regardless of the variables included in the model, within the class of linear models, the VAR setup in Equation (2.1) results, theoretically, in the best linear predictor (cp. Hamilton 1994). Even in the case of a non-linear specification of the conditional expectation of returns or volatility, the VAR provides the best, MSE minimizing linear approximation to the non-linear conditional expectation function. What we strive for is a conservative analysis based on a tested and carefully constructed data basis. To that end, the linear VAR approach provides a robust framework.

By transforming the variables to logarithmic differences, we also cover a non-linear relationship between the variables. We also tested non-logarithmic differences and obtained qualitatively similar results. Based on our analysis, we find no systematic or general way to forecast returns with Google search queries. On a daily and weekly frequency, we find predictability of cryptocurrencies' volatility by means of Google SVIs. Comparing weekly, daily and hourly data, we conclude that the lower the frequency the more effective is the addition of Google's SVI.

Kristoufek (2013) and Panagiotidis, Stengos and Vravosinos (2018a) document that the distinction whether the Bitcoin price is above or below its past moving average is important. To check this conjecture, we construct a dummy variable which indicates the state of the price following the description in Kristoufek (2013). It turns out that such a variable is in general not statistically significant and does not lead to a significant reduction of the out-of-sample $RMSE$ in our application.

However, the assumption of a linear model might be too restrictive. To check whether the forecast precision can be improved by a non-linear forecasting technique, we considered

the fit of an autoregressive local polynomial regression (referred to as LOESS regression; cp. Cleveland, Grosse and Shyu 1992). On any frequency, the results of the LOESS regression do not improve the forecasting precision compared to the VAR models that include Google's SVIs. Only for the return models without SVIs, the LOESS regression produces in general a slightly smaller $RMSE$ compared to the corresponding Model 0. Still, the $RMSE$ remains huge.

2.6 Summary

Based on the assumption that Google search queries proxy the interest of retail investors, we conduct a forecast analysis of cryptocurrency returns and volatility, similar to the studies of Kristoufek (2013), Dimpfl and Jank (2016), or Urquhart (2018). We consider weekly, daily and hourly sampling frequencies. Overall, we find that the in-sample fit of the models is good irrespective of the frequency, while the out-of-sample performance is mixed. For returns, the predictive ability is negligible and even though the inclusion of SVIs improves the forecast, there is still a giant forecast error left. This result partially contradicts the findings of Kristoufek (2013), but is in line with Aalborg, Molnár and de Vries (2019) who also conclude that Bitcoin returns are not predictable.

Volatility is in general better predictable which is first and foremost rooted in its strong persistence. However, we find that the inclusion of SVIs leads to a significantly better in-sample fit of the models on all considered frequencies. This is in contrast to Urquhart (2018) who cannot reject the hypothesis that Google's search volume does not Granger cause Bitcoin realized volatility. The out-of-sample prediction results are more mixed. It turns out that on an hourly basis, the best performing model is a pure autoregressive model of volatility. On a daily or weekly frequency, however, the inclusion of SVIs leads to a non-negligible reduction of the forecast error. The daily analysis for the cryptomarket is, therefore, in line with research on stock market indicies (e.g. Dimpfl and Jank (2016)) who find SVIs to help predict volatility of the Dow Jones index.

Irrespective of the considered sampling frequency, the time period analyzed has an impact on the results. This is a particular issue when conducting only an in-sample test for model fit and Granger causality. We show for our daily data that there are extended periods for which Granger causality could be established, but that there are more periods when it would have to be rejected. This issue is more important for returns than for volatility.

Overall we conclude that the inclusion of Google's search volume indices can help to predict cryptocurrency volatility, but does not help to predict returns. In this respect, cryptocurrencies are much more similar to equity than currency markets – a profane insight into cryptocurrency markets, yet another call from reality.

Chapter 3

Review of Infinitesimal Stochastic Operators on Markov Chains

In Chapter 4, heavy use of the operator algebra, described in Baez and Biamonte (2018), is made. In order to facilitate the reading, I shortly review the key concepts of the operator algebra related to Markov chains. Briefly, I also review the connection to moment generating functions and characteristic functions. These are called functional bases in Weber and Frey (2017) in the context of path integrals. However, in Weber and Frey (2017), this connection is not made explicit.

3.1 Probability Generating Functions

A nice tool for theoretical considerations on probability distributions are (probability) generating functions. They facilitate the handling of discrete random variables that can only assume values in $\mathbb{N}_0$.

A fairly good introduction can be found in Feller (1957, 1968) in Chapter XI. First of all, I start with Feller's definition of a generating function (p.264): 'Let $a_0, a_1, a_2 \ldots$ be a sequence of real numbers. If

$$A(s) = a_0 + a_1 s + a_2 s^2 + \ldots$$

converges in some interval $-s_0 < s < s_0$, then $A(s)$ is called the generating function of the sequence $\{a_j\}$.'

The main take-away from this definition is that with such a generating function, one is able to collect some sequence of numbers $\{a_i\}$ $\forall i \in \mathbb{N}_0$ in a handy object $A(s)$. Each element of the sequence $\{a_i\}$ is connected to the respective index i. The variable s has no significance and can take on any value and may also be complex. What is of relevance in this context is the convergence of the entire series. For this purpose, we may restrict the support of s to some interval in order to guarantee its convergence. The variable s is simply used to

make the collection possible. All that needs to hold is that the sequence $\{a_i\}$ generated in this manner does not take on arbitrarily large values, but is bounded by some finite value. If the sequence $\{a_i\}$ is bounded, then the entire series converges in some interval at least for $|s| < 1$.

One useful application of this object emerges in the context of discrete probability functions. Assume we have a discrete random variable which can assume non-negative values $\{0, 1, 2, \ldots\}$, we can collect the probability masses associated with the respective values in a probability generating function

$$P(z) = \sum_{n=0}^{\infty} \psi_n z^n.$$

Since for every probability distribution $\sum_n \psi_n = 1$ must hold, $P(1) = 1$. This means that probability generating functions converge absolutely at least for $|s| \leq 1$.

Furthermore, we can construct a generating function for the distribution function of the discrete variable from this probability generating function by

$$\begin{aligned} Q(s) &= q_0 + q_1 s + q_2 s^2 + \ldots \\ &= (p_1 + p_2 + p_3 + p_4 + \ldots) + (p_2 + p_3 + p_4 \ldots)s + (p_3 + p_4 + \ldots)s^2 + \ldots \\ &= (1 - p_0) + (1 - p_0 - p_1)s + (1 - p_0 - p_1 - p_2)s^2 \ldots \\ &= \frac{1 - P(s)}{1 - s}. \end{aligned}$$

This generating function converges at least for $|s| < 1$ as the coeffients are less than one (cp. Feller 1957, 1968, p.265).

For example, consider a Poisson distributed random variable X with rate parameter λ. Then we know that the weights of a probability generating function (PGF) should look like $\psi_n = e^{-\lambda}\frac{\lambda^n}{n!}$. Hence, the PGF is given by

$$P(z) = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}z^n = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda z)^n}{n!} = e^{\lambda(z-1)}.$$

Hence, the generating function for the distribution function is

$$Q(z) = \frac{1 - e^{\lambda(z-1)}}{1 - z}.$$

An interesting feature of PGFs is that they are related to the Moment Generating Function (MGF) and the Characteristic Function (CF). The MGF is the PGF evaluated at $z = e^t$.

So in the case of the Poisson distributed variable above, the MGF is given as

$$MGF(t) = P(e^t) = e^{\lambda(e^t-1)}.$$

The CF is the PGF evaluated at $z = e^{it}$ or the MGF evaluated at the imaginary axis $MGF(it) = CF(t)$, i.e.,

$$CF(t) = P(e^{it}) = e^{\lambda(e^{it}-1)}.$$

Last but not least, from the MGF also the the cumulant generating function can be constructed

$$CGF(t) = \ln MGF(t).$$

3.2 Time Evolution with Generating Functions

Probability Generating Functions can be a powerful tool when thinking about the time evolution of stochastic systems. Baez and Biamonte (2018) present how these functions can be used to describe the dynamics of stochastic systems. They make use of two operators: the creation operator a^+ and the annihilation operator a^-. Both operators act on the state of some system that can assume only values in $\mathbb{N}_0$.[19]

The probability distribution of such a system can be described in the form of a power series, which is the just discussed PGF in Section 3.1, i.e.,

$$\Psi(z) = \sum_{i=0}^{\infty} \psi_i z^i. \tag{3.1}$$

One may as well consider i as the count of balls in a box, neatly lined up. The balls appear and disappear randomly, and are distinguishable. ψ_i is the probability that we find the box filled with i balls.

If we know the system to be in one state n with certainty, then the weights $\psi_i = 0 \forall i \neq n$ except for $i = n$ where $\psi_n = 1$. The probability distribution, represented with this power series would be the Dirac Delta function and the associated PGF is given by

$$\Psi(z) = z^n.$$

[19] Weber and Frey (2017) show that other domains like $\mathbb{Z}$ are possible as well. The definition of the basis function, i.e., the object on which they act, determines the support. In our case, we stick to the generating functions discussed in Section 3.1.

If we now put one extra ball in the box, then we can apply the creation operator to $\Psi(z)$:

$$a^+\Psi(z) = z^{n+1}.$$

After the creation operator has acted (with certainty) on the certain state with n balls, i.e. in PGF representation z_n, we are, again with certainty, in a new state with $n+1$ balls, i.e., in PGF representation z^{n+1}. Thus, the creation operator can be represented by the multiplication with $a^+ = z$ in the PGF-context. Similarily, if we remove one ball from the box, then we can apply the annihilation operator to $\Psi(z)$ as

$$a^-\Psi(z) = nz^{n-1}.$$

After the annihilation operator has acted (with certainty) on the certain state with n balls, i.e., in PGF representation z^n, we find the box (with certainty) in one of n states with $n-1$ balls, i.e., in PGF representation z^{n-1}. Note that the state with $n-1$ is not unique. This is because, there are n ways to remove one ball from the box. In the PGF representation, the annihilation operator could be represented as the derivative with respect to z.

In order for the creation and annihilation operator to be infinitesimal stochastic operators they need to preserve total probability. Since the annihilation and creation operators are the transitions in the system, all transitions have to sum up to zero in order to preserve total probability across all states. So, in this example for the creation operator the Dirac Delta function is moved from state n to $n+1$. If probability mass of 1 is created at state $n+1$, a probability mass of 1 needs to be subtracted from state n. Thus, the infinitesimal stochastic annihilation operator is given by

$$(a^+ - 1)\Psi(z) = z^{n+1} - z^n.$$

In the case of the annihilation operator, total probability is preserved by the following infinitesimal stochastic operator:

$$(a^- - a^+a^-)\Psi(z) = nz^{n-1} - nz^n.$$

3.3 Univariate Counting Processes

Equipped with this notation, consider a counting process, as for example the number of customers that have visited a store at a given day. In the morning of this day, we start with $n = 0$. The respective power series is given by

$$\Psi(z) = 1.$$

If now the customers arrive at rate α in the store then the instantaneous change of the PGF would be

$$\frac{\partial \Psi(z)}{\partial t} = \alpha(a^+ - 1)\Psi(z) = \alpha(z-1). \tag{3.2}$$

Generalizing this and collecting the operator as well as the rate into an operator H, we find that this is an ordinary differential equation of the form

$$\frac{\partial \Psi(z)}{\partial t} = H\Psi_0(z), \tag{3.3}$$

where H is called the Hamiltonian of the system and $\Psi_0(z)$ is the initial state of the system. Equation (3.3) is called the Master Equation. Every Markov processes is governed by such a Master Equation. In the financial literature, H is often represented in the form of migration matrices (cp. Lando and Skødeberg 2002). Its matrix representation can be estimated via the Aalen-Johanson estimator (Aalen and Johansen 1978).

Given an initial state $\Psi_0(z)$, Equation (3.3) solves to

$$\Psi(z) = e^{Ht}\Psi_0(z).$$

In the counting process example, the solution, thus, is

$$\Psi(z) = e^{\alpha(a^+-1)t}\Psi_0(z) = e^{\alpha(z-1)t}.$$

Recalling that a^+ represents multiplication by z and the definition of the expansion of the exponential function, this gives

$$\Psi(z) = e^{\alpha(z-1)t} = \sum_{j=0}^{\infty} \exp(-\alpha t)\frac{(\alpha)^j}{j!}z^j = \sum_{j=0}^{\infty} \psi_j z^j,$$

where the weight ψ_j is the probability that j customers have visited at a given day. The resulting PGF represents the Poisson distribution.

3.4 Extension to Integer Numbers

The extension to integer numbers requires a change in the basis function. Weber and Frey (2017) show that there exist several possible basis functions. In order to represent probability distributions on the set of integer numbers, it is useful to change the variable

to $z = e^s$. Recall the relation of the PGF to the MGF in this case from above:

$$\Psi(e^s) = \sum_{n=0}^{\infty} \psi_n e^{sn}, \tag{3.4}$$

where the index is set to n so that it is not confused with the imaginary unit i. The infinitesimal stochastic creation and annihilation operators have to be exchanged with $a^+ = \exp(s) - 1$ and $a^- = \exp(-s) - 1$.

Now consider the situation in which a diner is managed. There is a stack of plates for customers to use. The stack can be reduced by 1 when a customer is served food. It can also be increased by 1 when the dishwasher has cleaned a plate. At each point in time the number of available plates can be evaluated. If no plates are available, but customers are waiting to be served, there is a lack of plates, i.e., a negative number of plates.

The rate at which the dishwasher is cleaning the plates is on average relatively constant at around $\alpha = 3$ plates per minute. However, the rate at which customers arrive in the diner depends on the time of the day and can be described by $\beta(t) = 12(sin(t) + 1)$. Clearly, the number of plates follows a so-called birth-death process. Each day the stack of plates is prefilled with 100 available plates. The process can be sampled using the Gillespie algorithm presented in Chapter 4. Figure 3.1 shows a possible simulation of the system.

Using the above operator algebra, we know that the Hamiltonian of the system is given by

$$H = 3(\exp(s) - 1) + 12(\sin(t) + 1)(\exp(-s) - 1).$$

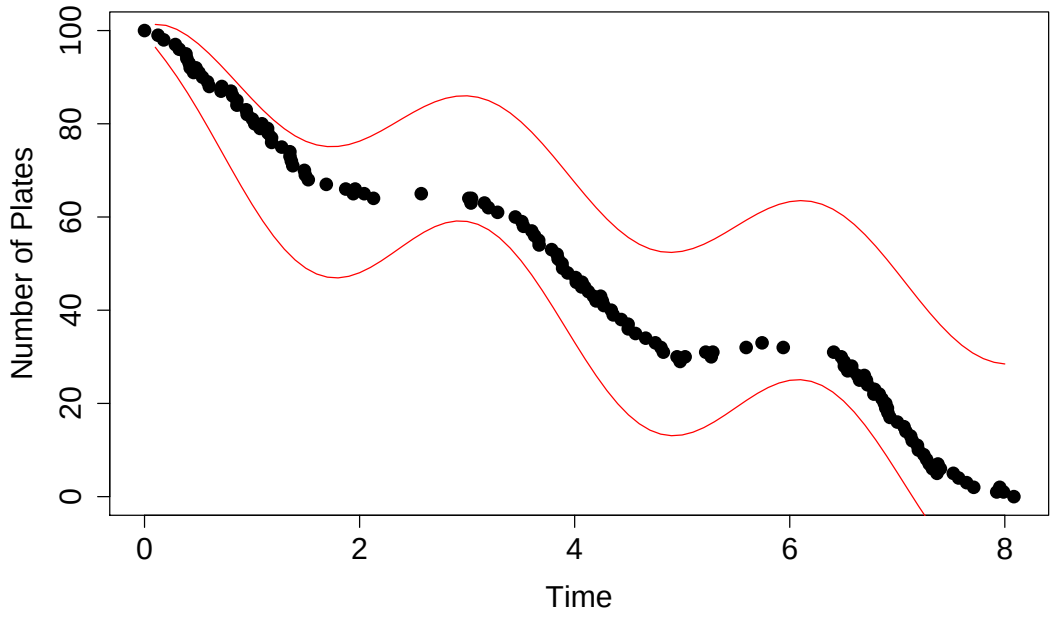
Hence, given the initial condition of $n_0 = 100$ the Master equation solves to

$$\begin{aligned} MGF(s) &= \Psi(e^s) \\ &= \exp\left(\int_0^t 3(\exp(s) - 1) + 12(\sin(t) + 1)(\exp(-s) - 1)dt + sn_0\right) \\ &= \exp\left((e^s - 1)3t + (e^{-s} - 1)(-12(\cos(2t) - 1 - t) + 100s\right) \end{aligned} \tag{3.5}$$

which is the moment generating function of the number of plates. Forming derivatives of this function with respect to s yields the moments of the distribution. This is also how the red bounds were calculated for Figure 3.1: They represent twice the standard deviation from the mean of the distribution. By expanding exponential functions and rearranging terms, it can be shown that the number of plates follows a Skellam-distribution.

Figure 3.1: Simulation of the Diner Example

The graph shows a simulation of the diner example described in the text. The red lines represent twice the distance from the mean of the theoretical Skellam distribution. They have been calculated by differencing the derived moment generating function in Equation (3.5).



3.5 Extensions and the SIRDS-Model

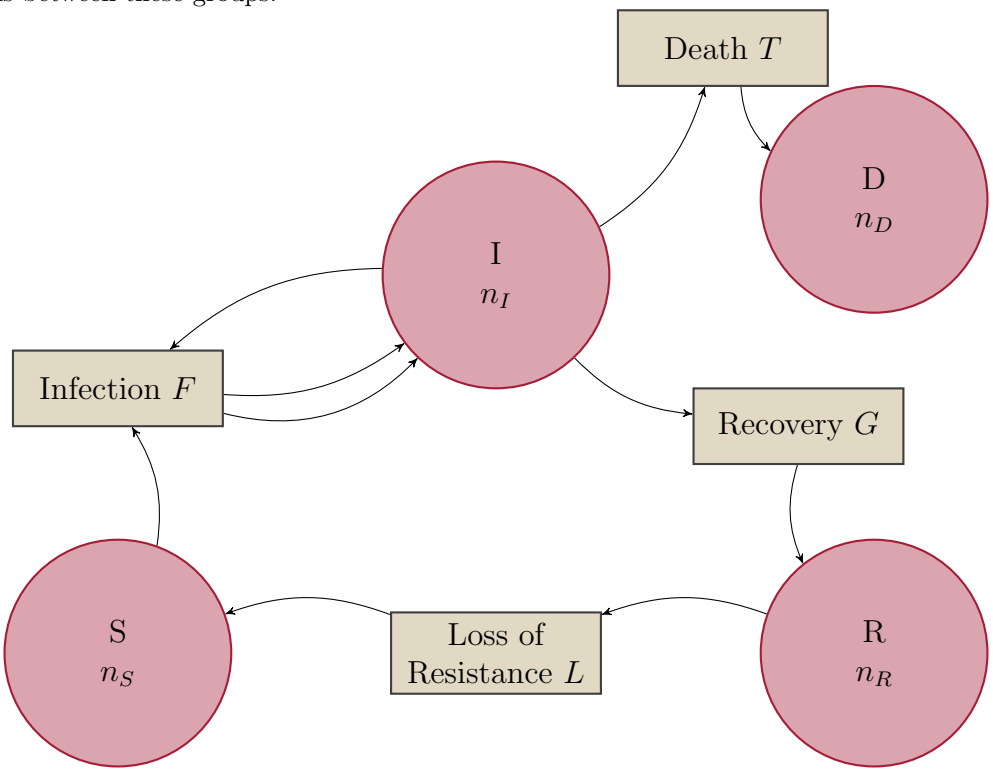
The concept of generating functions for discrete variables can be extended to continuous variables. Examples can be found in Weber and Frey (2017). In Weber and Frey (2017), the functional basis as well as the reference to PGF, CF and MGF is generalized by the usage of Dirac notation with bra and ket vectors. A ket vector represents a mixed or pure state of an observable, much like the PGFs discussed above. For example, if a random variable with support on the natural numbers is with certainty in the pure state n and we choose the power series representation (the PGF) above as the basis, the ket vector $|n\rangle$ could also be represented as z^n. One could also think of the state as an infinite vector filled with zeros and a single one at the $n+1$ element. To each representation of such a state, belongs a linear functional, the bra vector. Bra and ket vectors are the basis of a Hilbert space and its dual. For the PGF representation above, the functional $\langle m|$ acts on an element of the Hilbert space as follows

$$\langle m|\,\Psi = \frac{\partial_z^m}{m!}\Psi(z)\bigg|_{z=0}.$$

For the infinite vector representation, the transpose of the infinite vector would be the corresponding bra vector (up to some normalizing scale factor).

Figure 3.2: The Petri Net of the SIRDS-Model

The following figure illustrates the SIRDS-model. The round nodes represent four groups within a population: The susceptical S, the infected I, the resistant R and the deceased D. The boxes signify transitions between these groups.



So far I have only discussed transitions that increase or decrease a discrete variable by one unit. The infinitesimal operator can be generalized to transitions that increase a variable by m and decrease it by n. The method can also be extended to multivariate systems with transitions between several in- and output variables. To illustrate this extension, I discuss the SIRDS-model which is in the current Covid-19 pandemic often used to model and forecast the spread of the virus.

There are four groups within a population: the susceptical S who may potentially be infected with the virus; the confirmed infections I; the recovered R who have become temporarily resistant; and the deceased D. I use the MGF basis function to derive the Hamiltonian of the system. In this example, I consider the following transitions: death T, infection F, gaining resistance G by recovery and losing the resitance L. The SIRDS model can be illustrated using a petri net; which is presented in Figure 3.2.

The partial Hamiltonians for each of these transitions are given by

$$
\begin{aligned}
T &= D^+ I^- - I^+ I^-, \\
F &= \left(I^+\right)^2 S^- I^- - S^+ S^- I^+ I^-, \\
G &= R^+ I^- - I^+ I^-, \\
L &= S^+ R^- - R^+ R^-.
\end{aligned}
$$

With the corresponding rates $\alpha_T, \alpha_F, \alpha_G, \alpha_L$, the system solves to

$$
\begin{aligned}
\Psi(z) = \exp\Bigg(\int \; & \alpha_T \left((e^{s_D} - 1)(e^{-s_I} - 1) - (e^{+s_I} - 1)(e^{-s_I} - 1) \right) \\
& + \alpha_F \left((e^{s_I} - 1)^2 (e^{-s_I} - 1) - (e^{s_S} - 1)(e^{-s_S} - 1)(e^{s_I} - 1)(e^{-s_I} - 1) \right) \\
& + \alpha_G \left((e^{s_R} - 1)(e^{-s_I} - 1) - (e^{+s_I} - 1)(e^{-s_I} - 1) \right) \\
& + \alpha_L \left(((e^{s_S} - 1)(e^{-s_R} - 1) - (e^{+s_R} - 1)(e^{-s_R} - 1)) \right) dt \Bigg) \Psi_0(z).
\end{aligned}
$$

If the coefficients were time independent, the MGF could be directly determined in a closed form solution otherwise the integral across the time dependent rate functions has to be considered. Having the MGF at our disposal, one can derive the time varying moments of the dynamic system either in a closed form solution or numerically. Evaluating the MGF on the imaginary axis, one can also derive the characteristic function of the time varying joint distribution of the four groups. With a reversed Fourier transform, the joint probability function can then be determined – at least numerically.

3.6 Summary

In this brief chapter, I presented the mathematical tools used in Chapter 4. Especially for continuous time counting processes, the approach presented above is an interesting method to determine the (non-stationary) dynamics and distributions of such processes. However, the higher the number of variables, the more difficult it is to actually find a closed form solution and the more expensive become numerical algorithms to determine the solutions.

Chapter 4

A Stochastic Description of the Limit Order Book to Forecast Intraday Returns[20]

Ever since Glosten (1994) raised the question whether 'the electronic open limit order book [was] inevitable', limit order books (LOBs) have become the most important way of trading, resulting in more than 75% of exchanges around the world (including the recently sprouting cryptocurrency exchanges) using order driven systems and relying (at least partially) on limit order books (Jain 2003). Nevertheless, 'no comprehensive and realistic models (either statistical or economic) exist' (Hasbrouck 2007, Ch. 12, p. 118) which describe the deep-rooted mechanisms of limit order markets in their entirety. While Hasbrouck's statement is more than 10 years old and the literature has made significant progress, a dynamic model which comprehensively describes the interaction of individual orders and is able to incorporate the (strategic) behavior of market participants has, to the best of our knowledge, not yet been developed. Gould, Porter, Williams, McDonald, Fenn and Howison (2013) provide an overview about research on the dynamics of the LOB and identify (roughly speaking) two branches of research. One of them originates in the field of physics and focuses mainly on idealized models to describe statistical features of the LOB system, focusing on dynamic order flows. The second branch is rooted in the economics literature which tends to treat order flows as static. According to Gould et al. (2013), economists primarily focus on the (strategic) behavior of traders, but neglect the dynamical structure. Of course, this reduction is too simplistic and there are multiple attempts to combine strategic behavior and the order book dynamics. For example, Parlour (1998) develops a stylized, dynamic model for the LOB and strategic order placement. Nevertheless, only few models approach the subject by heuristically incorporating statistical regularities observed in market microstructure and incorporate trader interaction based on these statistical observations. A notable exception is Hautsch and Huang (2012) who use high-frequency cointegrated vector autoregressive models to shift the spotlight on the order flow of incoming orders and therewith, based on empirical analysis, draw attention

[20] This chapter is based on Bleher et al. (2020) available at SSRN https://ssrn.com/abstract=3589763 and arXiv https://arxiv.org/abs/2004.11953.

to the intersection of LOB mechanics and strategic behavior of market participants. They show that the revelation of trading intentions through limit order placements affects the LOB states. Large (2007) shows how trades of different size (marketable incoming orders) affect future states of the LOB. Paulsen (2014) derives macroscopic limiting models for a microscopic LOB system in which bid, ask, and transaction prices drive the dynamics of the order flow. These limiting models serve as first order approximations of the stochastic processes that describe the system.

This chapter aims to link models of statistical dynamic order flow and those of strategic interaction. Adapting ideas from Paulsen (2014) and inspired by Baez and Biamonte (2018), we relate the LOB dynamics to the modeling of reaction networks (developed by Baez and Pollard 2017) and present a simple but comprehensive description of the microscopic order book mechanisms. Based on operator algebra, we construct the LOB dynamics bottom up from the elementary events of the book, namely the entry and exit of orders which enables us to model the LOB system by a Markov process. Furthermore, the Hamiltonian, i.e., the operator which governs the time dynamics of the system, can be constructed from these elementary events. Similar to Paulsen (2014), our approach allows to incorporate both the dynamics of the fundamental LOB mechanisms as well as the (strategic) behavior of market participants. Using the event log of the first quarter of 2004 from XETRA, we heuristically develop, simulate and empirically evaluate the implications of our operator formulation. By describing the interaction of individual orders in a purely statistical fashion, the state space of the order book system is worked out which depends only on the rates of arriving and canceled orders as well as on the current state of the book. Based on a limited set of key variables, we show that the return dynamics can be approximated by a linear model. This variable set also allows to linearly forecast returns, arrival rates, and other measures such as order book imbalance and liquidity. Compared to prediction errors reported in the literature such as Zhou, Pan, Hu, Tang and Zhao (2018), forecasts based on these variables reduce the root mean squared prediction error ($RMSPE$) drastically by a factor of 1/10. The in-sample R^2 as well as the R^2 of a Mincer and Zarnowitz (1969) type regression for the out-of-sample predictions show an extraordinary fit for intraday return forecasts.

Similar to our approach, Cont, Stoikov and Talreja (2010) explore the idea of modeling the order book as a Markov process depending on the rates of arrivals and cancellations. They work out closed form solutions for the probability of an increase in the midprice, execution of an order at the best bid price (before a change of the best ask price), and execution of both a buy and a sell order at the best quotes before the price moves. Cont and de Larrard (2013) show that such a model can also be used to calculate the distribution of the duration between transactions. Unfortunately, as Cont and de Larrard (2012) state, these models are based on several assumptions which empirical research has shown

to be incorrect (Bouchaud, Mézard and Potters 2002, Hasbrouck 2007). In particular, the time intervals between order arrivals and cancellations are neither independent nor exponentially distributed and orders are not equally sized. Another model which is based on the Markov properties of the order book has been developed by Daniels, Farmer, Gillemot, Iori and Smith (2003) and extended by Smith, Farmer, Gillemot, Krishnamurthy et al. (2003). However, the model imposes broad restrictions on the functional structure, the parameters and the assumptions about the stochastic processes which govern the LOB dynamics. Again, some of their assumptions like equal order size or balanced order flow, order placement with uniform probability, among others are 'too simple to be literally true' (Smith et al. 2003), but the resulting insights provide a useful foundation for LOB modeling.

The chapter proceeds as follows: Section 4.1 sets out the description of a typical LOB by an algebra of operators. Simultaneously, we present some selected empirical characteristics of the order book which guide our model development. Section 4.2 presents the XETRA order book data. In Section 4.3, we report the results of a simulation study where we identify key factors/drivers of the order flow which determine the statistical distribution of what will happen in the order book and also when and where it happens on short time horizons. Finally, Section 4.4 holds an empirical analysis of the XETRA LOB and Section 4.5 concludes this chapter.

4.1 The Model

The limit order book is the place where traders' orders meet. These orders carry information about a trader's willingness to accept a certain price, the limit price, in exchange for the chosen number of instruments, or vice-versa. The price level at which two orders are matched is called *the reference price*. Within the LOB, we distinguish buy (ask) order and sell (bid) orders. Perceiving these two order types as species which populate price and order size levels inspired the use of the mathematical tools presented in Baez and Biamonte (2018). At any point in time the exchange keeps track of all orders within the LOB. Aside from the market side, the location within the LOB is defined by three key components: the limit price up to which the trader wants to buy or sell, the number of securities, and the time when the order arrives in the LOB.

If traders require immediacy, they rely on *market orders* which can be thought to have an infinite (bid order) or zero (ask order) limit price depending on the market side they were issued from. These orders are matched immediately and, normally, do not reside in the order book for an extended amount of time. They enter the LOB at the best price level of all limit orders currently residing in the LOB on the opposite market side.

The majority of orders, however, are designed to remain in the book for some time at their specified limit price level. These are called *limit orders* which (as we will show in Section 4.2) make up for roughly 90% of all orders in the XETRA LOB while only 3% are market orders. Limit orders have a well defined location in the price dimension. If their designated price location is behind the best price level of all limit orders which currently reside in the LOB on the other side of the market, they are matched (partially) before they can reach their designated limit price level. The smallest populated price level on the sell side of the market is the *best ask* and the highest price level on the bid side is the *best bid*. We will refer to one or the other as the *best quote*.

There are generally further order types that are only submitted to the market if certain conditions are met, for example *stop orders*, which are inserted in the LOB contingently on the reference price reaching or falling below a certain threshold price, or *XETRA BEST orders*. But once such orders enter the LOB, they are effectively equivalent to market or limit orders. As conditional orders can be perceived as more sophisticated versions of limit or market orders, they can in principle be incorporated within the framework presented below. For the purpose of this chapter, we restrict our considerations, thus, to plain market and limit orders.

4.1.1 The LOB Algebra

In the following, we describe order creation and cancellation in the LOB as determined by the rules of a typical order book. For this purpose, we borrow the so-called *Dirac* or *Bra-Ket Notation* from physics, where the state of a system is denoted by a *ket* $|\psi\rangle$. This notation was already introduced in Section 3.5, in the previous chapter. In Section 4.1.2, we will discuss the underlying notion of a state in detail. For now, we can refer to any possible configuration of the order book with $|\psi\rangle$. Even further, we can also assign weights (probabilities) to each of these possible configurations and refer to such a weighted bundle of pure states by $|\psi\rangle$. Nevertheless, we start off with a very concrete state: the empty order book (or vacuum) $|0\rangle$. From this vacuum state, more complicated order book states are created by successively acting on it with creation and annihilation operators. As we will see below, the rules of the LOB induce certain commutation relations in the algebra of these operators. It will be convenient to also introduce the notation

$$|0\rangle = |0| \tag{4.1}$$

which represents an empty ledger.

Rule 1a (Ask Order Submission)**.** *Traders can submit a limit ask order of quantity q at a specified price level k. The order is represented by a creation operator $a^+_{k,q}$ that acts on the order book state* from the right.[21]

For example, if n ask orders are residing in the book, each with its associated limit price k_i and size q_i, $i \in 1, \ldots, n$, the order book state is given by

$$|0|a^+_{k_1,q_1} \ldots a^+_{k_n,q_n}. \tag{4.2}$$

Rule 1b (Bid Order Submission)**.** *Traders can submit a limit bid order of quantity q at a specified price level k. The order is represented by a creation operator $b^+_{k,q}$ that acts on the order book state* from the left*.*

Analogously, for m bid orders residing in the book with specified limit prices k_j and sizes q_j with $j \in 1, \ldots, m$, the following string of operators describes the current state:

$$b^+_{k_m,q_m} \ldots b^+_{k_1,q_1}|0|. \tag{4.3}$$

Note, so far, the rules only describe the successive submission of orders. In particular, we do not yet have a rule that would allow us to reorder the queue of creation operators. Put differently, creation operators generally do not commute: $a^+_{k,q}a^+_{s,p} \neq a^+_{s,p}a^+_{k,q}$. As a result, the strings of creation operators of ask and bid type are time-ordered.

Rule 2a (Ask Order Cancellation)**.** *Traders can cancel a previously submitted ask order. An ask order cancellation is represented by an annihilation operator $a^-_{k,q}$ which acts from the right and satisfies*

$$|0|a^+_{k,q}a^-_{k,q} = |0|. \tag{4.4}$$

Clearly, the probability of a cancellation must be zero if there is no order in the book. This means that when an annihilation operator acts on the empty order book, it generates a state with probability mass zero:

$$|0|a^-_{k,q} = 0. \tag{4.5}$$

[21] This choice will become relevant in the context of price-time priority, see Rule 3.

There is a standard argument, that there is always one more possibility to create and then delete an object than deleting and then creating one. In terms of the operators this argument is represented by the commutation relation[22]

$$[a^+_{k,q}, a^-_{k,q}] = 1$$

where $[A, B] := AB - BA$ denotes the commutator of two operators. In fact this relation directly follows from (4.4) and (4.5)

$$|0|[a^+_{k,q}, a^-_{k,q}] = |0|(a^+_{k,q}a^-_{k,q} - a^-_{k,q}a^+_{k,q}) = |0|a^+_{k,q}a^-_{k,q} - |0|a^-_{k,q}a^+_{k,q} = |0|.$$

Furthermore, since the cancellation of an order $a^+_{k,q}$ does not influence other orders $a^+_{s,p}$, we also have

$$[a^+_{s,p}, a^-_{k,q}] = 0,$$

whenever $s \neq k$ and $p \neq q$. We can summarize these algebraic relations as follows:

$$[a^+_{k,q}, a^-_{s,p}] = \delta_{sk}\delta_{pq}, \tag{4.6}$$

where δ_{ij} is the Kronecker-Delta, defined on an index set $\mathcal{I} \ni i, j$ by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else.} \end{cases}$$

In fact, these commutation relations are usually viewed as the defining properties of creation and annihilation operators.

Rule 2b (Bid Order Cancellation)**.** *Traders can cancel a previously submitted bid order. A bid order cancellation is an annihilation operator $b^-_{k,q}$ that acts from the left. By analogy with the ask cancellations, it satisfies*

$$b^-_{k,q}|0| = 0, \tag{4.7}$$

$$[b^-_{k,q}, b^+_{j,p}] = \delta_{kj}\delta_{qp}. \tag{4.8}$$

In comparison to the commutation relation of ask orders, the order of annihilation and creation operators is reversed since bid orders act from the left. When there are several identical limit orders, i.e., orders with the same price level and quantity, we can distinguish

[22] In physics, these commutation relations are known as *canonical commutation relations.*

their position in the order queue by means of their time stamp. In contrast, for cancellation orders, an observer cannot predict which of the identical limit orders is supposed to be canceled. The algebraic formalism captures this uncertainty: up to normalization, the commutation relations lead to a stochastically mixed state that contains each possible cancellation, for example

$$(|0|a^+_{k,q}a^+_{r,s}a^+_{k,q})a^-_{k,q} = |0|a^+_{k,q}a^+_{r,s} + |0|a^+_{r,s}a^+_{k,q}.$$

Remark. We also introduce the convention that arrivals and cancellations with size $q = 0$ are equivalent to the identity operator. This is motivated by the fact that such arrivals and cancellations in practice do not exist. However, if they existed, they would render the current LOB state unchanged:

$$a^+_{k,0} = a^-_{k,0} = b^+_{k,0} = b^-_{k,0} = 1.$$

Rule 3 (Price-Time Priority)**.** *Orders are organized according to price-time priority.*

The order book state is the result of successive order submissions and the corresponding string of operators is strictly ordered by time. Hence, we get a *price-time ordering* by rearranging the operators into groups with identical price level whilst maintaining the time ordering within each group. This can be achieved by letting ask and bid orders commute whenever they have different price levels $k \neq s$:

$$[a^+_{k,q}, a^+_{s,p}] = 0, \tag{4.9}$$

$$[b^+_{k,q}, b^+_{s,p}] = 0. \tag{4.10}$$

Using these relations, the order book state can always be written in the price-time ordered form

$$|\psi\rangle = \; b^+_{k_1,q_1} \cdots b^+_{k_n,q_n} \; |0| \; a^+_{k_{n+1},q_{n+1}} \cdots a^+_{k_{n+m},q_{n+m}}, \tag{4.11}$$

where $k_i \leq k_{i+1}$. Whenever $k_i = k_{i+1}$, the order nearer to $|0|$ was submitted first.

Given a LOB state in price-time ordered form, the priority of an order is encoded by its distance to $|0|$, where orders closer to $|0|$ have higher priority.

Rule 4 (Order Matching)**.** *Two orders from different market sides* permit a transaction *if they have highest priority and the bid price is bigger or equal to the ask price. When the LOB executes orders that permit a transaction, the quantities are matched up as far as*

possible and unmatched quantities remain in the book. We write $\overbrace{b_{k,q}|0|a_{s,p}}$ *for a pair of executed orders, such that for* $k \geq s$ *the matching procedure is captured by*

$$\overbrace{b^+_{k,q}|0|a^+_{s,p}} = \begin{cases} |0|a^+_{s,p-q} & \text{if } q > p \\ |0| & \text{if } q = p \\ b^+_{k,q-p}|0| & \text{if } q < p \end{cases} \tag{4.12}$$

or as an algebraic relation of creation operators

$$\overbrace{b^+_{k,q}a^+_{s,p}} = \theta(q-p)\ b^+_{k,q-p} + \theta(p-q)\ a^+_{s,p-q},$$

where $\theta(x)$ *is the Heaviside step function*

$$\theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{else.} \end{cases}$$

Recall for the case $q = p$ that orders of size 0 are equivalent to the identity operator.

Since incoming bid (ask) orders always act on a state from the left (right), we need to commute them through older orders until they reach their designated position in the price-time ordered queue. This means orders automatically 'walk the book'[23] until they reach their destined price level k. Along its walk, an order may encounter orders of the other market side and will then be executed as described by Rule 4. It follows that *market orders* are described by creation operators $a^+_{k=0,q}$ and $b^+_{k=\infty,q}$, which will walk all the way through the book until they are completely executed.

Let us stress that the price level k of an order $a^+_{k,q}$ is not necessarily the *transaction price* at which the order will be executed. Instead, the transaction price is usually determined by the price level of the 'settled order' that is encountered by the 'walking order'. The 'settled order', however, may depend on the trading mode (see Section 4.1.5).

Also note that we did not yet specify *when* orders are executed and when a transaction will take place. The reason is that such rules do not add further algebraic relations. The question when orders are matched is not relevant for the description of the current state of the book. However, it is relevant for the time dynamics, i.e., if we examine time series of order book states. The question is whether matching occurs after each event, as continuous trading dictates, or whether matching is only conducted hypothetically after each event to

[23] In the LOB literature, 'walking the book' usually refers to an arriving, marketable order that is executed against several orders on the opposite market side. We borrow this notion of the walking order and extend it. In our case, every order 'walks through the book', however, only marketable orders encounter orders on their way to their destined limit price level.

produce indicative prices like in the pre-auction phase or at the end of an auction. These different modes matter for the evolution of the book. A detailed discussion of transactions and related issues follows in Section 4.1.5.

4.1.2 State Space and the Probability Generating Function

As we have seen in Section 4.1.1, for any given time the state of the order book is given by a price-time ordered string of creation operators

$$|\psi\rangle = \; b^+_{k_1,q_1} \cdots b^+_{k_j,q_j} \; |0| \; a^+_{k_{j+1},q_{j+1}} \cdots a^+_{k_n,q_n},$$

where $k_i \leq k_{i+1}$ for all $i \in S = \{1, \ldots, n\}$. Each operator in this string is specified by its market side $m \in \{a, b\} = \mathcal{M}$, the price level $k \in \mathcal{K} \subset \mathbb{R}$, and the order size $q \in \mathcal{Q} \subset \mathbb{R}$. Here $\mathcal{K}$ and $\mathcal{Q}$ are the price and quantity levels at which orders can be submitted to the LOB. These typically are discrete subsets of $\mathbb{R}$. In principle, price and quantity levels can become arbitrarily large, but in practice one can introduce a cutoff for both prices and quantities at a large enough value. As a result, $\mathcal{K}$ and $\mathcal{Q}$, can be thought of as finite sets.

It is convenient to introduce a partial ordering $\leq$ on creation operators via the following set of relations:

$$b^+_{k_1,q_1} \leq b^+_{k_2,q_2} \leq a^+_{k_3,q_3} \leq a^+_{k_4,q_4} \quad \Longleftrightarrow \quad k_1 \leq k_2 \leq k_3 \leq k_4.$$

Then any price-time ordered string of creation operators is equivalent to a monotonically increasing map

$$\begin{aligned} z : \; & S \to \mathcal{M} \times \mathcal{K} \times \mathcal{Q} \\ & i \mapsto z_i = \left(m_i^+\right)_{k_i,q_i} \qquad \text{s.t. } \; z_i \leq z_{i+1} \end{aligned}$$

from some finite set $S \subset \mathbb{N}$ to the set of creation operators. We denote the associated states by

$$|z\rangle := z_1 \cdots z_j |0| z_{j+1} \cdots z_n.$$

The collection $\{|z\rangle\}$ fully describes the possible configurations of the order book at any given moment. We may refer to these states as *pure states*. Note that S, $\mathcal{M}$, $\mathcal{K}$, and $\mathcal{Q}$ are countable sets, so the set $\{|z\rangle\}$ is countable as well.

Clearly, any prediction of the future state of the order book must be probabilistic. So, while pure states are potentially observable, *mixed states* are not. Mixed states are composed of several pure states in which each pure state is weighted with some probability mass.

For this reason, mixed and pure states are elements of the vector space spanned by the pure states $|z\rangle$:

$$\mathcal{H} = \left\{ |\psi\rangle = \sum_{|z\rangle \in \mathcal{H}} p(z)\,|z\rangle \quad \middle| \quad p(z) \in \mathbb{R} \right\}.$$

In reality, the order book must be in a *stochastically normalized* state, i.e., a state $|\psi\rangle = \sum_{|z\rangle} p(z)\,|z\rangle$, where $0 \leq p(z) \leq 1$ for all $|z\rangle$ and $\sum_{|z\rangle \in \mathcal{H}} p(z) = 1$. In this case, (the coefficient) $p(z)$ is the probability that the state $|z\rangle$ will be realized.

Remark. The normalized states are a closed subset of the full vector space $\mathcal{H}$. In particular, they do not form a vector space themselves. It is often easier to work with unnormalized vectors and rescale the result to a normalized state at the end.

In Section 4.1.3, we describe the time evolution of an initial state $|z_0\rangle$ at time t_0. We will see that (by construction) time evolution produces a state

$$|\psi(t); z_0, t_0\rangle = \sum_{z \in \mathcal{H}} p(z, t|z_0, t_0)\,|z\rangle\,. \tag{4.13}$$

In the literature, this object is usually referred to as *generalized probability generating function* (see for example Weber and Frey 2017, Section 2). For brevity, we often write $|\psi(t)\rangle$ and drop the reference to the conditional nature of the generalized probability generating function.

It is customary to denote linear functionals by 'bra' vectors $\langle\psi| \in \mathcal{H}^*$ and introduce the dual basis $\langle z|$, which satisfies

$$\langle z'|z\rangle = \delta_{z',z}.$$

The conditional probability to find a state $|z\rangle$ at time t in $|\psi(t); z_0, t_0\rangle$ is then given by

$$p(z, t|z_0, t_0) = \langle z|\psi(t)\rangle\,. \tag{4.14}$$

4.1.3 Time Evolution

In this section, we introduce dynamics to the order book, i.e., we explain how an order book state evolves over time. Throughout this section, we closely follow Baez and Biamonte (2018), where the general theory of stochastic time evolution is laid out in great detail.

The future state of the order book arises from acting on an initial state with the order operators introduced in Section 4.1.1. This means that we are automatically in the situation of a Markov process.

The only issue is that the rate (probability) of incoming orders can depend on the history of the order book. It is, however, not sensible to assume that the entire history of the order book affects the properties of arrival and cancellation rates as old configurations of the LOB are usually not relevant for the decision process of market participants. They usually seek to maximize the probability of order execution based on the current state of the order book and possibly a very narrow history of preceding order book configurations. This only implies that arrival rates may be dependent on several preceding states of the LOB, which is not in contradiction to the Markov property per se. It would only mean that the process governing the LOB dynamics might be of Markov order higher than 1. Theoretically, by appropriately extending the state space, every Markov process of finite order can be expressed as a Markov process of order one. Thus, we assume that the dynamics of the order book follow a Markov process of order 1.

As a continuous Markov process, the order book satisfies the Master Equation (cp. van Kampen 1992, Weber and Frey 2017). In our notation, the Master Equation is given by

$$\frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle , \tag{4.15}$$

where the so-called Hamiltonian operator H encodes all information on the transition probabilities between order book states.

A solution of the Master Equation is provided by a *stochastic time evolution operator* $U(t, t_0)$ via

$$|\psi(t)\rangle = U(t, t_0) |\psi(t_0)\rangle .$$

If the Hamiltonian is time-independent, the time evolution operator is remarkably easy:

$$U(t, t_0) = e^{H(t-t_0)}. \tag{4.16}$$

If the Hamiltonian is time dependent, the time evolution operator can similarly be written as

$$U(t, t_0) = \exp\left(\int_{t_0}^{t} H(\tau) d\tau \right), \tag{4.17}$$

but the evaluation of this expression is typically more involved.

