University of Economics, Prague
Faculty of Informatics and Statistics
Department of Econometrics

DOCTORAL THESIS

Financial High-Frequency Data

Vladimír Holý

Supervisor: prof. RNDr. Ing. Michal Černý, Ph.D.
Study Field: Econometrics and Operational Research
Prague 2019

Title: Financial High-Frequency Data

Author: Mgr. Vladimír Holý

Supervisor: prof. RNDr. Ing. Michal Černý, Ph.D.

Abstract: In finance, stock prices, exchange rates and commodity prices are recorded with each transaction or bid/ask offer resulting in intraday high-frequency data. Such time series have a very fine time scale (e.g. seconds or even fractions of seconds). High-frequency data have several specifics distinguishing them from low-frequency data (e.g. daily time series). The observations of prices are irregularly spaced, the values of prices are discrete and the price process consists of both bid and ask side. The latter two specifics are often captured by the market microstructure noise which conceals the information about the theoretical efficient price process. Over the past 20 years, a vast number of econometrical methods have been developed to address such market microstructure issues and to understand dynamics of financial markets. We focus on three aspects of high-frequency data analysis. First, we model durations between successive transactions using the autoregressive conditional duration model in a discrete framework with a special attention to split transactions. Second, we estimate and forecast quadratic variation of the price process by a variety of methods robust to the market microstructure noise. Third, we model prices as the Ornstein–Uhlenbeck process contaminated by the market microstructure noise with an application to the pairs trading strategy. In the empirical part of the thesis, we analyze selected stocks traded on the NYSE and NASDAQ exchanges.

Keywords: High-Frequency Data, Efficient Price, Market Microstructure Noise, Trade Durations, Autoregressive Conditional Duration Model, Quadratic Variation, Integrated Variance, Ornstein–Uhlenbeck Process, Pairs Trading

AMS Classification: 60J60, 62F10, 62G05, 62M10, 91B24

JEL Classification: C22, C41, C51, C58, G10

Název: Finanční vysokofrekvenční data

Autor: Mgr. Vladimír Holý

Vedoucí: prof. RNDr. Ing. Michal Černý, Ph.D.

Abstrakt: Ceny akcií, směnné kurzy a ceny komodit jsou ve financích zaznamenávány s každou transakcí nebo změnou bid/ask nabídky. Intradenní vysokofrekvenční časové řady mají velmi jemnou časovou škálu (např. sekundy nebo dokonce zlomky sekund). Vysokofrekvenční data mají několik výrazných specifik, které je odlišují od dat s nízkou frekvencí (např. od denních časových řad). Pozorování jsou nepravidelně rozmístěná, hodnoty cen jsou disktrétní a samotný proces cen se skládá z bid a ask stran. Poslední dva rysy jsou často zachyceny mikrostrukturním šumem, který zakrývá informace o teoretickém procesu eficientní ceny. Během posledních 20 let byla vyvinuta řada ekonometrických metod pro řešení takových problémů mikrostruktury trhů a pro pochopení dynamiky finančních trhů. Zaměřujeme se na tři aspekty analýzy vysokofrekvenčních dat. Za prvé, modelujeme doby mezi jednotlivými transakcemi pomocí diskrétního autoregresivního modelu podmíněných durací s ohledem na split transakce. Za druhé, odhadujeme a předpovídáme kvadratickou variaci procesu cen různými metodami robustními k mikrostrukturnímu šumu. Za třetí, modelujeme ceny jako Ornstein–Uhlenbeckův proces kontaminovaný mikrostrukturním šumem s aplikací pro pairs trading strategii. V empirické části práce analyzujeme vybrané akcie obchodované na burzách NYSE a NASDAQ.

Klíčová slova: Vysokofrekvenční data, Eficientní cena, Mikrostrukturní šum, Durace mezi transakcemi, Autoregresní model podmíněné durace, Kvadratická variace, Integrovaná variance, Ornstein–Uhlenbeckův proces, Párové obchodování

AMS Klasifikace: 60J60, 62F10, 62G05, 62M10, 91B24

JEL Klasifikace: C22, C41, C51, C58, G10

- Preface -

In my Ph.D. thesis, I study statistical and econometric methods analyzing financial high-frequency data from both theoretical and empirical point of view. The main contributions of the thesis are the following.