We assume that other variables, like news from outside the order book, may impact the rates of incoming orders. However, these variables are pre-determined outside the mechanism of the order book. This may lead to time dependent arrival and cancellation rates. In the system description, this would mean that the Hamiltonian is time dependent. Nevertheless, a time-independent approximation of such a system may still serve as a

good approximation if the time intervals are small enough. In Section 4.4, we will take an indeterministic approach towards those other variables and regard them as predetermined outside the book. Again, this is not in contradiction to the Markov property of the LOB system which is the key assumption for the Master Equation (4.15) to hold. At this point, it is also interesting to note that beyond the model presented in this chapter, the LOB may be an open Markov process, which can be described by relying on a compositional model framework – in the sense of Baez and Pollard (2017) – and would allow to incorporate trader behavior.

Given the above considerations, the dynamics of the order book are fully described by the choice of a Hamiltonian H. Baez and Biamonte (2018) show how H can be constructed from *infinitesimal stochastic operators* which describe the elementary transitions that can take place in a system.

In the LOB, there are four possible transitions for each price level k and each quantity q:

$$\begin{aligned}
\text{entry of an ask order} \quad & E^A_{k,q} = \left(a^+_{k,q} - 1\right) \\
\text{entry of a bid order} \quad & E^B_{k,q} = \left(b^+_{k,q} - 1\right) \\
\text{cancellation of an ask order} \quad & C^A_{k,q} = \left(a^-_{k,q} - N^A_{k,q}\right) \\
\text{cancellation of a bid order} \quad & C^B_{k,q} = \left(b^-_{k,q} - N^B_{k,q}\right)
\end{aligned}$$

where $a^\pm_{k,q}$, $b^\pm_{k,q}$ are the creation and annihilation operators of Section 4.1.1. The number operators $N^A_{k,q} = a^+_{k,q}a^-_{k,q}$ and $N^B_{k,q} = b^+_{k,q}b^-_{k,q}$ return the number of active bid and ask orders on price level k and of quantity q when they act on a pure LOB state (see Section 4.1.4).

Remark. Creation and annihilation operators are not infinitesimal stochastic operators. This is why there are additional terms $(-1, -N)$ in the operators corresponding to order entry and cancellation.

As mentioned above, the Hamiltonian of the LOB is a combination of elementary transitions

$$H = \sum_k \sum_q E^A_{k,q}\alpha_A(k,q) + E^B_{k,q}\alpha_B(k,q) + C^A_{k,q}\omega_A(k,q) + C^B_{k,q}\omega_B(k,q), \tag{4.18}$$

where each transition is weighted by its arrival rate α or cancellation rate ω, respectively. Also note that the arrival rates need to be scaled such that the time evolution operator $U(t, t_0)$ is indeed stochastic and maps one stochastically normalized state to another.

Generally, the arrival and cancellation rates in a LOB are observed to be time dependent. Intraday patterns of order flow have been documented for example by Biais, Hillion and Spatt (1995). Even for the very recent development of international Bitcoin markets, in which trading is possible 24/7 Eross, McGroarty, Urquhart and Wolfe (2019) document activity patterns related to the opening and closing of major markets. The observed

clustering of transactions in time can be conceived as the result of time dependent arrival and cancellation rates. These are usually modeled using Autoregressive Conditional Duration (ACD) models (see Engle and Russell 1998, Fernandes and Grammig 2006). We may treat the arrival and cancellation rates, especially on small and intermediate time scales, as mainly determined by the state of the order book, in the sense that the distributions of the rates across k and q depend on the current state of the order book – for example via current best bid/ask or the spread. With this dependence on the current state, we allow for a quite general feedback mechanism between the current state of the order book and arrival and cancellation rates. If the state of the LOB changes by an event, the arrival and cancellation rates may change subsequently as well.

We investigate both static and dynamic specifications of arrival rates. In Section 4.3, we will investigate static distributions using empirical unconditional frequencies and a uniform as well as a theoretical discrete Gaussian exponential (DGX) distribution for arrival and cancellation rates across price levels. The latter can be justified heuristically by the characteristics found in our data as described in Section 4.2, in particular Figure 4.2. We will, in one simulation scenario, also allow the parameters of the assumed DGX distribution, for arrival and cancellation rates across relative price levels, to depend on the spread. In Section 4.4, we measure the arrival rates during fixed non-overlapping time intervals and therewith allow them to vary over time.

We also incorporate in our empirical analysis in Section 4.4 the idea of conditional autoregressive arrival and cancellation rates and include lagged terms of arrival rates, moments of the spread and the distance to the opposite best quote. Sampling the LOB data on different time intervals, i.e., taking snapshots of the current state at different frequencies (for example 1, 2, and 5 minutes), allows to relate the moments of the relative integer distance d_l (as defined in Equation (4.28) in Section 4.2) and the quantity of incoming and canceled orders q to price changes and other observables of the system. Empirical tests of these implications can be found in Section 4.4.

For now, we focus on the conceptual implications of these empirical findings and on how they affect the set up of the time evolution of the LOB system. Thus, we denote the arrival rate of an order at price k and quantity q as $\alpha_M(k,q)$, $M \in \mathcal{M}$. Since the distance to the opposite market side d and the prevalent spread Δ depend on the current state of the order book, the arrival rates must be considered to be operators. When $\alpha_M(k,q)$ acts on a pure state $|z\rangle$, it returns an arrival rate which depends on the values of d and Δ that are realized in the state $|z\rangle$:

$$\alpha_M(k,q;z) = \langle z|\alpha_M(k,q)|z\rangle \qquad M \in \{A,B\}.$$

A similar operator yields the distribution of cancellation rates corresponding to the current state of the order book:

$$\omega_M(k,q;z) = \langle z|\omega_M(k,q)|z\rangle \qquad M \in \{A,B\}.$$

In the Hamiltonian given in Equation (4.18) the operators $\alpha_M(k,q)$ and $\omega_M(k,q)$ act on the state first, thus determining the rate of the corresponding transition $E^M_{k,q}$ that acts on the state subsequently.

Remark. Since the Hamiltonian H is linear in the transition operators, it can be decomposed into smaller pieces that describe a subsystem of the LOB. For example, we can split up the Hamiltonian into ask and bid Hamiltonians

$$\begin{aligned} H &= H^A + H^B \\ H^M &= \sum_{k,q} E^M_{k,q}\alpha_M(k,q) + C^M_{k,q}\omega_M(k,q) \quad , \ M \in \mathcal{M}. \end{aligned}$$

Similarly, we could decompose H into the Hamiltonians for all price and quantity levels:

$$\begin{aligned} H &= \sum_{k,q} H_{k,q} \\ H_{k,q} &= E^A_{k,q}\alpha_A(k,q) + E^B_{k,q}\alpha_B(k,q) + C^A_{k,q}\omega_A(k,q) + C^B_{k,q}\omega_B(k,q). \end{aligned}$$

While these decompositions are convenient in calculations, they also allow a different view on the evolution of the book: In principle, one could argue that the time evolution should be based on (groups of) traders, whose order submissions and cancellations can be described by Hamiltonians H_g where the index g may indicate a group of traders or individual traders. The notion of particular groups can be found quite often in the literature. For example, Foucault et al. (2011) group traders into institutional and individual traders, whereas Foucault, Kadan and Kandel (2005) distinguish patient and impatient traders. These subsystems sum up to an *effective Hamiltonian* $H_{\text{eff}} = \sum H_g$ which will necessarily be of the form (4.18). The only difference is that now the rates α and ω become population parameters in a fundamental model about traders. In this chapter, we refrain from modeling traders and instead estimate effective arrival rate distributions from LOB data.

However, there is surely a trader induced clustering or autocorrelation in arrival rates which we cannot ignore. There are also patterns induced by general business activity throughout the day. Additionally, when submitting orders to the LOB, traders often care about the probability that their submitted orders are executed in due time. There is a trade-off between immediacy and a slightly delayed order execution. The probability that an order is executed is directly linked to the arrival rates of orders in the LOB. Thus, traders may incorporate the history in their decision process, i.e., when, at which

limit price, and with which quantity they want to submit their orders to the LOB and again induce autocorrelation into arrival and cancellation rates. The decomposition of the Hamiltonian, as discussed above, would allow for an explicit model and cover such a scenario. In general, the model above does not exclude the notion of autocorrelation in the arrival rates. Especially in Section 4.4, however, we take a more indeterministic view in that we allow prior arrival and cancellation rates of ask or bid orders to proxy current arrival and cancellation rates. The idea that prior arrival rates determine current rates is also the guiding notion for the ACD literature mentioned above.

4.1.4 Observables

A specific configuration $|\psi\rangle$ of the order book contains an enormous amount of information. Usually, the focus lies on selected descriptive quantities which can be extracted from the order book at any state. We will call these quantities *observables* and describe them by the action of an operator O on pure order book states $|z\rangle$. The value of O for a given state $|z\rangle$ can be calculated as

$$O(z) = \langle z|O|z\rangle\,.$$

More generally, given a state $|\psi\rangle$, the νth conditional moment of the observable O is given by[24]

$$\mathbb{E}[O^\nu;\psi] = \sum_{|z\rangle\in\mathcal{H}} \langle z|\, O^\nu\, |\psi\rangle\,. \tag{4.19}$$

Similarly, we can calculate the expected value of sums and products of distinct operators. This gives rise to covariance and correlation measures, e.g.,

$$\mathrm{Cov}(O_1,O_2) = \sum_{|z\rangle\in\mathcal{H}} \langle z|\,(O_1-\mathbb{E}[O_1])(O_2-\mathbb{E}(O_2))\,|\psi\rangle\,.$$

Combined with the time evolution of an initial state $|\psi_0\rangle$, we obtain the moments of an observable's probability distribution at time t $(t>t_0)$ as

$$\mathbb{E}[O^\nu;\psi(t)] = \sum_{|z\rangle} \langle z|\, O^\nu e^{H(t-t_0)}\, |\psi_0\rangle = \sum_{|z\rangle} \langle z|\, O^\nu\Big(1+H(t-t_0)+\frac{1}{2}H^2(t-t_0)^2+\dots\Big)\,|\psi_0\rangle\,. \tag{4.20}$$

[24] In quantum mechanics, a similar relation holds, known as the Born rule $\langle\Psi|\,\hat{O}^\nu\,|\Psi\rangle$. Since we work with stochastic probabilities (and not with quantum mechanical amplitudes), $\langle\Psi|$ needs to be replaced by the sum over all dual basis vectors $\langle z|$.

Note that the expected value in Equation (4.20) is a conditional expectation. It is conditional on the state $|\psi_0\rangle$ at time t_0. Later on, in Section 4.4, to make this conditioning clear, we will denote conditional moments with $\mathbb{E}_{t_0}[O^\nu]$. The following example illustrates the rationale behind the formula in Equation (4.20). Consider the operator β_B which extracts the value of the best bid order in a state $|z\rangle$ (cf. Section 4.1.4). Furthermore, assume that at $t_0 = 0$, the initial state is given in price-time ordered form by

$$|\psi_0\rangle = b^+_{k_1,q_2} b^+_{k_2,q_3} |0| a^+_{k_3,q_1}.$$

Clearly, since $k_1 < k_2 < k_3$, the best bid is currently at price level k_2. According to (4.20), the expected value of β_B at time t is given by an infinite sum. To begin with, we consider terms for which the state does not change during the time period $\Delta t = t - t_0$. These include of course the identity operator 1 in Equation (4.20). But H is built from terms of the form $\alpha(p,q)(b^+_{p,q} - 1)$, so there are additional contributions at any order in H^k. Together, they contribute the following term to the expected value of the best bid price only for price level k_2:

$$k_2 \left(1 - \sum_{k,q} \alpha(k,q)\Delta t - \sum_{k,q} \omega(k,q)\Delta t - \dots \right).$$

The expression in parenthesis is of course nothing but the probability that the state will not change within Δt.

For other price levels, these contributions are different. Therefore, we next investigate the linear terms in H which describe the entry of a single order. There are only three cases in which the best bid changes. First, we may observe an entry of a limit bid order with a price level in between best ask and best bid. Its contribution to the expected value is $\sum_{k_2<k<k_3,q} k\alpha_B(k,q)\Delta t$. Second, the order residing on k_2, the current best bid price level, may be canceled. In this case the contribution to the expected value is $k_1\omega(k_2,q_2)\Delta t$. Third, an arrival of an ask order above the best bid which exhausts the best bid order's quantity may arrive. In this case we get $\sum_{k\geq k_2, q\geq q_2} k_3\alpha_A(k,q)\Delta t$. In all other cases of incoming limit orders the value of β_B remains at k_2.

A similar analysis is possible at order H^2. This would entail interaction terms of two orders entering the book: $a^+_{k,q}\alpha_A(k,q)b^+_{\ell,r}\alpha_B(\ell,r)$. There are again several different cases which depend on the type, price level and quantity of the incoming orders, each contributing differently to the expected value. More generally, at order H^n one encounters the probabilities that n orders enter the book and influence the best bid during time period Δt. For short time periods, such higher order contributions are negligible compared to the linear contributions because they depend on products of arrival rates, which are typically very small. However, for long time periods the powers Δt^n will eventually dominate.

The example above illustrates an important property of the model. According to Equation (4.20), the moments of observables depend solely on the (current) distributions $\alpha_M(k,q)$ and $\omega_M(k,q)$. These distributions, if they vary sufficiently slowly, may be measured or modeled from the event stream of the book. Therefore, a testable hypothesis implied by our model is whether the distributional moments of k (or equivalently d_l for that matter) and q (calculated by perceiving $\alpha_M(k,q)$ and $\omega_M(k,q)$ as their underlying probability functions) can be used to predict the expected value of observables, including price changes and inter-transaction duration.

In the following, we present a selection of observables, which are important for our analysis.

Number and Volume Operators

A basic observable is the number of active orders on price level k with size q. It can be described for the bid and ask side by the *number operators*

$$N^B_{k,q} = b^+_{k,q} b^-_{k,q},$$
$$N^A_{k,q} = a^+_{k,q} a^-_{k,q}.$$

These operators can be utilized to extract several other observables. In particular, the total number of active orders on price level k of ask or bid type $M \in \mathcal{M} = \{A, B\}$

$$N^M_k = \sum_q N^M_{k,q},$$

the *quantity* of active orders on price level k and the *total quantity* on each market side $M \in \mathcal{M}$

$$Q^M_k = \sum_q q N^M_{k,q},$$
$$Q^M = \sum_k Q^M_k,$$

or the *volume* of active orders at price level k and the *total volume* on each market side

$$V^M_k = k Q^M_k,$$
$$V^M = \sum_k V^M_k.$$

There are also operators that describe a global aspect of the configuration of an order book state $|z\rangle$, e.g. the best bid and best ask prices β_M, $M \in \mathcal{M}$. In the following, let the

state of the LOB be

$$|z\rangle = \; b^+_{k_1,q_1} \; \dots \; b^+_{k_j,q_j} \; |0| \; a^+_{k_{j+1},q_{j+1}} \; \dots \; a^+_{k_n,q_n}.$$

Then the *best bid* and *best ask* operators act on $|z\rangle$ as follows:

$$\begin{aligned}\beta_B\,|z\rangle &= k_j\,|z\rangle\,,\\ \beta_A\,|z\rangle &= k_{j+1}\,|z\rangle\,.\end{aligned}$$

Note that k on the right hand side is not an operator but the price level associated with the best quote. Combining the two, one obtains the *spread* operator Δ and *mid price operator* β_{mid} as

$$\begin{aligned}\Delta &= \beta_A - \beta_B,\\ \beta_{\text{mid}} &= \tfrac{1}{2}\big(\beta_B + \beta_A\big).\end{aligned}$$

Order Book Imbalance

These operators also allow to extract more complicated measures from the book like the order book imbalance, for example. It is a relevant quantity for order execution and of special interest to practitioners that design and develop trading algorithms (see, e.g., Bechler and Ludkovski 2015, Lipton, Pesavento and Sotiropoulos 2014, Cartea, Jaimungal and Penalva 2015). In general, the literature relies on two measures to quantify order book imbalance. First, Lipton et al. (2014) use the total number of ask and bid orders in the market and calculate order book imbalance IQ as the relative deviation of standing ask and bid orders as

$$IQ = \frac{Q^A - Q^B}{Q} = \frac{Q^A - Q^B}{Q^A + Q^B} = \frac{Q^A}{Q^A + Q^B} - \frac{Q^B}{Q^A + Q^B}. \tag{4.21}$$

Second, Bechler and Ludkovski (2015) use the volume of active orders. Their measure of market imbalance IV is given by

$$IV = \frac{V^A - V^B}{V} = \frac{V^A - V^B}{V^A + V^B} = \frac{V^A}{V^A + V^B} - \frac{V^B}{V^A + V^B}. \tag{4.22}$$

Liquidity

Harris (2003) defines liquidity as 'the ability to trade large size quickly, at low cost, when you want to trade.' According to the same source, the notion of liquidity incorporates four dimensions: immediacy of trade execution for a given size, depth, width, and resilience of

the market. Therefore, the spread itself is used frequently as a liquidity measure in the literature.

There are multiple approaches to measure liquidity and we rely on the exchange liquidity measure (XLM) which is based on the concept of implementation shortfall, introduced by Gomber and Schweickert (2002). It covers three dimensions of liquidity: depth, width, and immediacy. The XLM (also known as XETRA Liquidity Measure) is composed of liquidity measures for the ask side (XLM_A) and the bid side of the market (XLM_B),

$$XLM = XLM_A + XLM_B, \tag{4.23}$$

where

$$XLM_A = 10,000 \frac{\frac{\sum_k^\infty V_k^A}{\sum_k Q_k^A} - \beta_{\text{mid}}}{\frac{\sum_k^\infty V_k^A}{\sum_k Q_k^A}}, \tag{4.24}$$

$$XLM_B = 10,000 \frac{\beta_{\text{mid}} - \frac{\sum_k^\infty V_k^B}{\sum_k Q_k^B}}{\frac{\sum_k^\infty V_k^B}{\sum_k Q_k^B}}. \tag{4.25}$$

The XLM depends on the volume weighted price which can be realized immediately on each side of the market for a round trip order with a certain volume $\bar{V}$, i.e., simultaneously submitting marketable ask and bid orders with a total volume of $\bar{V}$. In other words, the XLM measures the cost of a round trip order (in basis points).

4.1.5 Transactions

Up to now, we have deferred the discussion of transactions since, strictly speaking, they are not necessary to set up the order book states. In this section we first discuss the trading modes of the XETRA order book and explain how one can augment the LOB states to also record information about transactions. This will allow us to study the transaction price and transaction rates, which were so far not available in the order book state.

The XETRA order book is organized as continuous trading augmented by opening-, intraday-, and closing-auctions. Before stating the rules for these modes, we make a small change in notation: Instead of the symbol $|0|$ for the empty book, we record via $|T_{k,q;t}|$ the last price k, quantity q, and time t at which a transaction occurred.

Rule 5a (Continuous Trading)**.** *Assume an incoming order is assigned highest priority and is such that it permits a transaction with its partner on the opposite market side. Then the orders will be executed at the price of the partner that was already residing in the market*

and a transaction of the matched-up quantity will be issued at this price. For an arriving ask order, this results in

$$\left(\cdots b^+_{k,q} \, |0| \, \cdots \right) a^+_{s,p} = \cdots \overbrace{b^+_{k,q} \, | \, T_{k,\min(q,p);t} \, |} \, a^+_{s,p} \cdots \, , \tag{4.26}$$

while for an arriving bid order, we have

$$b^+_{k,q} \left(\cdots |0| \, a^+_{s,p} \, \cdots \right) = \cdots \overbrace{b^+_{k,q} \, | \, T_{s,\min(q,p);t} \, |} \, a^+_{s,p} \cdots \, . \tag{4.27}$$

Rule 5b (Auction)**.** *Auctions consist of an outcry/call phase, during which incoming orders are collected and ordered by price-time priority as usual, but are* not executed. *The exchange may provide an indicative pricing to market participants i.e., the price level at which the current order book state would settle if the call phase were to end immediately. Upon closing of the call phase the transaction price is determined according to the principle of highest traded volume. Subsequently, orders of highest priority are executed iteratively at the previously determined transaction price. The transaction is recorded at the transaction price and with the total traded quantity. A description of the matching procedure like in Equation* (4.26) *and Equation* (4.27) *is possible for concrete situations. The principle of highest traded volume makes a general formulation exhausting and is not particularly illuminating. Therefore, we omit a general formulation at this point.*

The rules above are illustrated in Figure 4.1. We can now introduce the *transaction price*, *transaction quantity*, and *transaction volume operators*, which extract the corresponding numbers from the last recorded transaction. Let the state of the book be

$$|z\rangle = z_1 \ldots z_i \, | \, T_{k,q;t} \, | \, z_{i+1} \ldots z_n \, .$$

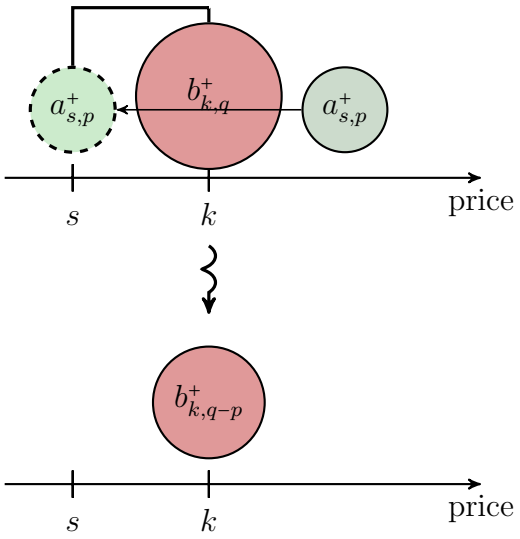
The operators are then defined as follows

$$\begin{aligned} T_K \, |z\rangle &= k \, |z\rangle \, , \\ T_Q \, |z\rangle &= q \, |z\rangle \, , \\ T_V \, |z\rangle &= kq \, |z\rangle \, . \end{aligned}$$

Furthermore, we can extract the time at which the last transaction occurred via $T_t \, |z\rangle = t \, |z\rangle$. This is the basis for an important observable, the *inter-trade duration* $T_{\Delta t} = t_2 - t_1$, i.e., the time between two transaction. In our current setup, $T_{\Delta t}$ cannot be expressed as an operator which only acts on the current state, while in practice we can calculate the time intervals from remembering earlier transactions' time stamps.

Figure 4.1: Transaction Matching

The figure illustrates the matching $\overbrace{b_{k,q}|0|a_{s,p}}$ for $s \leq k$ and $q > p$.



4.2 Data

The data used in the present chapter were provided by Deutsche Börse in 2004 and have been previously used by Grammig, Heinen and Rengifo (2004). They consist of all recorded order book events of the XETRA system[25] for trading of the 30 stocks that constituted the German stock market index, DAX between January 2 and March 26, 2004. Additionally, Deutsche Börse provided the open order positions in their books as of January 1, 2004, 12 pm. The data allow for the full recovery of the order book. Over the three months period, 228,275,832 events were recorded. Additionally, 2,282 initial positions are available at the beginning of the period. The data cover order arrivals, (partial) matches, changes and cancellations. The XETRA trading system allows for limit and market orders. It is also possible to mix the two standard order types (market and limit orders) with a market-to-limit order (MTL). An MTL order is filled on the best limit price in the book, either fully or partially. If an MTL order is matched only partially on the best ask or bid price, it enters the LOB on the best limit price with the remaining order size. Additionally, iceberg orders (ICE) are allowed for which only a fraction of the total volume chosen by the issuer is displayed to market participants.

Market and limit orders can also have a specified stop price. They are then called stop orders. Different from the limit price, i.e., the price upon which the trader is willing to trade, the stop price specifies a price level from which onwards the trader is willing to submit an order. Hence, if the reference price exceeds (in case of bid order) or undercuts (ask order) the stop price, the order is inserted in the book.

[25] XETRA Release 7.0

Furthermore, the XETRA system also allowed so called XETRA BEST execution orders during the sample period at hand. The BEST execution orders are matched against incoming market or crossing limit orders at a price level just before the currently prevailing best ask and bid prices. In that way they introduce an extra hidden layer of occupied price levels in front of the prevailing best ask and bid prices in the LOB. XETRA BEST orders can be market or limit orders. Market-to-limit orders, iceberg orders and stop-orders cannot be submitted as XETRA BEST orders. Also, if XETRA BEST orders are not executed they do not enter the LOB event log at all. This happens if the associated price level is better than the current best quote. If they are matched in a trade, they are recorded as the counterparty of a transaction. If they are submitted as limit orders and the associated price level is worse than the best quote, they enter as regular limit orders. In the latter case, we cannot distinguish regular limit orders from XETRA BEST orders in our data.

For all orders validity constraints can be set. Users may specify a termination date up to which the order is valid. Without such a restriction, orders are valid for 90 days. Iceberg orders, however, are only good for the day. Traders may also specify whether orders are valid only for specific trading phases such as auctions, or during which auction the order shall be valid. Orders with such a restricted validity reside in the book and become active during the trading phase for which they are valid.

With regard to execution restrictions, the XETRA system allows for the following two specifications for market, limit or MTL orders. First, the Fill-or-Kill order (FOK) is either filled entirely or canceled. FOK orders are only recorded as entries to the book if successfully filled. If no immediate filling is possible, FOK orders are canceled without notification within the LOB event log. Second, the Immediate-or-Cancel order (IOC) is filled as far as possible upon entry, or canceled. Similar to the FOK order, a record of the order is only entered in the LOB event log in case of a successful (partial) filling.

All these different order types and restrictions can be incorporated in the time evolution set out in Section 4.1 by introducing different types of arrival and cancellation rates for the related events as elements of the Hamiltonian. These order types, however, do not affect the generality of the algebra set up in Section 4.1.1.

Table 4.A.1 in Section 4.A provides an overview of the distribution of the events in the LOB log. It presents the total number of submitted limit, market, iceberg, and MTL orders along with their relative occurrences on the bid and ask side.

In empirical investigations of LOB data, it is frequently observed that the distributions of arrival and cancellation rates show a relatively stable connection with the current distance $d = |\beta_M - k|$ of the respective price level k to the best active price level on the opposite market side β_M, $M \in \mathcal{M} = \{A, B\}$ (cp. Bouchaud et al. 2002). We find a similar

phenomenon in the XETRA data. We set the distance d of all crossing, arriving orders to 0 and define the logarithmic relative integer distance as

$$d_l = \log\left(100 \max(d, 0) + 1\right). \tag{4.28}$$

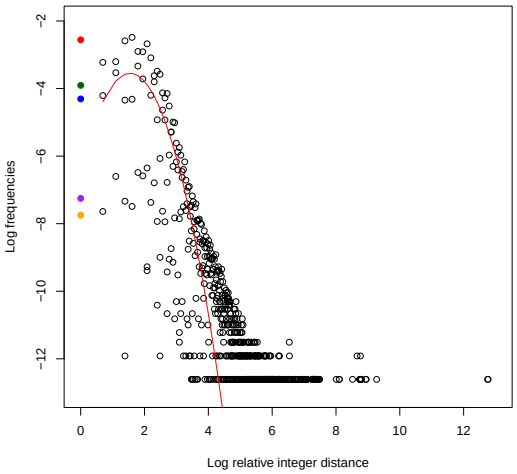
Taking logarithms exposes the heavy tails of the distribution across price levels and its similarities with the DGX-distribution (see Section 4.C for a short description and Bi, Faloutsos and Korn (2001) for further details). For our simulation study in Section 4.3, we use the DGX distribution to model order arrivals across d_l. Figure 4.2 displays the empirical logarithmic frequencies of order arrivals and cancellations at the logarithmic relative integer distance d_l of their limit order price to the best ask or best bid price, respectively. The red line indicates a DGX distribution (truncated at 1). The logarithmic frequencies of different types of marketable orders (limit, stop, iceberg and market orders) are displayed separately at $d_{l0} = 0$ in Figure 4.2.

In our data, we also observe a small correlation between d_l and the prevailing spread $\Delta = v_A - v_B$ in logarithms. Figure 4.3 presents scatter plots of d_l against the logarithmic integer spread $\log(100\Delta)$. For large spreads there is a stronger correlation between events that are mainly concerned with price levels around the best limit price of the same market side. When the spread is small, it seems that events occur more evenly spread out up to as much away as EUR 4 from the best limit price of the opposite market side. We also note that events on the bid side are less dispersed across price levels than events on the ask side as the natural limit price for a bid order is a price level of 0. For ask orders, theoretically, no such limit exists.

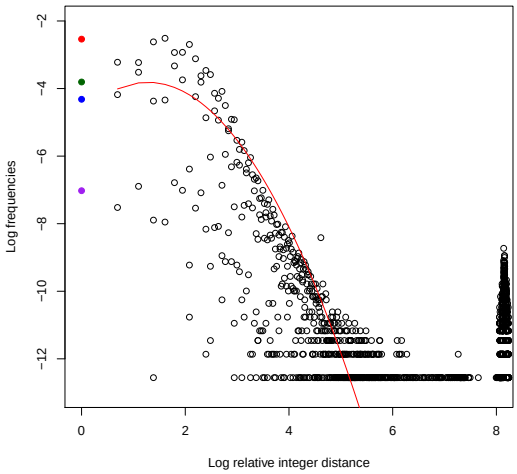
With respect to order size, we observe no clear pattern between order size and the logarithmic relative integer distance in the data, only a very small negative correlation can be noted. Budget restrictions of market participants would suggest that order sizes further away from the opposite best quote on the ask side bring order sizes down with growing price levels. For the bid side, the inverse argument should hold, nevertheless, a small negative correlation can also be observed for the bid side. However, as the correlations are low, the price level and quantity, at least in logs, may justify the approximating assumption of stochastic independence used in several scenarios of the simulation study in Section 4.3. Table 4.1 lists the correlations between the order volume q associated with a certain event and the logarithmic integer distance to the best opposite quote for the MEO stock. As can be seen, the approximating assumption that the two variables q and d_l are independent does not mirror reality exactly. Nevertheless, we will make the assumption several times in this chapter as it keeps the estimation and simulation manageable.

Figure 4.2: Frequency of Order Arrivals

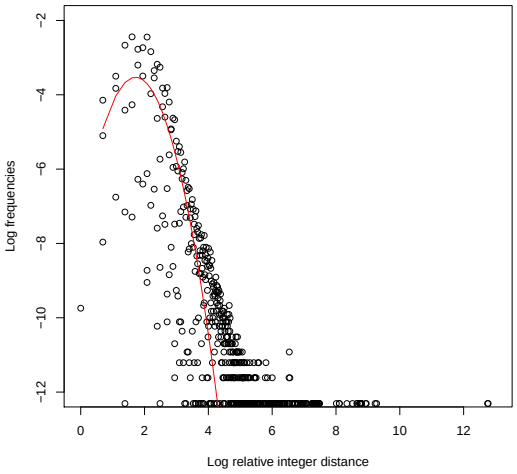
The graphs show the logarithmic frequency of arriving (a and b) or canceled (c and d) limit orders (including stop and iceberg orders) for the stock MEO, respectively, on their logarithmic relative integer distance to the best bid or ask prices. The logarithmic relative integer distance is defined as $d_l = \log(100 \max(d, 0) + 1)$. The red line is the logarithmic probability of a truncated discrete Gaussian exponential (DGX) distribution for $d_l > 0$, i.e., $d > 1$ (as described in Section 4.C). The theoretical value for $d_l = 0$ or $d = 1$ is intentionally ignored for the fitting of the parameters of the DGX distribution. At $d_l = 0$ the logarithmic frequencies of several types of marketable orders are displayed. The blue point represents the logarithmic frequency of market orders. Also, the log frequency of marketable limit orders at (red) and behind (green) the best quote is shown, as well as marketable iceberg orders in purple, and marketable stop orders in orange. Crossing cancellations occur in the XETRA event log when the orders are deleted before they are matched.



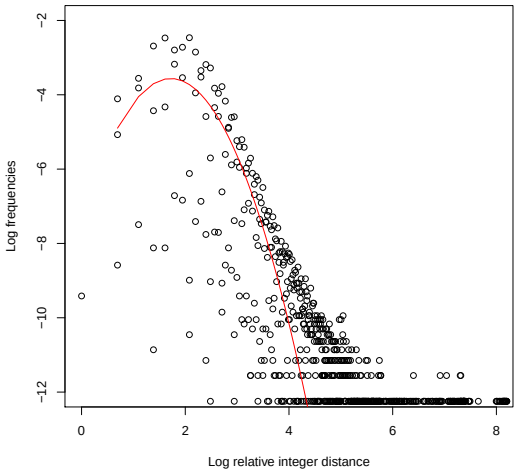
(a) Ask Arrivals



(b) Bid Arrivals



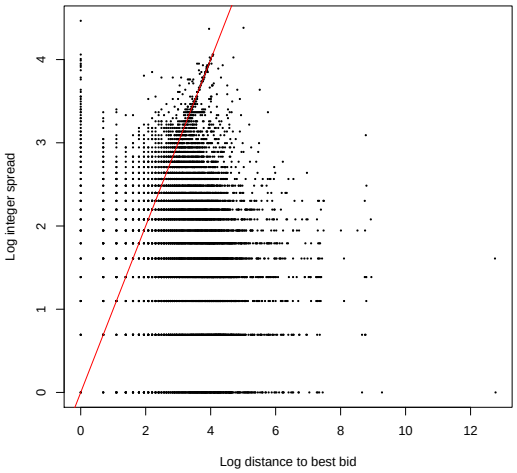
(c) Ask Cancellations



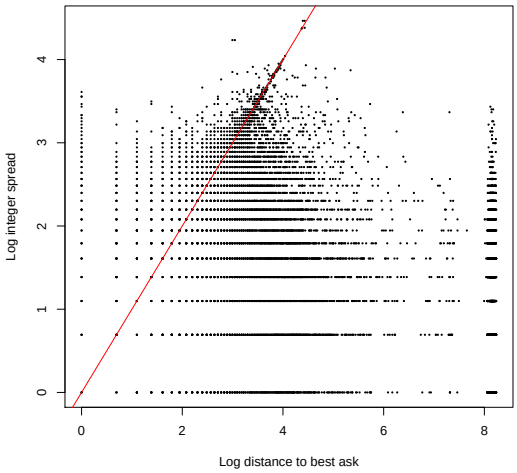
(d) Bid Cancellations

Figure 4.3: Relation Between Spread and Relative Price Distance

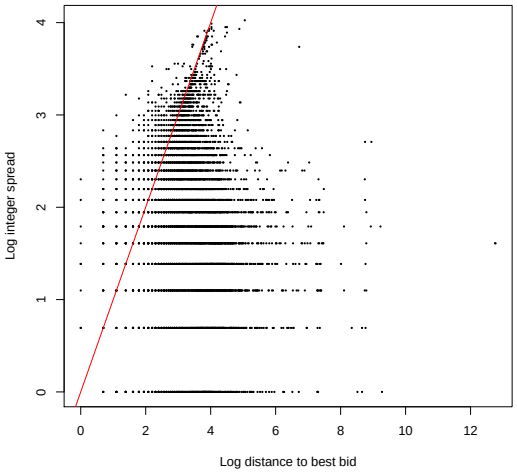
The graphs present the logarithmic relative integer distance to the best bid or ask price of orders arriving (a and b) or being canceled (c and d) related to the stock ALT together with the prevailing logarithmic integer spread. The logarithmic relative integer distance is defined as $d_l = \log(100 \max(d, 0) + 1)$ whereas the logarithmic integer is $\log(100\Delta)$ with Δ being the prevailing spread at arrival or cancellation. Even though cancellations smaller than the spread seem counter intuitive, they occur when orders are immediately canceled right after their insertion into the event log. The red line indicates the bisecting line.



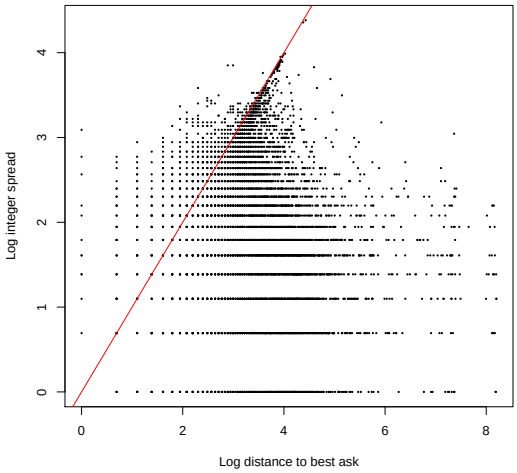
(a) Ask Arrivals



(b) Bid Arrivals



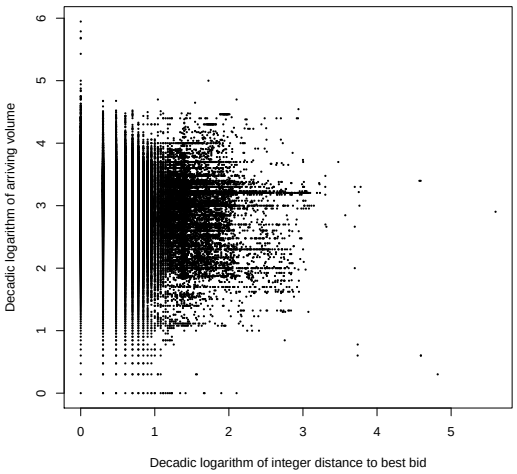
(c) Ask Cancellations



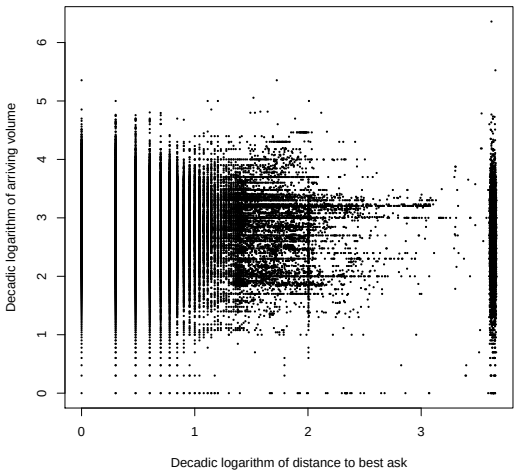
(d) Bid Cancellations

Figure 4.4: Relation Between Order Size and Relative Price Distance

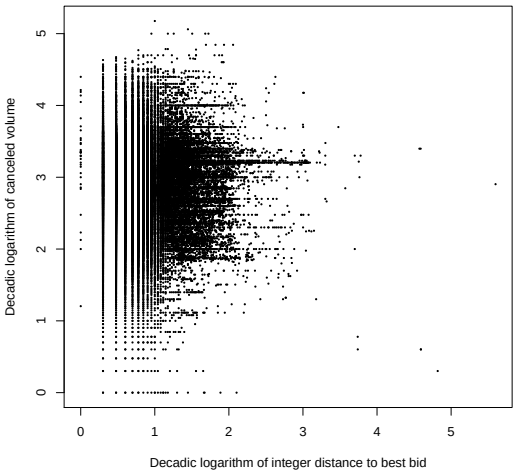
The graphs show the decadic logarithmic relative integer distance to the best bid or ask price of arriving (a and b) or canceled (c and d) orders related to the stock BAS against the decadic logarithm of the size. The decadic, logarithmic, relative integer distance is defined as $d_l = \log_{10}(100 \max(d, 0) + 1)$. For the arriving orders, the order size depicted is the original order size, while for the canceled orders, the order size depicted is the actually canceled order size, not the original size at entry.



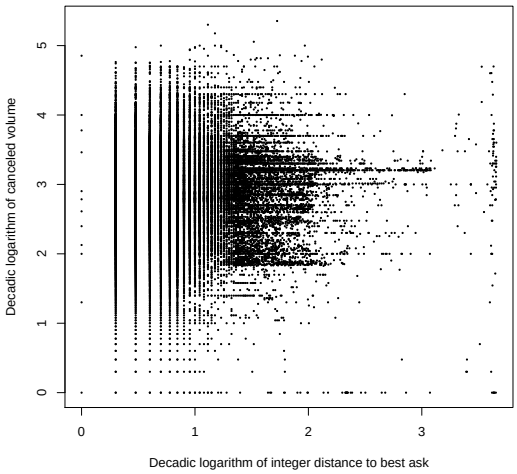
(a) Ask Arrivals



(b) Bid Arrivals



(c) Ask Cancellations



(d) Bid Cancellations

Table 4.1: Correlations Between d_l and q

The table lists the correlation coefficients r^2 between the logarithmic integer distance to the opposite best quote d_l as defined in Equation (4.28) together with standard errors calculated as s.e. = $\sqrt{(1-r^2)/(n-2)}$ where n is the number of observations. The approximately standard normally distributed z-score= $(r^2)/(\text{s.e.})$ is reported as well, complemented by its p-value. All values reported concern the MEO stock.

Event	r^2	s.e.	z-score	p-value
ask arrivals	-0.1517	0.0060	-25.2	<0.001
bid arrivals	-0.2090	0.0058	-35.9	<0.001
ask cancellations	0.0291	0.0062	4.7	1
bid cancellations	0.0359	0.0061	5.87	1

4.3 Simulation

The simulation algorithm presented in this section generates order book events by generating answers to the repeatedly asked question: When will what happen where next?

4.3.1 The Simulation Algorithm

In order to simulate the LOB, several assumptions have to be imposed on the functional structure of the arrival and cancellation rates included in the Hamiltonian H of the model developed in Section 4.1. The following section presents the stochastic simulation algorithm (SSA) developed by Gillespie (1977) which can be used to simulate an artificial history of the order book. The SSA is a direct consequence of the model and allows the exact simulation of the system.

Note that the assumptions made about the immanent functional structure of the arrival and cancellation rates are the crucial ingredients of the model. We therefore explore possible calibrations of our model with the subsequent simulation study. Our goal is not to fit the simulation results to an observed LOB history as closely as possible. Instead, the simulation offers insights into the sensitivity of the order book dynamics, especially the transaction price dynamics, to changes in the structure of arrival and cancellation rates.

The starting point of the SSA is based on the probability that within the next interval $\delta\tau$ no event occurs which we can denote in our notation as

$$P_0(\delta\tau) = \sum_{z\in\mathcal{H}} \langle z| \exp(\operatorname{diag}(H)\delta\tau) |\Psi(t_0)\rangle \,, \tag{4.29}$$

where the diagonal elements in H are obtained by

$$\text{diag}(H)\delta\tau = \langle z| H\delta\tau |z\rangle = -\sum_{k,q,M} \alpha_M(k,q;z)\delta\tau - \sum_{k,q,M} \omega_M(k,q;z)\delta\tau.$$

This is the negative sum of the rates of all possible events conditional on the book being in state $|z\rangle$.

Gillespie (1977) shows how to formulate this probability for some event μ to happen during the interval τ without the operator algebra. The probability that an order arrives during the interval $d\tau$ is $r_\mu d\tau$, where r_μ is the rate corresponding to the event. In our case, r_μ may be some rate from the set of arrival or cancellation rates, $\alpha_M(k,q)\delta\tau$ or $\omega_M(k,q)\delta\tau$. In fact, we may label all possible events with integer numbers and let μ be a specific integer denoting a specific event. Setting $\tau = \delta\tau + d\tau$, the probability that given the state $|\Psi(t_0)\rangle$ at time t_0 the next reaction μ will happen during the next interval of τ, denoted $P(\tau,\mu)$, can be written as the product of the probability that nothing will happen during $\delta\tau$ and the probability that μ will happen during $d\tau$:

$$P(\tau,\mu) = P_0(\delta\tau) r_\mu d\tau. \tag{4.30}$$

From Equation (4.30), Gillespie (1977) deduces that the probability that nothing happens during τ, can be formulated as

$$P_0(\tau) = P_0(\delta\tau)\left(1 - \sum_{\nu\neq\mu} r_\nu d\tau\right). \tag{4.31}$$

Noting that $\tau = \delta\tau + d\tau$ by definition, bringing all terms involving P_0 to the left hand side, dividing both sides by $d\tau$ and taking limits for $\delta\tau \to 0$, yields a differential equation that is solved by setting

$$P_0(\tau) = \exp\left(-\sum_{\nu} r_\nu \tau\right). \tag{4.32}$$

Substituting Equation (4.32) into Equation (4.31), the probability that μ will happen during the next time interval τ is given by

$$P(\tau,\mu) = r_\mu \exp(-r_0\tau) = r_\mu \sum_{z\in\mathcal{H}} \langle z| \exp(\text{diag}(H)\tau) |\Psi(t_0)\rangle, \tag{4.33}$$

where in our case $r_0 = \sum_{k,q,M} \alpha_M(k,q;z) + \sum_{k,q,M} \omega_M(k,q;z)$.

From Equation (4.33), we may randomly generate the pair (τ,μ), i.e., the time when an event occurs τ and which event will happen μ. As we have set up the rates as price and size specific, by generating the event μ we also specify the price location and the size

which are affected by the event. By noting that Equation (4.33) determines an exponential distribution with scale parameter r_0, we can first sample τ by drawing u_1 from a uniform distribution $\mathcal{U}(0,1)$ and calculating

$$\tau = \frac{1}{r_0} \log\left(\frac{1}{u_1}\right).$$

Having determined when an event occurs, we may now ask the question what will happen. By numerically specifying the rates for all possible events r_ν and drawing a second realization u_2 from a uniform distribution $\mathcal{U}(0,1)$, we may find the integer μ by solving

$$\sum_{\nu=1}^{\mu-1} \frac{r_\nu}{r_0} < u_2 \leq \sum_{\nu=1}^{\mu} \frac{r_\nu}{r_0}$$

for μ. In other words, by drawing u_1 and u_2, we can simulate an answer to the question *when* something will happen with u_1 and, with u_2, *what* as well as *where* it will take place. In fact, we also draw a third realization from a uniform distribution u_3, to answer the question what size is affected (see Section 4.B for details). Having drawn an event and it's characteristics, the current state of the system can be updated. This may change the rates r_ν and their sum r_0. Note that by sampling the events in this fashion, the events are conditionally independent. They may not be independent as the rates are conditional on the current state (and under the assumption of a higher Markov order also on finitely many previous states) of the LOB.