- A new framework for modeling zero values of trade durations caused by split transactions is presented. It is based on the generalized autoregressive score model with zero-inflated discrete distributions. The main advantage of this approach is its ability to determine the ratio of zero values caused by split transactions and zero values caused by simultaneous but independent transactions.
- New estimators of the Ornstein–Uhlenbeck process contaminated by the market microstructure noise are proposed. For equidistant data, method of moments, maximum likelihood method and reparametrization to ARMA(1,1) process are utilized. The proposed maximum likelihood estimator can also be used in the case of irregularly spaced tick data.

Besides these two main results, the thesis contains further minor contributions.

- The extensive theoretical and empirical literature dealing with financial high-frequency data is reviewed. The main focus is on the duration analysis and the volatility analysis. Tools for financial high-frequency data analysis in statistical software R are also reviewed.
- The issue of rounding is briefly visited in the context of financial data. It is shown that the rounding error is a significant part of the market microstructure noise. This finding is based on asymptotically uniform distribution of the rounding error.
- For the estimation of quadratic variation, an interval approach motivated by the bid-ask spread and discretness of prices is presented. It is proven, however, that the quadratic variation is not identifiable under this interval setting.
- Various non-parametric estimators of quadratic variation robust to the market microstructure noise are compared in a simulation study. Forecasting models for estimated quadratic variation are also compared in an empirical study.
- An intraday pairs trading strategy utilizing ultra-high-frequency data is presented. It is based on the Ornstein–Uhlenbeck process and the mean-variance optimization. The empirical study shows that this strategy is highly profitable for the right choice of mean and variance constraints.

The thesis is based on the two following articles. The article of Blasques, Holý and Tomanová (2018) contributes to duration analysis by modeling durations in a discrete GAS framework allowing for excessive zero durations corresponding to split transactions. This paper is used in Chapter 3. The article of Holý and Tomanová (2018) extends parametric analysis of financial high-frequency data by estimating Ornstein–Uhlenbeck process contaminated by the market microstructure noise with an application to intraday pairs trading strategy. This paper is used in Chapter 5.

The thesis is also based on the following conference proceedings. The conference proceedings of Holý (2016) investigate the impact of the market microstructure noise on the quadratic variance. This paper is used in Section 4.2.1. The conference proceedings of Holý (2017a) review the literature about the market microstructure noise. This paper is used throughout Chapter 1. The conference proceedings of Holý (2017b) propose a method for estimating quadratic variation by least squares. After the publication of this paper, however, I found that the least squares estimator has already been proposed by Nolte and Voev (2012). This paper is used in Section 4.2.5. The conference proceedings of Holý (2017c) compare forecasting accuracy of various quadratic variation models. This paper is used in sections 4.3 and 4.4.2. The conference proceedings of Holý (2017e) formulate various quadratic variation estimators as a quadratic form. This paper is used in Section 4.2. The conference proceedings of Holý (2018a) review the literature about the impact of high-frequency data. This paper is used throughout Chapter 1. The conference proceedings of Holý (2018b) investigate properties of the rounding error in high-frequency data. This paper is used in Section 2.2.4. The conference proceedings of Holý (2018c) reviews functions in R related to high-frequency data analysis. This paper is used in Appendix C. The conference proceedings of Holý and Černý (2017) compare various quadratic variation estimators in a simulation study. This paper is used in sections 4.2 and 4.4.1. The conference proceedings of Holý and Sokol (2018) deal with the quadratic variation under interval uncertainty. This paper is used in sections 2.2.5 and 4.1.2. Some passages are rewritten from the above mentioned papers while others are kept in their original form. The notation and applications are unified while the introductions are omitted.

The work on the thesis was supported by the Internal Grant Agency of the University of Economics, Prague Project No. F4/63/2016 (Analysis of Financial High-Frequency Data: Estimates in the Presence of Market Microstructure Noise), F4/58/2017 (Modern Methods of Uncertainty in Statistical and Optimization Models), F4/93/2017 (Transfer of Information on Financial Markets During Turbulences: Asymmetric Dependency Measures), F4/21/2018 (Analysis of High-Frequency Data and Data Stream) and by the Czech Science Foundation Project No. P402/12/G097 (DYME – Dynamic Models in Economics).

I wish to thank Michal Černý for his guidance, Petra Tomanová for collaboration on the key papers about zero trade durations and the Ornstein–Uhlenbeck process, Ondřej Sokol for discussions about all the ideas related to the interval analysis, Francisco Blasques for the insight into the asymptotic theory of GAS models, Tomáš Cipra for useful comments, Alena Holá for proofreading and Kateřina Koudelková for her support.

I declare that this thesis and the work presented in it are my own except for the shared authorship of the indicated parts. The literature and supporting materials are mentioned in the bibliography.

January 9, 2019, Prague Vladimír Holý

- Contents -

- Introduction -

The analysis of intraday stock prices, foreign exchange rates and commodity prices is an important aspect of quantitative finance. In this work, high-frequency data analysis is approached from both theoretical and empirical perspective. The thesis is organized as follows.

Chapter 1 reviews the extensive high-frequency data literature. First, the specifics of high-frequency data are discussed. Next, a vast number of analytical methods dealing with these specifics is presented. Finally, the impact of these methods in answering financial questions is assessed.

Chapter 2 establishes basic concepts on which the rest of the thesis is based. Practical issues of data cleaning, transformation and aggregation are presented. Theoretical models for times of observations, price process and market microstructure are formulated as well.

Chapter 3 analyzes durations between successive transactions. Traditional autoregressive conditional duration models based on continuous distributions are reviewed. A discrete model based on the zero-inflated negative binomial distribution with the general autoregressive score specification is proposed. Asymptotic properties of the maximum likelihood estimator are discussed. It is shown in an empirical study that the proposed model performs superior to the traditional continuous models as it is able to capture excessive zero values in duration data caused by split transactions.

Chapter 4 analyzes non-parametric volatility of the price process. Quadratic variation is defined and its properties analyzed in the stochastic calculus framework and interval framework as well. Several non-parametric estimators of quadratic variation commonly used in the literature are presented. Models for forecasting quadratic variation are also presented. All methods are compared in an empirical study of daily volatility.

Chapter 5 analyzes the price process using parametric methods. It is assumed that the prices follow the Ornstein–Uhlenbeck process. Several estimators of the Ornstein–Uhlenbeck process robust to the market microstructure noise for both equidistant and irregularly spaced data are proposed. The benefits of the proposed noise-robust estimators over traditional biased estimators are illustrated in an empirical study of the pairs trading strategy.

The thesis is supplemented by the following appendices. Appendix A describes the stock market and data used in the empirical analysis. Appendix B presents interesting statistics about the high-frequency data literature. Appendix C reviews capabilities of statistical computing software R in financial high-frequency data analysis. Appendix D reminds some lesser-known special functions in mathematics.

- Chapter 1 -

High-Frequency Data

In finance, Engle (2000) coined the term *ultra-high-frequency data* referring to irregularly spaced time series recorded at the highest possible frequency corresponding to each transaction or change in bid/ask offer. Financial high-frequency time series include stock prices, foreign exchange rates and commodity prices. The availability of these high-frequency data allows econometricians to construct more precise models while facing some new challenges.

Over the past 20 years, many scientific articles have been devoted to study financial high-frequency data and have proposed methods to utilize them correctly. This growing interest in financial high-frequency data is illustrated in Figure 1.1 using data from Scopus (Elsevier, 2019). For a better insight into the so-called high-frequency literature, see Appendix B. Contrary to the rising attention from the scientific community, there are still several myths surrounding high-frequency data within the financial industry. Among these myths are the following [1].

- *"High-frequency data are just a lot of low-frequency data."* This is simply not true as high-frequency data have distinct properties due to market microstructure specifics such as irregularly spaced observations, price discreteness and bid-ask spread.

- *"High-frequency data can be analyzed by econometric methods designed for low-frequency data."* As high-frequency data possess several market microstructure specifics, it follows that econometric and statistical analysis of financial high-frequency data requires special methods dealing with such specifics. For example, random times between observations are modeled in duration analysis while price discretness and bid-ask spread are captured by the market microstructure noise in volatility analysis.