In order to simulate the LOB dynamics, we have to specify the rates of all possible events and how they depend on the current state. In our case, all possible events comprise the order arrivals and order cancellations. Thus, we have to find a functional form for the respective rates α_M and ω_M. In our specifications presented in Section 4.B, we let α_M and ω_M be functions of the quantity q and the price level k (or more precisely of the integer distance to the opposite best quote d_l). For simplicity, we will assume that all rates $\alpha_M(k,q)$ and $\omega_M(k.q)$ are separable in k and q such that

$$\alpha_M(k,q) = \alpha_{1,M}(k)\alpha_{2,M}(q) \quad \text{and} \quad \omega_M(k.q) = \omega_{1,M}(k)\omega_{2,M}(q).$$

As the rates are proportional to the probability distribution of arriving (or canceled) orders across price and size, this means that the size of arriving (or canceled orders) is stochastically independent of the price level they concern. In Figure 4.4, it can be seen that for lower distances to the opposite best quote the size of arriving and canceled order is equally spread out across possible size levels. A clear relationship between the price level and the size level is not visible. In the absence of such a clear relationship, we find the approximating assumption that the size and price level are stochastically independent justifiable.

We decompose the arrival rates further by setting the general intensity of events for each market side $\bar{r}_{0,M,i}$ to the average event rate over the entire sample of stock i, where $\bar{r}_{0,M,i}$ is defined as

$$\bar{r}_{0,M,i} = \sum_{k,q,j} \alpha_{M,i}(k,q) + \omega_{M,i}(k,q)$$

which is calculated as the number of events on one market side divided by the total number of events. Note that since we have several order types, the arrival rates may be split into market orders as well as marketable and non-marketable limit orders. The empirical frequencies for $\bar{r}_{0,M,i}$ are reported in the last column in Table 4.A.1.

Hence, the arrival and cancellation rates for limit orders can be described by the partitioning of the average event rate $\bar{r}_{0,M,i,j,a}$ across price levels k and order sizes q:

$$\begin{aligned} \alpha_M(k,q) &= \bar{r}_{0,M,i,j,a}\; p_{K,M}(k;\boldsymbol{\theta}_{M,a})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,a}), \\ \omega_M(k.q) &= \bar{r}_{0,M,i,L,c}\; p_{K,M}(k;\boldsymbol{\theta}_{M,c})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,c}), \end{aligned} \tag{4.34}$$

where $\bar{r}_{0,M,i,j,a}$ is the rate for an order of type j (market or limit order) for stock i to arrive and $\bar{r}_{0,M,i,L,c}$ is the rate for a limit order (i.e., $j = L$) to be canceled. $p_{K,M}(k;\boldsymbol{\theta}_{M,a})$ denotes the discrete probability mass function of order arrivals across the integer price levels k given some parameter set $\boldsymbol{\theta}_{M,a}$ and similarly $p_{Q,M}(q;\boldsymbol{\phi}_{M,a})$ is the discrete probability mass function of order arrivals or cancellations across order sizes. The index a indicates the parameters for order arrivals, c the parameters for cancellations. The index M denotes the market side.

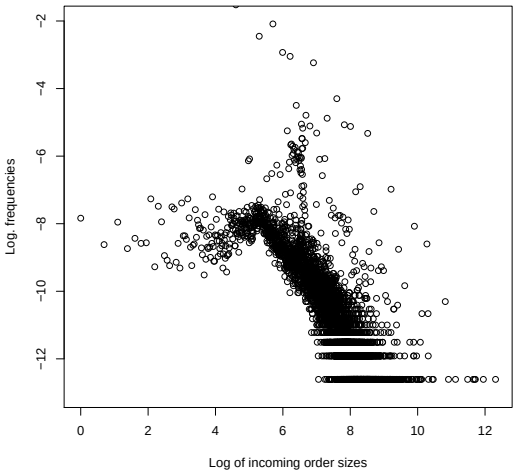
In our simulation, we consider three theoretical probability mass functions for $p_{K,M}(\cdot)$: The uniform distribution (uni), a discrete log-normal distribution with fixed parameters (fix), and a discrete log-normal distribution which depends on the prevailing spread (dyn). For $p_{Q,M}(\cdot)$ we only consider a power law distribution (pow). The power-law distribution captures the heavy tails of the volume distribution. The distribution of order size and the heavy tails can be seen in Figure 4.5 which depicts the frequencies of order arrivals and cancellations.

Section 4.B lists all the functional specifications as well as a description on how market orders are incorporated in the distributional setup. Iceberg orders, stop orders or fill-or-kill restrictions are neglected in the simulation study, as the events marked by these order types only make up for less than 1% of all events in our data set.

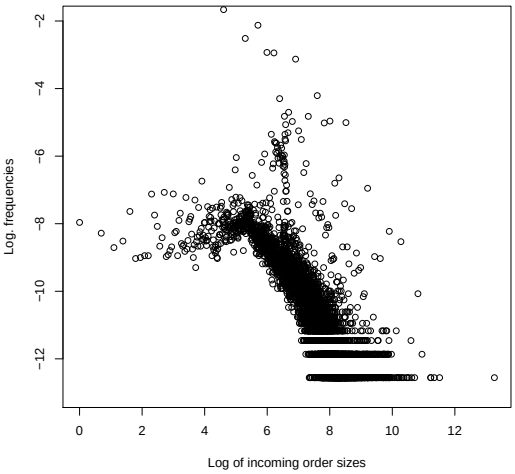
Additionally, we also investigate cases in which $p_{K,M}(\cdot)$ and $p_{Q,M}(\cdot)$ are described by the empirical univariate frequency distributions in our sample across k and q, respectively (emp). We also utilize the joint frequency distribution of the observed pairs (k,q) in one scenario (emp,emp). Note that although we use the empirical frequencies, the rates are

Figure 4.5: Distribution of Logarithmic Order Size

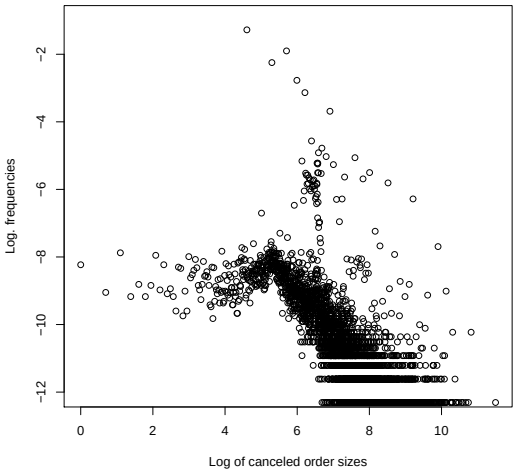
The figure presents the logarithmic frequencies of logarithmic order sizes of the MEO stock for arriving (a and b) or canceled (c and d) orders. For incoming orders, the logarithm of the original order size is used, whereas for order cancellations, the actually canceled remaining order size is utilized.



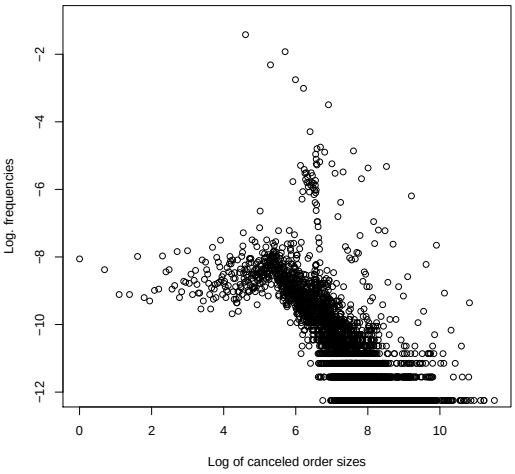
(a) Ask Arrivals



(b) Bid Arrivals



(c) Ask Cancellations



(d) Bid Cancellations

Table 4.2: Mean and Standard Deviation of Simulated Price Changes

For each of the 30 stocks, the time series mean and standard deviation of the logarithmic transaction price changes (in event time) across the 200 simulations has been calculated. The table reports the average across the 30 means $\bar{\mu}$ and standard deviations $\bar{\sigma}$ multiplied by 10^3. In the last row, the average across the observed time series means and standard deviations of the logarithmic transaction price changes are reported.

Initial Position:		January 2, 2004				March 31, 2004			
		Opening Auction		Midday Auction		Opening Auction		Midday Auction	
Scenario		$\bar{\mu}$	$\bar{\sigma}$	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\mu}$	$\bar{\sigma}$
uni	emp	-0.12	3.82	-0.02	3.05	-0.06	3.82	-0.04	3.17
uni	pow	-0.11	3.17	0.00	2.57	-0.03	3.29	-0.01	2.67
fix	emp	0.01	0.70	0.01	0.69	0.01	0.71	0.01	0.70
fix	pow	0.01	0.70	0.01	0.68	0.01	0.69	0.01	0.69
dyn	emp	0.00	1.21	0.00	1.17	-0.00	1.21	0.00	1.19
dyn	pow	0.00	0.87	-0.00	0.85	0.00	0.87	-0.00	0.86
emp	emp	-0.01	3.43	-0.01	3.43	-0.04	2.67	-0.04	2.89
emp	pow	-0.04	2.37	-0.02	1.48	-0.03	1.53	-0.03	1.56
observed		0.02	0.67	0.01	0.58	0.00	0.67	-0.00	0.49

fixed over the entire simulation run. Thus, in the scenario 'emp', no dynamic feedback between the state of the book and the arrival and cancellation rates is introduced.

For all combinations of these distributional specifications (in total 8 scenarios[26]) for each stock, we simulate 200 realizations of LOB evolutions over half a trading day (4 hours). The state at the beginning of our sample, i.e., after the opening auction on January 2, 2004 at 9h00 CET, serves as a starting point for the simulation.

4.3.2 Discussion of the Simulation Results

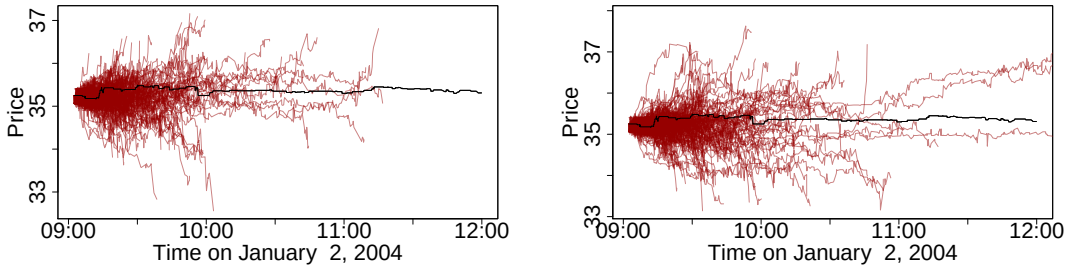
We first turn to the results from the uni scenario which are depicted in Figure 4.6. In this scenario, many simulation runs ended with an empty order book. We also see that the variance of transaction prices, which is induced by the uniform distribution, is rather high, especially, when using the empirical volume distribution. This is reported in Table 4.2 which shows the average mean and standard deviation of the simulated transaction price changes across all simulation runs.

Note that for these simulations, the average event rates on each market side (which is denoted $\bar{r}_{0,M,i,j,\cdot}$ in Equation (4.34) in Section 4.B) are the same as in the case of the fixed and dynamic arrival and cancellation rates. We may associate the uniform distribution across price levels with somewhat uninformed traders who, regardless of the price, randomly

[26] The scenarios are (uni,pow), (uni,emp), (fix,pow), (fix,emp), (dyn,pow), (dyn,emp), (emp,pow) and (emp,emp) where the list of pairs utilizes the introduced abbreviations and states the distribution across k in the first coordinate and the one across q in the second coordinate.

Figure 4.6: Scenario: Uniformly Distributed Arrival and Cancellation Rates

The graphs show 200 simulated paths of transaction prices (in red) using the scenario in which the arrivals and cancellations of orders follow a uniform distribution. The starting point of each simulation is the LOB position of the MEO stock on January 2, 2004. The true history of transaction prices during the first half of that day are depicted in black. In 4.6a, the empirical order size distribution is taken to generate the samples. In 4.6b, a power law is assumed to generate order sizes. Paths that end earlier than 12h00 result in an empty order book.



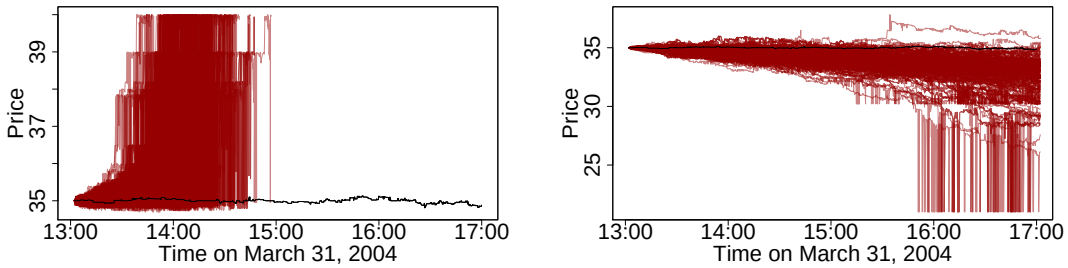
(a) Empirical **(b)** Power Law

submit orders in the vicinity of the current best quote. With the uniform distribution, the mean and variance of the price level is rather high compared to the DGX specifications in other simulation scenarios as well as the empirically observed equivalents. Throughout all simulation scenarios, we see that a higher mean and variance in the distribution across price levels are related to a higher variance in transaction prices. This result is interesting since, as noted before, arrival rates are linked to trader's behavior. So, if traders are uniformed on how the asset should be valued and constantly shift their valuation with no clear tendency and/or if traders are indifferent between immediate execution and delayed execution, transaction prices become highly volatile.

Second, the results for the fixed DGX distribution across price levels are presented in Figure 4.8. We can see that the power law very rarely induces large jumps in transaction prices due to the extremely large order sizes that are possible under this distributional scheme. In general, however, differences between the volume distributions are not obvious, neither in the 'uni' scenario nor in the 'fix' scenario. The average of the time series means and standard deviations for the simulations with the fixed DGX distributions as provided in Table 4.2 are close to the empirical ones. One very interesting result for the 'fix' scenario concerns the parameters of the DGX distribution. The distributional parameters μ and σ are almost identically defined: The parameter μ of the DGX distribution for incoming and canceled orders on the bid side has been specified slightly higher (at $\mu_{B,a} = 1.766$ and $\mu_{B,c} = 1.674$) than the one for the ask side ($\mu_{A,a} = 1.726$ and $\mu_{A,c} = 1.620$) to match estimated parameters from empirically observed frequencies. However, this small difference, does not seem to have any effect. In order to analyse the effect, we ran the simulation of the 'fix' scenario only for the stock MEO again with adjusted parameters: Leaving the

Figure 4.7: Special Case: Fixed DGX Distribution with an Imbalance in Arrival Rates

The graphs show 200 simulated paths of transaction prices (in red) using the scenario in which the arrival and cancellations of orders follow a fixed DGX distribution across price levels. However, an imbalance in the distribution of arrival rates is inserted as bid orders arrive densely in vicinity to the best ask price level. The starting point of each simulation is the LOB position for the MEO stock on March 31, 2004 after the midday auction at 13h00. The true history of transaction prices for the first half of that day is depicted in black. In 4.8a, the empirical order size distribution is taken to generate the samples. In 4.8b, a power law is assumed to generate order sizes.



(a) Empirical **(b)** Power law

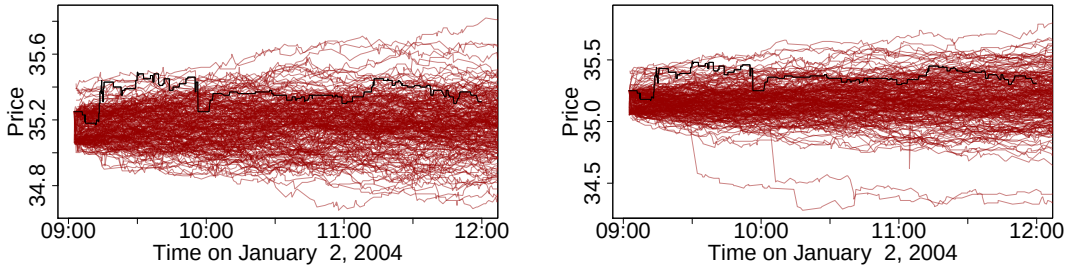
distribution of the cancellation rates untouched, only the parameters of DGX distribution across arrival rates are altered to $\mu_{B,a} = 0.1$ and $\mu_{A,a} = 2$ as well as $\sigma_{B,a} = 0.2$ and $\sigma_{A,a} = 1$. Therewith, the distribution of bid order arrivals is much more dense around the best ask price. The result can be seen in Figure 4.7.

The behavior of the transaction prices also depends crucially on the cancellation rates as well as the volume distribution. In the case where order size is distributed according to a power law distribution, order sizes on both market sides are identically distributed. Furthermore, the power law distribution generates a lot of small orders while large orders are very rare. At the same time, the initial position on the bid side contains several large orders close to the best bid price (which are more likely to be canceled). The ask side consists of several medium sized orders close to the best ask, while the large orders rest deep in the book. So, the frequent small orders inserted at or close to the best ask price are not able to move the market upwards permanently due to the medium sized orders sitting in the book on the ask side. At the same time large orders at the front of the bid side (from the initial position) are canceled frequently. One rare large ask order generated by the power law, thus, is able to move the bid price quite a lot. The longer the simulation is running, the more likely it is for a large ask order to occur and the more likely it is that the large orders at the top of the bid side are already canceled. This makes it easier for the ask side to move the best ask down and therefore transaction prices deteriorate.

In the empirical distribution, the volume distribution on the bid side dominates the volume distribution of the ask side. Thus, bid orders inserted into the book are larger in size than

Figure 4.8: Scenario: Fixed DGX Distribution for Arrival and Cancellation Rates

The graphs show 200 simulated paths of transaction prices (in red) using the scenario in which the arrival and cancellations of orders follow a fixed DGX distribution across price levels. The starting point of each simulation is the LOB position for the MEO stock on January 2, 2004. The true history of transaction prices for the first half of that day are depicted in black. In 4.8a, the empirical order size distribution is taken to generate the samples. In 4.8b, a power law is assumed to generate order sizes.



(a) Empirical **(b)** Power Law

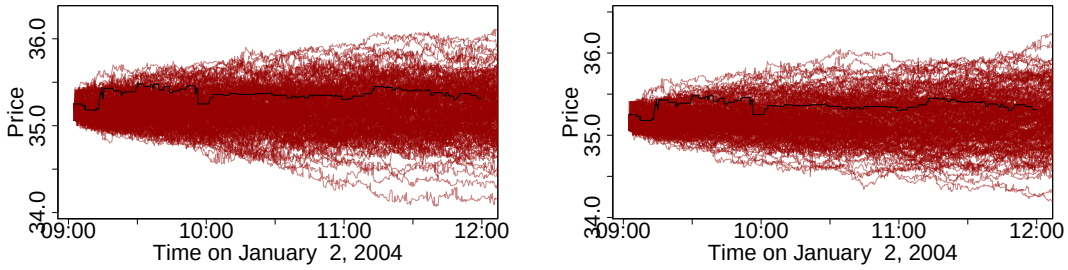
inserted ask orders. Since bid orders are inserted close to or at the best ask price level, the bid side moves the market soon after the simulation starts upwards. This leads to a hefty upward drift of transaction prices in each simulated transaction path and empty order books on the ask side. Hence, we can conclude that limit order distributions with a high probability mass in the vicinity of the best quote in one market side push transaction prices in the direction of the opposite market side if the order size distribution allows for frequent medium sized orders.

Third, the simulated transaction prices resulting from the scenario with dynamical shifting and scaling DGX distributions which depend on the prevailing spread, are shown in Figure 4.9. In this scenario the moments of the DGX distribution depend on the prevailing spread. The dynamical adjustment of the arrival and cancellation rates across price levels is balanced, So, the mean and standard deviation of the DGX distribution across price levels is the same on both market sides. Furthermore, the large jumps induced by the power law distribution are still rare, but rather pronounced. These jumps are, however, not sufficient to cause an increase in volatility. In fact, the scenarios with a power law distribution exhibit on average a slightly smaller volatility which might be due to the fact that the order size is rather small. For the power law distributed order size, the time series mean and standard deviation of the dynamically adjusting simulation scenario are close to the time series mean and standard deviation of real observed logarithmic transaction changes as presented in Table 4.2. The scenario with the empirical order size distribution is too volatile.

Fourth, using the unconditional empirical frequency distributions as well as the empirical rates $\bar{r}_{0,M,i,j,\cdot}$, the results presented in Figure 4.10 deviate from empirical stylized facts in

Figure 4.9: Scenario: Dynamical DGX Distribution for Arrival and Cancellation Rates

The graphs show 200 simulation paths of transaction prices (in red) using the scenario in which the arrival and cancellations of orders follow a dynamical DGX distribution across price levels. In the dynamical DGX distributions the parameters μ and σ are functions of the prevailing integer spread. The starting point of each simulation were the LOB positions for the MEO stock on January 2, 2004. The true history of transaction prices for the first half of that day are depicted in black. In 4.9a, the empirical order size distribution is taken to generate the samples. In 4.9b, a power law is assumed to generate order sizes.



(a) Empirical **(b)** Power Law

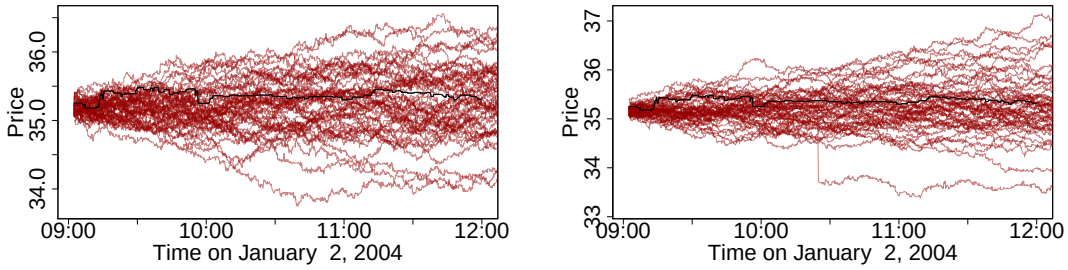
that the resulting paths are much more volatile. The imbalance between ask and bid order arrivals and cancellations exhibited by some stocks, together with the observation that on average ask orders arrive closer to the best bid than bid orders to the best ask, drags transaction prices on average slightly down (see Table 4.2). The smaller distance to the best quote, on average, of ask orders can be seen in Figure 4.2 as well as in Figure 4.11. It seems that in our sample, sellers tend to seek quicker order execution by placing their orders close to the buy side. Buyers on the other hand, test their fortunes and patiently wait for a good deal to occur deeper in the book. This can be clearly seen in Figure 4.11: For the same average spread, arriving bid orders are placed on average further away from the opposite market side than arriving ask orders.

4.4 Empirical Analysis

In our empirical analysis, we focus on three variables that we deem most important to market participants.The first is the logarithmic return which could be achieved based on a buy-and-sell strategy in subsequent intervals. The second is the return of a sell-and-buy strategy. The third one is the exchange liquidity measure (XLM) as introduced in Section 4.1.4, calculated with a round-trip of EUR 100.000. To implement them, we sample the data in intervals of fixed length Δt (1, 2, 5, 10, 15, 30, 45, 50, 60, 120, 240 minutes). Let t denote the last point in time of some arbitrary interval and $t-1$ the last point in the previous interval. Then the logarithmic returns of a buy-and-sell ($\Delta p_{t,b}$) and

Figure 4.10: Scenario: Empirical Frequency Distribution for Arrival and Cancellation Rates

The graphs show 200 simulation paths of transaction prices (in red) using the scenario in which the arrival and cancellations of orders follow a dynamical DGX distribution across price levels. In the dynamical DGX distributions the parameters μ and σ are functions of the prevailing integer spread. The starting point of each simulation were the LOB position for the MEO stock on January 2, 2004. The true history of transaction prices for the first half of that day are depicted in black. In 4.10a, the empirical order size distribution is taken to generate the samples. In 4.10b, a power law is assumed to generate order sizes.



(a) Empirical **(b)** Power Law

a sell-and-buy strategy ($\Delta p_{t,s}$) are given as

$$\Delta p_{t,b} = \log(\beta_{A,t}) - \log(\beta_{B,t-1}),$$
$$\Delta p_{t,s} = \log(\beta_{B,t}) - \log(\beta_{A,t-1}).$$

For the XLM the last observation in the respective interval is taken.

We have shown in Equation (4.20) that the moments of some observable O of the LOB system can be expressed as

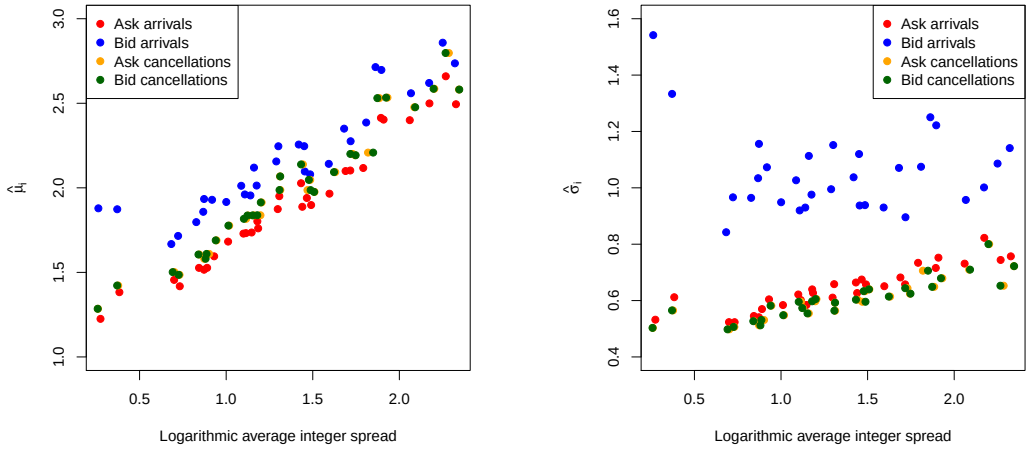
$$E_{t_0}[O^\nu] = \sum_{|z\rangle} \langle z| O^\nu e^{H(t-t_0)} |\psi_0\rangle .$$

We can perceive the right hand side of this equation as an intricate function of the arrival and cancellation rates. These rates depend on the event rates $\bar{r}_{0,M,i,j,e}$ of arrivals and cancellations of the various order types, the relative logarithmic integer price level d_l, the order size q, the spread Δ, and possibly other variables. Thus, we can formulate a linear approximation for the expectation of any observable in the moments of exactly these variables.

By repeated Taylor series approximations of the terms in Equation (4.34) in the variables d_l, q, and Δ around their respective mean, collecting terms, taking expectations and neglecting terms with an order higher than four (see Section 4.D), we get for the expected value of some observable a linear approximation of the form

Figure 4.11: Estimated DGX Parameters and Average Spread

The scatter plot depicts the average spread of DAX components in the sampling interval (first quarter of 2004) during continuous trading of the XETRA order book against the estimated parameters of a DGX-distribution fitted to the unconditional frequencies of order arrivals (and cancellations) across price levels. The parameters were estimated using the log-likelihood method described in Bi et al. (2001).



(a) $\hat{\mu}_i$ against average spread

(b) $\hat{\sigma}_i$ against average spread

$$
\begin{aligned}
\mathbb{E}_{t_0}[O_{i,t}] = \gamma_{0,i} + \Bigg(\delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v]\Bigg) \times \\
\sum_{M,j,e} \rho_{0,M,i,j,e}\mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}] \times \\
\Bigg(\kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e}[d_{l,t}] + \\
\sum_{v=2}^{4} \kappa_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e}[(d_{l,t} - \mu_{d_l,t})^v]\Bigg) \times \\
\Bigg(\xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e}[q_{t,M,j,e}] + \\
\sum_{v=2}^{4} \xi_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e}[(q_{t,M,j,e} - \mu_{q,t})^v]\Bigg) + \varepsilon_i, \qquad (4.35)
\end{aligned}
$$

where $\mu_{x,t} = \mathbb{E}_{t_0,M,i,j,e}[x_t]$. Note that the indices in the subscript of the expected values indicate their conditioning set as outlined in Section 4.1.4. Hence, the tuple of subscripts (t_0, M, i, j, e) indicates that the conditional expectation is formed with information available at time t_0 for stock i given the market side M, order type j, and event type e.

As we are not interested in $\gamma_{0,i}$, $\rho_{0,M,i,j,e}$, $\kappa_{M,i,j,e,v}$, $\xi_{M,i,j,e,v}$, and $\delta_{i,v}$, we can collect all possible products in Equation (4.35) in the parameters $\gamma_{0,i}, \gamma_{1,i}, \ldots \gamma_{1,R}$. In total, this

yields 1.971 parameters. Hence, the estimation of the parameters is sensibly feasible using ordinary least-squares up to non-overlapping intervals with a length of 15 minutes. For non-overlapping 15 minute intervals, we can get 2.164 observations from the 64 trading days in our sample. Increasing the interval length beyond 15 minutes makes the use of overlapping intervals and rolling variable calculation necessary. While this is in principle feasible, we restrict the analysis for the specification in Equation (4.35) to non-overlapping intervals and sampling frequencies below 15 minutes.

To increase the length of the intervals, we use three alternative specifications which entail less parameters. In the first alternative, the moments of the spread are only included additively:

$$\begin{aligned}\mathbb{E}_{t_0}[O_{i,t}] = \gamma_{0,i} + \delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v] + \\ \sum_{M,j,e} \rho_{0,M,i,j,e}\mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}] \times \\ \left(\kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[d_{l,t}] + \right. \\ \left. \sum_{v=2}^{4} \kappa_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(d_{l,t} - \mu_{d_l,t})^v]\right) \times \\ \left(\xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[q_{t,M,j,e}] + \right. \\ \left. \sum_{v=2}^{4} \xi_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(q_{t,M,j,e} - \mu_{q,t})^v]\right) + \varepsilon_i. \end{aligned} \tag{4.36}$$

This specification entails the estimation of 399 parameters. This enables us to estimate the model on interval lengths of up to one hour.

In the second alternative, the moments of the spread and the moments of the order size are included additively:

$$\begin{aligned}\mathbb{E}_{t_0}[O_{i,t}] = \gamma_{0,i} + \delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v] + \\ \sum_{M,j,e} \rho_{0,M,i,j,e}\mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}] \times \\ \left(\kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[d_{l,t}] + \right. \\ \left. \sum_{v=2}^{4} \kappa_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(d_{l,t} - \mu_{d_l,t})^v]\right) + \\ \xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[q_{t,M,j,e}] + \\ \sum_{v=2}^{4} \xi_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(q_{t,M,j,e} - \mu_{q,t})^v] + \varepsilon_i \; . \end{aligned} \tag{4.37}$$

This reduces the number of parameters to 95 and makes interval lengths of up to 4 hours possible.

In the third alternative, we employ a completely additive structure:

$$\begin{aligned}\mathbb{E}_{t_0}[O_{i,t}] = \gamma_{0,i} + \delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v] + \\ \sum_{M,j,e} \rho_{0,M,i,j,e}\mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}] + \\ \kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[d_{l,t}] + \\ \sum_{v=2}^{4} \kappa_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(d_{l,t} - \mu_{d_l,t})^v] + \\ \xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e,t}[q_{t,M,j,e}] + \\ \sum_{v=2}^{4} \xi_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e,t}[(q_{t,M,j,e} - \mu_{q,t})^v] + \varepsilon_i \; . \end{aligned} \tag{4.38}$$

This reduces the number of parameters which have to be estimated further to 43.

For each of the four specifications in Equations (4.35) to (4.38), we use a formulation in which the moments on both sides of the equations are estimated contemporaneously, i.e., at the same time t. Naturally, this is only possible in-sample. Additionally, we also investigate specifications of Equations (4.35) to (4.38) in which the moments on the right hand side are estimated at $t-1$ to describe the expectation of the observable on the left hand side. This formulation allows for out-of-sample evaluation of the model in terms of predictive power assuming that the moments in one interval anchor the moments of the subsequent interval. We evaluate our model across several measures in- and out-of-sample.

4.4.1 In-Sample Analysis

For the in-sample evaluation, we consider the adjusted and unadjusted R^2 as well as the root mean squared error ($RMSE$). Furthermore, following Zhou et al. (2018), we also use the direction prediction accuracy (DPA) defined as

$$DPA = \frac{100}{T} \sum_{t=1}^{T} \frac{\max\left(0, \Delta p_{\cdot,t} \cdot \Delta \hat{p}_{\cdot,t}\right)}{\Delta p_{\cdot,t} \cdot \Delta \hat{p}_{\cdot,t}}. \tag{4.39}$$

The in-sample results are depicted in Figures 4.12 to 4.15. In each figure, the contemporaneous models are depicted in subfigures (a) and (b) for returns and (e) for the XLM measure while the results for the specifications which use only past information to model the current state are presented in subfigures (c) and (d) for returns and (f) for the XLM measure. Every line in the graph represents the results for one stock. The highlighted thicker line is the average of the respective measure across all stocks. There are a few noteworthy results.

The DPA as well as the R^2 measures (adjusted and unadjusted) allow us to reject the hypothesis that the contemporaneous as well as the lagged models have no significance in explaining the data. Our results suggest that the extensive model in Equation (4.35) overfits the data with growing sampling frequency and, hence, less observations. This can be seen by the drastically increasing R^2 and DPA values when the sampling frequency is increased. The adjusted R^2 should account for this effect, and indeed remains rather stable. However, when the degrees of freedom of the model become sparse, the adjusted R^2 is not able to correct the full extent of the overfitting.

Nonetheless, the high values at the highest frequencies indicate that the contemporaneous model describes high-frequency returns very well. As can be seen in Figures 4.12 to 4.15, for all measures, the contemporaneous specification in Equation (4.35) (blue) turns out to be superior to the specifications in Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) , i.e., it has a higher direction prediction accuracy, better fit in terms of higher R^2 values, and results in a lower root mean squared error.

Subfigures (c), (d), and (f) in Figures 4.12 to 4.15 present the results using the lagged specifications of Equations (4.35) to (4.38), i.e., they compare their out-of-sample forecast performance. Recall that these specifications rely heavily on the assumption that the arrival rates stay constant for some (very) short time horizons. Therefore, the results regarding the performance of the different models turns out to be different compared to the in-sample evaluation above. Now, the specifications in Equations (4.37) and (4.38), which use far less interaction terms and have a rather small number of parameters, capture the dynamics of returns almost as well as the other two specifications in Equations (4.35)

and (4.36) which becomes apparent, for example, when looking at the DPA in Figure 4.12. Only considering the adjusted R^2 (Figure 4.13), the specification in Euqation (4.35) has a better performance. It also slightly decreases the $RMSE$.

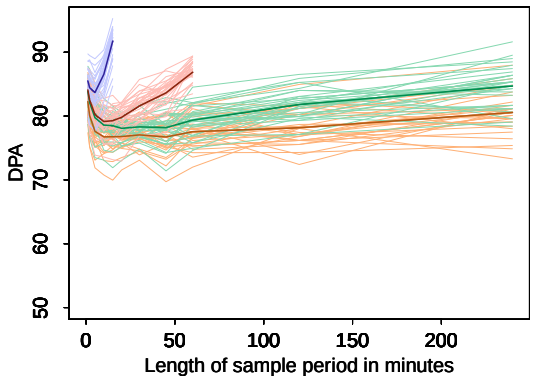
For the XLM measure, the performance of the highly parameterized models is better. In our opinion, this is due to the following two reasons. First, the XLM measure changes with each event, so that on a high frequency the volatility of the XLM measure is higher than the volatility of returns. The second reason is rooted in the definition of the XLM measure as a fraction of observables which in general requires a higher polynomial degree to arrive at a sensible approximation. The more complex models in Equations (4.35) and (4.36) are, thus, better suited to provide such an approximation. Note also that we take the last observation for the XLM within an interval which is highly volatile. From a trading perspective, the average XLM over the interval may be better suited to describe liquidity of the market during the interval and might be a less volatile measure. Nevertheless, we choose the more volatile measure for our analysis. Our results should therefore pose a lower bound with respect to the modeling accuracy.

Nevertheless, for intraday returns, the relatively parsimonious models in Equations (4.37) and (4.38) perform remarkably well. Compared to other studies like Zhou et al. (2018) who use high-dimensional neural networks (without contemporaneous information), we find a decisively smaller $RMSE$. For example, the out-of-sample forecast error reported by Zhou et al. (2018) for their best model (GAN for minimizing forecast error loss and direction prediction loss with training sets with a length of 20 days and test sets with a length of 5 days) is 0.0079 with a DPA of 69% on a 1 minute interval. Even though the in-sample results in our case are not exactly comparable, we can note that for the models with lagged information on 1 minute intervals an average $RMSE$ of 0.0010 can be achieved together with a DPA slightly above 85% on average. We also see that, on high frequencies, the models with lagged information all result in similar $RMSE$ and DPA values. Again the results for R^2 show that overfitting is a problem for lower sampling frequencies. These only differ in R^2 (adjusted and unadjusted). Nevertheless, the size of the adjusted R^2 of around 15% on 1 minute intervals is remarkable.

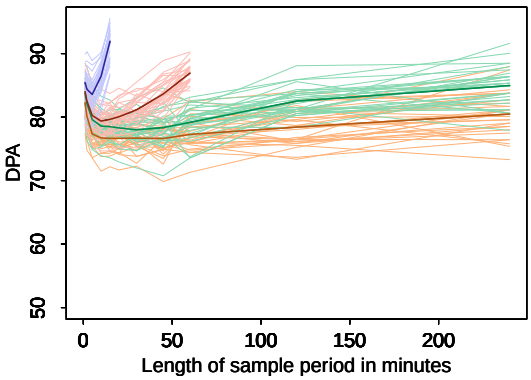
All models perform better at intervals of 45 minutes than on other frequencies. This is rooted in an above-average precision of the prediction of the return in the last interval of the day. When using 45 minutes intervals, the last interval of the day is only 30 minutes long. Hence, estimated moments based on the observations within the previous 45min interval are used to predict the last, shorter, 30min-long interval. If these shorter end-of-day intervals are not excluded from the data, the model performs better compared to other frequencies. This suggests another avenue for further research: Can the model performance be improved by using longer intervals to estimate sample moments to then make predictions on smaller time horizons. Put differently, is it in general beneficial to

Figure 4.12: In-Sample Direction Prediction Accuracy

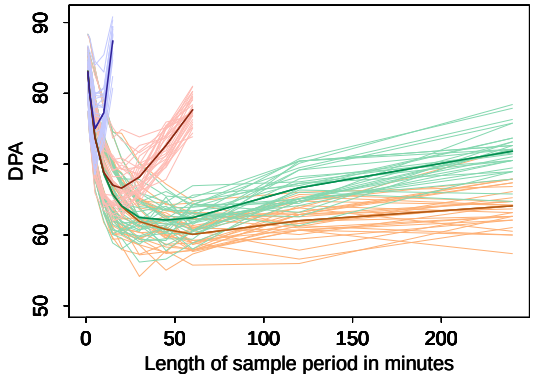
The figures report the in-sample direction prediction accuracy (DPA) (as defined in Equation (4.39)). The in-sample DPA is reported for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240.



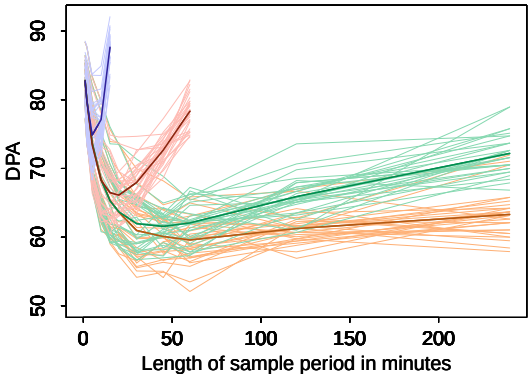
(a) Contemporaneous Buy



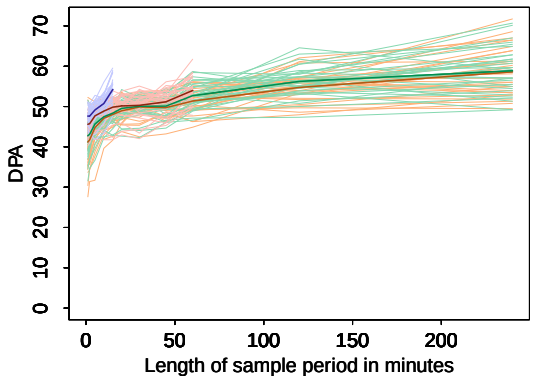
(b) Contemporaneous Sell



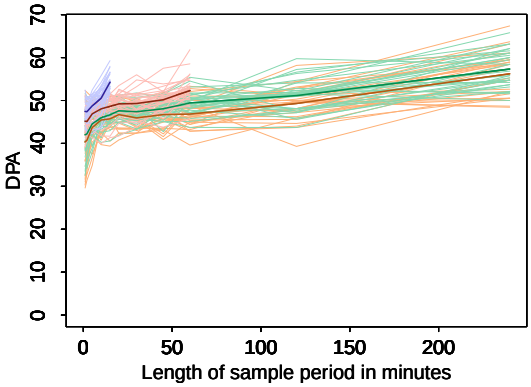
(c) Lagged Buy



(d) Lagged Sell



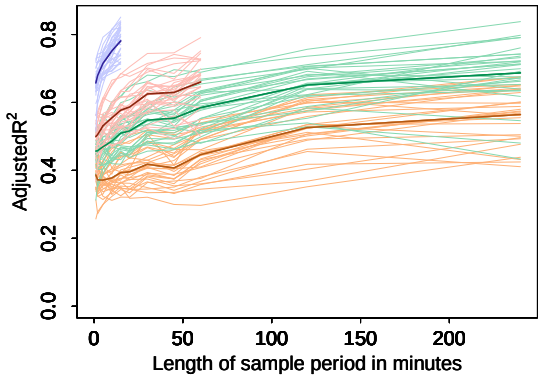
(e) Contemporaneous XLM



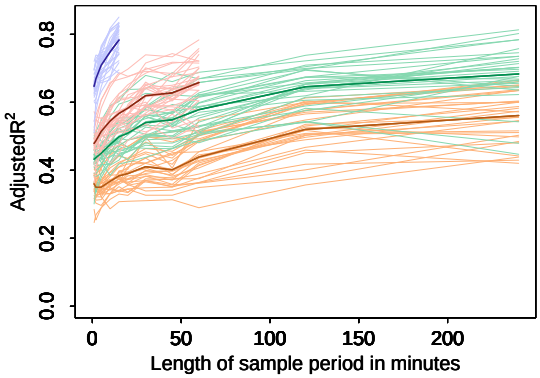
(f) Lagged XLM

Figure 4.13: In-Sample Adjusted R^2

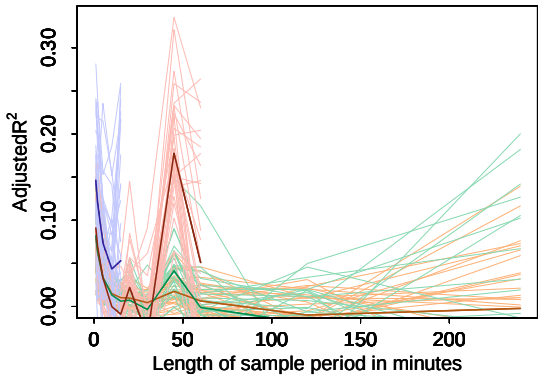
The figures below report the in-sample, adjusted R^2 for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240.



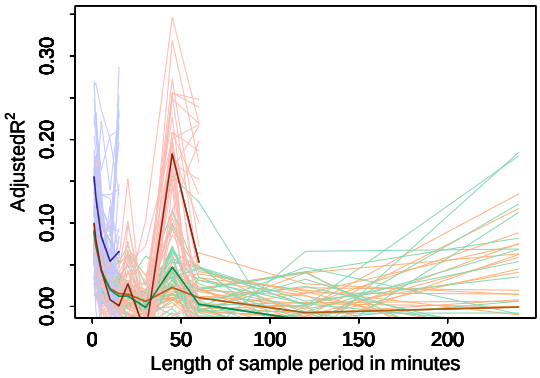
(a) Contemporaneous Buy



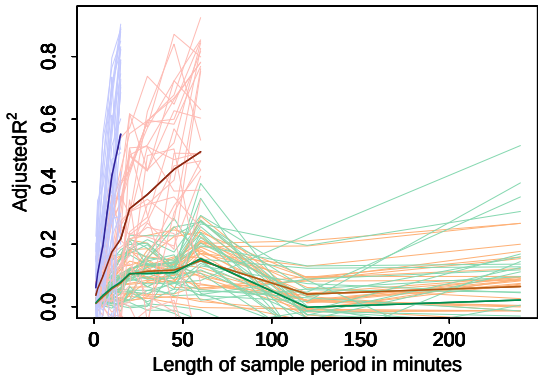
(b) Contemporaneous Sell



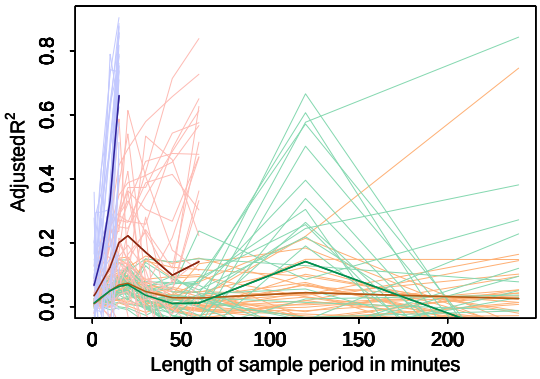
(c) Lagged Buy



(d) Lagged Sell



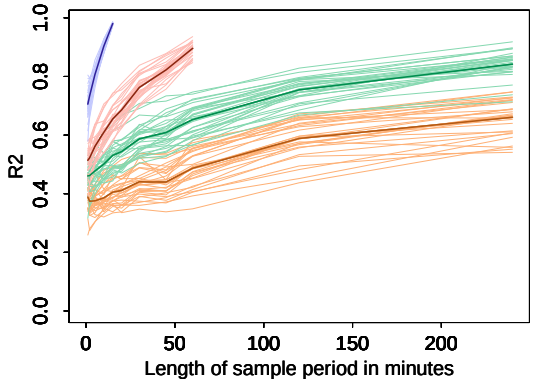
(e) Contemporaneous XLM



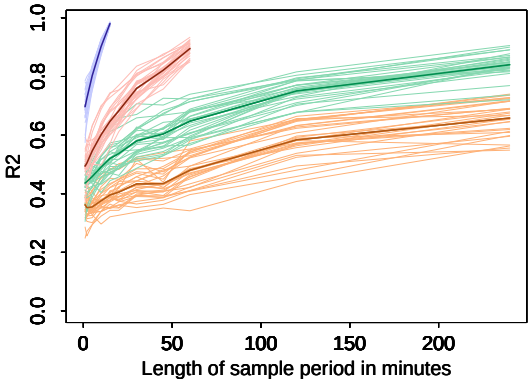
(f) Lagged XLM

Figure 4.14: In-Sample R^2

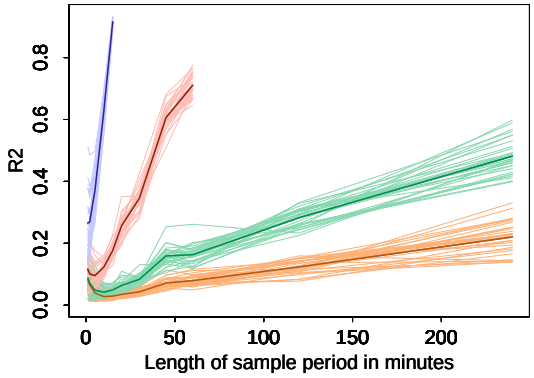
The figures report the in-sample R^2 for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240.



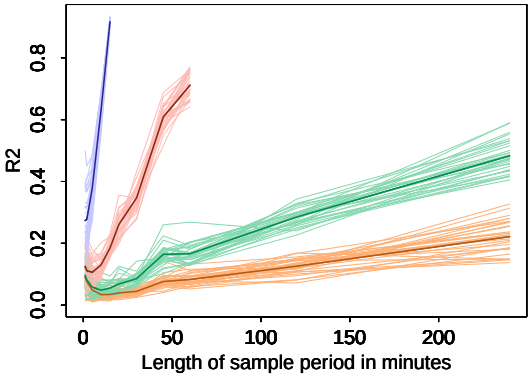
(a) Contemporaneous Buy



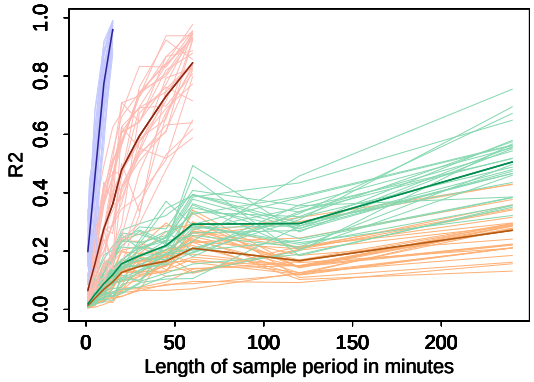
(b) Contemporaneous Sell



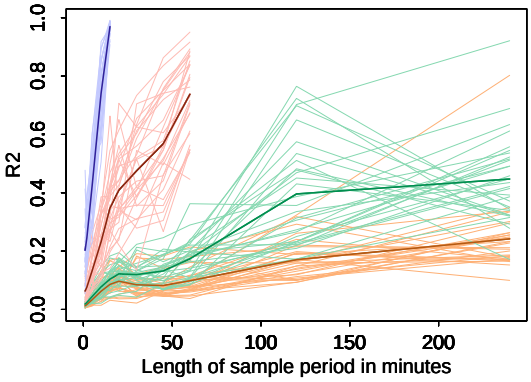
(c) Lagged Buy



(d) Lagged Sell



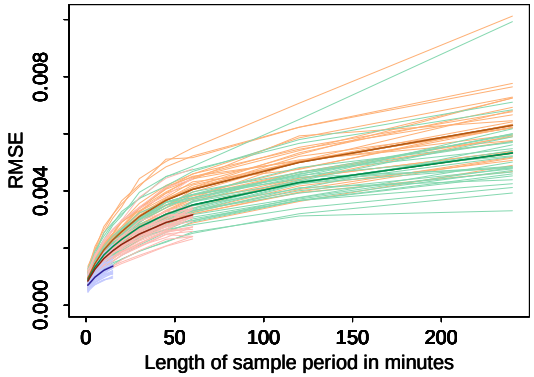
(e) Contemporaneous XLM



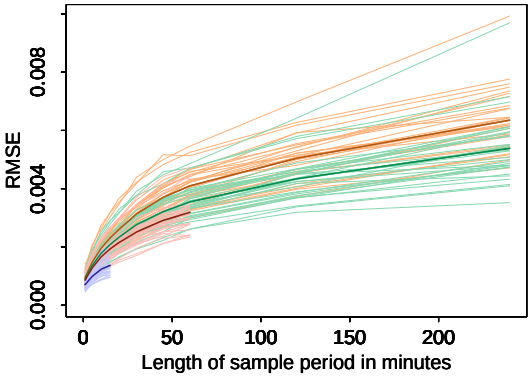
(f) Lagged XLM

Figure 4.15: In-Sample $RMSE$

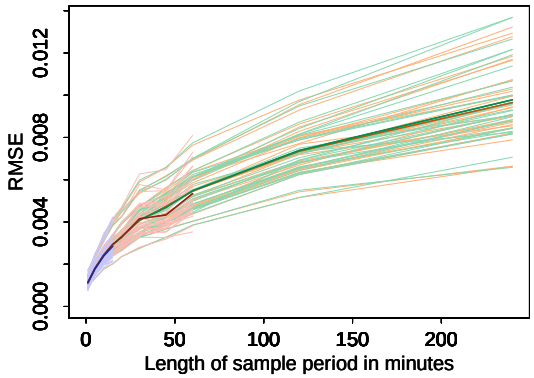
The figures report the in-sample $RMSE$ for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240.



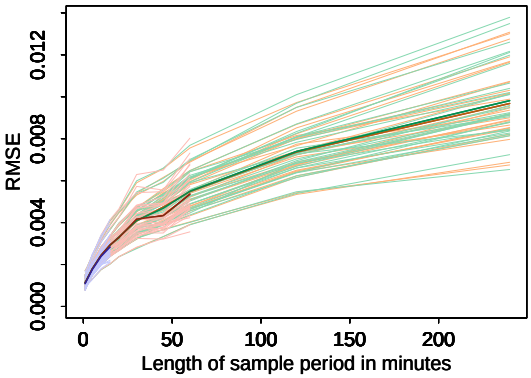
(a) Contemporaneous Buy



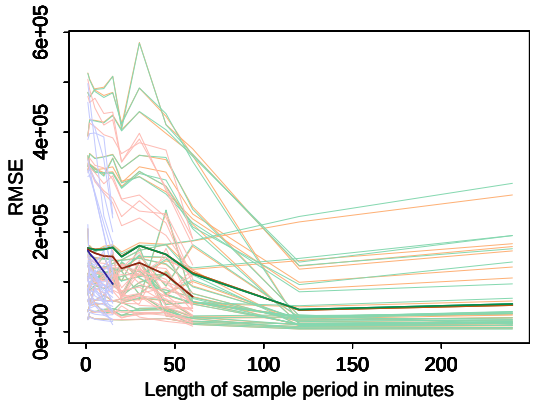
(b) Contemporaneous Sell



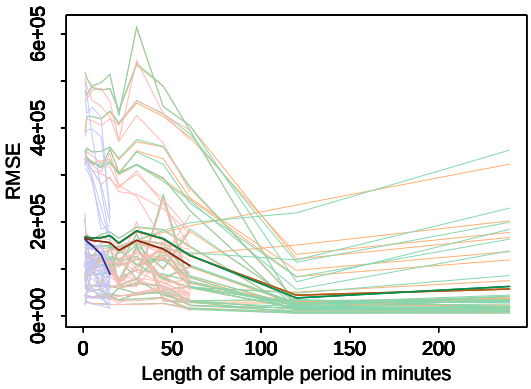
(c) Lagged Buy



(d) Lagged Sell



(e) Contemporaneous XLM



(f) Lagged XLM

calculate moments on the last hour of data to forecast the next ten minutes? The result for the 45min interval suggest there may be some benefits. However, answering the question would entail a rolling calculation of means on overlapping intervals which is beyond the scope of the current chapter.

4.4.2 Out-of-Sample Analysis

We also evaluate our models based on out-of-sample predictions using rolling windows. Depending on the size of the non-overlapping intervals, we vary the number of intervals included in the rolling window.[27] Table 4.3 presents the lengths of these windows for the frequencies considered.

In addition to the root mean squared prediction error ($RMSPE$) and the out-of-sample DPA, we also use the R^2 of a Mincer-Zarnowitz regression (Mincer and Zarnowitz 1969) to evaluate the model.

Table 4.3: Rolling Windows

The table lists the number of intervals (and thus observations) within the rolling windows used to fit the model. The third column lists the approximate number of trading days over which the rolling window is spanned. For each window, we conduct an out-of-sample prediction. In the last column, the potential total number of non-overlapping intervals, i.e., the number of available observations is reported. The actual number of observations for which an out-of-sample forecast is produced depends on the availability of the necessary moments for the estimation of Equations (4.35) to (4.38).

Frequency	Intervals	Days	Total
1 min	10.000	21	32.503
2 min	5.000	21	16.252
5 min	4.000	42	6.501
10 min	2.500	52	3.251
15 min	1.500	47	2.167
20 min	750	31	1.626
30 min	500	31	1.084
45 min	500	47	723
60 min	300	38	542
120 min	150	38	271
240 min	100	50	136

[27] Recall that when we talk about using *non-overlapping intervals*, we mean the procedure to use the observations associated to the last interval of the rolling window to predict the observation of the next interval. Alternatively, one could use *overlapping-intervals*, i.e., update the observation at each new event to predict the next interval. However, this procedure is computationally very demanding and, therefore, out of the scope of the current chapter.

Results are reported in Table 4.4 for selected 1 and 5 minute intervals. As can be seen, for the return series, the precision of the out-of-sample prediction is remarkably high. On a 1 minute frequency, we are able to predict the direction of the next price change with an average accuracy at or over 80%, irrespective of the model. But even on the 5 minute frequency, the accuracy only falls slightly below 75%. For both buy and sell returns, we deem the R^2 rather high and the $RMSPE$ rather low given that we intend to predict financial returns on a ultra-high frequency.[28]

For the XLM, results are somewhat different. The precision of the forecast is poor which is in line with the in-sample results. Again, this is due to the high variability of the measure and its structure. The XLM changes with each event and is a highly nonlinear function in the arrival rates (see Equations (4.23) to (4.25)). Therefore, the linear approximation may be poor and the approximation for longer time horizons may be especially poor. In this line of argument, it is worth mentioning that the model with the highest complexity in our considerations (specified in Equation (4.35)) performs best in all evaluation measures.

The results for all stocks and all frequencies are presented in Figures 4.16 to 4.18. As we can see the smaller the interval, the better the forecasting ability of all our linear models. The extensive linear approximation in Equation (4.35) predicts the direction of the returns very well on small intervals. The R^2_{MZ} of the Mincer-Zarnowitz regression of above 2% for the sell strategy is above what we had expected for returns on ultra-high frequencies. On intervals longer than 10 minutes, the predictive ability of all three linear approximations is, however, poor. It can also be noted, that the sparse model formulations in Equations (4.37) and (4.38) perform just as well, or even better in some situations, than the heavily parameterized formulation in Equation (4.35) in the case of the return series. This is not true for the XLM. For the XLM, the more complex formulations in Equation (4.35) and (4.36) perform better in all measures. Especially, the constant $RMSE$ and the increasing DPA and R^2_{MZ} up to 5-min intervals are remarkable. The variance of the DPA in Figure 4.16c shows how noisy the XLM and the associated forecasts are and by how much the more complex model is able to reduce this variability.

[28] For the stock ADS, we also observe a 100% accuracy when predicting the direction of the next price change. Also the R^2 of the Mincer-Zarnowitz regression is around 50%. However, it needs to be mentioned that for this stock, only 19 out-of-sample predictions are made in total. Since, based on 1 minute intervals, some moments that enter the right hand side of Equations (4.35) to (4.37) cannot be calculated, we drop the observations for these intervals from our sample. In effect ADS has 10.019 valid observations on a 1 minute frequency. Due to such missing values also out-of-sample 1 minute results for DB1, FME and HEN3 are not reported since less than 10.000 observations are valid for these stocks. This is why we do not include the 1-min out-of-sample results for ADS in the figures.

Table 4.4: Out-of-Sample Results: 1 and 5 Minute Interval Forecasts

The table presents the out-of-sample results for the 1 and 5 minute intervals for the stocks ALV and FME. The model alternatives in Equation (4.35), Equation (4.36), Equation (4.37) and Equation (4.38) are referred to in the rows A1 - A4 respectively. For each model, the $RMSPE$, the fit of the Mincer-Zarnowitz regression (R^2_{MZ}) and the direction prediction accuracy (DPA) are reported, both for the returns of the buy strategy $\Delta p_{t,b}$ and the ones of the sell strategy $\Delta p_{t,s}$. ALV is the most liquid stock with the highest number of events in the sample period while FME is one of the less liquid stocks.

		1min (ALV)			**5min (FME)**		
	Model	$\Delta p_{t,s}$	$\Delta p_{t,b}$	XLM	$\Delta p_{t,s}$	$\Delta p_{t,b}$	XLM
$RMSPE$	A1	0.00103	0.00104	146989	0.00202	0.00199	216247
	A2	0.00095	0.00096	161087	0.00183	0.00174	220908
	A3	0.00094	0.00094	164625	0.00180	0.00170	224286
	A4	0.00094	0.00094	165198	0.00179	0.00170	224677
R^2_{MZ}	A1	0.0101	0.0071	0.2546	0.0357	0.0150	0.1076
	A2	0.0131	0.0099	0.0802	0.0307	0.0316	0.0372
	A3	0.0146	0.0128	0.0396	0.0321	0.0378	0.0079
	A4	0.0145	0.0137	0.0213	0.0323	0.0393	0.0050
DPA	A1	77.02	77.75	47.11	74.66	72.96	51.58
	A2	80.22	80.45	46.56	78.28	79.19	41.18
	A3	80.67	80.80	45.94	80.32	79.75	34.16
	A4	80.72	80.92	46.32	80.54	79.86	32.13

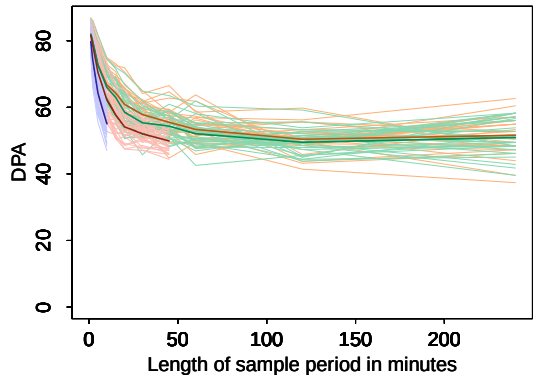
4.5 Summary

In this chapter, we have shown that the limit order book (LOB) can be described as a continuous Markov process. The description is based on the operator algebra which we borrow from physics. Our model closely describes the reality of the order book and identifies the arrival and cancellation rates as the key ingredients of the book's dynamics. Via a simulation study, we show that the distribution of order arrival rates across price levels determines the shape of the book and, as a consequence, the transaction price evolution. By varying the type and shape of arrival and cancellation rates across prices and volume, we find that the moments of price levels and quantity levels of incoming and canceled orders are important determinants for the evolution of the book.

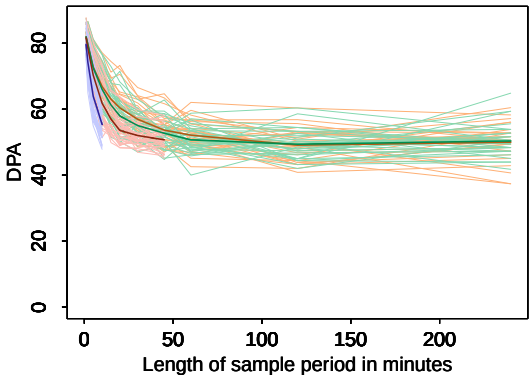
In an empirical study which is based on a linearized version of our model, we estimate

Figure 4.16: Out-of-Sample Direction Prediction Accuracy

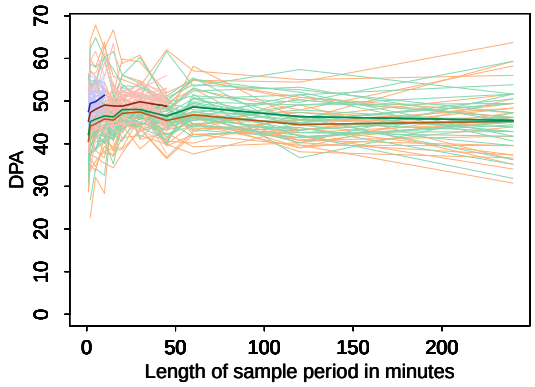
The figures report the out-of-sample direction prediction accuracy (DPA) (as defined in Equation (4.39)). The out-of-sample DPA is reported for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240 minutes, estimated with a rolling window one-step ahead forecast. The respective window lengths are listed in Table 4.3.



(a) Buy



(b) Sell



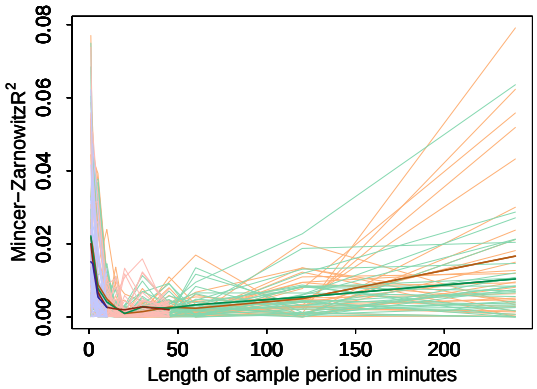
(c) XLM

three different specifications on non-overlapping intervals of various frequencies. As we have the entire record of the XETRA order book for 3 months at our disposition, we can include a large number of parameters in the estimation such that an evaluation of the model becomes feasible. In-sample, all considered models exhibit a good fit in terms of R^2, $RMSE$ and direction prediction accuracy (DPA). Our fully parameterized model seems to overfit the data on lower frequencies. Nevertheless, when using only past information, the values for the adjusted R^2 range for the minute-by-minute intervals around or over 10% whereas the direction is correctly predicted in around 70% of all cases.

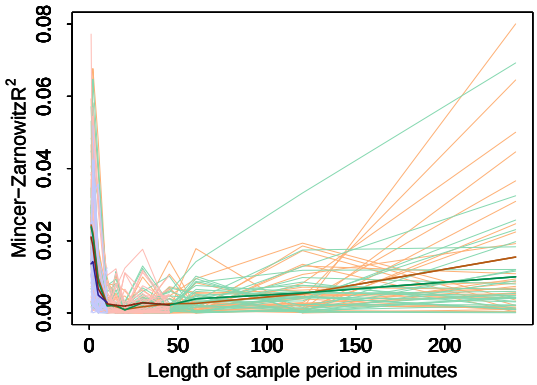
To evaluate the robustness of our results we also conduct an out-of-sample test of the model. We use one-step-ahead forecasts on various frequencies and evaluate the accuracy with the R^2_{MZ} of a Mincer-Zarnowitz regression, the DPA as well as the $RMSPE$. We find

Figure 4.17: Out-of-sample Mincer-Zarnowitz R^2_{MZ}

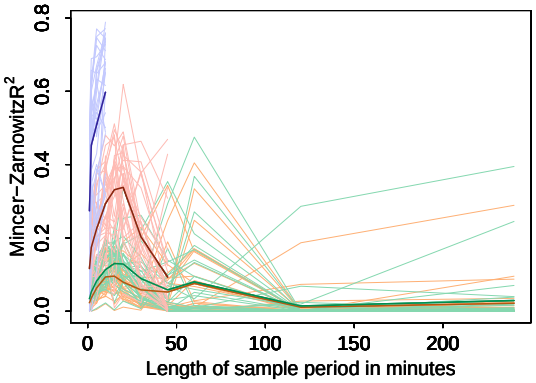
The figures report the R^2_{MZ} of the Mincer-Zarnowitz regression Mincer and Zarnowitz (1969) based on the out-of-sample one-step ahead rolling window forecast. The R^2_{MZ} is reported for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red) , Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240 minutes. The respective window lengths are listed in Table 4.3.



(a) Buy



(b) Sell



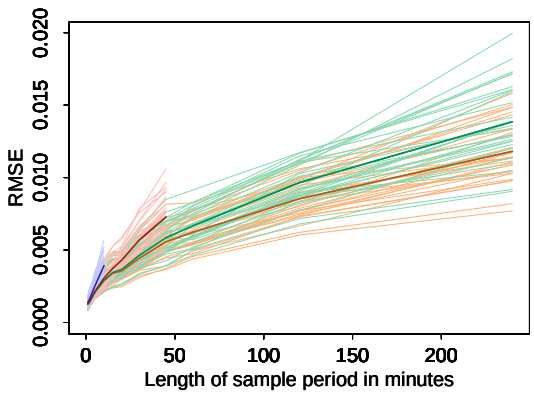
(c) XLM

that our model predicts 1 minute returns very well. Again, we find an unexpected fit for ultra-high returns as the R^2_{MZ} is generally above 2%. In addition, on low frequencies the DPA is around 80%, we can predict the directional change of the next return very well. The time varying estimates for the parameters as well as the short forecasting horizon make the return prediction astonishingly accurate. We also try to predict liquidity at the end of each interval with the XLM measure. The measure cannot be forecast well for longer time intervals with adequate accuracy. Also on very short time horizons, the best fitting model is barely able to predict the direction of the next change of the XLM in more than 50% of the cases. This has to do with the very volatile nature of the XLM and how the XLM is defined in the first place.

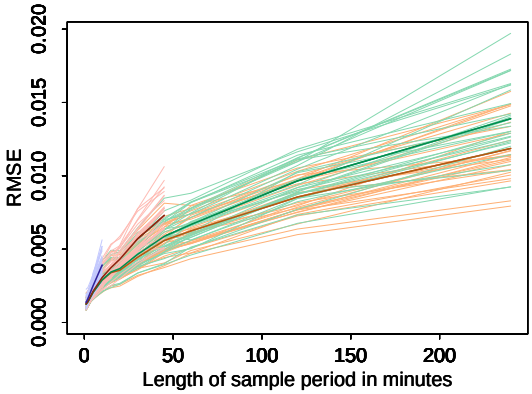
On the basis of the event log of XETRA for the first quarter of 2004, we have, nevertheless,

Figure 4.18: $RMSPE$

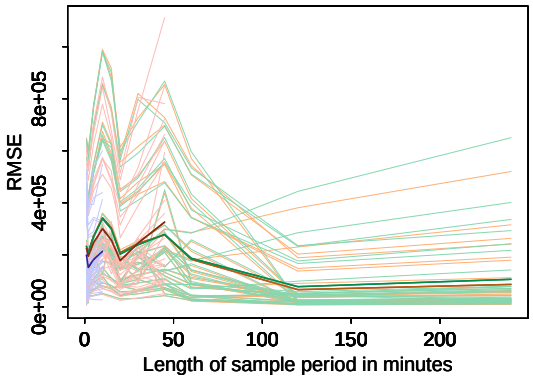
The figures below report the out-of-sample $RMSPE$ for the estimated model equations specified in Equation (4.35) (blue), Equation (4.36) (red), Equation (4.37) (green) and Equation (4.38) (orange) for the sampling frequencies 1, 2, 5, 10, 15, 20, 30, 45, 60, 120 and 240. The predictions are based on the out-of-sample one-step ahead rolling window forecast. The respective window lengths are listed in Table 4.3.



(a) Buy



(b) Sell



(c) XLM

shown that our model describes the LOB data well, both in- and out-of-sample. The data requirements are rather high as knowledge about price and quantity levels of incoming and canceled orders are required. This sort of data is usually not available. Even though returns may be predicted, market impact of actual trading strategies as well as order costs may hamper profitability of a trading strategy based on our model. Nevertheless, we are convinced that our empirical analysis provides the lower limits of forecast accuracy, as we have made several approximating decisions in the course of this chapter. In addition, for time horizons beyond 1 minute other variables may possibly help to predict returns or any other measure in the order book.

Appendix

4.A Distribution of Events

Table 4.A.1: Number of Events Related to Order Type and Market Side

The table lists the total number of events related to limit (#L), market (#M), iceberg (#I), and market-to-limit (#T) orders for each stock in our sample. In the columns %L, %M, %I, and %T the percentage of events occurring on the sell (S) and buy (B) market side is tabled for each order type. The last column (%Total) reports the share of all events for each market side.

Ticker	#L	#M	#I	#T		%L	%M	%I	%T	%Total
ADS	1,129,682	16,976	4,749	1,969	B	48.4	55.1	49.7	41.4	48.5
					S	51.6	44.9	50.3	58.6	51.5
ALT	1,094,414	16,785	9,552	1,412	B	46.4	50.1	60.6	39.9	46.6
					S	53.6	49.9	39.4	60.1	53.4
ALV	4,237,243	68,446	39,416	2,105	B	48.3	55.3	44.4	51.2	48.4
					S	51.7	44.7	55.6	48.8	51.6
BAS	2,585,776	31,450	36,082	1,885	B	49.5	50.7	58.7	45.3	49.7
					S	50.5	49.3	41.3	54.7	50.3
BAY	2,199,894	34,187	29,943	1,721	B	49.6	50.9	42.9	46.5	49.5
					S	50.4	49.1	57.1	53.5	50.5
BMW	2,087,167	28,557	34,625	1,707	B	48.6	60.5	47.0	45.0	48.7
					S	51.4	39.5	53.0	55.0	51.3
CBK	1,676,325	23,601	24,753	1,413	B	49.8	48.7	44.2	41.7	49.7
					S	50.2	51.3	55.8	58.3	50.3
CONT	1,130,309	15,866	9,641	1,419	B	48.5	49.2	51.0	43.6	48.6
					S	51.5	50.8	49.0	56.4	51.4
DB1	936,959	16,205	16,381	1,315	B	48.5	50.8	49.2	39.4	48.5
					S	51.5	49.2	50.8	60.6	51.5
DBK	3,339,752	48,314	47,215	2,278	B	49.3	46.2	50.9	45.8	49.3
					S	50.7	53.8	49.1	54.2	50.7

Ticker	#L	#M	#I	#T		%L	%M	%I	%T	%Total
DCX	2,711,327	44,301	57,847	2,003	B	50.1	41.8	56.0	49.5	50.1
					S	49.9	58.2	44.0	50.5	49.9
DPW	1,001,394	26,360	28,571	1,468	B	47.6	50.1	50.5	40.2	47.8
					S	52.4	49.9	49.5	59.8	52.2
DTE	2,349,138	87,942	49,129	3,581	B	49.7	51.8	48.1	39.0	49.7
					S	50.3	48.2	51.9	61.0	50.3
EOA	2,701,672	35,484	34,695	2,106	B	51.2	54.3	53.2	47.9	51.3
					S	48.8	45.7	46.8	52.1	48.7
FME	801,834	11,156	4,159	1,223	B	48.9	50.4	56.2	39.8	49.0
					S	51.1	49.6	43.8	60.2	51.0
HEN3	1,101,152	11,405	3,565	1,492	B	47.1	51.7	39.3	39.5	47.1
					S	52.9	48.3	60.7	60.5	52.9
HVM	1,482,520	29,616	41,955	1,392	B	50.4	54.1	50.9	41.6	50.5
					S	49.6	45.9	49.1	58.4	49.5
IFX	1,594,470	50,125	63,497	1,584	B	48.8	54.1	42.4	40.7	48.7
					S	51.2	45.9	57.6	59.3	51.3
LHA	1,169,415	23,026	28,737	1,570	B	48.8	48.6	50.3	42.2	48.8
					S	51.2	51.4	49.7	57.8	51.2
LIN	1,157,591	13,496	7,384	1,807	B	48.3	48.6	63.2	46.0	48.4
					S	51.7	51.4	36.8	54.0	51.6
MAN	1,023,998	14,952	15,252	1,697	B	47.3	49.0	60.2	39.1	47.5
					S	52.7	51.0	39.8	60.9	52.5
MEO	1,144,291	16,064	15,028	1,460	B	48.8	52.0	52.8	39.3	48.9
					S	51.2	48.0	47.2	60.7	51.1
MUV2	2,896,094	46,036	37,068	1,908	B	49.2	57.0	43.7	45.4	49.3
					S	50.8	43.0	56.3	54.6	50.7
RWE	2,061,625	31,746	35,398	2,014	B	51.6	44.6	53.3	47.5	51.5
					S	48.4	55.4	46.7	52.5	48.5
SAP	2,800,569	36,907	20,332	1,530	B	49.6	49.4	57.1	42.9	49.7
					S	50.4	50.6	42.9	57.1	50.3
SCH	1,312,153	24,385	17,908	1,439	B	48.3	50.3	51.6	41.6	48.4
					S	51.7	49.7	48.4	58.4	51.6
SIE	3,444,640	58,410	54,186	2,172	B	48.3	49.5	53.8	48.0	48.4
					S	51.7	50.5	46.2	52.0	51.6
TKA	1,130,506	23,019	19,060	1,797	B	48.5	52.8	45.2	41.1	48.5
					S	51.5	47.2	54.8	58.9	51.5
TUI	970,118	21,965	15,737	1,269	B	48.7	53.8	40.5	40.4	48.7
					S	51.3	46.2	59.5	59.6	51.3

Ticker	#L	#M	#I	#T		%L	%M	%I	%T	%Total
VOW	1,966,460	26,478	42,894	1,715	B	48.6	49.3	53.0	44.4	48.7
					S	51.4	50.7	47.0	55.6	51.3
TOTAL	55,238,488	933,260	844,759	52,451	B	49.1	51.2	50.2	43.5	49.1
					S	50.9	48.8	49.8	56.5	50.9

4.B Simulation Specification

In order to simulate the order book, several probabilities and other conventions have to be specified. Therefore, we go through the terms in Equation (4.34) and present how we have chosen to specify $\alpha_M(k,q)$ and $\omega_M(k,q)$. For convenience, recall Equation (4.34) as

$$\alpha_M(k,q) = \bar{r}_{0,M,i,j,a}\; p_{K,M}(k;\boldsymbol{\theta}_{M,a})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,a})$$
$$\omega_M(k.q) = \bar{r}_{0,M,i,L,c}\; p_{K,M}(k;\boldsymbol{\theta}_{M,c})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,a}).$$

Figure 4.B.1 illustrates the components of Equation (4.34).

Recall that we choose three theoretical scenarios for the distributions across price levels $p_{K,M}(\cdot)$: First, the uniform distribution (uni), second, a discrete log-normal distribution with fixed parameters (fix), and third, a discrete log-normal distribution with dynamic parameters where the parameters depend on the prevailing spread (dyn). For the distribution across order sizes, we only consider one theoretical specification: a power law distribution. Additionally, we also consider the unconditional empirical frequencies of incoming and canceled orders as observed in the first quarter of 2004, both across price and size levels.

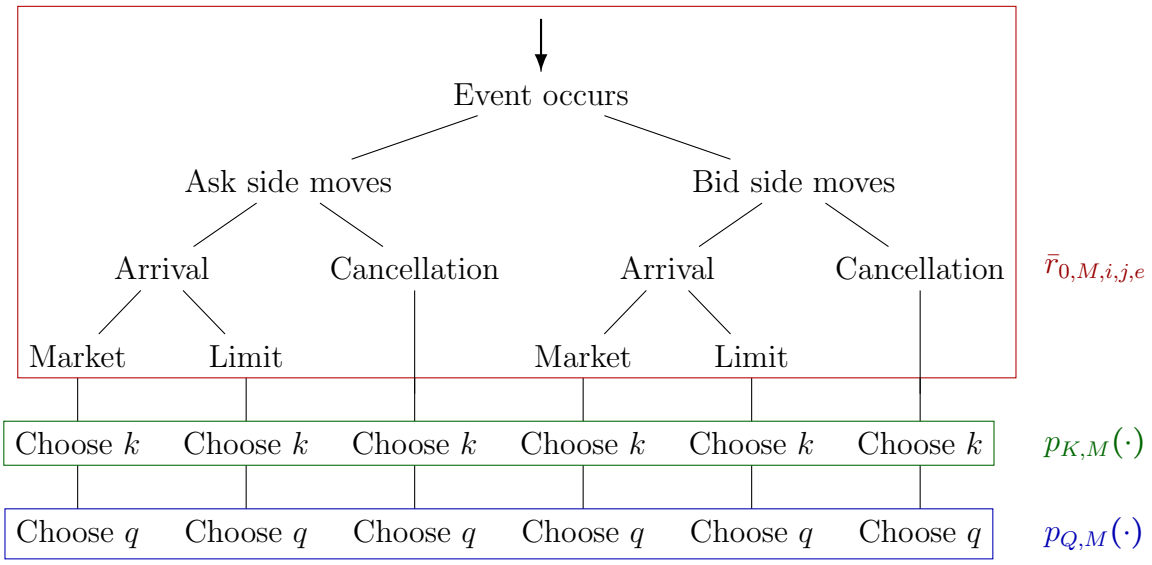
4.B.1 Rates of Order Types $\bar{r}_{0,M,i,j,e}$

The first element of Equation (4.34) is $\bar{r}_{0,M,i,j,e}$, the rate for an arrival ($e = a$) or a cancellation ($e = c$) of order type j on market side M for stock i. We first need to specify the order types that we include in the simulation. In Figure 4.2, we have depicted limit orders and market orders across relative integer distances to the best quote to show that there is a somewhat stable distribution across price levels when the best quote is used as a fix point. At the zero level, we have plotted the marketable orders split up into different types.

Table 4.B.2 shows the percentages of the different types of marketable orders in detail. In general, approximately 10% of all incoming orders (cancellations excluded) are marketable. In fact, about half of those marketable orders are arriving on the best quote i.e., with $d = 0$. Around a quarter is due to market orders with no limit price $d < -\infty$ and another

Figure 4.B.1: Simulation Event Tree

The Figure depicts a decision tree to visualize the components of Equation (4.34). Taking the subtree marked by the red box, each of the leafs that originate the red box has a different rate $\bar{r}_{0,M,i,j,e}$ where the subscript e refers either to a an order arrival or c an order cancellation. Also each of the nodes inside the blue and green box has a different $p_{K,M}(\cdot)$



quarter are marketable limit orders i.e., with $d < 0$. Marketable iceberg and stop order are tiny in comparison. The inverse of the last column are the limit orders that are submitted before the best quote. As depicted in Figure 4.B.1, in our simulation scenarios, we only treat market and limit orders separately. We do not distinguish iceberg and stop orders, since they are market and limit orders with some additional features. Thus, when we use the unconditional empirical frequencies for $\bar{r}_{0,M,i,j,e}$, we calculate

$$\bar{r}_{0,M,i,j,e} = \frac{n_{M,i,j,e}}{\Delta T},$$

where $n_{M,i,j,e}$ is the number of arrivals ($e = a$) or cancellations ($e = c$) of order type j on market side M for stock i observed during the entire first quarter of 2004. ΔT refers to the total trading time during this period. In our case, ΔT is specified to be 64 trading days. As we restrict the simulation to continuous trading, we only include events during the 8h28m of continuous trading to calculate the frequencies. In our sample, option settlement is conducted in three dates. On these three days, further 3 minutes have to be subtracted from the continuous trading phase. In total, we have $(64 \cdot (8 + 28/60) \cdot 60 - 3 \cdot 3) \cdot 60 = 1{,}950{,}180s$ of continuous trading time in our sample. Note that we can also decompose the unconditional empirical rates according to

$$\bar{r}_{0,M,i,j,e} = \frac{n_{\cdot,i,\cdot,\cdot}}{\Delta T} \cdot \frac{n_{M,i,\cdot,\cdot}}{n_{\cdot,i,\cdot,\cdot}} \cdot \frac{n_{M,i,j,\cdot}}{n_{M,i,\cdot,\cdot}} \cdot \frac{n_{M,i,j,e}}{n_{M,i,j,\cdot}}, \tag{4.40}$$

Table 4.B.1: Event Rates for Order Types

The table lists the order arrival and cancellation rates imposed in the scenarios 'dyn', 'fix' and 'uni'. The separation between marketable limit orders is only used for the 'uni' scenario. In the scenarios 'dyn' and 'fix', we only distinguish between market orders (incl. marketable limit orders) and limit orders. The rates have are given in the unit [orders/second].

Order Type		Market Side	Rate
Limit Order (non-marketable)	Arrival	ask	0.12
		bid	0.12
	Cancellation	ask	0.10
		bid	0.10
Limit Order (marketable)	Arrival	ask	0.0025
		bid	0.0025
Market Order	Arrival	ask	0.0025
		bid	0.0025

where n refers to a number of events and the indices specify which characteristic is relevant for counting. $n_{\cdot,i,\cdot,\cdot}$ means that only the index i (referring to the event concerning stock i) is relevant to determine the number of events. Categories marked with a $\cdot$ in the index are summed over. In other words, $n_{\cdot,i,\cdot,\cdot}$ denotes the number of events concerning stock i. In the empirical scenarios, all elements of Equation (4.40) can be observed. In theory, we can craft theoretical scenarios to investigate, ceteris paribus, the sensitivity of the LOB dynamics to changes in just one conditional frequency in Equation (4.40). In this chapter, we choose to focus on the sensitivity of the order book dynamics to changes in the distribution across price and quantity levels.

In the scenarios that entail a theoretical distribution, we do not use the empirical values observed in our sample. We also choose to focus on the distribution of arrival rates across price and size levels. Thus, we set the values summarized in Table 4.B.1. The rates are specified in the unit [orders/second]. They approximately mirror the observed values in reality, but we fix them to parity, so that the two sides of the market are symmetric and balanced.

One peculiarity in the theoretical scenarios 'fix' and 'dyn' is that we treat marketable limit orders below or above the best quote as market orders. Marketable limit orders on the best-quote, i.e., with $d = 0$, are modeled together with the rest of the limit orders as they approximately seem to fit into the discrete logarithmic distributions across price levels (cp. Figure 4.2). In the scenario 'uni', we separate the market orders and the marketable limit orders (strictly) below or above the best quote up to $d = -10$.

Table 4.B.2: Marketable Orders by Type

The table reports the share of marketable orders of all incoming orders in percentages across all stocks in the XETRA data. The column %L($d < 0$) shows the share of marketable limit orders behind the best quote, whereas the column %L($d = 0$) gives the share of all marketable limit orders directly at the best ask or bid. The column %M contains the percentages of market orders. %I tables the share of marketable iceberg orders and %T those of stop orders. The column %all is the total share of all marketable orders.

Ticker	**Buy/Sell**	**%L($d < 0$)**	**%L($d = 0$)**	**%M**	**%I**	**%T**	**%all**
ADS	S	5.40	1.73	1.50	0.14	0.02	8.79
	B	5.70	1.97	1.93	0.00	0.03	7.69
ALT	S	6.61	1.61	1.69	0.03	0.05	9.99
	B	7.54	2.06	1.76	0.00	0.06	9.67
ALV	S	5.09	2.59	1.99	0.02	0.05	9.74
	B	5.73	3.07	2.58	0.00	0.04	8.85
BAS	S	6.67	2.11	1.47	0.03	0.07	10.35
	B	6.63	2.15	1.46	0.00	0.10	8.88
BAY	S	7.15	1.93	1.90	0.01	0.10	11.08
	B	7.76	2.27	1.92	0.00	0.09	10.13
BMW	S	6.87	1.87	1.22	0.03	0.13	10.12
	B	6.97	2.10	2.00	0.00	0.12	9.19
CBK	S	6.10	1.34	1.81	0.01	0.11	9.37
	B	6.28	1.56	1.59	0.00	0.09	7.93
CONT	S	6.50	1.37	1.57	0.02	0.07	9.53
	B	6.75	1.53	1.35	0.00	0.06	8.34
DB1	S	7.14	1.93	1.98	0.03	0.13	11.21
	B	7.66	2.08	1.98	0.00	0.14	9.88
DBK	S	7.01	2.90	2.07	0.04	0.06	12.08
	B	7.37	2.95	1.65	0.00	0.08	10.40
DCX	S	7.90	2.49	2.50	0.02	0.12	13.02
	B	7.85	2.52	1.56	0.00	0.15	10.53
DPW	S	8.24	1.80	3.10	0.04	0.21	13.40
	B	9.56	1.85	3.22	0.00	0.25	11.67
DTE	S	12.56	3.20	5.02	0.21	0.13	21.12
	B	13.09	3.34	5.43	0.00	0.11	16.54
EOA	S	6.97	2.42	1.51	0.04	0.06	11.00
	B	6.65	2.37	1.69	0.00	0.06	9.09
FME	S	5.34	1.56	1.37	0.01	0.03	8.31
	B	5.58	1.72	1.22	0.00	0.04	7.34
HEN3	S	4.20	1.31	0.98	0.05	0.01	6.56
	B	4.44	1.68	1.03	0.00	0.01	6.13

Ticker	Buy/Sell	%L($d < 0$)	%L($d = 0$)	%M	%I	%T	%all
HVM	S	8.94	1.76	2.23	0.01	0.17	13.12
	B	9.32	1.93	2.59	0.00	0.17	11.41
IFX	S	11.40	2.29	3.91	0.05	0.37	18.01
	B	12.96	2.94	4.54	0.00	0.30	16.20
LHA	S	8.36	1.42	2.51	0.05	0.19	12.53
	B	8.67	1.50	2.28	0.00	0.20	10.37
LIN	S	5.59	1.19	1.34	0.08	0.03	8.23
	B	5.83	1.30	1.09	0.00	0.05	7.18
MAN	S	7.55	1.41	1.63	0.10	0.11	10.81
	B	8.06	1.53	1.50	0.00	0.12	9.71
MEO	S	7.75	2.01	1.35	0.04	0.07	11.22
	B	7.91	2.22	1.33	0.00	0.09	10.22
MUV2	S	6.41	2.98	1.72	0.02	0.07	11.20
	B	6.90	3.32	2.39	0.00	0.06	10.29
RWE	S	7.67	2.16	2.26	0.04	0.11	12.24
	B	7.17	2.04	1.45	0.00	0.12	9.33
SAP	S	5.57	2.53	1.67	0.01	0.04	9.83
	B	5.79	2.58	1.61	0.00	0.05	8.43
SCH	S	7.57	1.67	2.31	0.03	0.09	11.67
	B	8.32	1.99	2.32	0.00	0.11	10.41
SIE	S	7.31	2.61	2.22	0.03	0.08	12.25
	B	8.08	3.02	2.25	0.00	0.10	11.20
TKA	S	7.54	1.89	2.42	0.10	0.14	12.09
	B	8.12	1.64	2.68	0.00	0.13	9.89
TUI	S	6.82	1.50	2.62	0.01	0.16	11.12
	B	7.67	2.07	2.96	0.00	0.15	9.88
VOW	S	8.55	2.85	1.55	0.01	0.16	13.13
	B	9.09	3.16	1.43	0.00	0.19	12.44

4.B.2 Order Distribution Across Price Levels $p_{K,M}(\cdot)$

For the probability distribution of order arrivals across price levels specified in the factor $p_{K,M}(\cdot)$, we distinguish three theoretical scenarios and one scenario using unconditional empirical frequencies.

Uniform Distribution (uni)

The easiest approach to define the arrival rates across price levels is a uniform distribution. In this scenario, we assume that the arrivals of orders are concentrated on the first 90 integer price levels before the best quote of the opposite market. Additionally, marketable limit orders are also allowed to cross the best quote up to 10 price levels. In essence, this means that the arrivals of bid and ask orders are concentrated on 100 price levels around the best quote of the opposite market where the arrival rate on each price level is 0.0012 orders per second.

For the cancellations, we distribute the probability for an order cancellation uniformly among the occupied price levels.

Fixed Probability Distribution (fix)

Empirical frequencies of (non-marketable) limit orders across relative price levels exhibit pronounced probability mass at the tails of the distribution. For the distribution across price levels, in the scenario 'fix', we use a discrete Gaussian exponential distribution (DGX) as presented by Bi et al. (2001). The distribution is especially useful in cases where the random variable to be modeled is discrete and has pronounced probability mass at the tails. It is particularly interesting that the DGX reduces to the generalized Zipf distribution when $\mu \to -\infty$. Thus, it is flexible enough to incorporate situation where the probability distribution is a straight line in log-log-plots and cases in which it exhibits some curvature. A short summary of the DGX distribution is given in Section 4.C.

In the simulation scenario with a fixed probability distribution, we choose to set the values as outlined in Table 4.B.3. The values are the empirical mean and standard deviation across incoming orders of a random sample over several stocks. Note that the mean of arrivals is slightly higher on the bid side of the market, i.e., orders are more likely to arrive deeper in the book. Also the variance of order arrivals is higher. The same holds for cancellations. So while there are more arrivals deeper in the book, slightly more orders deep in the book are also canceled.

Table 4.B.3: Parameters of Probability Distribution Across k

Order Type		Market Side	μ	σ
Limit Order (non-marketable)	Arrival	ask	1.726301	0.674654
		bid	1.765909	0.711773
	Cancellation	ask	1.619866	0.620127
		bid	1.674366	0.650024

Dynamical Probability Distribution (dyn)

Similar to the case in which we use a DGX distribution with fixed parameters μ and σ, in the simulation scenario with a dynamical distribution across price levels, we also use the DGX distribution as the fundamental distribution. However, in this case we specify the parameters of the distribution to depend on the prevailing spread. The functional relationship we use is the following:

$$\mu(\Delta) = \log(100 \cdot \Delta) + \frac{1}{2},$$
$$\sigma(\Delta) = \sqrt{20 \cdot \log(100 \cdot \Delta)} + 1.2.$$

The functional relation is inspired by a scatter plot of $\hat{\mu}_i$ and $\hat{\sigma}_i$ estimated on the unconditional frequencies of order arrivals (and cancellations) across price levels against the average spread Δ_i for each stock i.[29] This scatter plot is depicted in Figure 4.11. Note also, that we have switched the scale of standard deviation and expectation as observed in the data on purpose. In that way, we hope to get an impression on how an increase in variance and a decrease of the mean may affect the characteristics of the order book evolution. The 'dyn' scenario is theoretically also motivated by the quest to study the sensitivity of the LOB system to feedback reactions between the state of the book and traders' order submission behavior.

Empirical Distribution (emp)

We also simulate one scenario where we take the empirical frequencies observed across price levels into account. The empirical log-frequencies are depicted in Figure 4.2.

[29] We are aware of the fact that the expectation of the functional relationship between DGX parameters and spread is not the same as a function for the log-likelihood in dependence of the expectation of the spread i.e., $\mathbb{E}_{t_0}[\mu(\Delta)] \neq \mu(\mathbb{E}_{t_0}[\Delta])$.

4.B.3 Order Distribution Across Size Levels $p_{Q,M}(\cdot)$

For the distribution across volume, we employ two different specifications. In one specification, we use a power law distribution. Even though, in our data at hand, we find that a power law does not match the volume distribution. This can be seen by sheer eyeballing of Figure 4.5. Nevertheless, the good fit of the power law distribution to describe order sizes has been shown in various articles (Bouchaud et al. 2002, Gopikrishnan, Plerou, Gabaix and Stanley 2000, Maslov and Mills 2001).