- *"High-frequency data are useful just for high-frequency trading."* High-frequency trading naturally demands high-frequency data. Financial decisions and operations on lower frequencies, however, do benefit from high-frequency information as well. For example, portfolio optimization carried out on a daily basis can utilize more precise estimation of daily volatility based on intraday price movements.

The goal of this chapter is to refute these myths, present the key ideas in financial high-frequency data analysis and review related literature. For a comprehensive overview of high-frequency methods, we refer to the book of Hautsch (2011) focusing on analysis of durations and liquidity as well as the book of Aït-Sahalia and Jacod (2014) focusing on analysis of volatility, higher moments and jumps.

[1] Similar myths were discussed by Stephanie Toper during the 7th Annual Stevens Conference on High Frequency Finance and Analytics, Hoboken, November 3–5, 2016.

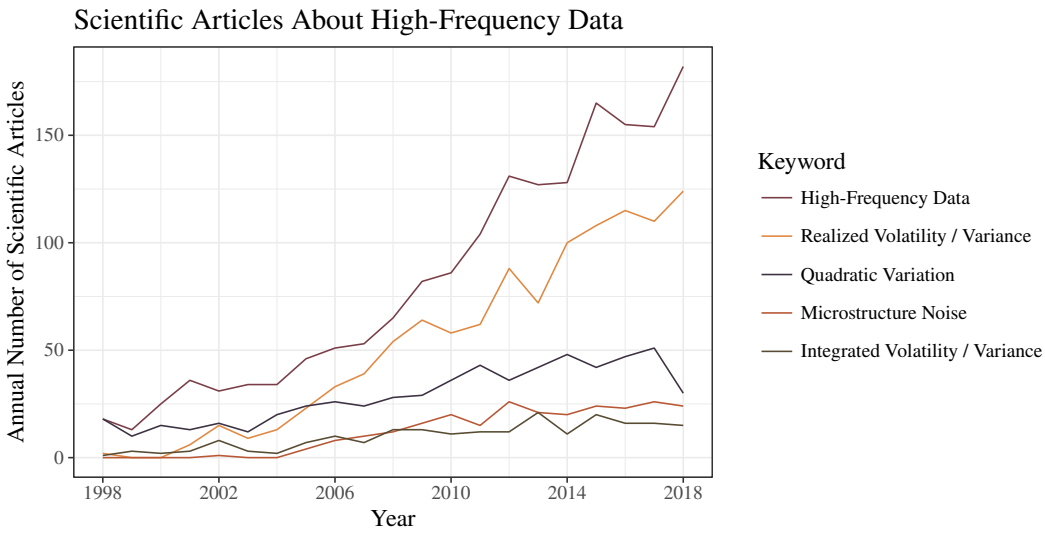

Figure 1.1: The annual number of scientific articles containing a specific term in the title, abstract or keywords.

1.1 Specifics of High-Frequency Data

We review specifics of high-frequency data from both practical and theoretical point of view. Let us illustrate differences between low-frequency and high-frequency data. Daily closing prices in Figure 1.2 can be treated within the traditional time series framework with discrete times and continuous values. Intraday tick prices (also known as ultra-high-frequency data) illustrated in Figure 1.3 and Figure 1.4, however, require a special treatment. As they come in a huge quantity and are irregularly spaced, it is natural to model prices by a process with continuous time or to model prices and times of observations simultaneously. Other specifics of high-frequency prices which need to be addressed in mathematical models include discretness of price values and bid-ask spread.

1.1.1 Data Considerations

An unpleasant specific of financial high-frequency data is a huge amount of recording errors. The reason for this is the velocity and volume at which high-frequency data are recorded as argued by Falkenberry (2002). The remedy is a careful data cleaning procedure. Such procedures are described by Brownlees and Gallo (2006) in the context of duration modeling and by Barndorff-Nielsen et al. (2009) in the context of volatility estimation. We discuss data cleaning procedures in more detail in Section 2.1.1.

A particular question is what to do with observations with the same timestamp. Most of the literature resort to merging transactions occuring at the same time into a single value. However, this leads to a significant data loss. Recently, some papers advocate keeping multiple observations with the same timestamp in the dataset. For example, Liu et al. (2018c) examine their effect on integrated variance estimation while Blasques, Holý and Tomanová (2018) suggest to model them using the *zero-inflated conditional autoregressive duration (ZIACD) model*. We thoroughly explore this issue in Chapter 3.