The probability mass function of a power law distribution where the smallest value of the support is 1, is theoretically defined as

$$p(x;\lambda) = (\lambda - 1)x^{\lambda-1} \forall x \in \mathbb{N}.$$

We fix the parameter in all simulations at $\lambda = 1.6$ which is close to empirically observed values.

According to Clauset, Shalizi and Newman (2009, Appendix D), given a random number $u \in [0,1]$, we can generate an integer realization $\tilde{x}$ from the power law distribution by calculating

$$\tilde{x} = \lfloor (1 - 0.5)(1 - u)^{-1/\lambda} + 1/2 \rfloor - 1,$$

where $\lfloor \cdot \rfloor$ signify the floor operator which cuts off the decimal places of the argument.

Note that since we assume independence of $p_{Q,M}(\cdot)$ and $p_{K,M}(\cdot)$, and each arriving order surely has to be assigned a size, we simply generate a third realization from a uniform distribution $\mathcal{U}(0,1)$ to determine the volume. In other words, before randomly generating what size is affected with u_3, we answer the question *when* something will happen with u_1 and, with u_2, *what* as well as *where* (i.e., at which limit price level) it will happen, as described in Section 4.3.

Empirical Distribution (emp)

We also simulate one scenario where we use the empirical frequencies observed across quantity levels to simulate the LOB evolution. The empirical log-frequencies are depicted in Figure 4.5.

4.B.4 The Fully Empirical Scenario (emp,emp)

For the case where both the distribution across price levels and the distribution across price levels are sampled from the empirically observed frequencies, we take the joint frequencies (not the product of the marginal frequencies) to sample both size and price of an incoming or canceled order.

4.C Discrete Gaussian Exponential Distribution (DGX)

In the simulation, as described in Section 4.B, we use the DGX for the simulation of order arrivals and cancellations across price levels.

The probability mass function of the distribution can be defined according to Bi et al. (2001) as

$$p(x = k; \mu, \sigma) = \frac{A(\mu, \sigma)}{k} \exp\left(-\frac{(\log(k) - \mu)^2}{2\sigma^2}\right), \qquad \forall k \in \mathbb{N},$$

where the normalizing constant $A(\mu, \sigma)$ is defined as

$$A(\mu, \sigma) = \left\{\sum_{k=1}^{\infty} \frac{1}{k} \exp\left(-\frac{(\log(k) - \mu)^2}{2\sigma^2}\right)\right\}^{-1}.$$

In Figure 4.2, we use a slightly modified version of the DGX by truncating the distribution at 1 to show how the DGX can be fit to the data. The truncated DGX can be derived from the truncated (log-)normal distribution for continuous values and has the following probability distribution function:

$$p(x = k; \mu, \sigma) = \frac{1}{1 - \phi(-\frac{\mu}{\sigma})} \frac{A_T(\mu, \sigma)}{k\sigma} \exp\left(-\frac{(\log(k) - \mu)^2}{2\sigma^2}\right), \qquad \forall k \in \mathbb{N}.$$

The normalization factor $A_T(\mu, \sigma)$ is similarly defined by

$$A_T(\mu, \sigma) = \frac{1}{\sigma\left(1 - \phi(-\frac{\mu}{\sigma})\right)} \left\{\sum_{k=1}^{\infty} \frac{1}{k} \exp\left(-\frac{(\log(k) - \mu)^2}{2\sigma^2}\right)\right\}^{-1}.$$

The parameters μ and σ can be estimated using a maximum likelihood specification as described in Bi et al. (2001).

4.D Taylor Series Expansion of Linear Models

In this appendix the Taylor series expansion that justifies the model specifications in Equations (4.35) to (4.38) is derived.

Starting point of the derivation is the decomposition of the event rate in Equation (4.34). First, for the sake of brevity, we introduce the intensity of events at the price level k with affected order size q given a prevailing spread Δ

$$r_{M,i,j,e}(k,q \mid \Delta) = \alpha_{M,i,j}(k,q \mid \Delta) + \omega_{M,i,j}(k,q \mid \Delta),$$

where the index M denotes the market side (either bid or ask), i indicates the instrument, j denotes the order type (limit or market) and e denotes the event type (arrival or cancellation). The right hand side follows when writing arrival α and cancellation rates ω separately.

As we have seen in Section 4.1.3, the Hamiltonian H is directly constructed from the event rates. As a direct consequence, the conditional probability to find the system in state $|z\rangle$ given that it has been in $|z_0\rangle$ at time t_0 can be expressed as described in Equation (4.14) or in more detail as

$$\begin{aligned} p(z,t|z_0,t_0) &= \langle z \mid \psi(t)\rangle = \langle z| U(t,t_0)\,|\psi(t_0)\rangle = \langle z| \exp\Big(\int_{t_0}^{t} H(\tau)d\tau\Big)\,|\psi(t_0)\rangle \\ &= \langle z| \sum_{w=1}^{\infty} \frac{\left(\int_{t_0}^{t} H(\tau)d\tau\right)^{w}}{w!}\,|\psi(t_0)\rangle\,. \end{aligned} \tag{4.41}$$

As one can see, by construction the conditional distribution is polynomial in the (time integral over) arrival and cancellation rates, and, depending on the succession of orders in $|z\rangle$ (see e.g. Equation (4.9)), further combinatorical factors have to be introduced (which include the factorial $w!$). Only regarding the terms up to order one the conditional probability can be written as

$$p(z,t|z_0,t_0) = \langle z \mid \psi(t_0)\rangle + \langle z| \int_{t_0}^{t} H(\tau)d\tau\,|\psi(t_0)\rangle\,.$$

This first order approximation may fit the conditional probability for short time horizons well, however, for longer time horizons the interactions between order arrivals and cancellations may become the more important factor.

Nevertheless, with this approximation, we may view the conditional expected value of some observable O given the state of the order book at time t_0 of stock i at some future time $t > t_0$ as

$$\begin{aligned}\mathbb{E}_{t_0}[O_{i,t}] &= \sum_z \langle z \mid O \mid \psi(t_0)\rangle + \sum_z \langle z| O \int_{t_0}^{t} H(\tau) d\tau \,|\psi(t_0)\rangle \\ &= O_{i,t_0} + \sum_z \langle z| O \int_{t_0}^{t} \sum_q \sum_k \alpha_{M,i,j}(k,q) E_{M,i,j} + \omega_{M,i,j}(k,q) C_{M,i,j} d\tau \,|\psi(t_0)\rangle\,,\end{aligned}$$

where E and C denote the order entry and cancellation operators laid out in Section 4.1.3 and O_{i,t_0} is the realization of the observable at time t_0.

We may generalize this notion for the conditional expected value to some function that depends on the intensity of events, and, thus, again on order size q, the price level k and additionally further variables that determine arrival and cancellation rates, e.g. the spread. Making the very crude assumption that the expected value of some observable is linear in the intensity of events, we could formulate the approximation as

$$\mathbb{E}_{t_0}[O_{i,t}] \approx \gamma_{0,i} + \gamma_{1,i}\, \mathbb{E}[r_{M,i,j,e}(k,q \mid \Delta)].$$

Decomposing $r_{M,i,j,e}(k,q \mid \Delta)$ as done in Equation (4.34) and additionally assuming that the average intensity $\bar{r}_{0,M,i,j,a}(\Delta)$ is some function of the prevailing spread yields

$$\begin{aligned}\mathbb{E}_{t_0}[O_{i,t}] &\approx \gamma_{0,i} + \gamma_{1,i}\, \mathbb{E}[(\alpha_{M,i,j}(k,q \mid \Delta) + \omega_{M,i,j}(k,q \mid \Delta))] \\ &\approx \gamma_{0,i} + \gamma_{1,i}\, \mathbb{E}[\bar{r}_{0,M,i,j,a}(\Delta)\; p_{K,M}(k;\boldsymbol{\theta}_{M,a})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,a}) \\ &\qquad + \bar{r}_{0,M,i,L,c}(\Delta)\; p_{K,M}(k;\boldsymbol{\theta}_{M,c})\; p_{Q,M}(q;\boldsymbol{\phi}_{M,c})]. \qquad (4.42)\end{aligned}$$

Now, expanding each term by a Taylor series expansion around the respective expected value we have the following expansions for $p_{Q,M}(q;\boldsymbol{\phi}_{M,e})$

$$\begin{aligned}p_{Q,M}(q;\boldsymbol{\phi}_{M,e}) \approx\; & p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e}) + dp_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})(q - \mathbb{E}[q]) \\ & + \frac{d^2 p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})}{2}(q - \mathbb{E}[q])^2 + \ldots.\end{aligned}$$

Taking expectations with respect to $p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})$ yields an approximation of the expected value in moments of order 2 and higher

$$\mathbb{E}[p_{Q,M}(q;\boldsymbol{\phi}_{M,e})] \approx p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e}) + \frac{d^2 p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})}{2}\,\mathrm{var}[q] + \ldots.$$

Using then, again, the first order Taylor approximation of $p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})$ around 0 reintroduces the first moment

$$\mathbb{E}[p_{Q,M}(q;\boldsymbol{\phi}_{M,e})] \approx p_{Q,M}(0;\boldsymbol{\phi}_{M,e}) + dp_{Q,M}(0;\boldsymbol{\phi}_{M,e})\,\mathbb{E}[q] + \frac{d^2 p_{Q,M}(\mathbb{E}[q];\boldsymbol{\phi}_{M,e})}{2}\,\mathrm{var}[q] + \dots.$$

Thus, we may write the expected value of $p_{Q,M}(q;\boldsymbol{\phi}_{M,e})$ as a linear function of the moments

$$\mathbb{E}[p_{Q,M}(q;\boldsymbol{\phi}_{M,e})] \approx \xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\,\mathbb{E}[q] + \xi_{M,i,j,e,2}\,\mathrm{var}[q] + \dots. \tag{4.43}$$

Changing variables in $p_{K,M}(k;\boldsymbol{\theta}_{M,e})$ by considering the relative distance d_l to the opposing best quote instead of the absolute price level k, the same can be done for $\mathbb{E}[p_{K,M}(k;\boldsymbol{\theta}_{M,e})]$

$$\mathbb{E}[p_{K,M}(k;\boldsymbol{\theta}_{M,e})] \approx \kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\,\mathbb{E}[d_l] + \kappa_{M,i,j,e,2}\,\mathrm{var}[d_l] + \dots. \tag{4.44}$$

Last but not least, we may model the expected value of the event specific intensity by its expected value shifted by a event specific factor $\rho_{0,M,i,j,e}$ and the expected value of an event unspecific function $f(\Delta)$ solely dependent on the spread

$$\mathbb{E}[\bar{r}_{0,M,i,j,e,t}(\Delta)] = \mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}]\rho_{0,M,i,j,e}\,\mathbb{E}[f(\Delta)]. \tag{4.45}$$

Approximating $\mathbb{E}[f(\Delta)]$ by an expansion in moments as above yields

$$\mathbb{E}[f(\Delta)] = \delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v]. \tag{4.46}$$

Reinserting the Taylor expansions in Equations (4.43), (4.44) and (4.46) up to order four together with Equation (4.45) in Equation (4.42) yields Equation (4.35)

$$\begin{aligned}
\mathbb{E}_{t_0}[O_{i,t}] = \gamma_{0,i} + \left(\delta_{i,0} + \delta_{i,1}\mathbb{E}_{t_0}[\Delta_t] \; + \sum_{v=2}^{4} \delta_{i,v}\mathbb{E}_{t_0}[(\Delta_t - \mu_{\Delta,t})^v]\right) \times \\
\sum_{M,j,e} \rho_{0,M,i,j,e}\mathbb{E}_{t_0}[\bar{r}_{0,M,i,j,e,t}] \times \\
\left(\kappa_{M,i,j,e,0} + \kappa_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e}[d_{l,t}] + \right. \\
\left. \sum_{v=2}^{4} \kappa_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e}[(d_{l,t} - \mu_{d_l,t})^v]\right) \times \\
\left(\xi_{M,i,j,e,0} + \xi_{M,i,j,e,1}\mathbb{E}_{t_0,M,i,j,e}[q_{t,M,j,e}] + \right. \\
\left. \sum_{v=2}^{4} \xi_{M,i,j,e,v}\mathbb{E}_{t_0,M,i,j,e}[(q_{t,M,j,e} - \mu_{q,t})^v]\right) + \varepsilon_i.
\end{aligned}$$

Chapter 5

Estimation of Transfer Entropy and Other Relative Entropy Measures Based on Smoothed Quantile Regressions[30]

Detecting dependencies between different variables is important across scientific fields. Regardless of whether one is interested in finding a causal dependence or simply wants to improve a prediction, identifying amidst a plenitude of variables the few that contain important information for either of these goals is essential in empircal science. There are well established tools across sciences to gauge associations between several measures. The possibilities to discover structures in complex, multivariate, and often dynamic systems are, however, usually limited to linear approximations. Haber and Unbehauen (1990) or Giannakis and Serpedin (2001) provide an overview of the conventional methods.

Beyond these, relative entropy based measures (such as mutual information (Shannon 1948) or transfer entropy (Schreiber 2000)) identify these relationships among the random variables, but are able to capture more general and in particular also non-linear functional dependencies between them. Estimation of mutual information (MI) or transfer entropy (TE), however, requires density estimation which is either computationally complex or, when a discretization scheme is applied, to some extent arbitrary. In this chapter, I suggest a new approach to calculate relative entropy measures based on quantile regression with the aim to avoid binning and still keep the computational requirement to a minimum. In addition, the resulting measures are statistically testable as I work out the asymptotic distributions in a Generalized Method of Moments (GMM) framework. It turns out that the computational complexity is relatively low and the calculation of relative entropy measures are extendable to higher dimensions.

Relative entropy is a measure of divergence between two distributions of random variables, say X and Y (cp. Cover and Thomas 2005, pp. 19f). It is, thus, a measure of the inefficiency in describing X with Y. More precisely, it measures the inefficiency of assuming 'Y is distributed like X', when in reality it is not. Formally, relative entropy is defined as

[30] The work on this chapter is part of the project 'Robust estimation of time-varying moments, mutual information, and transfer entropy by means of quantile regression based density forecasts' (DI2160/3-1) and was funded by the Deutsche Forschungsgemeinschaft (DFG).

the expected value of the logarithmic likelihood ratio, i.e., the Kullback-Leibler distance (KL-distance). Especially in the time series context the estimation of TE, which is also a relative entropy measure, can be a powerful tool. For time series, TE is interpreted as a model-free measure describing the information flow between stochastic processes (cp. **?**).

TE as well as causation entropy (also known as conditional transfer entropy; Sun and Bollt 2014) generalize the notion of Granger causality for linear dynamic systems in the sense of predictive causality such that knowledge about one random variable helps to predict another one. The measure has been used in a wide variety of subjects, for example in biomedical engineering (Lee, Nemati, Silva, Edwards, Butler and Malhotra 2012, Zheng, Pan, Li, Luo, Wang and Liu 2017), ecological modeling (Oh, Kim, Lim and Kim 2018), economics (Dimpfl and Peter 2013, Sandoval, Mullokandov and Kenett 2015, Dimpfl and Peter 2019), neuroscience (Dimitrov, Lazar and Victor 2011, Amblard and Michel 2011, Vicente, Wibral, Lindner and Pipa 2011), or thermodynamics (Prokopenko, Lizier and Price 2013). Still, care has to be taken since the underlying assumption of TE is that the structure of relationships between the variables are pairwise or dyadic and can be reflected by a directed graph (James, Barnett and Crutchfield 2016).[31] Smirnov (2013) highlights further limitations such as low temporal resolution of the samples at hand on which TE is calculated or hidden variables. Nonetheless, I deem TE to be among the most empirically relevant applications of relative entropy measures.

Still, even in the case where the data generating process of a time series and the theoretical distributions of all random variables involved are known, the derivation of a closed form solution for TE is cumbersome. Therefore, in this chapter, I conceive TE as a form of conditional mutual information (CMI). I start the analysis with MI and then gradually move to CMI. As set out in Section 5.1, conceptually, all measures are constructed as KL-distances. Therefore, once I am in the position to estimate conditional and joint distributions from a data sample, the calculation of the various relative entropy based measures and the derivation of their asymptotic distribution is quite similar.

The prevailing approach to estimate relative entropy measures is based on conditional frequencies which are calculated using the assumption that the underlying random variables are discrete. In situations when this requirement is not met, several authors (**?**Sandoval et al. 2015, Behrendt and Prange 2019) rely on a discretization of the continuous data based

[31] A directed graph is a graph in which the nodes only have pairwise edges and the edge may have a direction. A graph that reflects polyadic interactions between nodes is a graph in which nodes can have more than one connection (polyadic). Such a graph is called an annotated hypergraph. Both, directed graphs and annotated hypergraphs, can be illustrated by academic articles and their authorship as described by Zhou, Huang and Schölkopf (2007). Assume that each article is a node of a graph. If one forms a link between two articles if they have at least one (co)author in common, one gets a dyadic graph. The information by how much an author contributed to an article, which is usually reflected in the ordering of the authors, is lost in such a graph. Also the number of articles one author has written is lost in such a graph. An annotated hypergraph is a graph in which this information is preserved. Usually such graphs are hard to draw and represented as sets.

on quantiles of the empirical distribution. In the respective applications, financial returns are grouped into three bins based on the idea that tail events (either positive or negative) have a higher information content than small observations in the center of the distribution. This discretization comes at the cost that in particular for autoregressive models of higher order, the number of observations required to fill all possible combinations is very high. Furthermore, while the selection of the quantiles is motivated economically, the specific choice (e.g. 5% or 10%) is still ad hoc. Introducing a symbol for each observation based on the location of each quantile in the empirical distribution, i.e., based on the estimated quantiles does not yield a consistent estimator for TE of the underlying continuous distributions since the discretization scheme affects the shape of the distribution of symbols as described by Kaiser and Schreiber (2002). Also, making the discretization scheme ever finer bears another challenge mentioned by Kaiser and Schreiber (2002): A discretized (coarse-grained) CMI (or TE) does not converge monotonically towards the continuous counter part for a finer graining . This is not a problem for MI, only for CMI. A strategy to remedy the situation would be consistent estimates for each point of the involved continuous probability densities. Also the (asymptotic) distribution of these estimates should be known. Given such estimates, the variance of relative entropy measures built from these estimates can be worked out.

Estimating the required joint and conditional densities via quantile regressions can remedy the situation in this manner. This is what I propose. I use quantile regression to directly estimate the required joint and conditional density to estimate relative entropy measures such as MI and TE calculation. I also provide the asymptotic distribution of the estimates and facilitate hypothesis testing. Thereby, discretization of the data (and the problems associated with it) are avoided. Furthermore, quantile regression allows a much more flexible calculation of TE as the conditional models can be specified in a richer fashion as opposed to joint frequencies which become more complex the more conditioning variables there are. The approach is similar to Baur and Dimpfl (2018a) who use quantile regression to estimate moments of a time series. Similar to their study, I rely on quantile regression to obtain the conditional probability density functions.

This chapter proceeds as follows. In Section 5.1 I present the theoretical basis of the estimation strategy. I outline how to estimate relative entropy measures based on conditional probability density functions obtained through quantile regression. In Section 5.2 I evaluate the methodology by a simulation study and Section 5.3 presents two small empirical examples. Section 5.4 concludes.

Table 5.1: List of Relative Entropy Measures with KL-Representation

The table lists the relative entropy measures discussed in this chapter. The concept of entropy, mutual information and conditional information go back to Shannon (1948) while their application to time series data can be referenced to Schreiber (2000).

Name		**Associated KL-Divergence**
Entropy	$h(X)$	$= D_{\mathrm{KL}}\left(\mathcal{U}(x) \| f_X(x)\right)$
Mutual Information	$I(X,Y,Z)$	$= D_{\mathrm{KL}}\left(f_{XYZ}(x,y,z) \| f_X(x) f_Y(y) f_Z(z)\right)$
Cond. Mutual Information	$I(X,Y\|Z)$	$= \int_{\mathbb{R}} D_{\mathrm{KL}}\left(f_{XY\|Z}(x,y \mid z) \| f_{X\|Z}(x \mid z) f_{Y\|Z}(y \mid z)\right) dz$
Lagged Mutual Information	$I(X_t, Y_{t-p})$	$= D_{\mathrm{KL}}\left(f_{X_t Y_{t-p}}(x_t, y_{t-p}) \| f_{X_t}(x_t) f_{Y_{t-p}}(y_{t-p})\right)$
Transfer Entropy	$T_{X\to Y}$	$= I(Y_t, X_{t-1:t-L} \| Y_{t-1:t-M})$

5.1 Method

In order to facilitate the analysis and to have a clear theoretical basis for the later simulation study, I present the estimation method first for MI and subsequently extend the analysis to TE as a special case of CMI. Relative entropy measures are all defined through different Kullback-Leibler distances. The Kullback-Leibler distance for two random variables X and Y is defined as

$$D_{\mathrm{KL}}\left(f_X \| f_Y\right) = \int_{\mathbb{R}} f_X(u) \log\left(\frac{f_X(u)}{f_Y(u)}\right) du$$

where $f_{\cdot}$ denotes the respective density function. The KL distance can equivalently be defined for discrete probability distributions by interchanging integration with summation. Actually, the Kullback-Leibler distance is not a distance measure between probability distributions, but merely a divergence measure. The KL-distance is not symmetric $D_{\mathrm{KL}}\left(f_X \| f_Y\right) \neq D_{\mathrm{KL}}\left(f_Y \| f_X\right)$ and it does not satisfy the triangle inequality. Table 5.1 lists the relative entropy measures referred to in this chapter in their Kullback-Leibler distance form.

In order to estimate these relative entropy measures, one needs estimates for the joint probability density function, possibly via conditional density functions, as well as the marginal densities.

To begin with, consider three random variables X, Y, and Z. Mutual information (cp. Shannon 1948) is then given as

$$
\begin{aligned}
I(X,Y,Z) &= \iiint_{\mathbb{R}^3} f_{X,Y,Z}(x,y,z) \log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right) dzdxdy \qquad (5.1)\\
&= \iiint_{\mathbb{R}^3} f_{X|Y,Z}(x \mid y,z) f_{Y|Z}(y \mid z) f_Z(z) \log\left(\frac{f_{X|Y,Z}(x \mid y,z) f_{Y|Z}(y \mid z)}{f_X(x)f_Y(y)}\right) dzdxdy.
\end{aligned}
$$

The estimation strategy is straightforward. The joint probability density function $f_{X,Y,Z}$ is never estimated. Instead, I partition the joint density into conditional density functions for which the corresponding quantile functions $Q_{X|Y,Z}(\boldsymbol{\tau} \mid \cdot)$ and $Q_{Y|Z}(\boldsymbol{\tau} \mid \cdot)$ can be estimated via quantile regression. $\boldsymbol{\tau}$ is the vector of desired probability levels. Furthermore, I use the representativity of the sample to estimate the integral in Equation (5.1) via a sample mean. Note that the integral can be read as the expected value of the logarithmic likelihood ratio. Thus, for the calculation of MI of K variables, I need to run $K-1$ quantile regressions to estimate the conditional densities in the numerator of the fraction inside the logarithmic term. Additionally, I need $K-1$ estimates for the unconditional densities in the denominator. These can be calculated using quantile regressions on a constant. All in all, I need to run $2K-2$ quantile regressions. In the case of CMI this number is reduced to 2 quantile regressions, one for the numerator and one for the denominator of the log-likelihood ratio. Once the quantile functions $Q_{\cdot}(\boldsymbol{\tau} \mid \cdot)$ are estimated, I use a locally weighted polynomial regression to fit the estimated conditional quantiles to the corresponding probability levels $\boldsymbol{\tau}$ at each point in the sample. The parameter estimates of the locally weighted polynomial regression are associated with the derivatives of the conditional distribution functions, i.e., the conditional probability densities. Equipped with estimates for the (conditional) probability densities, a sample mean estimator of MI can be constructed.

In this setting, MI is a function of the parameter estimates of several quantile regressions on the same data set. In order to facilitate statistical tests on the resulting measure of MI, an asymptotic theory for the parameter estimates' joint distribution is needed. This can be obtained by casting the quantile regression estimation problem into the GMM framework without actually using it to estimate the parameters.

To outline the details, the present section is structured as follows. First, in Section 5.1.1, I discuss the estimation of conditional and joint distributions using quantile regression. Second, in Section 5.1.1, I present the GMM framework that allows me to obtain a joint asymptotic theory for parameter estimates from several quantile regressions on the same set of data. Third, Section 5.1.1 explains the smoothing and actual estimation of densities from the fitted quantile functions obtained through quantile regression. The final calculation of MI and the resulting asymptotic theory is presented in Section 5.1.2.

5.1.1 Density estimation via Quantile Regression

Quantile Regression via General Method of Moments

The heart of quantile regression as introduced by Koenker and Bassett (1978) is the specification of the conditional quantile function $\boldsymbol{Q_y}(\tau_j|\boldsymbol{X}) = \boldsymbol{X'\theta}(\tau_j)$ where I denote with τ_j the conditional probability $\tau_j = P\left(\boldsymbol{Q_y}(\tau_j|\boldsymbol{X}) \leq Y|\boldsymbol{X}\right)$. In the quantile regression setup each conditional quantile τ_j is modeled as a linear combination governed by the parameters $\boldsymbol{\theta}$ of some set of regressors $\boldsymbol{X}$. I use the subscript j in order to indicate that τ_j is the j^{th} entry in the vector $\boldsymbol{\tau}$ in which I collect all probabilities for which the quantile regression shall be estimated. In order to estimate the parameters $\boldsymbol{\theta}_j$, the following minimization problem has to be solved:

$$F_j = \min_{\boldsymbol{\theta}\in\mathbb{R}} \sum_{i=1}^{N} \rho_{\tau_j}(y_i - \boldsymbol{x}_i'\boldsymbol{\theta}_j), \tag{5.2}$$

where $\rho_{\tau_j}(u) = u\left(\tau_j - \mathbb{1}(u < 0)\right)$ with $\mathbb{1}(\cdot)$ as the indicator function. $\rho_{\tau_j}(u)$ is called the check function and has a discontinuity at $u_0 = 0$. The discontinuity is due to the indicator function. The ultimate goal is to derive a joint asymptotic distribution for the parameter estimates of several quantile regressions on the same data sample. This would enable statistical tests based on functions of these parameter estimates. The parameter estimates obtained from separate quantile regressions on the same sample may be related. To account for these possible relations in the asymptotic joint distribution of the estimates, I translate each quantile regression estimation into a GMM estimation problem. While the actual parameter values are not estimated via GMM – the standard procedures and algorithms for their estimation established in the literature (see Koenker (2005) for a detailed discussion) are better suited for this task – I only base the estimation of the parameter estimates' joint variance-covariance matrix on the GMM theory. In order to do so, however, continuous moment conditions are required (cp. Hansen 1982, Assumption 2.3).

Similar to Engle and Manganelli (2004), I suggest to substitute the indicator function used in the check function in Equation (5.2) by the sigmoid function $\mathbf{1}(t) = \frac{1}{1+e^{Gt}}$. These functions are asymptotically equivalent as G goes to infinity, i.e.,

$$\lim_{G\to\infty} \frac{1}{1+e^{Gt}} = \mathbb{1}(t).$$

$\mathbf{1}$ is thus a continuous counterpart of $\mathbb{1}$ which closely mimics its behavior when G is sufficiently high. The derivatives of $\mathbf{1}$ are also well defined. As Engle and Manganelli (2004) note, already for $G = 10$, $\mathbb{1}(t)$ and $\mathbf{1}(t)$ are quite similar.[32] They also note that G

[32] Note that in this form the indicator function is nothing else than the Heaviside-Step-Function in u. The derivative of the Heaviside function is the Dirac Delta function and corresponds to the indicator

may be estimated as a parameter in the minimization problem. In their work, however, they fix it to $G = 10$. In the applications later on, I set $G = 100$. With this change, the derivatives of the indicator function are well defined and one can write the first order conditions of Equation (5.2) as

$$\frac{\partial F_j}{\partial \theta_k} = \sum_{i=1}^{N} x_{ik} \left(\tau_j - \frac{1}{1 + e^{Gu_i}} \right) - x_{ik} u_i G \frac{e^{Gu_i}}{(1 + e^{Gu_i})^2} \stackrel{!}{=} 0 \tag{5.3}$$

or, in vector notation,

$$\underset{K \times N}{\frac{\partial F_j}{\partial \boldsymbol{\theta}}} = \underset{K \times N}{\boldsymbol{X}'} \left(\tau_j - \underset{N \times 1}{\boldsymbol{v}} \right) - \boldsymbol{X}' \underset{N \times N}{\operatorname{diag}(\boldsymbol{w})} \underset{N \times 1}{\boldsymbol{u}} \stackrel{!}{=} \mathbf{0}, \tag{5.4}$$

where $\boldsymbol{v} = (1 + e^{G\boldsymbol{u}})^{-1}$ and $\boldsymbol{w} = G e^{G\boldsymbol{u}} (1 + e^{G\boldsymbol{u}})^{-2}$, and the exponential function $e^{(\cdot)}$ denotes elementwise operations on the vector $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_j$ rather than matrix exponentials.

Multiplying each line in the system of K equations given by Equation (5.4) by $\frac{1}{N}$ defines the requirements for the K empirical moment condition estimates. The theoretical lth moment function can be written as

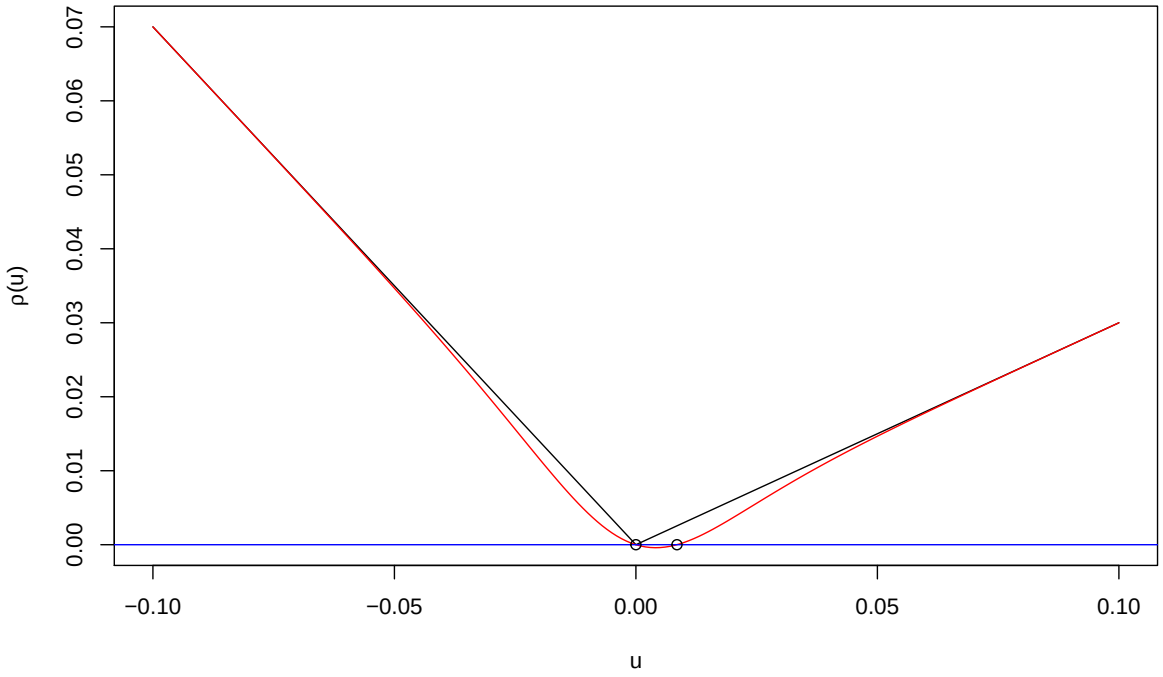
$$g_l(\tau_j, \boldsymbol{\theta}_j) = \mathbb{E}[X_l \left(\tau_j - \mathbf{1}(U < 0) \right) - X_l U \, \mathbf{1}(U = 0)].$$

The first order conditions of the quantile regression minimization problem put exactly satisfiable requirements on the empirical moment conditions. Note that formulating the problem in this manner violates one assumption for a consistent GMM estimator, namely that of global identification. In order to obtain a globally identified and consistent estimator, the moment conditions need to have a single root at the true parameter estimate (cp. Assumption 2.4 in Hansen 1982). When substituting the indicator function with the sigmoid function, this requirement does not hold even for large G. Figure 5.1 illustrates that the continuous counterpart of the check function has two roots, one at $u_1 = 0$ and another one at some $u_2 > 0$ (the exact location of u_2 is not relevant). As $G \to \infty$, the two roots get closer and eventually merge to one in the limit. However, in finite samples, the parameter estimates are in this setting not globally identified. For the application, this is not necessary as I am only interested in the limiting distribution of $\hat{\boldsymbol{\theta}}$. Instead, the quantile regression approach of Koenker and Bassett (1978) can be used to identify consistent estimates and use the GMM inference framework only to derive the joint limiting distribution.

function $\mathbb{1}(u = 0)$. Engle and Manganelli (2004) simply use a continuous approximation to the Heaviside-Step-function in the context of quantile estimation.

Figure 5.1: Check Function vs. Sigmoid Function

The black line plots the check function $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$ for $\tau = 0.3$. The red line plots the continuous counterpart of the check function, in which the indicator function $\mathbb{1}(u < 0)$ is substituted with the sigmoid function $\mathbf{1}(u) = \frac{1}{1+e^{Gu}}$ with $G = 100$. As one can see, the continuous function has two roots (black dots) one at $u = 0$ and another positive root, whereas the check function has only one root at $u = 0$.



Thus, the corresponding empirical means over the realizations of the moment condition functions can be written as

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} g_{1i}(\tau_j, \hat{\boldsymbol{\theta}}_j) &\stackrel{!}{=} \mathbf{0} \\ &\vdots \\ \frac{1}{N}\sum_{i=1}^{N} g_{Ki}(\tau_j, \hat{\boldsymbol{\theta}}_j) &\stackrel{!}{=} \mathbf{0}. \end{aligned}$$

Stacking the empirical moment conditions for each quantile τ_j, for which the parameters $\boldsymbol{\theta}_j$ shall be estimated, together in one large vector of moment conditions, the joint distribution of all estimated parameters $\hat{\boldsymbol{\theta}}$ for all $\tau_j \in \boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_Q)'$ can be derived to be, following standard asymptotic theory,

$$\sqrt{N}(\text{vec}(\hat{\boldsymbol{\theta}}) - \text{vec}(\boldsymbol{\theta})) \sim \mathcal{N}\left(0, \text{Avar}(\hat{\boldsymbol{\theta}})\right).$$

The estimator of the asymptotic variance $\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}})$, as derived by Hansen (1982), reads as follows:

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{D}}'\boldsymbol{W}\hat{\boldsymbol{D}})^{-1}\,\hat{\boldsymbol{D}}'\boldsymbol{W}\;\hat{\boldsymbol{S}}\;\boldsymbol{W}\hat{\boldsymbol{D}}\,(\hat{\boldsymbol{D}}'\boldsymbol{W}\hat{\boldsymbol{D}})^{-1},$$

where $\hat{S}$ is the asymptotic variance of the moment condition functions $\boldsymbol{g_N}(\boldsymbol{\tau},\boldsymbol{\theta})$ which can be estimated from the observed deviations from the theoretical moment conditions $\boldsymbol{g}_i(\boldsymbol{\tau},\boldsymbol{\theta})$. $\hat{\boldsymbol{D}}$ is the sample estimate of the derivative of the moment conditions with respect to the vector $\boldsymbol{\theta}_j$. The element d_{lk} in the l^{th} row and k^{th} column of $\boldsymbol{D}$ is the derivative of the l^{th} moment condition with respect to the k^{th} parameter in $\boldsymbol{\theta}_j$ and follows directly from the system of equations in (5.4). d_{lk} can be estimated by

$$d_{lk} = \frac{1}{N}\sum_{i=1}^{N} 2x_{ik}x_{il}u_iG^2\frac{e^{2Gu_i}}{(1+e^{Gu_i})^3} - x_{ik}x_{il}u_iG\frac{e^{Gu_i}}{(1+e^{Gu_i})^2}.$$

$\boldsymbol{W}$ is the weighting matrix. Choosing $\boldsymbol{S}^{-1}$ as the weighting matrix gives the variance-covariance estimate for the most efficient parameter estimates. I might as well choose $\boldsymbol{W}$ as the identity matrix $\boldsymbol{I}_{KQ(2K-1)}$ to calculate $\hat{\boldsymbol{S}}$. For $N \to \infty$ the two estimates are both asymptotically consistent as shown by Hansen (1982). In the concrete application the variance-covariance matrix of the moment-conditions $\boldsymbol{S}$ may be singular. This may be due to a situation where a high number of quantiles are estimated and the moment conditions are (almost) perfectly correlated. In such a situation, I use the generalized Moore-Penrose-inverse $\boldsymbol{S}^{+}$ as a weighting matrix. The variance estimate for $\hat{\boldsymbol{\theta}}$, in this case is [33]

$$\text{var}(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{D}}'\boldsymbol{S}^{+}\hat{\boldsymbol{D}})^{+}.$$

Smoothing of Estimated Quantiles and Density Estimation

For the calculation of the relative entropy based measures, one needs estimates of the density function. One can obtain such estimates from the finite first difference of the fitted values of the quantile function estimated via quantile regression as

$$\hat{f}_{Y|X}(y|x) \approx \frac{\tau_{j+1}-\tau_j}{\boldsymbol{Q_y}(\tau_{j+1}|\boldsymbol{X}) - \boldsymbol{Q_y}(\tau_j|\boldsymbol{X})}. \tag{5.5}$$

However, a common problem when estimating the quantile function via quantile regression is the crossing of the estimated quantiles. The latter would lead to negative estimates

[33] The extension of Carrasco and Florens (2000) for an overidentified GMM setup with a continuum of moment conditions for only one parameter is in the situation here not applicable. However, it may be extended to a continuum of parameters as suggested in their conclusion. The idea of Carrasco and Florens (2000) is based on the continuous equivalent of the Moore-Penrose pseudo inverse. I have tested the possibility of such an integral operator, however, found that the pseudo inverse $\boldsymbol{S}^{+}$ is not as computationally demanding as its continuous counterpart. The idea of Carrasco and Florens (2000) is nonetheless remarkable and in the opinion worth further research.

of the values of the density function at some points. Chernozhukov, Fernandez-Val and Galichon (2007) suggests to rearrange the quantile estimates in order to get a strictly increasing series of quantiles to circumvent the problem. A further problem is that the density estimate is always restricted to the sample range or, more precisely, to the estimated quantile range. For values outside this range no estimates can be found. To solve these two issues, I propose to smooth the fitted values from the quantile regression and adding ghost points at the tails of the distribution. This approach allows me to circumvent the problem of crossing quantiles and to extrapolate to points outside the sample range.

Therefore, the second step in the approach to estimate MI is to smooth the conditional distribution function described by the estimated quantiles. This can be done with a locally weighted polynomial regression (cp. Stone 1977, 1980, 1982, Cleveland 1979, Fan and Gijbels 1996) of the probabilities τ_j on the estimated quantiles $\boldsymbol{Q_y}(\tau_j|\boldsymbol{X})$. As the local fit of the smoothing can be heavily impacted by outliers, Cleveland (1979) suggests a locally estimated scatterplot smoothing (LOESS-regression) to make the local fit robust against outliers, while Cleveland and Grosse (1991) propose an efficient algorithm to calculate fitted values and parameters. For local polynomial regressions in general, Fan and Marron (1994) and Seifert, Brockmann, Engel and Gasser (1994) propose an even faster algorithm. In this case, the fitted values from the quantile regression are (usually) not very volatile and are already designed to lie around the curve of the conditional distribution function. So, in order to derive a clear asymptotic distribution and since I deem the problem of outliers not relevant in the application, the iterative procedure suggested by Cleveland (1979) is not used. Nonetheless, I have explored the method of Cleveland (1979) and have attempted different orders of polynomials (selected locally), in combination with various kernel functions together with the suggested relative nearest-neighbor bandwidth. I do not reproduce all the exploratory results in this chapter, as the outcomes of the attempts mirrored the well documented results of Fan and Gijbels (1996) which offer clear guidance on the choice of these parameters.

As argued by Fan and Gijbels (1996), the choice of the kernel function and the local selection of the order of the polynomial approximation is not as important as the locally optimal choice of the bandwidth. They show that the Epanechnikov kernel function minimizes the MSE at interior points. Thus, I choose this kernel function. The local selection of the order of the local polynomial regression, described by Fan and Gijbels (1996) introduces heavy computational requisites with no clear added value[34]. Ruppert and Wand (1994) find that if the polynomial order exceeds the order of the derivative by

[34] The selection of the optimal order of the polynomial approximation is based on an estimate of the conditional mean squared error. This is in itself computational expensive. Furthermore, in this context this estimate is potentially biased (see the discussion below) which may lead in small samples locally to strange order selections. Choosing globally a local polynomial approximation of order $p = 4$, yields in this context stable results.

an odd number, the bias is smaller compared to an even number. Therefore, since I want to use a first order approximation, I opt for a global polynomial order of $p = 4$. However, due to its importance, I endure the computational costs of selecting a local, approximately optimal, bandwidth h. I discuss the calculation in Section 5.1.1.

In summary the estimation approach consists in a first step of estimating Q quantiles via quantile regression using the procedure introduced by Koenker and Bassett (1978) on the data sample. In the second step, I estimate the density by smoothing the distribution function implied by the quantile estimates from the first step. For this purpose, I use a locally weighted polynomial regression of order $p = 3$ with an Epanechnikov kernel function with a dynamically chosen bandwidth parameter. The parameter estimate of the linear term in this local polynomial regression serves as an estimate for the conditional density.

More explicitly, to estimate the smooth density function from the quantile regression and in order to calculate the value of the (conditional) distribution function at $P = (\hat{\boldsymbol{\theta}}\mathbf{x}_0, y_0)$, I construct the local polynomial regression estimator as

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}_P' \boldsymbol{W}_P \boldsymbol{Z}_P)^{-1} \boldsymbol{Z}_P' \boldsymbol{W}_P \boldsymbol{\tau}, \tag{5.6}$$

where $\underset{Q\times 3}{\boldsymbol{Z}_P} = \left(\boldsymbol{\iota}, \hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0, (\hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0)^2\right)$ with $\boldsymbol{\iota}$ as the column vector of ones and $\boldsymbol{W}_P = \text{diag}\left(h^{-1}K(\frac{\hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0}{h})\right)$ where $K(\cdot)$ denotes the weight function and h is the chosen bandwidth. Note that the matrix $\boldsymbol{Z}_P'\boldsymbol{W}_P\boldsymbol{Z}_P$ may not be invertible if the parameter h is small and the kernel function selects less than 4 observations at each point to be included in the estimation. Thus, algorithmically, one may require that h must be widened in some situations in such a way that the matrix $(\boldsymbol{Z}_P'\boldsymbol{W}_P\boldsymbol{Z}_P)$ is of full rank. As specified in Equation (5.6) and recommended by Fan and Gijbels (1996) as well as by Ruppert and Wand (1994), I use a fourth order (locally weighted) polynomial regression to estimate the first derivative of the distribution function to reduce the bias of this estimate. While the first entry of the vector $\hat{\boldsymbol{\gamma}}$ gives the fitted value for the distribution function $\hat{\gamma}_0$, the second entry $\hat{\gamma}_1$ is a valid estimate for the density function at y_0. $2!\hat{\gamma}_2$ can provide an estimate for the derivative of the density function and $3!\hat{\gamma}_3$ for the third derivative.

Furthermore, in this chapter, the goal is to smooth a sufficiently high number of quantiles in order to arrive at a density estimate γ_1 even beyond the highest and lowest estimated quantile. However, using local polynomial smoothing outside of the support covered by the estimated quantiles produces heavy estimation errors. The solution, I propose, is to add H ghost points both at the lower end of the estimated quantile function with the coordinates $\Xi_{\underline{\omega},i} = (\min(\hat{\boldsymbol{\theta}}\boldsymbol{x}_0) - \omega i \Delta_q, 0)$ and again H ghost points at the upper end at $\Xi_{\bar{\omega},i} = (\max(\hat{\boldsymbol{\theta}}\boldsymbol{x}_0) + \omega i \Delta_q, 1)$, where ω denotes some chosen share of the range between the highest and the lowest estimated quantile $\Delta_q = \max(\hat{\boldsymbol{\theta}}\boldsymbol{x}_0) - \min(\hat{\boldsymbol{\theta}}\boldsymbol{x}_0)$ and $i \in \{1, \ldots, H\}$. I discuss the consequences of adding these extra points in Section 5.1.1.

Note, as I approximate the distribution function with a finite polynomial, $\hat{\boldsymbol{\gamma}}$ is entirely a function of the $\hat{\boldsymbol{\theta}}$. However, even in the case where the parameter estimates for $\hat{\boldsymbol{\theta}}$ would fall together with their true values $\boldsymbol{\theta}$, due to the finite polynomial approximation, a deterministic smoothing error still be observed.

To see this, consider the framework of local polynomial regression (cp. Fan and Gijbels 1996, Chapter 3) in which we have an unknown continuous function which can only be measured disturbed by some noise. In the case, the continuous function is the probability distribution function $F_{Y|X}(y \mid \boldsymbol{X} = \boldsymbol{x})$. Since we are now talking about concrete realizations or values of $\boldsymbol{X}$ that condition the distribution of Y, I specify these with the vector $\boldsymbol{x}$. The probability function is effectively defined by the uncountably infinite set of ordered pairs $F_{Y|X} = \{(\boldsymbol{x}, y, \tau) \in \mathbb{R}^k \times \mathbb{R} \times [0,1] \mid \tau = F_{Y|X}(y \mid \boldsymbol{X} = \boldsymbol{x})]\}$ where k is the dimension of the support of $\boldsymbol{X}$. With quantile regression estimates – given the linear structure of the conditional quantile function holds – the set of ordered pairs can be can be written as

$$F_{Y|X} = \{(\boldsymbol{x}, y, \tau) \in \mathbb{R}^k \times \mathbb{R} \times [0,1] \mid y_{[\tau]} = Q_y(\tau|\boldsymbol{X} = \boldsymbol{x})\}.$$

Recall that in the quantile regression set up, it is assumed that

$$y_{[\tau]} = Q_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) = \boldsymbol{\theta}(\tau)\boldsymbol{x}_0.$$

Estimating the parameters $\boldsymbol{\theta}(\tau)$ in a finite sample introduces an estimation error ε_τ. Thus, one can write

$$\begin{aligned} y_{[\tau]} &= \hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) + \varepsilon_\tau \\ &= \hat{\boldsymbol{\theta}}(\tau)\boldsymbol{x}_0 + \left(\boldsymbol{\theta}(\tau) - \hat{\boldsymbol{\theta}}(\tau)\right)\boldsymbol{x}_0. \end{aligned} \tag{5.7}$$

Using a Taylor expansion of $F_{Y|X}(y_{[\tau]} \mid \boldsymbol{X} = \boldsymbol{x}_0)$, I get

$$\tau = F_{Y|X}(y_{[\tau]} \mid \boldsymbol{X} = \boldsymbol{x}_0) = \sum_{k=1}^{\infty} \frac{1}{k!} F^{(k)}(y_0)\left(y_{[\tau]} - y_0\right)^k., \tag{5.8}$$

where $F^{(k)}(x_0) = \left.\frac{d^{(k)}F_{Y|X}}{dy^k}\right|_{y=y_0}$. Substituting y in Equation (5.8) with the estimated quantile function Equation (5.7) and extending X to be a real-valued multivariate random variable with $\boldsymbol{x}_0$ as one specific point in its support yields

$$\begin{aligned} \tau &= \sum_{k=0}^{\infty} \frac{1}{k!} F^{(k)}(y_0)\left(\hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) + \varepsilon_\tau - y_0\right)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} F^{(k)}(y_0) \sum_{j=0}^{k} \binom{k}{j} \left(\hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) - y_0\right)^j \varepsilon_\tau^{k-j}. \end{aligned}$$

Thus, for example for a third order approximation, I have

$$\begin{aligned}\tau = F^{(0)}(y_0) + b_{0\tau} \\ &+ \left(F^{(1)}(y_0) + b_{1\tau}\right)\left(\hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) - y_0\right) \\ &+ \left(\frac{F^{(2)}(y_0)}{2} + b_{2\tau}\right)\left(\hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) - y_0\right)^2 \\ &+ \left(\frac{F^{(3)}(y_0)}{6} + b_{3\tau}\right)\left(\hat{Q}_y(\tau|\boldsymbol{X} = \boldsymbol{x}_0) - y_0\right)^3 + \eta_\tau, \end{aligned} \tag{5.9}$$

where η_τ is the approximation error while the $b_{k\tau}$ are the unobserved biases and are polynomials of ε_τ and defined by

$$b_{k\tau} = \frac{\varepsilon_\tau^k}{k!} F^{(k+1)}\left(y_0 + \varepsilon_\tau\right). \tag{5.10}$$

In order to estimate the quantities $\gamma_k = \frac{F^{(k)}(y_0)}{k!}$, a weighted local polynomial regression can be conducted. Having the quantile regression estimates available, one can denote the quantile estimates centered around some point y_0 as $Z_\tau = \hat{Q}_y(\tau|X = \boldsymbol{x}_0) - y_0$. The objective function of the locally weighted polynomial regression then reads

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\Theta}{\operatorname{argmin}} \sum_{l=1}^{Q}\left(\tau_i - \sum_{k=0}^{p}\left(\tilde{\gamma}_k + b_{k\tau_i}\right) Z_i^k\right)^2 K_h(Z_i). \tag{5.11}$$

In matrix notation the first order conditions for this minimization problem can be written as

$$\boldsymbol{Z}'\boldsymbol{W}\left(\boldsymbol{\tau} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}} - (\boldsymbol{Z}\circ\boldsymbol{b})\,\boldsymbol{\iota}\right) \stackrel{!}{=} 0,$$

where $\circ$ denotes the element-wise (Hadamard) product and $\boldsymbol{\iota}$ is the $p\times 1$ vector of ones $\boldsymbol{\iota} = (1, 1, \ldots, 1)'$. Solving for $\hat{\gamma}$ yields

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{W}\left(\boldsymbol{\tau} - (\boldsymbol{Z}\circ\boldsymbol{b})\,\boldsymbol{\iota}\right) .$$

The conditional moment of the estimate, thus, is equal to

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\gamma}} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}_0, y_0] &= \left(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{W}\left(\boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\eta} - (\boldsymbol{Z}\circ\boldsymbol{b})\,\boldsymbol{\iota}\right) \\ &= \boldsymbol{\gamma} + \left(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{W}\left(\boldsymbol{\eta} - (\boldsymbol{Z}\circ\boldsymbol{b})\,\boldsymbol{\iota}\right) \\ &= \boldsymbol{\gamma} + \text{bias} .\end{aligned}$$

First, note that given $\boldsymbol{X}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}_0$ and y_0, the bias of the estimator is fully determined, but unobserved. Second, the parameter estimates $\hat{\gamma}$ are calculated from means across all τ i.e., the different quantile levels. Third, one can also recognize that the bias of $\hat{\boldsymbol{\gamma}}$ depends

on $\boldsymbol{b}$, which crucially depends on the derivatives of the distribution function $F(y_0)$, but, more decisively, on powers of the estimation error ε_τ introduced by the quantile regression estimates at each τ. Since $\varepsilon_\tau \to 0$ for $N \to \infty$ in the case where the assumptions of the quantile regression hold, the part of the bias $(\boldsymbol{Z} \circ \boldsymbol{b})\,\boldsymbol{\iota} \to 0$ as well. I conjecture that the even powers of ε_τ play a more dominant role in $(\boldsymbol{Z} \circ \boldsymbol{b})\,\boldsymbol{\iota}$. Depending on the shape of the distribution function, and thus the sign of its derivatives, the bias may be positive or negative. Fourth, the bias also depends on $\boldsymbol{\eta}$ which depends on both the approximation error of a finite Taylor expansion as well as on ε_τ. Recall that

$$\eta_\tau = \sum_{j=p+1}^{\infty} \left(\frac{F^{(j)}(y_0)}{j!} + b_{j\tau} \right) \left(\hat{Q}_y(\tau | \boldsymbol{X} = \boldsymbol{x}_0) - y_0 \right)^j .$$

If the bandwidth $h \to 0$, $Q \to \infty$ and $N \to \infty$, note, the parameter estimates are consistent as $\varepsilon_\tau \to 0$ and also $\eta_\tau \to 0$ at each τ, and, thus, $\mathbb{E}[\hat{\boldsymbol{\gamma}}] = \boldsymbol{\gamma}$ in large samples with a large number of quantiles estimated.

Since the estimate (including the bias) is fully determined by the quantile regression estimates $\hat{\boldsymbol{\theta}}$, the conditional variance of the parameter estimate is given by

$$\operatorname{var}[\hat{\boldsymbol{\gamma}} \mid \boldsymbol{X}, \hat{\theta}, \boldsymbol{x}_0, y_0] = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\, \boldsymbol{Z}'\boldsymbol{W}\ \boldsymbol{\Omega}\ \boldsymbol{W}\boldsymbol{Z}\, (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1},$$

where under the assumption of homoscedasticity $\boldsymbol{\Omega}$ can be expressed as

$$\boldsymbol{\Omega} = \boldsymbol{\eta}'\boldsymbol{\eta} - 2\boldsymbol{\eta}'\,(\boldsymbol{Z} \circ \boldsymbol{b})\,\boldsymbol{\iota} + \boldsymbol{\iota}'(\boldsymbol{b}'\boldsymbol{Z})^2\boldsymbol{\iota}.$$

If the quantile specific biases $b_{k\tau}$ for each of the k parameters in $\hat{\boldsymbol{\gamma}}$ as well as the quantile specific approximation error η_τ were observable in the sample, then the bias of $\hat{\boldsymbol{\gamma}}$ as well as its variance could be estimated by sample means. This is unfortunately not the case. However, note, the quantile specific as well as the usual bias of local polynomial regression estimates are added on top of each other.

Since γ_1 is estimated from a locally weighted polynomial regression with stochastic bandwidth (the procedure is described in Section 5.1.1), the asymptotic bias of such an estimator can be accounted for. For non-stationary strongly mixing processes, Masry and Fan (1997) establish the bias and asymptotic normality of local polynomial estimates obtained with non-stochastic bandwidths. Martins-Filho and Saraiva (2012) derive the bias as well as the asymptotic normality of a local polynomial regression estimate with stochastic bandwidth.

Translating their Corollary 3.1 into the notation, establishes the asymptotic normality of the estimate $\hat{\gamma}_1$ as

$$\sqrt{Qh^3}\left((\hat{\gamma}_1(y_0;\boldsymbol{\theta},h)-f(y_0))-\frac{F^{p+1}h^p}{(p+1)!}+h^p o_p(1)\right)\sim\mathcal{N}\left(0,\sigma_h^2\right), \tag{5.12}$$

where I have suppressed the dependence of h on the sample size Q, and $o_p(1)$ represents terms that converge in probability to 1 as $Q\to\infty$. Furthermore, σ_h^2 signifies the asymptotic variance derived by Martins-Filho and Saraiva (2012). Since in the local polynomial estimates are based on estimated quantiles from a quantile regression, I do not use the variance developed by Martins-Filho and Saraiva (2012).

In the absence of the quantities necessary to estimate the exact bias, I account for the bias $\frac{F^{p+1}h^p}{(p+1)!}+h^p$ derived by Masry and Fan (1997). The estimation of F^{p+1} is also necessary to calculate the variable bandwidth h in Section 5.1.1. As the bias and the bandwidth both depend on the estimation of F^{p+1} which in turn requires the estimation of a polynomial regression and a pilot bandwidth as set out in Section 5.1.1, I find that the bias correction is only necessary when the estimate for F^{p+1} surpasses a certain threshold. In this chapter, I use a fourth order local polynomial regression and find that if the value of the fifth derivative F^5 exceed the value of 23 the correction is necessary. The value of 23 has been established by graphically comparing the similarity between estimated and theoretical density plots. Surely, there is a more systematic approach to determine the value. This is left for the time being to future research. Nonetheless, some threshold value around 23 should be universally applicable to the smoothing of distribution functions, since the function values are always restricted to unit interval.

In order to account for the special situation where the local polynomial estimation is based on quantile regression results, I estimate the variance of $\hat{\gamma}_1$ via the procedure described in Section 5.1.1 which is based on asymptotic theory rooted in the GMM setup already described in Section 5.1.1. Before discussing the estimation of the variance of the parameter estimate, however, I set out the estimation of the local bandwidth h in Section 5.1.1.

Bandwidth Selection

The choice of the optimal bandwidth manifests the trade-off between variance and bias of the estimated parameter. Choosing the bandwidth too narrowly leaves out too many estimations and increases the variance of the density estimator. Selecting a bandwidth that is too wide reduces the variance of the estimator, while its bias is increased.

Since there is no clear guidance in the framework of Cleveland (1979), how the parameters are chosen optimally, I opt for the locally optimal bandwidth as worked out by Fan, Gijbels, Hu and Huang (1996). The locally optimal bandwidth requires several quantities to be

estimated. In this way, the bandwidth and kernel function are chosen approximately optimal in the sense that they approximately minimize the conditional mean squared error at each point $P = (\hat{\boldsymbol{\theta}}\mathbf{x}_0, y_0)$.

In this case, the choice of the bandwidth is not straightforward. While Fan et al. (1996) find asymptotic expressions for the conditional bias and the variance for parameter estimates of a general local polynomial expression to estimate the optimal bandwidth, in this context these quantities are biased. Nevertheless, these asymptotic expressions still hold approximately, especially when the sample size goes to infinity.

In a bivariate standard local polynomial setting with identically and independently distributed (iid) data points $(X_1, Y_1) \dots (X_n, Y_n)$ from a population (X, Y) at hand, Fan and Gijbels (1996) derive the optimal bandwidth at the point x_0 to estimate the vth derivative $m^{(v)}(x_0)$ with a local polynomial regression of order p as

$$h_{\text{opt}}(x_0) = C_{v,p}(K) \left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f(x_0)} \right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{5.13}$$

with

$$C_{v,p}(K) = \left[\frac{(p+1)!^2(2v+1) \int K_v^{*2}(t)dt}{(2p+1-v) \{\int t^{p+1} K_v^* dt\}^2} \right]^{1/(2p+3)}.$$

and $\sigma^2(x_0)$ being the conditional variance and $f(x_0)$ the design density at point x_0. The constants $C_{v,p}(K)$ only depend on the kernel function and are easily calculated.

In this context, I choose $p = 4$ and approximate the first derivative of the conditional distribution function, i.e., $v = 1$. Thus, I use the constant $C_{1,4} = 3.8565$. If I would have used $p = 6$ or $p = 8$ to estimate the first derivative, the constant would be larger by a factor of 1.4 and 1.8, respectively. One can also choose $p = 2$. This would yield a smaller constant by a factor of 0.6. Therefore, similar to Fan and Gijbels (1995), the resulting rule of thumb bandwidth estimator – presented in Equation (5.16) below – is multiplied by a factor of 1.39 in order to form a compromise between the various candidates for the constant factor.

A conundrum emerges when estimating the quantities in the numerator and denominator of Equation (5.13). The conditional variance $\sigma^2(x_0)$, the design density $f(x_0)$, and the derivative $m^{p+1}(x_0)$ can only be estimated (by local polynomial regression) when the optimal bandwidth is already known. One solution would be to conduct a grid search to find the minimum of the residual squares criterion (RSC) as discussed in Fan and Gijbels (1996). Another approach has been developed by Yu and Jones (1998) who resort to the optimal bandwidth obtained by minimization of the mean integrated squared error (MISE) as developed by Ruppert and Wand (1994) and refined in Ruppert, Sheather and Wand (1995). Both approaches come with heavy computational costs. With the assumption that

the underlying density does not have many high-frequency alternations, one can resort to the rule of thumb approach presented in Fan and Gijbels (1995, 1996), slightly modified to fit the situation. To obtain a approximately optimal bandwidth, I proceed as follows:

First, I fit a polynomial on the uncentered quantile estimates

$$\breve{\boldsymbol{\tau}} = \breve{\boldsymbol{Z}}\breve{\boldsymbol{\gamma}}$$

where $\breve{\boldsymbol{Z}} = \left(\mathbf{1}, \hat{\boldsymbol{\theta}}\mathbf{x}_0, (\hat{\boldsymbol{\theta}}\mathbf{x}_0)^2, \ldots, (\hat{\boldsymbol{\theta}}\mathbf{x}_0)^{p+3}\right)$ and $\breve{\boldsymbol{\gamma}} = (\breve{\gamma}_0, \breve{\gamma}_1, \ldots, \breve{\gamma}_{p+3})'$.

Second, from the residuals of the first step $\boldsymbol{\nu} = \boldsymbol{\tau} - \breve{\boldsymbol{\tau}}$, I calculate an estimate for the conditional variance as

$$\breve{\sigma}^2(y_0) = \frac{1}{Q - p - 3 - 1} \sum_{l=1}^{Q} \boldsymbol{\nu}'\boldsymbol{\nu}. \tag{5.14}$$

Third, I use twice the standard deviation of $\hat{\boldsymbol{\theta}}\mathbf{x}_0$ as a preliminary bandwidth parameter h_p.

Fourth, using the results from the first and the third step, I obtain a rough estimate for the denominator of the optimal bandwidth the probability weighted derivative of order $p+1$ of the distribution function $F^{(p+1)}(y_0)$, as

$$\left\{F^{(p+1)}(y_0)\right\}^2 f(y_0) \approx \frac{1}{Q} \sum_{l=1}^{Q} \left(\sum_{j=0}^{2} \frac{(p+1+j)!}{j!} \breve{\gamma}_{p+1+j} \left(\boldsymbol{\theta}_l \mathbf{x}_0\right)^j \right)^2 K_{h_p}\left(\frac{\boldsymbol{\theta}_l \mathbf{x}_0 - y_0}{h_p} \right) \tag{5.15}$$

where $K_{h_p} = K(\cdot)/h_p$.

In a fifth last step, I combine these results to obtain the rule of thumb estimate for the optimal local bandwidth

$$h(y_0) = 1.39 \cdot C_{1,4}(K) \left[\frac{\breve{\sigma}^2(y_0)}{\{F^{(p+1)}(y_0)\}^2 f(y_0)} \right]^{1/11} Q^{-1/11}. \tag{5.16}$$

Variance of the Density Estimate

In the following, I denote the estimate for the density $f_{Y|X}(y_0)$ as $\hat{f}(\hat{\boldsymbol{\theta}})$. This is the estimate stemming from the local polynomial regression of order p with stochastic bandwidth h on the estimated quantiles using the parameter estimates $\hat{\boldsymbol{\theta}}$ which is already corrected for the asymptotic bias $\frac{\hat{F}^{p+1}h^p}{(p+1)!} + h^p$.

Given the quantile regression assumptions hold, for $N \to \infty$ the parameter estimates from the quantile regression converge $\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}$ and $\varepsilon_\tau \to_p 0 \ \ \forall \tau$. Since the parameters $\hat{\boldsymbol{\theta}}$ fully determine the estimates $\hat{f}(\hat{\boldsymbol{\theta}})$, by the continuous mapping theorem, it follows that for

$N \to \infty$ and $Q \to \infty$ as well as $0 < h \to 0$, the parameter estimates $\hat{f}(\hat{\boldsymbol{\theta}})$ converge in probability to the conditional density function $f_{Y|X}$. Note that the rate of convergence of h and Q need to ensure that $h^3 Q \to \infty$.[35]

This means only when both N and $h^3 Q$ approach infinity, I have both a good fit of the theoretical conditional quantile function and a good fit for the conditional density estimate via the smoothing. Therefore N and $h^3 Q$ need to be chosen sufficiently large in order for $\hat{f}(\hat{\boldsymbol{\theta}})$ to approximate the true value.

Since the asymptotic variance of the quantile regression parameter estimates $\hat{\boldsymbol{\theta}}$ has been worked out previously in the literature (see Koenker 2005), one can determine the distribution of the density function estimate $\hat{f}$.

The question how $\operatorname{Avar}\left(\hat{f}\left(\boldsymbol{\theta}\right)\right)$ can be approximated remains. This can be answered with the mean value theorem (cp. Hayashi 2000). For this purpose, the estimation of the derivative of $\hat{f}(\hat{\boldsymbol{\theta}})$ with respect to $\hat{\boldsymbol{\theta}}$ is needed. Thus, for some sufficiently large fixed Q, letting $N \to \infty$, in the limit, one can write

$$\lim_{N\to\infty} \sqrt{Qh^3} \frac{\left(\hat{f}(\hat{\boldsymbol{\theta}}) - f\right)}{\operatorname{vec}(\hat{\boldsymbol{\theta}}) - \operatorname{vec}(\boldsymbol{\theta})} \approx \left.\frac{\partial \hat{f}}{\partial \bar{\boldsymbol{\theta}}}\right|_{\bar{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}.$$

where f is the true value. Note that I have utilized the factor $\sqrt{Qh^3}$ to stabilize the convergence of the derivative of $\hat{f}(\boldsymbol{\theta})$. This is the same stabilizing factor as in Equation (5.12). Then, for Q and N sufficiently large and $\left.\frac{\partial \hat{f}}{\partial \bar{\boldsymbol{\theta}}}\right|_{\bar{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}} = \mathbf{H}$, I have the approximation

$$\sqrt{Qh^3}\left(\hat{f}(\hat{\boldsymbol{\theta}}) - f\right) \approx \operatorname{vec}\left(\mathbf{H}\right)'\left(\operatorname{vec}(\hat{\boldsymbol{\theta}}) - \operatorname{vec}(\boldsymbol{\theta})\right).$$

Since $\sqrt{N}\left(\operatorname{vec}(\hat{\boldsymbol{\theta}}) - \operatorname{vec}(\boldsymbol{\theta})\right)$ is asymptotically normal distributed, I multiply both sides by $\sqrt{N}$. It follows that the left hand side is asymptotically normal distributed, as well, i.e.,

$$\sqrt{NQh^3}\left(\hat{f}(\hat{\boldsymbol{\theta}}) - f\right) \sim \mathcal{N}\left(0, \operatorname{vec}\left(\mathbf{H}\right)' \operatorname{Avar}(\hat{\boldsymbol{\theta}}) \operatorname{vec}\left(\mathbf{H}\right)\right). \tag{5.17}$$

Equation (5.17) now allows for an approximation of the asymptotic variance of $\hat{f}\left(\boldsymbol{\theta}\right)$ by

$$\operatorname{Avar}\left(\hat{\gamma}_1\left(\boldsymbol{\theta}\right)\right) \approx \frac{1}{NQh^3} \operatorname{vec}\left(\mathbf{H}\right)' \operatorname{Avar}(\hat{\boldsymbol{\theta}}) \operatorname{vec}\left(\mathbf{H}\right), \tag{5.18}$$

where the calculation of the elements of the $Q \times K$ matrix $\boldsymbol{H}$ is worked out in Section 5.B.

[35] The rate of convergence of the local polynomial estimate for the first derivative has been worked out under fairly general conditions with stochastic bandwidth by Martins-Filho and Saraiva (2012).

Since $\hat{f}(\hat{\boldsymbol{\theta}})$ is a density, for some rare points where $\hat{f}(\hat{\boldsymbol{\theta}}) < 0$, I have to make the following additional transformation which ensures that the density estimate is non-negative

$$f(\hat{\boldsymbol{\theta}}) = \max\left(0, \hat{\gamma}_1(\hat{\boldsymbol{\theta}})\right).$$

Even though this case is rare, since the insertion of the ghost points almost eliminates its occurrence, at some rare instances the transformation is necessary. These instances are limited to density estimates outside of the range observed in the sampled data. However, at this point a note of caution is in order. To apply the mean value theorem, one needs the continuous mapping theorem to hold (e.g. Hayashi 2000). However, $f(\hat{\boldsymbol{\theta}})$ has a discontinuity at 0 and the maximum function is a non-continuous transformation. Thus, variance estimates outside the support of the sample are prone to error. The ghost points artificially reduce uncertainty for these estimates, as, for one, they increase the number of observations, and second, they occur at a rather regular frequency. Thus, the flooring of the density estimates by means of ghost points renders the presented asymptotic theory prone to error, especially at the tails of the distribution.

While the distributional issues are challenging, the practical results of the procedure work nicely. Figure 5.2 shows the results for two bivariate density estimates. I see that in the center of the distribution the densities are slightly underestimated while at the tails the converse is true and the density is slightly overestimated. However, for 1000 observations and only 100 quantiles estimated in each regression, I deem the fit satisfactory.

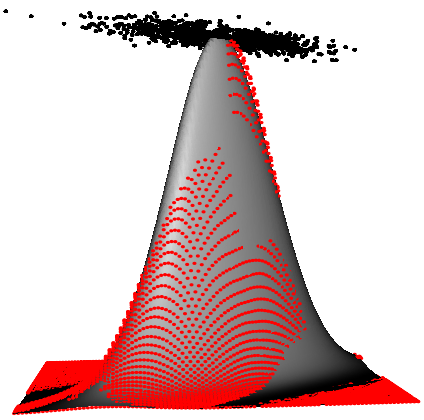
Limitations

The limits of the procedure become apparent when looking at different distributions where the conditional quantile functions are not linear and/or not continuous. For this purpose, I consider samples of 1000 observations from a distribution on a 2 dimensional spherical shell. Figure 5.2c shows the corresponding sample, the estimate and the theoretical density. As can be seen the density is not continuous. Hence, the jump from 0 to the density value $(\pi^2(r_o - r_i))^{-1}$ (where r_i is the radius of the inner circle and r_o is the radius of the outer circle) cannot be picked up by the estimation procedure properly. By forcing the linear structure of quantile regression on the figure, the procedures somewhat tries to square the circle, a proverbial impossibility. Thus, the estimated densities assume high values at the 'edges of the circle'.

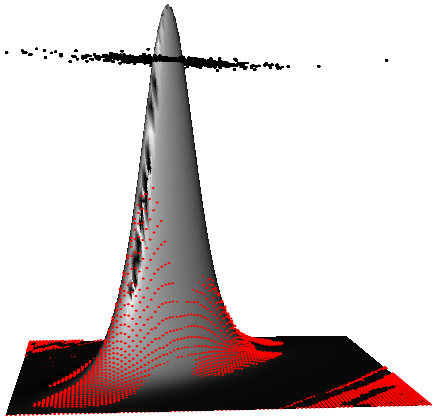
The geometrical figure of the spherical shell is extendible to n dimensions. I make use of this in the simulation study for MI in Section 5.2.2 below. The distribution on a spherical shell ensures non-linearity between all involved random variables, while simultaneously exhibiting zero correlation between all related variables.

Figure 5.2: Bivariate Normal and t-Distribution Estimates

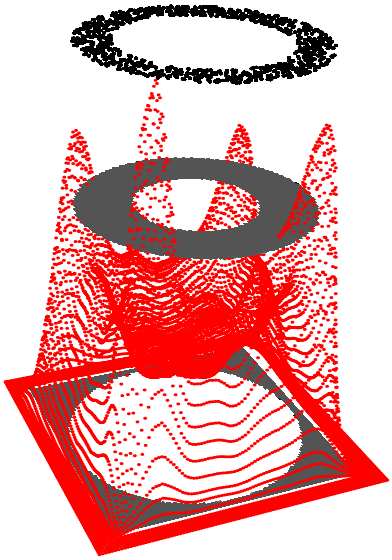
The graph shows the theoretical joint density function (gray surface) of a bivariate normal distribution in Subfigure 5.2a and of a t-distribution in Subfigure 5.2b. Simultaneously, the respective density estimates based on smoothed quantile regressions (red points) are depicted. The black cloud of points at the top of the density function is the sample of 1000 points on which the density estimates are based. For both density estimates $Q = 100$ quantiles were estimated via quantile regression. Estimated quantiles were fitted with a local polynomial regression of order 4 with stochastic bandwidth and an Epanechnikov-kernel function. In order to plot the surface, the estimated density was evaluated on a grid with $M = 100$ points in each dimension.



(a) Multivariate Normal Distribution



(b) Multivariate t-Distribution



(c) Uniform Distribution on a Spherical Shell

While the approximation for the joint density obtained through the estimates may not be satisfactory for other purposes, for the estimation of MI in a practical setting they may be suitable anyway. The research question is often whether two or more variables are unrelated. The null hypothesis for $MI = 0$ would most probably, also be rejected in this setting with the estimated joint density. I consider such a case in the simulation study in Section 5.2.2.

However, an exact estimate for MI is not possible for such a scenario. The class of densities for which the quantile regression approach presented in this chapter is feasible, is limited to the situation where the quantiles of the variables are globally linearily related.

At this point, one could think of a potentially problematic situation in which there are several stochastically dependent variables and all estimated conditional quantiles are parallel to the axis of the conditioning variable. However, if this is the case the variables are by definition independent. Therefore, linear quantile regression is in principle suited to detect non-zero MI and/or stochastic dependence among sets of random variables. However, the approach is not exact concerning the concrete value of MI. A similar argument holds for CMI and TE.

If the limitation for an exact estimate needs to be overcome, one could also use at each point the quantile estimates of a locally weighted quantile regression. The further development of this approach, however, is beyond the scope of this chapter and left to future research.

5.1.2 Asymptotic Theory for Relative Entropy Measures

As discussed in the introduction to this chapter and listed in Table 5.1, relative entropy measures are constructed as Kulback-Leibler divergences. The discussion in Section 5.1.2 focuses first on the asymptotics of MI as a show case of the method. However, the simulation study for MI will also be accompanied by another simulation study. I also look at the specifics of TE as another example for a relative entropy measure and a special case of CMI in Section 5.1.2. While discretization approaches are deemed appropriate for the estimation of MI, for TE coarse-graining methods have some unsatisfying characteristics (see Kaiser and Schreiber 2002). Therefore, even though I first focus on MI (and leave time series considerations aside), I deem the asymptotic behavior and estimation technique for TE as the major contribution of this chapter.

Mutual Information

While the concept of MI is defined in general for K variables, it is useful to limit the discussion here to the three variable case in order to alleviate notational complexities. A

limitation to two variables would be possible, however, for the extension to conditional MI (and TE as a special case), I use the three variable case in which MI is defined as

$$I(X,Y,Z) = \mathbb{E}\left[\log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right)\right] \tag{5.19}$$

$$= \int_{x\in X}\int_{y\in Y}\int_{z\in Z} f_{X,Y,Z}(x,y,z)C(x,y,z)dzdxdy$$

$$= \int_{x\in X}\int_{y\in Y}\int_{z\in Z} f_{X,Y,Z}(x,y,z)\log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right)dzdxdy. \tag{5.20}$$

This definition allows for two approaches to calculate MI. The first one calculates the sample mean equivalent of Equation (5.19) over all observations (x_i, y_i, z_i) with $i \in \{1,\dots,N\}$.

$$\hat{I}(X,Y,Z) = \mathbb{E}\left[\log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\hat{f}_{X,Y,Z}(x_i,y_i,z_i)}{\hat{f}_X(x_i)\hat{f}_Y(y_i)\hat{f}_Z(z_i)}\right) \tag{5.21}$$

This approach uses the representativity of the sample to circumvent the integration over a grid of artificial support points. Which leads to the other calculation approach: the numerical integration suggested by the integral formulation in Equation (5.20). This entails the integration across a grid that covers the sample space across the various dimensions. If I chose M grid points then for three variables X,Y and Z, the integral would need to be evaluated at M^3 grid points. The number of points grows exponentially with each dimension, making the calculation practically already for a small number of variables infeasible. The expectation formulation is much more feasible.

For MI, another possibility of estimation seems interesting: One can estimate the involved densities by means of kernel density techniques. For the concept of conditional MI and TE, where the conditional densities are directly needed in the calculation, however, the quantile regression approach emerges much more naturally as can be seen in Section 5.1.2 as it strongly limits the computational resources. So even though the approach may not be the most flexibel to estimate MI, I use the method of smoothed quantile regression estimates for MI estimation as a show case in order to make the concept more accessible. Also, the estimates come with standard errors and are testable.

For the purpose of calculating the MI contributions via conditional densities, one can rewrite the joint density as

$$f_{X,Y,Z}(x,y,z) = f_{X|Y,Z}(x|y,z)f_{Y|Z}(y|z)f_Z(z).$$

This makes the summand for the ith observation in Equation (5.21)

$$C(x_i,y_i,z_i) = \log\left(\frac{\hat{f}_{X|Y,Z}(x_i|y_i,z_i)\hat{f}_{Y|Z}(y_i|z_i)}{\hat{f}_X(x_i)\hat{f}_Y(y_i)}\right).$$

This decomposition provides the basis for the quantile regression based approach. All constituents of the joint density can be estimated from the parameter estimates of the respective quantile regression such that

$$\hat{f}_{X,Y,Z}(x,y,z) = \hat{\gamma}_1(\hat{\boldsymbol{\theta}}_{X|Y,Z})\hat{\gamma}_1(\hat{\boldsymbol{\theta}}_{Y|Z})\hat{\gamma}_1(\hat{\boldsymbol{\theta}}_Z).$$

Note, the summed up terms in the sample mean estimate $\hat{I}(X,Y,Z)$ can again be conceived as functions of the parameter estimates from the various quantile regressions.

Also recall that each of the density estimates in C may converge at a different rate to the true density value. This is due to the result of Martins-Filho and Saraiva (2012) for the asymptotic convergence of local polynomial regression parameters estimated with stochastic bandwidths. The result for the local polynomial regression parameter associated to the first order derivative is reproduced for convenience in Equation (5.12). Therefore, in order to derive an asymptotic distribution of a test statistic, I need to align the convergence of the density estimates. For the construction of a test statistic, I therefore divide each of the density estimates by the square root of the third power of its bandwidth. In effect, since MI is formulated in logarithms this is equivalent to subtracting a constant term from the TE estimate. For the so normalized term, a function of normalized density estimates, I conjecture that $\sqrt{Q}$-normality is sustained. The simulation results in section Section 5.2.3 underpin this conjecture.

For the construction of the test statistic, I treat the bandwidths as somewhat fixed and independent of the regression estimates $\hat{\boldsymbol{\theta}}$. The delta method for the adjusted contributions, thus, can be derived to be

$$\begin{aligned}\lim_{N\to\infty} \sqrt{Q}\frac{\hat{C}(\hat{\boldsymbol{\theta}}) - C - C^*}{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}} &= \lim_{N\to\infty} \frac{1}{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}\sqrt{Q}\log\left(\frac{h^{-1}_{X|Y,Z}(\hat{f}_{X|Y,Z} - f_{X|Y,Z})h^{-1}_{Y|Z}(\hat{f}_{Y|Z} - f_{Y|Z})}{h^{-1}_{X}(\hat{f}_{X} - f_{X})h^{-1}_{Y}(\hat{f}_{Y} - f_{Y})}\right) \\ &= \left.\frac{\partial \hat{C}_{x_i,y_i,z_i}}{\partial \bar{\theta}_{lm}}\right|_{\bar{\theta}_{lm}=\hat{\theta}_{lm}},\end{aligned}$$

where the correcting term $C^* = \frac{3}{2}\log\left(\frac{h_{X|Y,Z}h_{Y|Z}}{h_X h_Y}\right)$ is used.

In order to calculate the variance of the MI estimate, not only the variance of each summand is needed, but also the covariances among all of the summands. Again, the asymptotic convergence results for the density estimates developed in Section 5.1.1 are of importance.

Knowing the limiting distribution of $\hat{\boldsymbol{\theta}}$, one can work out the limiting distribution of each summand $\hat{C}(x,y,z)$ and the covariance between any two summands $C_{\mathbf{i}} = C(x_i,y_i,z_i)$ and $C_{\mathbf{j}} = C(x_i,y_i,z_i)$ using the delta method. Everything that is needed are the gradients of the summands with respect to the parameter estimates $\hat{\boldsymbol{\theta}}$.

Collecting the derivatives of the i^{th} contribution C_i with respect to $\boldsymbol{\theta}$ in a matrix $\boldsymbol{\Upsilon}_\mathbf{i}$, the elements of $\boldsymbol{\Upsilon}_\mathbf{i}$ can be written as

$$\left.\frac{\partial \hat{C}_{x_i,y_i,z_i}}{\partial \bar{\theta}_{lm}}\right|_{\bar{\theta}_{lm}=\hat{\theta}_{lm}} = \begin{cases} \frac{\partial \hat{f}_{X|Y,Z}}{\hat{f}_{X|Y,Z}} & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{X|Y,Z} \\ \frac{\partial \hat{f}_{Y|Z}}{\hat{f}_{Y|Z}} & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{Y|Z} \\ 0 & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{Z} \\ -\frac{1}{\hat{f}_Y}\,\partial \hat{f}_Y & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{Y} \\ -\frac{1}{\hat{f}_X}\,\partial \hat{f}_X & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{X} \end{cases}$$

where the elements $\partial \hat{f}$ can be replaced with the respective elements of $\boldsymbol{H}$ and $\hat{f}$ with the density estimates.

Note that when the representativity of the sample is not used and MI is estimated by integrating over the estimated joint density function, the derivatives of the contributions $C_\mathbf{i}$ need to be extended by additional terms.

Knowing $\boldsymbol{\Upsilon}_\mathbf{i}$ and its variance, the limiting distribution of the contribution can be approximated using the delta method (cp. Oehlert 1992, Hayashi 2000) which leads to

$$\hat{C}_\mathbf{i}\left(\hat{\boldsymbol{\theta}}\right) + C^* \sim \mathcal{N}\left(C_\mathbf{i}, \frac{1}{QN}\operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{i}\right)' \operatorname{Avar}(\hat{\boldsymbol{\theta}}) \operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{i}\right)\right). \tag{5.22}$$

In order to calculate MI, the conventions $0 \log\left(\frac{0}{0}\right) = 0$, $0 \log\left(\frac{0}{f_Y}\right) = 0$ as well as $\log\left(\frac{f_X}{0}\right) = \infty$ need to be introduced (cp. Cover and Thomas 2005).[36] Therewith, the covariance between $\hat{C}_\mathbf{i}\left(\hat{\boldsymbol{\theta}}\right)$ and $\hat{C}_\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)$ may be approximated by (cp. Klein 1953)

$$\operatorname{cov}\left(\hat{C}_\mathbf{i}\left(\hat{\boldsymbol{\theta}}\right), \hat{C}_\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)\right) = \frac{1}{QN}\operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{i}\right)' \operatorname{Avar}(\hat{\boldsymbol{\theta}}) \operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{j}\right). \tag{5.23}$$

Based on the estimation of MI by the sample mean equivalent of Equation (5.21), the variance of the MI estimate can be computed as

$$\begin{aligned} \operatorname{var}\left(\hat{I}_{X,Y,Z}\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\operatorname{cov}\left(\hat{C}_i\left(\hat{\boldsymbol{\theta}}\right), \hat{C}_j\left(\hat{\boldsymbol{\theta}}\right)\right) \\ &= \frac{1}{QN}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{N}\operatorname{vec}\left(\boldsymbol{\Upsilon}_i\right)' \operatorname{Avar}(\hat{\boldsymbol{\theta}})\frac{1}{N}\operatorname{vec}\left(\boldsymbol{\Upsilon}_j\right) \\ &= \frac{1}{QN}\left[\frac{1}{N}\sum_{i=1}^{N}\operatorname{vec}\left(\boldsymbol{\Upsilon}_i\right)\right]' \operatorname{Avar}(\hat{\boldsymbol{\theta}})\left[\frac{1}{N}\sum_{j=1}^{N}\operatorname{vec}\left(\boldsymbol{\Upsilon}_j\right)\right]. \end{aligned} \tag{5.24}$$

[36] Note that during implementation, one can choose to numerically represent infinity by a sufficiently large value. However, I choose to exclude such points in the calculation, since dragging these values through all calculations results in numerical instabilities.

This approach also leads me to conjecture that the limiting distribution of $\hat{\text{var}}\left(\hat{I}_{X,Y,Z}\right)$ is a χ^2 distribution. For the application, however, I am only interested in the distribution of $\hat{I}_{X,Y,Z}$.

Transfer Entropy

Similar to Equations 5.19 and 5.20, transfer entropy as a special case of conditional MI in a time series context with time ordered random variables can be constructed as

$$\begin{aligned} T_{X\to Y} &= I(Y_t, X_{t-1} \mid Y_{t-1}) \\ &= \mathbb{E}\left[\Theta_t\right] = \mathbb{E}\left[\log\left(\frac{f_{Y_t|X_{t-1},Y_{t-1}}(y_t \mid x_{t-1}, y_{t-1})}{f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1})}\right)\right] \\ &= \iiint_{\mathbb{R}^3} f_{Y_t,X_{t-1},Y_{t-1}}(y_t, x_{t-1}, y_{t-1}) \log\left(\frac{f_{Y_t|X_{t-1},Y_{t-1}}(y_t \mid x_{t-1}, y_{t-1})}{f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1})}\right) dy_t dx_{t-1} dy_{t-1} \end{aligned}$$

Note that only two quantile regressions are necessary to calculate the measure. Given stationary and ergodic time series for y_t, x_t and z_t, one can approximate TE via a sample mean

$$\hat{T}_{X\to Y} = \mathbb{E}\left[\log\left(\frac{f_{Y_t|X_{t-1},Y_{t-1}}(y_t \mid x_{t-1}, y_{t-1})}{f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1})}\right)\right] \approx \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{f_{Y_t|X_{t-1},Y_{t-1}}(y_t \mid x_{t-1}, y_{t-1})}{f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1})}\right) \tag{5.25}$$

The derivative of each summand is then given by

$$\left.\frac{\partial\hat{\Theta}_t}{\partial\bar{\theta}_{lm}}\right|_{\bar{\theta}_{lm}=\hat{\theta}_{lm}} = \begin{cases} \frac{\partial\hat{f}_{Y_t|X_{t-1},Y_{t-1}}}{\hat{f}_{Y_t|X_{t-1},Y_{t-1}}} & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{Y_t|X_{t-1},Y_{t-1}} \\ -\frac{\partial\hat{f}_{Y_t|Y_{t-1}}}{\hat{f}_{Y_t|Y_{t-1}}} & \text{if } \hat{\theta}_{lm} \in \hat{\boldsymbol{\theta}}_{Y_t|Y_{t-1}} \end{cases}$$

The variance of $\hat{T}_{X\to Y}$ may be estimated analogously to MI.

5.2 Simulation Studies

In order to check the theoretical arguments and limiting distributions above, I conduct several simulation studies. First, in Section 5.2.1, I analyze the asymptotic distribution of the estimate γ_1 for the conditional density $f_{X|Y}(X = x|Y = y)$ as derived in Section 5.1.1. Second, I investigate the asymptotic behavior of relative entropy estimates derived in Section 5.1.2 such as estimates for the MI as well as TE.

5.2.1 Simulation of Conditional Density Estimates

In this subsection, the behavior of the conditional densities estimates is discussed when the number of observations N is varied. For this purpose and to keep it brief, I exclusively sample in this section from the bivariate normal distribution of X and Y with the parameters

$$\mu_{XY} = \begin{pmatrix} -0.4 \\ -2.5 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma}_{XY} = \begin{pmatrix} 2.18 & 1.38 \\ 1.38 & 0.93 \end{pmatrix}. \tag{5.26}$$

I refer to this specific bivariate distribution as Φ_{XY} for the rest of this section and $\Phi_{X|Y}$ is the conditional distribution of X given Y.

To roughly illustrate the support of the distribution, Figure 5.3 shows a random sample of 1,000 data points drawn from this bivariate normal distribution. The blue lines indicate the conditional quantiles of Y at the 0.01, 0.1, 0.5, 0.9 and 0.99 level for each value of X, estimated via quantile regression. The numbered red dots are located at the intersection of 5 unconditional quantiles of X and Y. The coordinates of these red points are also listed in Table 5.2.

Figure 5.3: Location of Simulation Points S_i

The figure depicts a random sample of 1000 points drawn from the bivariate normal distribution Φ_{XY}. The blue lines indicate the conditional quantiles at the 0.01, 0.1, 0.5, 0.9 and 0.99 level estimated via quantile regression. The points are located at combinations of theoretical quantiles of the univariate marginal distributions of X and Y tabled in Table 5.2.

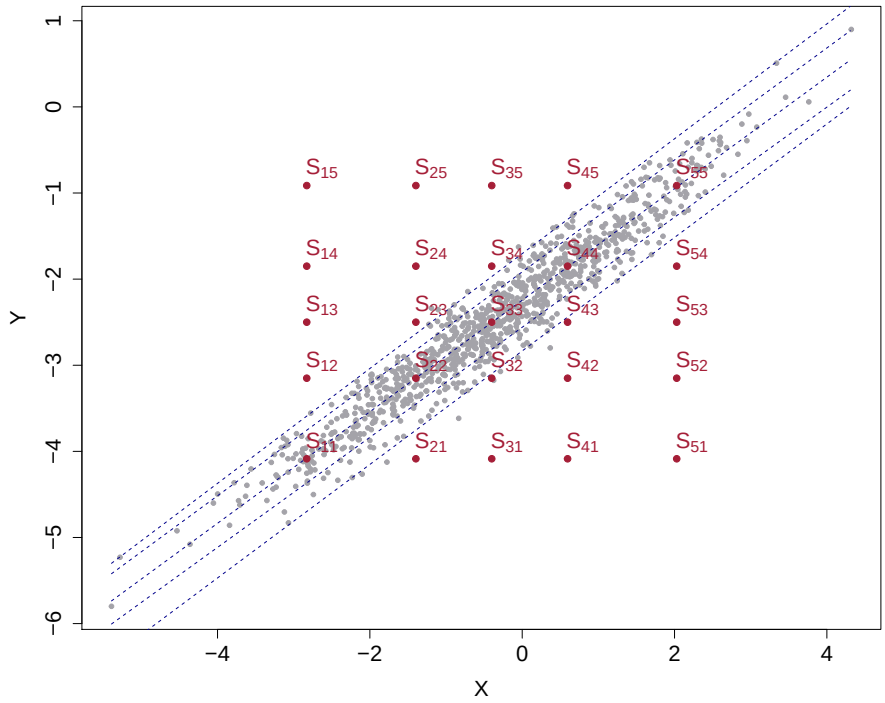


Table 5.2: Theoretical Quantiles of Φ_{XY}

The table lists the coordinates of selected points within the bivariate normal distribution Φ_{XY}. The points are located at combinations of theoretical quantiles of the univariate marginal distributions of X and Y. Figure 5.3 further illustrates their location within a random sample of 1000 draws from Φ_{XY}.

		Quantile Y				
		0.05	0.25	0.5	0.75	0.95
Quantile X	0.05	S_{11}:(-2.83,-4.09)	S_{12}:(-2.83,-3.15)	S_{13}:(-2.83,-2.5)	S_{14}:(-2.83,-1.85)	S_{15}:(-2.83,-0.91)
	0.25	S_{21}:(-1.4,-4.09)	S_{22}:(-1.4,-3.15)	S_{23}:(-1.4,-2.5)	S_{24}:(-1.4,-1.85)	S_{25}:(-1.4,-0.91)
	0.5	S_{31}:(-0.4,-4.09)	S_{32}:(-0.4,-3.15)	S_{33}:(-0.4,-2.5)	S_{34}:(-0.4,-1.85)	S_{35}:(-0.4,-0.91)
	0.75	S_{41}:(0.6,-4.09)	S_{42}:(0.6,-3.15)	S_{43}:(0.6,-2.5)	S_{44}:(0.6,-1.85)	S_{45}:(0.6,-0.91)
	0.95	S_{51}:(2.03,-4.09)	S_{52}:(2.03,-3.15)	S_{53}:(2.03,-2.5)	S_{54}:(2.03,-1.85)	S_{55}:(2.03,-0.91)

In a first step, I look at the conditional density function at the 0.05-, 0.5- and 0.75-quantile of Y. The theoretical density function at $Y = y[\tau]$, i.e., the τ-quantile of Y, is given by the density function of a normal distribution with parameterization $\mathcal{N}\left(\mu_X + \frac{\sigma_X}{\sigma_Y}\rho_{XY}(y[\tau] - \mu_Y), (1-\rho_{XY}^2)\sigma_X^2\right)$, where ρ_{XY} is the correlation coefficient, σ denotes the standard deviation and μ the mean; subscripts indicate the corresponding random variable.

Figure 5.4 shows density estimates for $f_{X|Y}$ at 200 equidistant points across the range of X when Y is fixed to some quantile of Y. The sample size is varied in the rows of Figure 5.4. One can see that the applied local polynomial approximation of order $p = 4$ produces a good fit.

Recall that for a given number of N sample points, one can choose the number of quantiles to be estimated (Q), the number of ghost points (H) inserted outside the range of the taken sample (at the upper as well as at the lower tail) and the share of the distance at which the first and every subsequent ghost point is inserted (ω). The density estimates below have been calculated with the parameters $H = 100$, $\omega = 0.05$, $G = 100$ for varying numbers of observations. Since I am interested in density estimates as a result of smoothing quantile estimates, a large number of quantiles improves the result of the smoothing. Therefore, I set the minimal number of estimated quantiles to $Q = 100$. If the number of observations is larger than 200, one may set the number of quantiles to $Q = N/2$, in order to use the additional information available in the data. Theoretically, the bias of the density estimate – especially at the mode of the distribution – may be reduced when more quantiles are available. Producing density estimates with increasing numbers of quantiles comes, however, at a significant computational cost and the additional accuracy is limited. I also ran simulations setting the number of quantiles according to the rule

$$Q = \min\left(\max\left(100, \frac{N}{2} - 1\right), 1000\right). \tag{5.27}$$

Hence, the number of quantiles is capped, for $N > 2000$ to $Q = 1000$. If this rule is applied the density estimates for $N > 100$ exhibit a better fit. This result is expected as more information is used to construct the estimates. However, the distribution of the test statistic under the correct null hypothesis is much more biased. Results can be seen in Section 5.C

Developing a more systematic approach to derive the optimal number of quantiles, the optimal number of ghost points as well as the optimal distance at which to insert these ghost points is, for the time being, left for future research.

The choices have been obtained by extensive tests. I found that the choice of the bandwidth has the highest impact on all results, density estimates as well as test statistics. Therefore, I have implemented the data driven bandwidth selection as presented above.

The estimation of the density estimates' variance brings me to the simulation of test statistic and statistical inference. For this purpose, I have simulated a standard Z-test statistic that follows from the asymptotic distribution of the density estimate $\hat{f}_1$ described in Equation (5.12). At each point S_i listed in Table 5.2 the conditional density $\varphi_{X|Y}(X = x_i|Y = y_i)$ (where x_i and y_i are the coordinates of S_i) can be estimated as $\hat{f}(S_i, \boldsymbol{\theta}_{X|Y})$ with the variance $\sqrt{\text{var}(\hat{f}(S_i, \boldsymbol{\theta}_{X|Y}))}$. If $\hat{f}(S_i, \boldsymbol{\theta}_{X|Y})$ was unbiased and the variance was consistently estimated, then a z-score test statistic would be standard normally distributed

$$t_D = \frac{\hat{f}(S_i, \boldsymbol{\theta}_{X|Y}) - \varphi_{X|Y}(X = x_i|Y = y_i)}{\sqrt{\text{var}(\hat{f}(S_i, \boldsymbol{\theta}_{X|Y}))}} \sim \mathcal{N}(0, 1)\,.$$

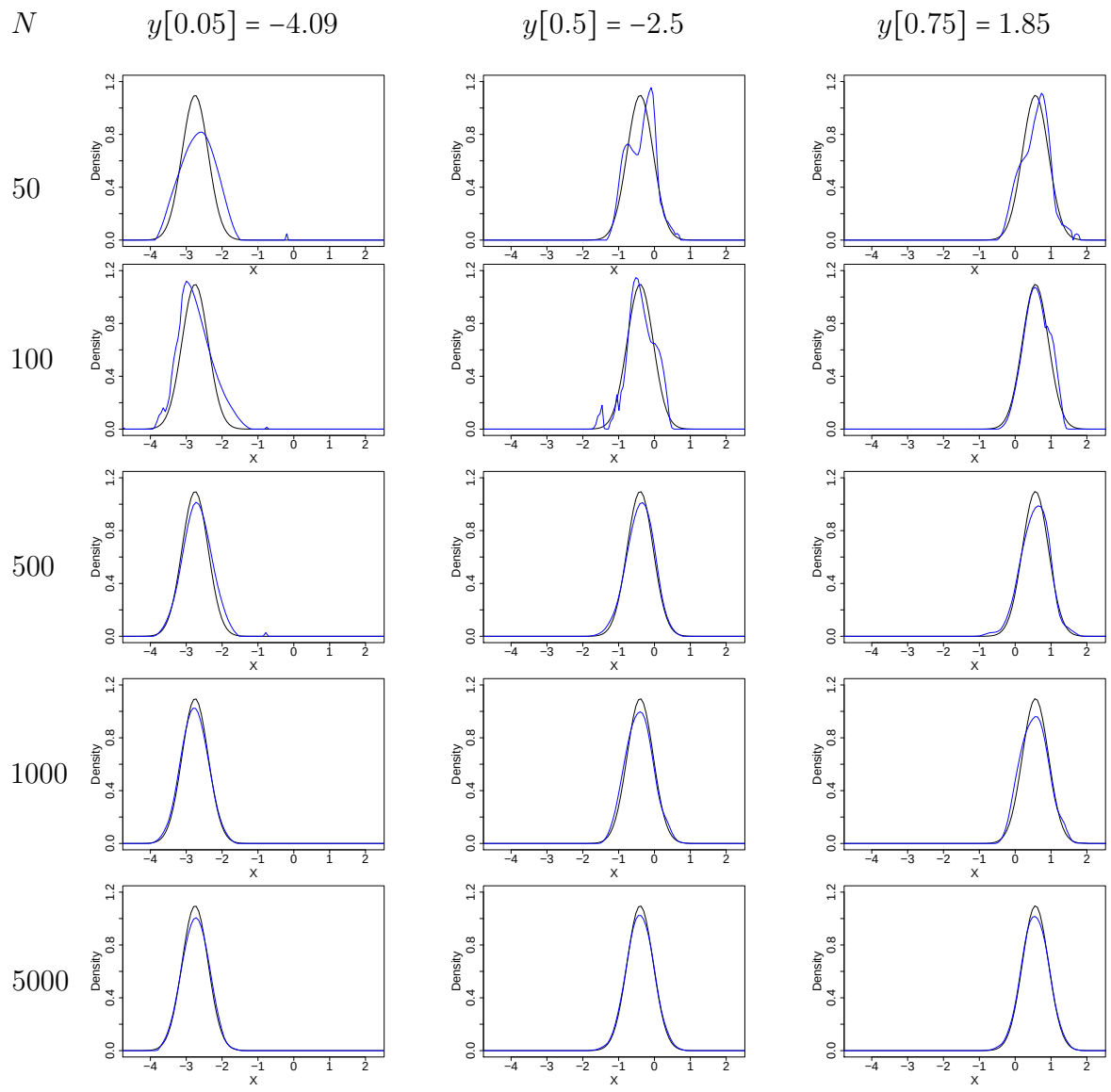
From the discussions in Section 5.1.1, it is already known that the estimates are somewhat biased. However, also the question remains whether the variance can sensibly be estimated as discussed in Section 5.1.1. Nonetheless, in order to assess the distributional properties of the estimate, I simulate this test statistic 5.000 times for selected points S_i listed in Table 5.2.

Looking at the location of the sample points S_i in Figure 5.3, it can be sensibly expected that the density estimates outside the sample support, especially those at the upper left and lower right corner, are zero and their variance also has to approach zero. The test-statistics for the densities at these points, thus, can be expected to be degenerate. Thus, these points are not simulated.

More interesting are the test statistics at the points that lie on the anti-diagonal in the point grid, i.e., S_{12}, S_{23}, S_{43}, S_{22}, S_{33} and S_{55}. These points are all in the center of the conditional densities. However, S_{33} has a very dense neighborhood whereas S_{12} and S_{43} are at the sparse rim of the distribution.

Figure 5.4: Density Estimates

The panel shows density estimates for various sample sizes N at various quantiles $y[\tau]$ of the bi-variate normal distribution Φ_{XY} described by the moments given in Equation (5.26). The solid blue line depicts the estimate using a polynomial smoothing of order $p = 4$. The solid black line shows the theoretical distribution.

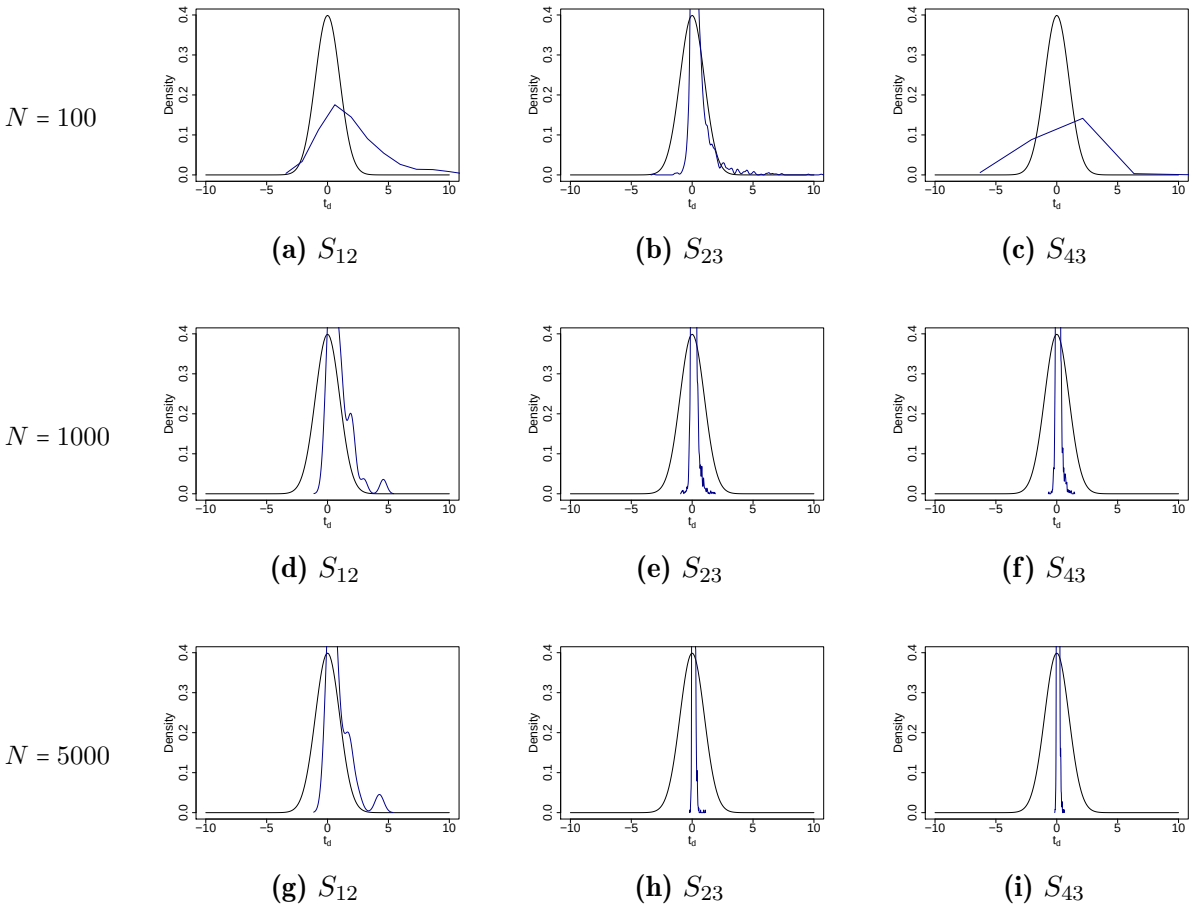


The distribution of the test statistic at these points are shown in Figure 5.5 and 5.6. I use kernel density estimates over 5000 test statistics calculated on varying sample sizes with local polynomials of order $p = 4$ to illustrate the distribution. The black solid line shows the theoretical standard normal distribution for comparison.

One can see in Figure 5.5, which presents the test-statistic simulations at the off-center points, and in Figure 5.6, which covers the center points, several important characteristics of the estimates. In the center of the conditional distributions at the points S_{22}, S_{33} and S_{55} the estimates systematically underestimate the true density, i.e., the estimate is slightly biased negatively. The distribution has similarities to a truncated normal

Figure 5.5: Test Statistics for Conditional Densities (Off-Center)

The panel shows kernel density estimates of 5000 simulated test statistics (blue solid) at the respective points S_{ij} of Φ_{XY} listed in Table 5.2. For this figure, only points that are not in the center of the distribution, where the data density is low, were selected. For the graphs in each row, the sample size N is fixed. Each statistic is calculated with local polynomials of order $p = 4$ and with a fixed number of quantiles $Q = 100$. The black solid line shows the density of a standard normal distribution for comparison.



(a) S_{12} **(b)** S_{23} **(c)** S_{43}

(d) S_{12} **(e)** S_{23} **(f)** S_{43}

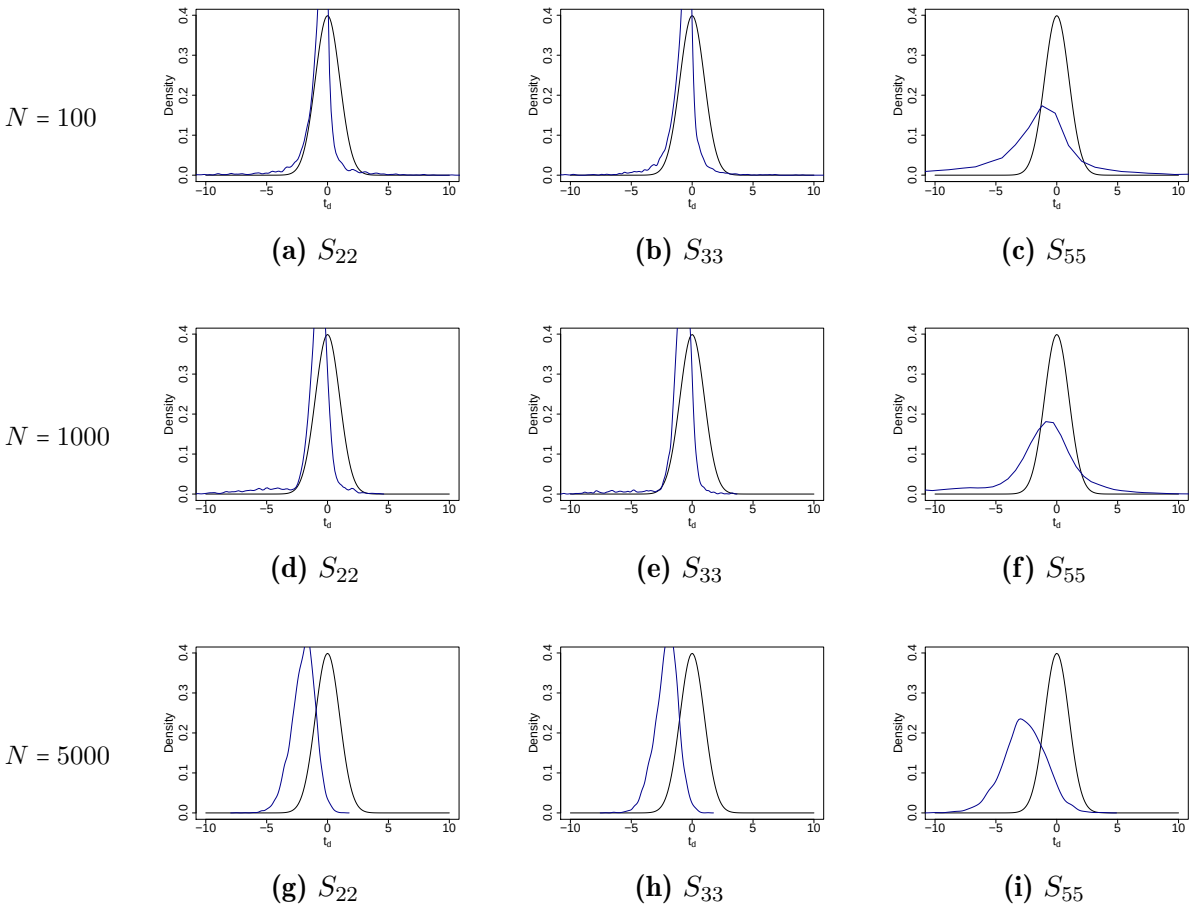
(g) S_{12} **(h)** S_{23} **(i)** S_{43}

distribution – truncated at the upper tail. Taking such a truncation into account, the variance of the estimate would be, in the very center of the joint distribution at S_{33} estimated quite accurately. At the outter rim of the distribution, at S_{55} the variance is slightly overestimated.

As can be seen in Figure 5.5, at the off-center points S_{12}, S_{23} and S_{43} the truncation of the density estimates at 0 becomes visible. At S_{12}, the estimates are all zero. Therefore, the distribution of the test statistic at this point is degenerate. The test statistic distributions at S_{23} are asymmetric and have similarities to truncated normal distributions – truncated at the lower tail.

Figure 5.6: Test Statistics for Conditional Densities (Center)

The panel shows kernel density estimates of 5000 simulated test statistics (blue solid) at the respective points S_{ij} of Φ_{XY} listed in Table 5.2. For this figure, only points that are in the center of the distribution, where the data density is high, were selected. For the graphs in each row, the sample size N is fixed. Each statistic is calculated with local polynomials of order $p = 4$ and with a fixed number of quantiles $Q = 100$. The black solid line shows the density of a standard normal distribution for comparison.



(a) S_{22} **(b)** S_{33} **(c)** S_{55}

(d) S_{22} **(e)** S_{33} **(f)** S_{55}

(g) S_{22} **(h)** S_{33} **(i)** S_{55}

With the exception of the distribution at S_{23}, the estimates seem unbiased or exhibit a neglectable positive bias. Taking into account that the estimates are in-fact truncated at zero and the similarity to the truncated normal distribution is expected, the variance estimates seem to accurately reflect the true variance of the estimates.

In conclusion, depending on the location on the conditional density curve, the bias of the estimate for the conditional density is either negative (closer to the center) or positive (closer to the tails). Also, the less dense the neighborhood of the sampling point is, i.e., the further we move to the rim of the distribution, the more pronounced becomes the location dependent bias for the reported estimates. The effect on relative entropy measure estimates is not clear, as the slight bias in different directions at different locations of the conditional density distribution maybe netted out. Nevertheless, these findings also have some importance when analyzing the relative entropy measure estimates.

5.2.2 Simulation of Mutual Information

In order to check the above derivations of the asymptotic distributions, I repeatedly simulate a z-score test statistic of the form

$$t_I = \frac{\hat{I} - I_0}{\sqrt{\text{var}(\hat{I})}} \sim \mathcal{N}(0,1)\,.$$

Especially the null hypothesis $H_0 : I = 0$, i.e., the hypothesis that the involved variables are stochastically independent, is interesting for practical research. The ex-ante expectation would be, that if the estimate as well as the estimate for its variance are consistently estimated, the test-statistic is standard normally distributed. However, as can be seen in Section 5.2.1, the density estimates are slightly biased in different directions in different locations of the distribution. In order to estimate $\hat{I}$ according to the procedure set out in Section 5.1.2 using the techniques from Section 5.1, I use local polynomials of order 4 with estimated, stochastic variable bandwidth. I fix the number of quantiles to $Q = 100$, the number of ghost points to 100 and the parameter is also set to $G = 100$.

The MI estimate is calculated on samples from different distributions. First, I consider 2-, 3- and 5-dimensional normal distributions. In each case, I generate two samples. One from an independent and one from a dependent joint normal distribution. The realisation of the random variables collected in the matrix $\mathbf{Y}$ are generated in the form

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is a $k \times N$ matrix while each element is independently drawn from a standard normal distribution. $\boldsymbol{\mu}$ is a $k \times 1$ vector that determines the mean of the simulated distribution and the $k \times k$ matrix $\boldsymbol{B}$ scales the joint distribution and determines the variance. If $\boldsymbol{B}$ is a diagonal matrix, the k variables are independent.

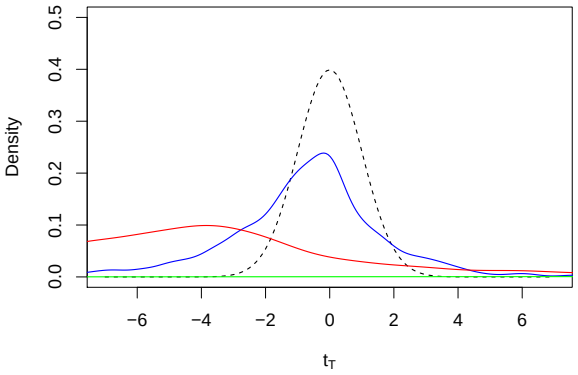
Second, I also explore the behavior of the method for realisations from uniform distributions on spherical shells in 2, 3 or 5 dimensions. For this purpose, I use the method proposed by Marsaglia (1972) where one draws k Euclidean coordinates $\mathbf{X} = (x_1, x_2, \ldots x_k)'$ from a standard normal distribution. To generate the realisations within a spheric shell, one then calculates

$$\mathbf{Y} = \left(u\left(1 - \frac{r_i}{r_o}\right) + \frac{r_i}{r_o}\right)^{\frac{1}{k}} \frac{r_o \mathbf{X}}{\sqrt{x_1^2 + x_2^2 + \ldots x_k^2}}$$

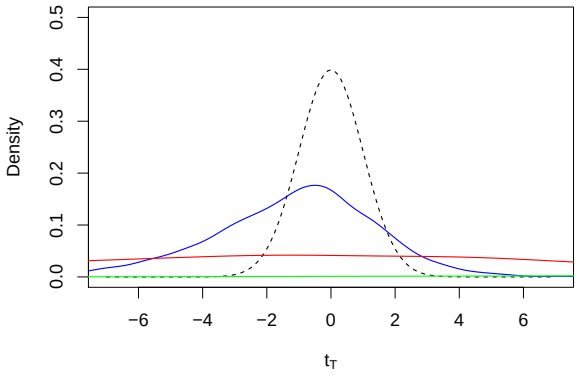
where u is a uniformly distributed random number $u \sim \mathcal{U}(0,1)$. Furthermore, r_i is the radius of the inner sphere and r_o the radius of the outer sphere of the shell. In the simulations, I fix the inner radius to $r_i = 5$ and the outer radius to $r_o = 10$.

Figure 5.7: Simulated Test Statistics for Mutual Information

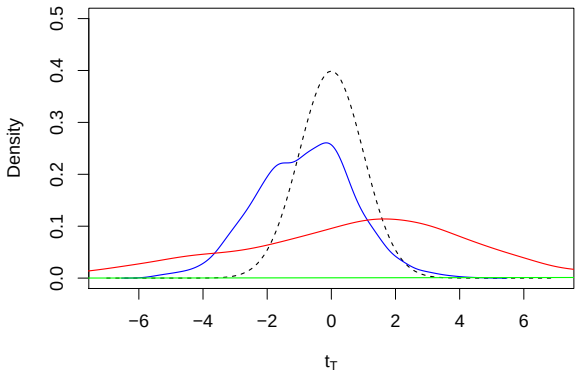
The Figure 5.7a, 5.7b and 5.7c show the kernel density estimates of 2000 simulated test statistics t_I for different distributions of 2, 3 and 5 random variables, respectively. The variables are either drawn from a uniform distribution of a n-dimensional sphere (red lines) or they are jointly normally distributed and independent (blue lines) or dependent (green lines). Hence, the blue line represents the kernel density of the independent case for which $MI = 0$ is the true value while for the jointly normal but dependent and the n-sphere case the true MI value is not 0.



(a) 2 Dimensions



(b) 3 Dimensions



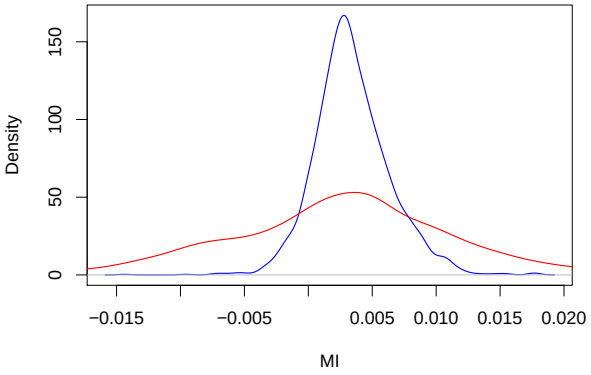
(c) 5 Dimensions

The simulation results are presented in the form of plotted kernel densities of the test statistics in Figure 5.7 and the MI values in Figure 5.8. As one can see from the distribution of 2000 simulated test statistics, under the null hypothesis of no MI the distribution is not standard normal. The estimated variance is too small and the MI estimates are slightly biased positively. Depending on the number of variables involved, the distribution has either a slight negative or a slight positive skewness. Also the test statistic does not keep its size across dimensions. Different numbers of variables result in different quantiles.

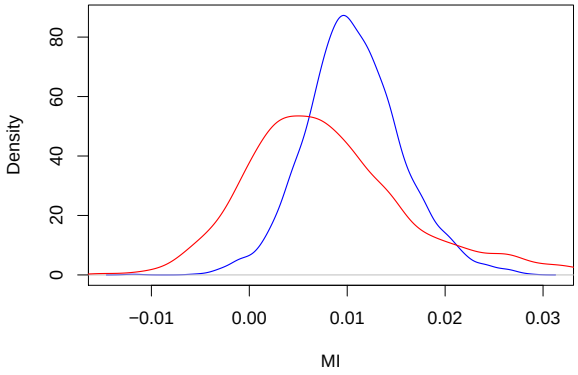
Nevertheless, one can distinguish the case in which no MI is present from the case where there is a connection between the variables by a very conservative rule of thumb. If the test statistic is below -6 or above 4, one could reject the null hypothesis of no MI at least on a two-sided 10%-significance level.

Figure 5.8: Simulated Values for Mutual Information

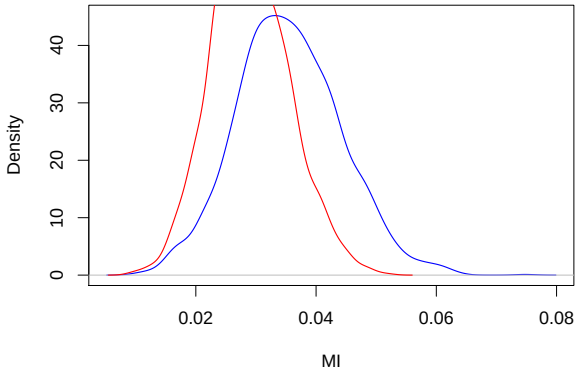
The Figure 5.7a, 5.7b and 5.7c show the kernel density estimates of 2000 simulated estimates for mutual information for different distributions of 2, 3 and 5 random variables, respectively. The variables are either drawn from a uniform distribution of a n-dimensional sphere (red lines) or they are jointly normally distributed and independent (blue lines) or dependent (green lines). Hence, the blue line represents the kernel density of the independent case for which $MI = 0$ is the true value while for the jointly normal but dependent and the n-sphere case the true MI value is not 0.



(a) 2 Dimensions



(b) 3 Dimensions



(c) 5 Dimensions

The detection of no MI is not be possible based on the value of the MI alone, as can be seen in Figure 5.8. While in Figure 5.8 the case for the dependent multivariate normal distribution is outside the plotted range (theoretically a green line), the MI values for the variabels uniformly distributed on a spherical shell are close to zero and exhibit a similar distribution as the case of independent normal variables (blue line). The associated variance of the estimate and the constructed test statistic can help to better distinguish dependent from independent random variables. Nevertheless, in order to improve the variance estimates and map the behavior of the test-statistic across different numbers of observations, further research is necessary.

5.2.3 Simulation of Transfer entropy

To assess the asymptotic theory for TE, I simulate three systems of time series and estimate $T_{X\to Z}$ in each of the three systems. Again, I consider a z-score test statistic of the form

$$t_T = \frac{\hat{T}_{X\to Z} - T_{X\to Z}}{\sqrt{\text{var}(\hat{T}_{X\to Z})}} \sim \mathcal{N}(0,1)\,.$$

Again the null hypothesis $H_0 : T_{X\to Z} = 0$, i.e., the hypothesis of no predictive ability of X on Z, is interesting for practical research and, thus, the simulations solely concentrate on the simulation of this hypothesis. If the estimate as well as the estimate for its variance are unbiased, the test-statistic is standard normally distributed.

I estimate $T_{X\to Z}$ according to the procedure set out above with local polynomial regression of order 4 on quantiles estimated via quantile regression. I fix the number of quantiles to $Q = 100$, the number of ghost points to 100 and the parameter is also set to $G = 100$.

For the simulation, I consider three time series whose specifications are inspired by Papana, Kyrtsou, Kugiumtzis, Diks et al. (2013). The first is a system of stationary AR(1) processes in which the TE is obviously 0:

$$\begin{aligned} y_t &= 0.5y_{t-1} + \varepsilon_{1t}, \\ x_t &= 0.7x_{t-1} + \varepsilon_{2t}, \\ z_t &= 0.3z_{t-1} + \varepsilon_{3t}, \end{aligned} \tag{5.28}$$

where $\varepsilon_{it} \sim \mathcal{N}(0,1)\ \forall i \in \{1,2,3\}$. I refer to this system in the following as the independent case.

The second is a nonlinear auto-regressive system in which a linear dependence between x_{t-1} and z_t is somewhat hidden

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.001\sqrt{|2 - y_{t-1}|} - 0.1\exp(0.1y_{t-1}^2) + 0.4\varepsilon_{1t}, \\ x_t &= 0.9x_{t-1} + 0.001\sqrt{|2 - x_{t-1}|} - 0.1\exp(0.1x_{t-1}^2) + 0.5y_{t-1} + 0.4\varepsilon_{2t}, \\ z_t &= 0.4z_{t-1} + 0.001\sqrt{|2 - y_{t-1}|} - 0.1\exp(0.1y_{t-1}^2) + 0.5y_{t-1} + 0.3x_{t-1} + 0.4\varepsilon_{3t}, \end{aligned} \tag{5.29}$$

where again $\varepsilon_{it} \sim \mathcal{N}(0,1)\ \forall i \in \{1,2,3\}$. Since I investigate whether the linear relation between z_t and x_{t-1} is detected by the estimated TE measure, I refer to this system in the following as the linear case. Clearly, there is also a non-linear relation between y_{t-1} and z_t, however this has no consequence for the linearity of the relationship between x_{t-1} and z_t.

The third system is a nonlinear auto-regressive system. Again, I focus on the connection between x_{t-1} and z_t which in this case is a nonlinear one.

$$\begin{aligned} y_t &= 3.4 y_{t-1}(1 - y_{t-1})^2 \exp(-y_{t-1}^2) + 0.4\varepsilon_{1t}, \\ x_t &= 3.4 x_{t-1}(1 - x_{t-1})^2 \exp(-x_{t-1}^2) + 0.5 y_{t-1} x_{t-1} + 0.4\varepsilon_{2t}, \\ z_t &= \min(\max(x_{t-1}^5, -2), 2)(1 - z_{t-1})^2 \exp(-z_{t-1}^2) + 0.03 z_{t-1} y_{t-1}^2 + 0.4\varepsilon_{3t}, \end{aligned} \tag{5.30}$$

and as before $\varepsilon_{it} \sim \mathcal{N}(0,1)\ \forall i \in \{1,2,3\}$. I refer to this system as the non-linear case.

The results of the simulation are presented in form of kernel densities of the simulated test statistics and in Figure 5.10. As can be seen for the simulated systems, the kernel density estimate of the test statistic under the null hypothesis of no TE is close to, but not identical with, the density of a standard normal distribution. The test statistic is negatively biased while the actual TE values are biased positively. Also the test statistic exhibits more outliers and has a somewhat higher probability mass at the tails.

Nonetheless, in order to distinguish the case of no TE from $X \to Z$ from a case where X is able to help predict Z, both the test statistic as well as the raw TE values are helpful. From the simulations, as a rule of thumb, one can state that a test statistic outside the range between -2.3 and 1.9 would with a probability of less than 10% be associated to a situation where no TE is present.

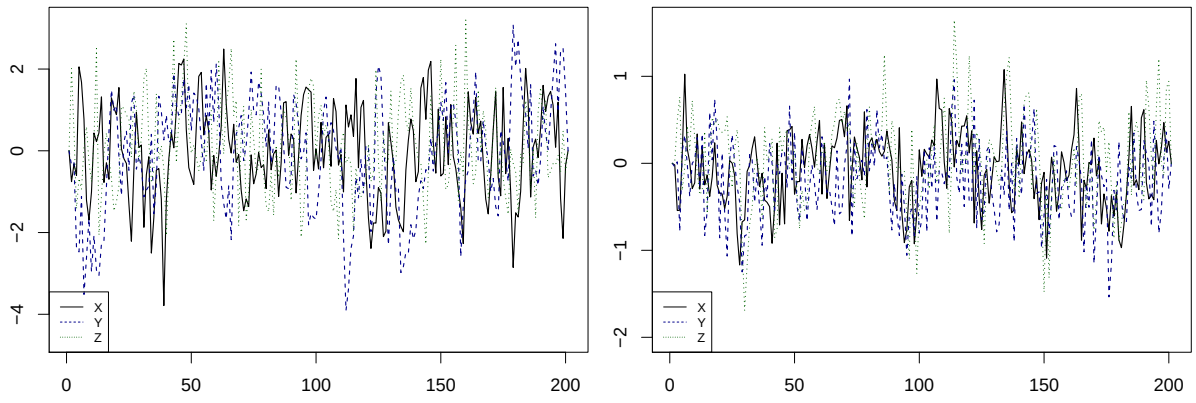
Nevertheless, in the case of TE, it seems the TE values alone are more distinct in distinguishing the case of some predictive power of X on Z from the case of no TE. However, no clear guidance is possible at this point in whether TE values of below or above ± 0.02 can be conceived as a clear signal of non-zero TE. In order to improve the estimates for the variance of the TE estimator and analyse the behavior of the TE test-statistic for different numbers of observations, further research is needed.

5.3 Empirical Applications

In this section, two applications of TE in empirical finance are presented. In the first application, I analyze the same dataset as Dimpfl and Peter (2013) on information flows between the market for Credit Default Swaps (CDS) and bond markets. The second application focuses on the impact of the financial crisis on transatlantic information flows between stock indices, using the same dataset as Dimpfl and Peter (2014). While the TE estimates of both studies are based on a symbolic encoding, i.e., a discrete binning of the return time series, I use the quantile regression methodology presented in Section 5.1 to investigate whether TE can be identified for the entire support of the data.

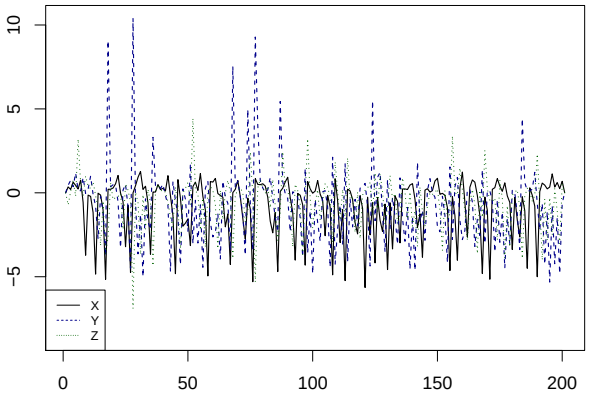
Figure 5.9: Simulated Time Series for TE Estimation

The graph depicts samples of 100 observations from the systems of auto-regressive time series under consideration. On the horizontal axis is the time index and on the vertical axis the value of the time series.



(a) Independent

(b) Linear



(c) Non-Linear

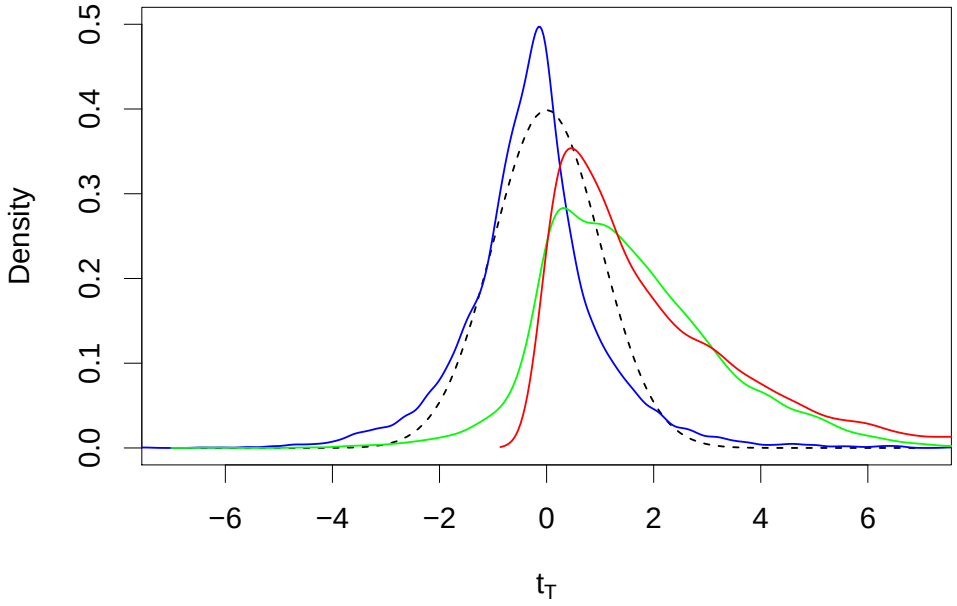
5.3.1 Credit Default Swaps (CDS) and Bond markets

Using discretized time series, Dimpfl and Peter (2013) employ symbolic transfer entropy to measure the information flow between CDS and bond markets. For this purpose they consider the difference between the yield of a certain bond and the currently risk free rate as mainly attributed to the credit risk, i.e, the risk that the issuer of the bond will fail to pay the outstanding commitments and default. The difference is also called the credit spread (CS). Inflation risk, liquidity risk, the risk that a holder of the bond has to reinvest because the contract is terminated somehow earlier than agreed or other risks are not considered dominant and are not controlled for.

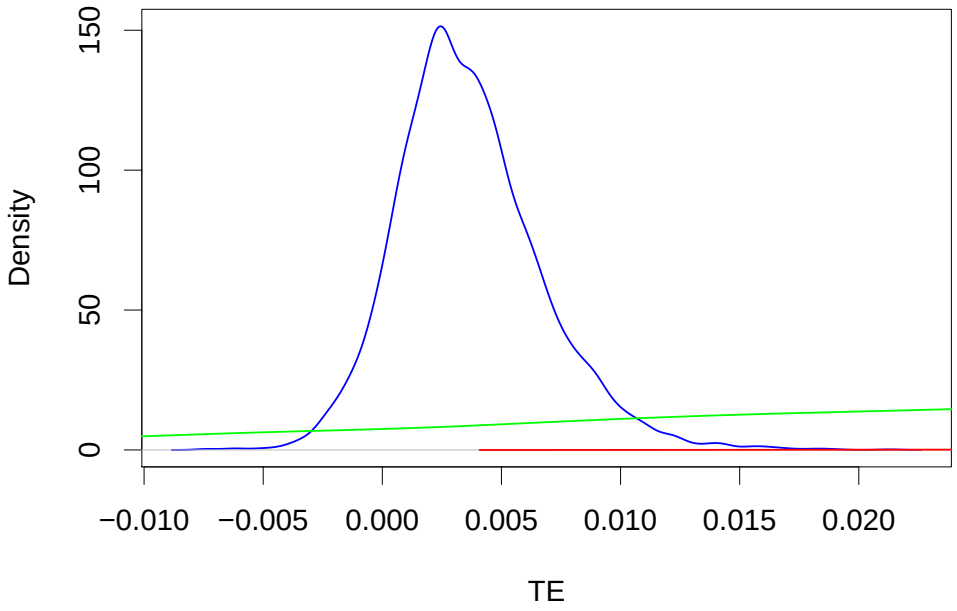
As another measure for the default risk of a certain issuer, the authors consider the CDS premium. With CDS the credit risk associated to the bond issuer can be traded, if such CDS are available for the specific issuer. Since deviation of the CDS premium from the

Figure 5.10: Simulated Transfer Entropy

The top figure shows the kernel density estimates of 5000 simulated test statistics t_T for the three time series systems in Equations (5.28) (the independent case), (5.29) (the linear case) and (5.30) (the non-linear case) are depicted. The blue line represents the kernel density of the independent case for which $TE = 0$ is true. The red and the green line show the densities for the linear and non-linear systems, respectively. For both, $TE \neq 0$. The coloring applies as well to the bottom graph. It shows the kernel density of the actual estimated TE values.



(a) Kernel Density Estimations of Test Statistic



(b) Kernel Density Estimations of TE Values

CS may present an arbitrage opportunity, CDS premium and the CS are usually modeled in the literature as co-integrated process in a Vector Error Correction Model (VECM).

Dimpfl and Peter (2013) encode the times series of changes in the CDS premium as well as the time series of changes in the CS in three discrete categories. All observations below the estimated unconditional 5%-quantile belong to the first bin and are encoded as 1 and observations between the 5%- up to the 95%-quantile belong to the second bin and are encoded with 2. All other observations above the unconditional 95%-quantile are encoded as 3 in the third bin. With these strongly coarse-grained time series, they find TE in both directions. Hence, they find that knowing in which bin the change of CS was yesterday helps to forecast the bin in which the CDS premium will be today, and vice-versa.

With the method developed in this chapter, I reanalyze the same time series Dimpfl and Peter (2013) used and explore whether their result can be generalized to the entire, continuous support of the CS and CDS premium. The results are presented in Table 5.3. As one can see, I find that in general neither changes in the CDS premium help in predicting changes in the CS nor is today's CS helpful in predicting tomorrows CDS. This is not in direct opposition to the results of Dimpfl and Peter (2013), since their time series was encoded. On the contrary, it rather hints to the informativeness of tail events. However, their results do not generalize to the continuous support of the underlying random variables.

5.3.2 Transatlantic Information Flows

Based on one minute intraday returns of the European, blue chip stock market indices the German DAX30, the British FTSE50 and the French CAC40, Dimpfl and Peter (2014) analyse the TE of these markets with the American S&P500 index. For the DAX, the data sample spans the years 2003 until 2010. For the FTSE and the CAC, the years 2006 until 2010 are covered. The S&P data are available for both periods. With regard to the financial crisis 2008, they subdivide their sample into a pre-crisis, crisis and post-crises period. Based on their data set I only estimate the TE for the entire sample and omit the partition of the sample with regard to the financial crisis. For all periods as well as for the entire sample, Dimpfl and Peter (2013) find significant symbolic transfer entropy. Knowing in which bin the return one minute ago was, helps to predict the current one-minute return.

In contrast to Dimpfl and Peter (2014), I not discretize the return series of the indices into three bins. I use the techniques developed in this chapter to explore whether the results for the encoded time series of Dimpfl and Peter (2014) can be generalized to the entire continuous support of the respective return series and whether TE between the markets can be detected.

Table 5.3: Results: Transfer Entropy CDS and CS

The table shows the estimated transfer entropy values in the second and fourth column. The corresponding test statistics are reported in the first and third column. The simulation study above indicates that values below -2.3 and above 1.9 are rather rare and can be considered as significant deviations from the null hypothesis.

	$t_{T,\text{CDS}\to\text{CS}}$	$\hat{T}_{\text{CDS}\to\text{CS}}$	$t_{T,\text{CS}\to\text{CDS}}$	$\hat{T}_{\text{CS}\to\text{CDS}}$
Allianz	-0.0015	-0.0004	0.0327	0.0231
BASF	0.0070	-0.0017	0.1954	-0.0062
Bayer	0.0989	0.0055	0.0028	0.0019
BMW	-0.0512	0.0063	0.2204	0.0085
Carrefour	-0.1443	-0.0237	0.0001	-0.0067
Deutsche Telekom	0.0002	0.0140	0.0322	0.0122
Electricité de France	-0.0016	0.0020	-0.0000	0.0006
Enel	0.0000	0.0196	-0.0192	-0.0015
Fortum Oyi	0.0000	0.0071	0.0030	0.0088
France Télécom	0.0000	0.0128	0.1419	0.0055
GDF Suez	-0.0000	0.0019	-0.1117	-0.0053
Iberola	0.0636	0.0102	0.0509	0.0111
Koninklijke KPN	-0.0282	0.0117	0.8598	0.0076
LVMH	-0.0000	0.0076	0.3342	0.0142
Metro	0.5125	0.0099	0.1878	0.0098
ArcelorMIttal	-0.0000	0.0138	**-7.7769**	0.0069
National Grid	-0.1304	-0.0014	0.1796	0.0079
Repsol	0.0027	0.0026	0.0402	0.0012
RWE	0.0116	-0.0006	0.4038	0.0214
St. Gobain	-0.0000	0.0018	-0.0110	0.0044
Solvay	-0.0418	-0.0003	0.0008	-0.0078
Banco Santander Central Hispano	0.1395	0.0078	0.3026	0.0095
Telefonica	0.0186	-0.0068	0.2699	0.0165
Telecom Italia	0.0000	0.0472	0.2611	0.0098
Vattenfall	-0.0000	-0.0093	-0.0198	-0.0026
Veolia	-0.0000	0.0007	0.1663	-0.0077
VW	-0.0000	0.0100	-0.0011	0.0043

Table 5.4: Results: Transfer Entropy Transatlantic Information Flow

The table shows the estimated transfer entropy values in the second and fourth column. The corresponding test statistics are reported in the first and third column. The simulation study above indicates that values below -2.3 and above 1.9 are rather rare and can be considered as significant deviations from the null hypothesis.

	$t_{T,\text{EU}\to\text{US}}$	$\hat{T}_{\text{EU}\to\text{US}}$	$t_{T,\text{US}\to\text{EU}}$	$\hat{T}_{\text{US}\to\text{EU}}$
DAX	**52.3082**	0.0411	1.4214	-0.0006
CAC	-0.1721	0.0150	**5.0460**	-0.0003
FTSE	-1.8607	0.0081	0.7634	0.0038

The results are presented in Table 5.4. As one can see, only for the DAX, the TE value is both distinctly different from zero (with 0.041) and the value of the z-score is higher than the critical value obtained in the simulation (with 52.3). This is an interesting result, since it indicates that during the underlying sample period, one minute returns from the German DAX were able to help predict the one-minute returns of the S&P500. This highlights the importance of the German market.

In all other cases, the result is not as clear. Even though the z-score for the CAC indicates a reversed information flow from the S&P to the CAC the TE is with -0.0003 rather close to zero. Even though the sample size is large with more than 100.000 observations, further analysis of the test statistic for very large samples and values of TE close to zero need to be undertaken in order to ensure that the test statistics keeps its size with a growing sample size.

5.4 Summary

In this chapter I consider the estimation of MI based on density estimates obtained via quantile regression. This approach avoids the necessary binning of continuous data which is usually based on not data-driven choices of bin limits. Furthermore, I expect the approach to require less data in complex settings than required for estimation of conditional frequencies. Also, the computational intensity is lower compared to kernel density based estimation of MI. My results indicate that testing MI and TE estimates on up to five random variables from a sample with around 1.000 data points is sensible and possible. There are several issues left that need to be analyzed more closely. The behavior of the test statistic needs a more thorough examination. Namely, I have shown that the test statistic does not keep its size and power when the number of involved random variables is altered. Further research could focus on the question whether there is a missing factor that may stabilize the distribution across dimensions. The question, how an increasing

number of observations ameliorates the power of the test statistic, needs to be analyzed more closely.

In the empirical applications analyzed in Section 5.3, I have shown that results obtained with symbolic transfer entropy on encoded time series are different from a TE estimate that uses the entire continuous support of the underlying data. The symbolic transfer results based on the encoding of tail events, i.e., events below the 5%- or above the 95%-quantile do not generalize to the entire support of the time series at hand and may lead to the impression that some markets may contain predictive information for others when only the tail events may contain some predictive information. Note that also the case where a tail event in one time series is more often followed by an observation close to the median in the other time series than the unconditional probability would suggest (i.e., the probability without knowledge of the tail event would suggest), may also be classified as predictive information. If this were the case, this may serve as a further justification for an error correction model for the two time series.

All in all, I have shown that quantile regression is a powerful tool and can be used in combination with local polynomial regressions to estimate conditional density as well as relative entropy measures such as MI and TE. Especially, the estimation of TE via smoothed quantile regression estimates can add to the scientific toolbox across disciplines.

Appendix

5.A Calculation of γ_1

5.A.1 Second Order

As introduced by Fan and Marron (1994) and summarized in Fan and Gijbels (1996) (p. 95), the local polynomial estimator as defined in Equation (5.6) (for a second order polynomial smoothing) can be rewritten as

$$\begin{aligned}\hat{\gamma} &= (\boldsymbol{Z}_P'\boldsymbol{W}_P\boldsymbol{Z}_P)^+\boldsymbol{Z}_P'\boldsymbol{W}_P\boldsymbol{\tau} \\ &= \mathbf{S}^{-1}\mathbf{T} \\ &= \begin{pmatrix} S_0 & S_1 & S_2 \\ S_1 & S_2 & S_3 \\ S_2 & S_3 & S_4 \end{pmatrix}^{-1} \begin{pmatrix} T_0 \\ T_1 \\ T_2 \end{pmatrix}\end{aligned} \tag{5.31}$$

where

$$S_j = \sum_{l=1}^{Q} K_h \left\{ \sum_{m=1}^{K} \hat{\theta}_{lm} x_{0,m} - y_0 \right\} \left(\sum_{m=1}^{K} \theta_{lm} x_{0,m} - y_0 \right)^j \tag{5.32}$$

$$T_j = \sum_{l=1}^{Q} K_h \left\{ \sum_{m=1}^{K} \hat{\theta}_{lm} x_{0,m} - y_0 \right\} \left(\sum_{m=1}^{K} \theta_{lm} x_{0,m} - y_0 \right)^j \tau_l \tag{5.33}$$

where I have used the previous notation introduced in section 5.1.1. In order to simplify the notation I shortened the kernel function $K_h\{\cdot\} = K(\cdot/h)/h$. Also note that in the above equation, h depends on the choice of $\mathbf{x}_0$, the estimates $\hat{\boldsymbol{\theta}}$ and y_0. The inverse of $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \frac{1}{\det(\mathbf{S})} \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix}'$$

where

$$
\begin{aligned}
C_{11} &= S_2S_4 - S_3^2 \\
C_{12} &= -S_1S_4 + S_2S_3 \\
C_{13} &= S_1S_4 - S_2^2 \\
C_{21} &= -S_1S_4 + S_3S_2 \\
C_{22} &= S_0S_4 - S_2^2 \\
C_{23} &= -S_0S_3 + S_2S_1 \\
C_{31} &= S_1S_3 - S_1S_2 \\
C_{32} &= -S_0S_3 + S_1S_2 \\
C_{33} &= S_0S_2 - S_1^2
\end{aligned}
$$

Note that $C_{ij} = C_{ji}$. Also, note that for the calculation of γ_1 I only need the co-factors in the second column (or row) of the co-factor matrix. Thus, the estimate for γ_1 can be written as

$$
\begin{aligned}
\hat{\gamma}_1 &= \frac{\sum_{k=1}^{3} T_{k-1}C_{2k}}{\sum_{v=1}^{3} S_vC_{2v}} \\
&= \frac{-S_1S_4T_0 + S_3S_2T_0 + S_0S_4T_1 - S_2^2T_1 - S_0S_3T_2 + S_2S_1T_2}{S_0S_2S_4 + 2S_1S_2S_3 - S_1^2S_4 - S_0S_3^2 - S_2^3}
\end{aligned} \tag{5.34}
$$

5.A.2 Third Order

For the third order approximation, the general structure is maintained. However, the matrix $\mathbf{S}$ is then a 4×4 matrix. Also note that the entry at the i^{th} row and j^{th} column of $\mathbf{S}$ in general is S_{i+j-2}. Since I am only interested in the second entry of $\boldsymbol{\gamma}$ and only need the cofactors C_{2k}, to save unnecessary computations one can choose to use Laplace's expansion formula along the second row to calculate the determinant

$$
\det \mathbf{S} = \sum_{j=1}^{4} S_jC_{2j}
$$

where C_{2j} are the entries in the second row of the co-factor matrix $\mathbf{C}$ of $\mathbf{S}$. Again note that I have $C_{2j} = C_{j2}$.

$$
\begin{aligned}
C_{12} = C_{21} &= -\left(S_1S_4S_6 + S_2S_5S_4 + S_3S_3S_5 - S_3S_4S_4 - S_2S_3S_6 - S_1S_5S_5\right) \\
C_{22} &= \left(S_0S_4S_6 + S_2S_5S_3 + S_3S_2S_5 - S_3S_4S_3 - S_2S_2S_6 - S_0S_5S_5\right) \\
C_{32} = C_{23} &= -\left(S_0S_3S_6 + S_1S_5S_3 + S_3S_2S_4 - S_3S_3S_3 - S_1S_2S_6 - S_0S_5S_4\right) \\
C_{42} = C_{24} &= \left(S_0S_3S_5 + S_1S_4S_3 + S_2S_2S_4 - S_2S_3S_3 - S_1S_2S_5 - S_0S_4S_4\right)
\end{aligned}
$$

For the estimate of $\hat{\gamma}_1$, I get similar to the second order case

$$\hat{\gamma}_1 = \frac{\sum_{k=1}^{4} T_{k-1} C_{2k}}{\sum_{v=1}^{4} S_v C_{2v}}$$

5.B Calculation of the Derivative of γ_1 with Respect to θ_{lm}

5.B.1 Second Order

Deriving Equation (5.34) with respect to θ_{lm} for the two order case is given, generally, by

$$\frac{\partial \gamma_1}{\partial \theta_{lm}} = \frac{1}{\sum_{v=1}^{3} S_v C_{2v}} \sum_{k=1}^{3} \left(\partial T_{k-1} - \gamma_1 \partial S_k\right) C_{2k} + \left(T_{k-1} - \gamma_1 S_k\right) \partial C_{2k}$$

or in detail

$$\frac{\partial \gamma_1}{\partial \theta_{lm}} = \frac{\begin{pmatrix} -\partial S_1 S_4 T_0 - S_1 \partial S_4 T_0 - S_1 S_4 \partial T_0 + \partial S_3 S_2 T_0 + S_3 \partial S_2 T_0 + S_3 S_2 \partial T_0 \\ +\partial S_0 S_4 T_1 + S_0 \partial S_4 T_1 + S_0 S_4 \partial T_1 - 2 S_2 \partial S_2 T_1 - S_2^2 \partial T_1 \\ -\partial S_0 S_3 T_2 - S_0 \partial S_3 T_2 - S_0 S_3 \partial T_2 + \partial S_2 S_1 T_2 + S_2 \partial S_1 T_2 + S_2 S_1 \partial T_2 \end{pmatrix}}{S_0 S_2 S_4 + 2 S_1 S_2 S_3 - S_1^2 S_4 - S_0 S_3^2 - S_2^3}$$

$$- \gamma \frac{\begin{pmatrix} \partial S_0 S_2 S_4 + S_0 \partial S_2 S_4 + S_0 S_2 \partial S_4 \\ +2(\partial S_1 S_2 S_3 + S_1 \partial S_2 S_3 + S_1 S_2 \partial S_3) \\ -2 S_1 \partial S_1 S_4 - S_1^2 \partial S_4 - \partial S_0 S_3^2 \\ -2 S_0 S_3 \partial S_3 - 3 S_2^2 \partial S_2 \end{pmatrix}}{(S_0 S_2 S_4 + 2 S_1 S_2 S_3 - S_1^2 S_4 - S_0 S_3^2 - S_2^3)} \tag{5.35}$$

where I have omitted to extensively write the derivatives. Written more extensively, these derivatives are

$$\partial S_0 = \frac{\partial S_0}{\partial \theta_{lm}} = \partial K_h$$

$$\partial S_j = \frac{\partial S_j}{\partial \theta_{lm}} = \left(\sum_{m=1}^{K} \theta_{lm} x_{0,m} - y_0\right)^j \partial K_h + K_h \left\{\sum_{m=1}^{K} \hat{\theta}_{lm} x_{0,m} - y_0\right\} j \left(\sum_{m=1}^{K} \theta_{lm} x_{0,m} - y_0\right)^{j-1} x_{0m}$$

$$\partial T_j = \frac{\partial T_j}{\partial \theta_{lm}} = \partial S_j \tau_l$$

$$\partial K_h = \frac{\partial K_h}{\partial \theta_{lm}} = -\frac{1}{h^2} K \left\{\frac{\sum_{m=1}^{K} \hat{\theta}_{lm} x_{0,m} - y_0}{h}\right\} \frac{\partial h}{\partial \theta_{lm}} + \frac{1}{h} \left(\frac{x_{0m}}{h} - \frac{\sum_{m=1}^{K} \hat{\theta}_{lm} x_{0,m} - y_0}{h^2} \frac{\partial h}{\partial \theta_{lm}}\right) \partial K$$

$$\frac{\partial h}{\partial \theta_{lm}} = 1.39 \cdot C_{1,4}(K) \frac{Q^{-1/11}}{11} \left[\frac{\breve{\sigma}^2(y_0)}{\{F^{(p+1)}(y_0)\}^2 f(y_0)}\right]^{-10/11}$$
$$\frac{\frac{\partial \breve{\sigma}^2(y_0)}{\partial \theta_{lm}} \{F^{(p+1)}(y_0)\}^2 f(y_0) - \breve{\sigma}^2(y_0) \frac{\partial \{F^{(p+1)}(y_0)\}^2 f(y_0)}{\partial \theta_{lm}}}{\left(\{F^{(p+1)}(y_0)\}^2 f(y_0)\right)^2}$$

where ∂K is the derivative of the kernel function with respect to θ_{lm}. Depending on the choice of the kernel function $K\{\cdot\}$ the above derivatives have to be adjusted accordingly. For the derivatives of the preliminary estimate for the conditional variance $\frac{\partial \breve{\sigma}^2(y_0)}{\partial \theta_{lm}}$ recalling the definition from Equation (5.14) is helpful.

$$\breve{\sigma}^2(y_0) = \frac{1}{Q-p-3-1}\left(\boldsymbol{\tau}'\boldsymbol{\tau} - \boldsymbol{\tau}'\breve{\boldsymbol{Z}}\left(\breve{\boldsymbol{Z}}'\breve{\boldsymbol{Z}}\right)^{-1}\breve{\boldsymbol{Z}}'\boldsymbol{\tau}\right)$$

Denoting $\breve{\boldsymbol{T}} = \breve{\boldsymbol{Z}}'\boldsymbol{\tau}$ and $\breve{\boldsymbol{S}} = \left(\breve{\boldsymbol{Z}}'\breve{\boldsymbol{Z}}\right)^{-1}$ the derivative of $\breve{\sigma}^2(y_0)$ can be derived to be

$$\frac{\partial \breve{\sigma}^2(y_0)}{\partial \theta_{lm}} = -\frac{1}{Q-p-3-1}\left(\frac{\partial \breve{\boldsymbol{T}}'}{\partial \theta_{lm}}\breve{\boldsymbol{S}}^{-1}\breve{\boldsymbol{T}} + \breve{\boldsymbol{T}}'\breve{\boldsymbol{S}}^{-1}\frac{\partial \breve{\boldsymbol{S}}}{\partial \theta_{lm}}\breve{\boldsymbol{S}}^{-1}\breve{\boldsymbol{T}} + \breve{\boldsymbol{T}}'\breve{\boldsymbol{S}}^{-1}\frac{\partial \breve{\boldsymbol{T}}}{\partial \theta_{lm}}\right)$$

where $\frac{\partial \breve{\boldsymbol{S}}}{\partial \theta_{lm}}$ and $\frac{\partial \breve{\boldsymbol{T}}}{\partial \theta_{lm}}$ signify the derivatives of the entries of the vector $\breve{\boldsymbol{T}}$ and the matrix $\breve{\boldsymbol{S}}$. The entries are defined as in Equation (5.31) without the weighting of the kernel function. Their derivatives are similar to ∂S_j and ∂T_j above, again without the part that is due to the kernel function. The derivative for $\{F^{(p+1)}(y_0)\}^2 f(y_0)$ follows the same logic as the derivatives before.

5.B.2 Third Order

Again, the main structure of the derivative with respect to θ_{lm} remains the same. Thus, I get

$$\frac{\partial \gamma_2}{\partial \theta_{lm}} = \frac{1}{\sum_{v=1}^{4} S_v C_{2v}} \sum_{k=1}^{4} \left(\partial T_{k-1} - \gamma_1 \partial S_k\right) C_{2k} + \left(T_{k-1} - \gamma_1 S_k\right) \partial C_{2k}$$

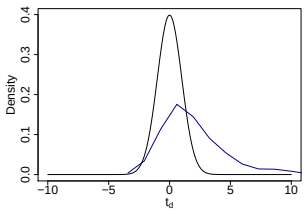
With the above, the detailed structure of $\partial \gamma_1$ is clear.

5.C Additional Graphs: Density Test Statistic Simulations

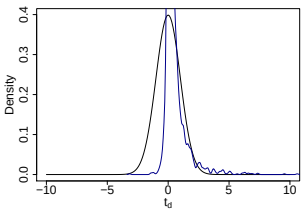
Figure 5.C.1: Test Statistics for Conditional Densities (Off-Center)

The panel shows kernel density estimates of 5000 simulated test statistics (blue solid) at the respective points S_{ij} of Φ_{XY} listed in Table 5.2. For this figure, only points that are not in the center of the distribution, where the data density is low, were selected. For the graphs in each row, the sample size N is fixed. Each statistic is calculated with local polynomials of order $p = 4$. The number of quantiles is chosen according to the rule in Equation (5.27). The black solid line shows the density of a standard normal distribution for comparison.

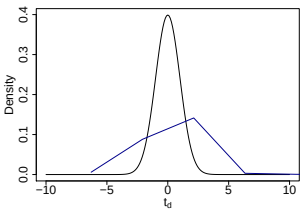
$N = 100$



(a) S_{12}

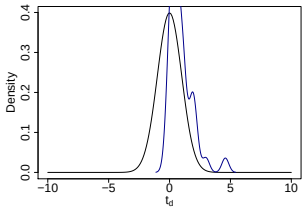


(b) S_{23}

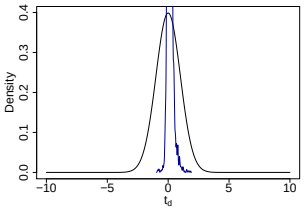


(c) S_{43}

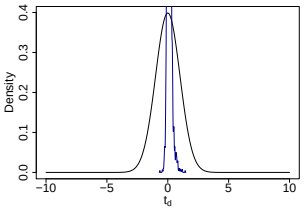
$N = 1000$



(d) S_{12}

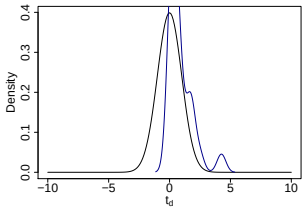


(e) S_{23}

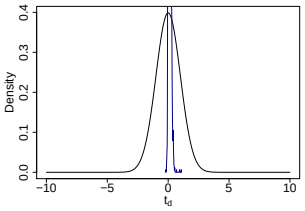


(f) S_{43}

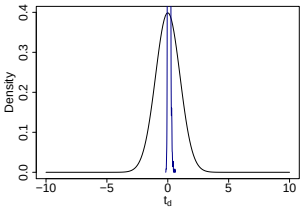
$N = 5000$



(g) S_{12}



(h) S_{23}

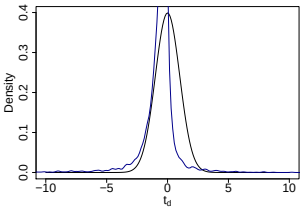


(i) S_{43}

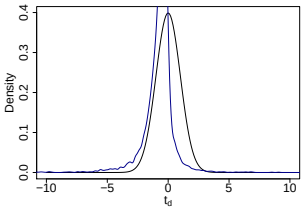
Figure 5.C.2: Test Statistics for Conditional Densities (Center)

The panel shows kernel density estimates of 5000 simulated test statistics (blue solid) at the respective points S_{ij} of Φ_{XY} listed in Table 5.2. For this figure, only points that are in the center of the distribution, where the data density is high, were selected. For the graphs in each row, the sample size N is fixed. Each statistic is calculated with local polynomials of order $p = 4$. The number of quantiles is chosen according to the rule in Equation (5.27). The black solid line shows the density of a standard normal distribution for comparison.

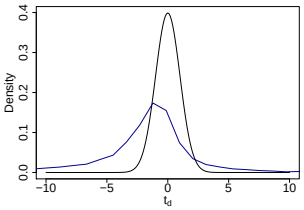
$N = 100$



(a) S_{22}

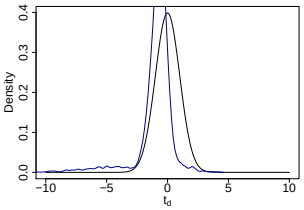


(b) S_{33}

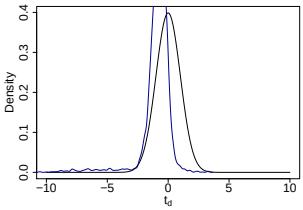


(c) S_{55}

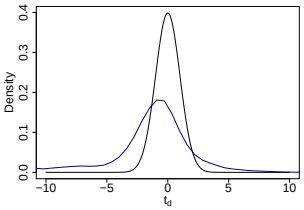
$N = 1000$



(d) S_{22}

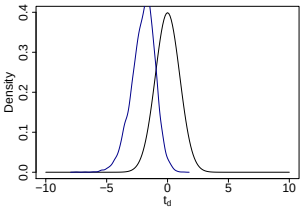


(e) S_{33}

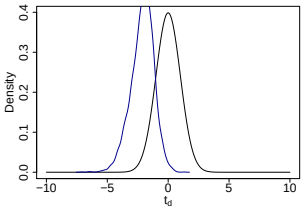


(f) S_{55}

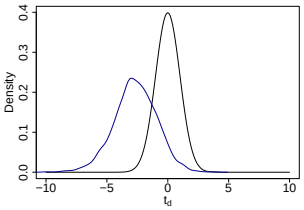
$N = 5000$



(g) S_{22}



(h) S_{33}



(i) S_{55}

Conclusion

This dissertation presents a bouquet of studies that all follow the guiding questions: what information does matter when prices are formed and how do you separate the relevant from the irrelevant?

Starting with an eagle-eyed perspective, the question whether the search behavior of Google users is helpful for the prediction of price movements in cryptocurrency markets. This question is tackled in Chapter 2. First, for this purpose and in order to enable further analysis, a regression based algorithm is developed in Chapter 1. With this algorithm, consistent and coherent time series of Google's search volume index (SVI) can be constructed for arbitrary time frames. The algorithm is evaluated against current data downloaded from Google recently for other frequencies, as well as checked against vintage data, downloaded prior to 2011. Back then Google provided a more comparable version of its search volume index. The regression based algorithm enables users also to compare different SVIs for different search-terms with each other. Therefore, the limitations of Google's SVI when it comes to sample length and sample width are overcome. Using this algorithm, an index of prices searched online, the IPSO, is constructed. In Chapter 1, I demonstrate that the IPSO improves monthly inflation forecasts to the US and the Euro Area. Especially, the clear results for household loan growth in the Euro Area are remarkable.

With the algorithm developed in Chapter 1, the helpfulness of Google's SVI on cryptocurrency returns and volatility is analyzed in Chapter 2. In this chapter, it is demonstrated that in these cryptocurrency markets, where a majority of individual investors are active, Google search volume does help to predict volatility, but fails to provide additional information for the prediction of returns. For this purpose, data on several frequencies was analyzed. While other studies in the literature were restricted to monthly or weekly data with a limited sample, this study was able to analyze not only weekly, but also daily and even hourly return and search volume data.

Chapter 2 and Chapter 1 show that the aggregate internet search behavior of individuals contains predictive information for market price movements, both for individual assets like cryptocurrencies as well as for inflation on a macroeconomic level. Chapter 4 takes a different perspective by focusing on the microstructure of financial markets and the

mechanics of the limit order book (LOB). By using elements of operator algebra, which usually is heavily used in Quantum Mechanics, a bottom-up approach is taken to develop a model for the limit order book that describes reality as closely as possible. In order to introduce the used concepts, Chapter 3 provides a short introduction into the fundamental ideas of these techniques. Using stochastic operators, the model allows for a very high degree of complexity with very few assumptions. While previous models in the literature resort to a variety of simplifying assumptions, the model shows in a realistic setting that order flow is at the heart of price fluctuations. It allows for most of the complexities encountered in reality. In this way, it ties the price formation process to the individual decision of order submission. Harvesting the information contained in the distribution of order flow across relative price levels contemporaneously describes price movements well and is helpful in predicting price movements on frequencies up to 10 minutes. While the model describes the relation between arrival rates and price movements to be a non-linear one, the empirical specifications used to forecast returns are all linearized approximations. There is plenty of opportunity to improve upon the specifications and the results. Machine learning techniques may be able to better grasp the inherent non-linearity and improve the forecast performance. Also on a theoretical level there is room for improvement in this current state, the chapter does not provide the basis functions of the developed operator algebra. These functions may provide additional insights; by specifying them, a closed form solution for joint distributions or the time varying moments of prices, volume, liquidity measures, and other observables of the limit order book may come into reach. In this regard, Chapter 4 only provides a rudimentary basis for further research endeavors. Nevertheless, it provides a new approach and adds a new perspective to the existing literature on limit order books.

The last study, Chapter 5, again takes a step back. In the preceding chapters questions are analyzed like: Does Google's SVI help predict price movements in cryptocurrency markets? Does Google's SVI help predict inflation or consumption? Is information from the sheer mechanics of limit order books helpful in describing price movements? To answer all these questions linear approximations have been made to find answers. For each of the questions, also non-linear relations were suspected between the variables. In Chapter 4, such a non-linear relation was theoretically derived. From a generalizing point of view, all of these studies analyze whether some set of variables $\boldsymbol{X}$ contains significant information for the prediction of Y. In this sense, Chapter 5 provides a new approach to the research questions for all of the studies by providing a method to answer the question of predictive power of $\boldsymbol{X}$ on Y in a very general manner. Regardless of the functional form between the variables, the measure of transfer entropy, if distinct from zero, is an indication for predictive information in $\boldsymbol{X}$. Using smoothed quantiles estimated with the well established method of quantile regression, Chapter 5 provides first a way to estimate conditional densities in order to subsequently use these conditional density estimates to calculate

transfer entropy. Also the asymptotic properties of the transfer entropy measure are derived and test statistics worked out. Even though the test statistics provide some indication on whether transfer entropy is distinctly different from zero or not, further research has to be undertaken to examine and generalize their behavior in different settings. Further improvements with respect to the efficiency and consistency of the involved estimates are possible. Nonetheless, I apply the new method to estimate the continuous transfer entropy measure on datasets previously used in the literature on credit default swap premia and credits spreads on bond market in Chapter 5. Furthermore, in a second application, the transatlantic information flows between minute-by-minute returns are analyzed. In both cases the positive results based on coarse grained measures of symbolic transfer entropy can not be reproduced. Prior research found significant transfer entropy from the premia paid on credit default swap markets to the credit spread observed in bond markets and vice-versa. Previous literature also documents a significant information flow from leading European indices to the American S&P. In this literature, returns are encoded into symbols according to whether they belonged to the top or bottom 5% of observations or were part of the other 90% in the center of the return distribution. Chapter 5 shows that by omitting the encoding of the variables involved and estimating the continuous transfer entropy instead of the symbolic transfer entropy, leads to different results. In the light of these results, the interpretation of the symbolic transfer entropy is more nuanced, more intricate and more determined by the applied discretization scheme than the interpretation in the established literature previously expected.

All in all, this dissertation is a humble attempt to provide additional insights on a variety of issues related to statistics of financial markets. I hope the answers discussed and the methods developed in this dissertation may contribute also to other research projects and help to improve the understanding of price formation in financial markets.

Bibliography

Aalborg, H. A., Molnár, P. and de Vries, J. E.: 2019, What can explain the price, volatility and trading volume of bitcoin?, *Finance Research Letters* **29**, 255–265.

Aalen, O. O. and Johansen, S.: 1978, An empirical transition matrix for non-homogeneous markov chains based on censored observations, *Scandinavian Journal of Statistics* pp. 141–150.

Afkhami, M., Cormack, L. and Ghoddusi, H.: 2017, Google search keywords that best predict energy price volatility, *Energy Economics* **67**, 17–27.

Alabi, K.: 2017, Digital blockchain networks appear to be following Metcalfe's law, *Electronic Commerce Research and Applications* **24**, 23–29.

Alfarano, S., Lux, T. and Wagner, F.: 2008, Time variation of higher moments in a financial market with heterogeneous agents: An analytical approach, *Journal of Economic Dynamics and Control* **32**(1), 101–136.

Amblard, P.-O. and Michel, O. J.: 2011, On directed information theory and granger causality graphs, *Journal of Computational Neuroscience* **30**(1), 7–16.

Baez, J. C. and Biamonte, J. D.: 2018, *Quantum Techniques In Stochastic Mechanics*, World Scientific Publishing Company.

Baez, J. C. and Pollard, B. S.: 2017, A compositional framework for reaction networks, *Reviews in Mathematical Physics* **29**(09), 1750028.

Bank, M., Larch, M. and Peter, G.: 2011, Google search volume and its influence on liquidity and returns., *Financial Markets and Portfolio Management* **25**(3), 239–264.

Baur, D. G. and Dimpfl, T.: 2018a, A Quantile Regression Approach to Estimate the Variance of Financial Returns*, *Journal of Financial Econometrics* **17**(4), 616–644.

Baur, D. G. and Dimpfl, T.: 2018b, Asymmetric volatility in cryptocurrencies, *Economics Letters* **173**, 148–151.

Baur, D. G. and Dimpfl, T.: 2021, The volatility of bitcoin and its role as a medium of exchange and a store of value, *Empirical Economics* .

Baur, D. G., Hong, K. and Lee, A. D.: 2018, Bitcoin: Medium of exchange or speculative assets?, *Journal of International Financial Markets, Institutions and Money* **54**, 177–189.

Bechler, K. and Ludkovski, M.: 2015, Optimal execution with dynamic order flow imbalance, *SIAM Journal on Financial Mathematics* **6**(1), 1123–1151.

Behrendt, S. and Prange, P.: 2019, What are you searching for? on the equivalence of proxies for online investor attention, *Finance Research Letters* p. 101401.

Bi, Z., Faloutsos, C. and Korn, F.: 2001, The 'DGX' distribution for mining massive, skewed data, *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 17–26.

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

Bleher, J., Bleher, M. and Dimpfl, T.: 2020, The what, when and where of limit order books, *arXiv preprint arXiv:2004.11953* .

Bleher, J. and Dimpfl, T.: 2019, Knitting multi-annual high-frequency google trends to predict inflation and consumption, *Available at SSRN 3357424* .

Bouchaud, J.-P., Mézard, M. and Potters, M.: 2002, Statistical properties of stock order books: empirical results and models, *Quantitative Finance* **2**(4), 251–256.

Carrasco, M. and Florens, J.-P.: 2000, Generalization of GMM to a continuum of moment conditions, *Econometric Theory* pp. 797–834.

Cartea, Á., Jaimungal, S. and Penalva, J.: 2015, *Algorithmic and High-Frequency Trading*, Mathematics, Finance and Risk, Cambridge University Press.

Chen, H., De, P., Hu, Y. and Hwang, B.-H.: 2014, Wisdom of crowds: The value of stock opinions transmitted through social media, *The Review of Financial Studies* **27**(5), 1367–1403.

Chernozhukov, V., Fernandez-Val, I. and Galichon, A.: 2007, Improving estimates of monotone functions by rearrangement, *Technical report*, CEMMAP working paper.

Choi, H. and Varian, H.: 2012, Predicting the present with Google trends, *Economic Record* **88**(s1), 2–9.

Chronopoulos, D., Papadimitriou, F. and Vlastakis, N.: 2018, Information demand and stock return predictability, *Journal of International Money and Finance* **80**, 59–74.

Ciaian, P., Rajcaniova, M. and Kancs, d.: 2016, The economics of BitCoin price formation, *Applied Economics* **48**(19), 1799–1815.

Clark, T. E. and West, K. D.: 2006, Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis, *Journal of Econometrics* **135**(1), 155–186.

Clark, T. E. and West, K. D.: 2007, Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* **138**(1), 291–311.

Clauset, A., Shalizi, C. R. and Newman, M. E.: 2009, Power-law distributions in empirical data, *SIAM review* **51**(4), 661–703.

Cleveland, W. S.: 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American statistical association* **74**(368), 829–836.

Cleveland, W. S. and Grosse, E.: 1991, Computational methods for local regression, *Statistics and Computing* **1**(1), 47–62.

Cleveland, W. S., Grosse, E. and Shyu, W.: 1992, Local regression models, *in* J. Chambers and T. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, chapter 8, pp. 309–376.

Cochrane, J. H.: 2008, The dog that did not bark: A defense of return predictability, *The Review of Financial Studies* **21**(4), 1533–1575.

Cont, R. and de Larrard, A.: 2012, Order book dynamics in liquid markets: limit theorems and diffusion approximations, *Available at SSRN 1757861* .

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

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

Corbet, S., Lucey, B. and Yarovya, L.: 2018, Datestamping the bitcoin and ethereum bubbles, *Finance Research Letters* **26**, 81–88.

Cover, T. M. and Thomas, J. A.: 2005, *Elements of Information Theory*, 2nd edn, John Wiley & Sons.

Da, Z., Engelberg, J. and Gao, P.: 2015, The sum of all FEARS investor sentiment and asset prices, *Review of Financial Studies* **28**(1), 1–32.

D'Acunto, F., Malmendier, U., Ospina, J. and Weber, M.: 2019, Exposure to daily price changes and inflation expectations, *Technical report*, National Bureau of Economic Research.

Daniels, M. G., Farmer, J. D., Gillemot, L., Iori, G. and Smith, E.: 2003, Quantitative model of price diffusion and market friction based on trading as a mechanistic random process, *Physical Review Letters* **90**(10).

Dastgir, S., Demir, E., Downing, G., Gozgor, G. and Lau, C. K. M.: 2019, The causal relationship between bitcoin attention and bitcoin returns: Evidence from the copula-based granger causality test, *Finance Research Letters* **28**, 160–164.

Dimitrov, A. G., Lazar, A. A. and Victor, J. D.: 2011, Information theory in neuroscience, *Journal of Computational Neuroscience* **30**(1), 1–5.

Dimpfl, T. and Jank, S.: 2016, Can internet search queries help to predict stock market volatility?, *European Financial Management* **22**(2), 171–192.

Dimpfl, T. and Kleiman, V.: 2017, Investor pessimism and the german stock market: Exploring google search queries, *German Economic Review* (forthcoming).

Dimpfl, T. and Peter, F. J.: 2013, Using transfer entropy to measure information flows between financial markets, *Studies in Nonlinear Dynamics & Econometrics* **17**(1), 85–102.

Dimpfl, T. and Peter, F. J.: 2014, The impact of the financial crisis on transatlantic information flows: An intraday analysis, *Journal of International Financial Markets, Institutions and Money* **31**, 1–13.

Dimpfl, T. and Peter, F. J.: 2019, Group transfer entropy with an application to cryptocurrencies, *Physica A: Statistical Mechanics and its Applications* **516**, 543–551.

Efron, B.: 1982, *The jackknife, the bootstrap and other resampling plans*, SIAM.

Engle, R. F. and Manganelli, S.: 2004, Caviar: Conditional autoregressive value-at-risk by regression quantiles, *Journal of Business & Economic Statistics* **22**(4), 367–381.

Engle, R. F. and Russell, J. R.: 1998, Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica* **66**(5), 1127–1162.

Eross, A., McGroarty, F., Urquhart, A. and Wolfe, S.: 2019, The intraday dynamics of bitcoin, *Research in International Business and Finance* **49**, 71–81.

Fan, J. and Gijbels, I.: 1995, Adaptive order polynomial fitting: bandwidth robustification and bias reduction, *Journal of Computational and Graphical Statistics* **4**(3), 213–227.

Fan, J. and Gijbels, I.: 1996, *Local Polynomial Modelling and its Applications: Monographs on Statistics and Aapplied Probability*, Routledge.

Fan, J., Gijbels, I., Hu, T.-C. and Huang, L.-S.: 1996, A study of variable bandwidth selection for local polynomial regression, *Statistica Sinica* pp. 113–127.

Fan, J. and Marron, J. S.: 1994, Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics* **3**(1), 35–56.

Feller, W.: 1957, 1968, *An introduction to probability theory and its applications*, Vol. I, 3 edn, John Wiley & Sons.

Fernandes, M. and Grammig, J.: 2006, A family of autoregressive conditional duration models, *Journal of Econometrics* **130**(1), 1–23.

Foucault, T., Kadan, O. and Kandel, E.: 2005, Limit order book as a market for liquidity, *Review of Financial Studies* **18**(4), 1171–1217.

Foucault, T., Sraer, D. and Thesmar, D. J.: 2011, Individual investors and volatility, *Journal of Finance* **66**(4), 1369–1406.

Garcia, D. and Schweitzer, F.: 2015, Social signals and algorithmic trading of bitcoin, *Royal Society Open Science* **2**(9).

Garcia, D., Tessone, C. J., Mavrodiev, P. and Perony, N.: 2014, The digital traces of bubbles: feedback cycles between socio-economic signals in the bitcoin economy, *Journal of the Royal Society Interface* **11**(99), 20140623.

Garman, M. B. and Klass, M. J.: 1980, On the estimation of security price volatilities from historical data, *The Journal of Business* **53**(1), 67–78.

Geuder, J., Kinateder, H. and Wagner, N. F.: 2018, Cryptocurrencies as financial bubbles: The case of bitcoin, *Finance Research Letters* **in press**.

Giannakis, G. B. and Serpedin, E.: 2001, A bibliography on nonlinear system identification, *Signal Processing* **81**(3), 533–580. Special section on Digital Signal Processing for Multimedia.

Gillespie, D. T.: 1977, Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry* **81**(25), 2340–2361.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L.: 2009, Detecting influenza epidemics using search engine query data, *Nature* **457**, 1012–1014.

Glosten, L. R.: 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* **49**(4), 1127–1161.

Gomber, P. and Schweickert, U.: 2002, Der Market Impact: Liquiditätsmaß im elektronischen Wertpapierhandel, *Die Bank* **7**(2002), 485–489.

Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H. E.: 2000, Statistical properties of share volume traded in financial markets, *Physical Review E* **62**(4), R4493.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J. and Howison, S. D.: 2013, Limit order books, *Quantitative Finance* **13**(11), 1709–1742.

Grammig, J., Heinen, A. and Rengifo, E. W.: 2004, Trading activity and liquidity supply in a pure limit order book market. an empirical analysis using a multivariate count data model, *CORE Discussion Papers* (2004/58). Available at SSRN 676567.

Granger, C. W. J.: 1969, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**(3), 424–438.

Granger, C. W. and Newbold, P.: 1976, Forecasting transformed series, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 189–203.

Haber, R. and Unbehauen, H.: 1990, Structure identification of nonlinear dynamic systems – a survey on input/output approaches, *Automatica* **26**(4), 651–677.

Hamilton, J. D.: 1994, *Time Series Analysis*, Vol. 2, Princeton University Press Princeton.

Hansen, L. P.: 1982, Large sample properties of generalized method of moments estimators, *Econometrica* **50**(4), 1029–1054.

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

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

Hautsch, N. and Huang, R.: 2012, The market impact of a limit order, *Journal of Economic Dynamics and Control* **36**(4), 501–522.

Hayashi, F.: 2000, Econometrics. 2000, *Princeton University Press. Section* **1**, 60–69.

Hayes, A. S.: 2016, Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin, *Telematics and Informatics* .

Hornik, J., Satchi, R. S., Cesareo, L. and Pastore, A.: 2015, Information dissemination via electronic word-of-mouth: Good news travels fast, bad news travels faster!, *Computers in Human Behavior* **45**, 273–280.

Hyndman, R. J. and Khandakar, Y.: 2008, Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software* **26**(3), 1–22.

Jain, P. K.: 2003, Institutional design and liquidity at stock exchanges around the world, *Available at SSRN 869253* .

James, R. G., Barnett, N. and Crutchfield, J. P.: 2016, Information flows? A critique of transfer entropies, *Physical review letters* **116**(23), 238701.

Kaiser, A. and Schreiber, T.: 2002, Information transfer in continuous processes, *Physica D: Nonlinear Phenomena* **166**(1), 43–62.

Kirman, A.: 1993, Ants, rationality, and recruitment, *The Quarterly Journal of Economics* **108**(1), 137–156.

Klein, L. R.: 1953, A textbook of econometrics.

Koenker, R.: 2005, *Quantile Regression*, Cambridge University Press.

Koenker, R. and Bassett, Jr., G.: 1978, Regression quantiles, *Econometrica* **46**(1), 33–50.

Kristoufek, L.: 2013, Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the internet era, *Scientific Reports* **3**, 1–7.

Kristoufek, L.: 2015, What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis, *PloS one* **10**(4), e0123923.

Lando, D. and Skødeberg, T. M.: 2002, Analyzing rating transitions and rating drift with continuous observations, *Journal of banking & finance* **26**(2-3), 423–444.

Large, J.: 2007, Measuring the resiliency of an electronic limit order book, *Journal of Financial Markets* **10**(1), 1–25.

Lee, J., Nemati, S., Silva, I., Edwards, B. A., Butler, J. P. and Malhotra, A.: 2012, Transfer entropy estimation and directional coupling change detection in biomedical time series, *Biomedical Engineering Online* **11**(1), 19.

Li, X., Shang, W., Wang, S. and Ma, J.: 2015, A midas modelling framework for chinese inflation index forecast incorporating google search data, *Electronic Commerce Research and Applications* **14**(2), 112–125.

Lipton, A., Pesavento, U. and Sotiropoulos, M.: 2014, Trading strategies via book imbalance, *Risk* pp. 70–75.

Long, J. S. and Ervin, L. H.: 2000, Using heteroscedasticity consistent standard errors in the linear regression model, *The American Statistician* **54**(3), 217–224.

Lütkepohl, H. and Xu, F.: 2012, The role of the log transformation in forecasting economic variables, *Empirical Economics* **42**(3), 619–638.

Marsaglia, G.: 1972, Choosing a point from the surface of a sphere, *The Annals of Mathematical Statistics* **43**(2), 645–646.

Martins-Filho, C. and Saraiva, P.: 2012, On asymptotic normality of the local polynomial regression estimator with stochastic bandwidths, *Communications in Statistics-theory and Methods* **41**, 1052–1068.

Maslov, S. and Mills, M.: 2001, Price fluctuations from the order book perspective & empirical facts and a simple model, *Physica A: Statistical Mechanics and its Applications* **299**(1), 234–246.

Masry, E. and Fan, J.: 1997, Local polynomial estimation of regression functions for mixing processes, *Scandinavian Journal of Statistics* **24**(2), 165–179.

Massicotte, P. and Eddelbuettel, D.: 2018, *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.4.2.

Mincer, J. A. and Zarnowitz, V.: 1969, The evaluation of economic forecasts, *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, NBER, pp. 3–46.

Oehlert, G. W.: 1992, A note on the delta method, *The American Statistician* **46**(1), 27–29.

Oh, M., Kim, S., Lim, K. and Kim, S. Y.: 2018, Time series analysis of the antarctic circumpolar wave via symbolic transfer entropy, *Physica A: Statistical Mechanics and its Applications* .

Panagiotidis, T., Stengos, T. and Vravosinos, O.: 2018a, The effects of markets, uncertainty and search intensity on bitcoin returns, *International Review of Financial Analysis* .

Panagiotidis, T., Stengos, T. and Vravosinos, O.: 2018b, On the determinants of bitcoin returns: a lasso approach, *Finance Research Letters* .

Papana, A., Kyrtsou, C., Kugiumtzis, D., Diks, C. et al.: 2013, Partial symbolic transfer entropy, *University of Amsterdam* pp. 13–16.

Parkinson, M.: 1980, The extreme value method for estimating the variance of the rate of return, *Journal of business* pp. 61–65.

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

Pasteels, J. M., Deneubourg, J.-L. and Goss, S.: 1987, Self-organization mechanisms in ant societies. i. trail recruitment to newly discovered food sources, *From individual to collective behavior in social insects: les Treilles Workshop/edited by Jacques M. Pasteels, Jean-Louis Deneubourg*, Basel: Birkhauser, 1987.

Patton, A. J.: 2011, Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* **160**(1), 246–256.

Paulsen, M. C.: 2014, *Limit theorems for limit order books*, PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II.

Perlin, M. S., Caldeira, J. a. F., Santos, A. A. P. and Pontuschka, M.: 2017, Can we predict the financial markets based on google's search queries?, *Journal of Forecasting* **36**(4), 454–467.

Pfaff, B.: 2008, *Analysis of Integrated and Cointegrated Time Series with R*, second edn, Springer, New York. ISBN 0-387-27960-1.

Poon, J. and Dryja, T.: 2015, The bitcoin lightning network: Scalable off-chain instant payments, *Technical Report (draft)* .

Prokopenko, M., Lizier, J. T. and Price, D. C.: 2013, On thermodynamic interpretation of transfer entropy, *Entropy* **15**(2), 524–543.

Qadan, M. and Nama, H.: 2018, Investor sentiment and the price of oil, *Energy Economics* **69**, 42–58.

R: 2018, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Reid, F. and Harrigan, M.: 2011, An analysis of anonymity in the Bitcoin system, *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, IEEE, pp. 1318–1326.

Rochdi, K. and Dietzel, M.: 2015, Outperforming the benchmark: online information demand and reit market performance, *Journal of Property Investment & Finance* **33**(2), 169–195.

Rogers, L. C. G. and Satchell, S. E.: 1991, Estimating variance from high, low and closing prices, *The Annals of Applied Probability* pp. 504–512.

Ruppert, D., Sheather, S. J. and Wand, M. P.: 1995, An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association* **90**(432), 1257–1270.

Ruppert, D. and Wand, M. P.: 1994, Multivariate locally weighted least squares regression, *Ann. Statist.* **22**(3), 1346–1370.

Sandoval, J. L., Mullokandov, A. and Kenett, D. Y.: 2015, Dependency relations among international stock market indices, *Journal of Risk and Financial Management* **8**(2), 227–265.

Schreiber, T.: 2000, Measuring information transfer, *Physical Review Letters* **85**(2), 461–464.

Scott, S. L. and Varian, H. R.: 2015, Bayesian variable selection for nowcasting economic time series, *Economic analysis of the digital economy*, University of Chicago Press, pp. 119–135.

Seifert, B., Brockmann, M., Engel, J. and Gasser, T.: 1994, Fast algorithms for nonparametric curve estimation, *Journal of Computational and Graphical Statistics* **3**(2), 192–213.

Shannon, C. E.: 1948, A mathematical theory of communication, *Bell System Technical Journal* **27**(3), 379–423.

Shapiro, C. and Varian, H. R.: 1998, *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press.

Smirnov, D. A.: 2013, Spurious causalities with transfer entropy, *Physical Review E* **87**, 042917.

Smith, E., Farmer, J. D., Gillemot, L. s., Krishnamurthy, S. et al.: 2003, Statistical theory of the continuous double auction, *Quantitative Finance* **3**(6), 481–514.

Stephens-Davidowitz, S. and Varian, H.: 2014, A hands-on guide to Google data, *Technical Report* .

Stone, C. J.: 1977, Consistent nonparametric regression, *The annals of statistics* pp. 595–620.

Stone, C. J.: 1980, Optimal rates of convergence for nonparametric estimators, *The Annals of Statistics* **8**(6), 1348–1360.

Stone, C. J.: 1982, Optimal global rates of convergence for nonparametric regression, *The annals of statistics* pp. 1040–1053.

Sun, J. and Bollt, E. M.: 2014, Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings, *Physica D: Nonlinear Phenomena* **267**, 49–57. Evolving Dynamical Networks.

Urquhart, A.: 2018, What causes the attention of bitcoin?, *Economics Letters* **166**, 40–44.

U.S. Bureau of Economic Analysis: 2019, Real personal consumption expenditures [pcec96] retrieved from fred.

U.S. Bureau of Labor Statistics: 2019, All items in u.s. city average, all urban consumers, not seasonally adjusted (cuur0000sa0).

van Kampen, N. G.: 1992, *Stochastic processes in physics and chemistry*, Vol. 1, Elsevier.

Vicente, R., Wibral, M., Lindner, M. and Pipa, G.: 2011, Transfer entropy–a model-free measure of effective connectivity for the neurosciences, *Journal of Computational Neuroscience* **30**(1), 45–67.

Weber, M. F. and Frey, E.: 2017, Master equations and the theory of stochastic path integrals, *Reports on Progress in Physics* **80**(4), 046601.

Yang, D. and Zhang, Q.: 2000, Drift-independent volatility estimation based on high, low, open, and close prices, *The Journal of Business* **73**(3), 477–492.

Yelowitz, A. and Wilson, M.: 2015, Characteristics of bitcoin users: an analysis of google search data, *Applied Economics Letters* **22**(13), 1030–1036.

Yu, K. and Jones, M. C.: 1998, Local linear quantile regression, *Journal of the American Statistical Association* **93**(441), 228–237.

Zhang, W., Wang, P., Li, X. and Shen, D.: 2018, Quantifying the cross-correlations between online searches and bitcoin market, *Physica A: Statistical Mechanics and its Applications* .

Zheng, L., Pan, W., Li, Y., Luo, D., Wang, Q. and Liu, G.: 2017, Use of mutual information and transfer entropy to assess interaction between parasympathetic and sympathetic activities of nervous system from HRV, *Entropy* **19**, 489.

Zhou, D., Huang, J. and Schölkopf, B.: 2007, Learning with hypergraphs: Clustering, classification, and embedding, *Advances in neural information processing systems*, pp. 1601–1608.

Zhou, X., Pan, Z., Hu, G., Tang, S. and Zhao, C.: 2018, Stock market prediction on high-frequency data using generative adversarial nets, *Mathematical Problems in Engineering* **2018**(ID 4907423).