Another issue regarding data is the sampling of the price process. Hansen and Lunde (2006) identify three commonly used sampling schemes – *tick time sampling*, *calendar time sampling* and *business time sampling*. Aït-Sahalia and Mykland (2003), Oomen (2005, 2006) and Fukasawa (2010a,b) investigate the effects of sampling schemes on estimation of integrated variance. Dong and Tse (2017a) utilize business time sampling to test the semimartingale hypothesis of the log-price process. We present various sampling schemes in Section 2.1.3.

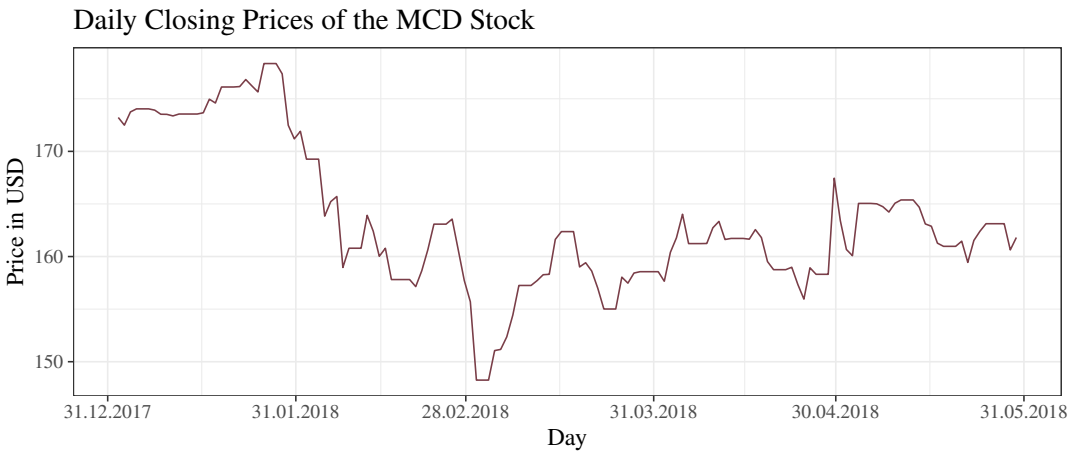

Figure 1.2: Daily closing prices of the MCD stock from January, 2018 to May, 2018.

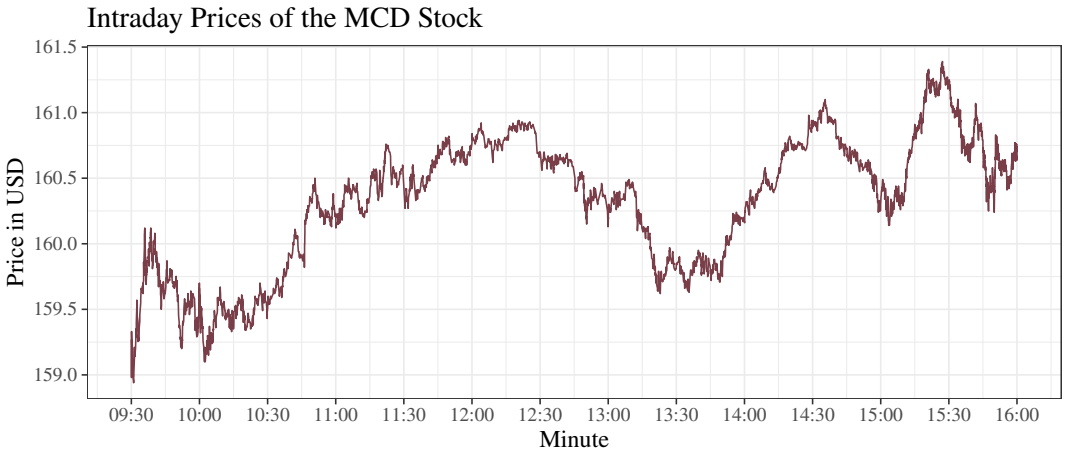

Figure 1.3: Intraday prices of the MCD stock during trading hours on February 22, 2018.

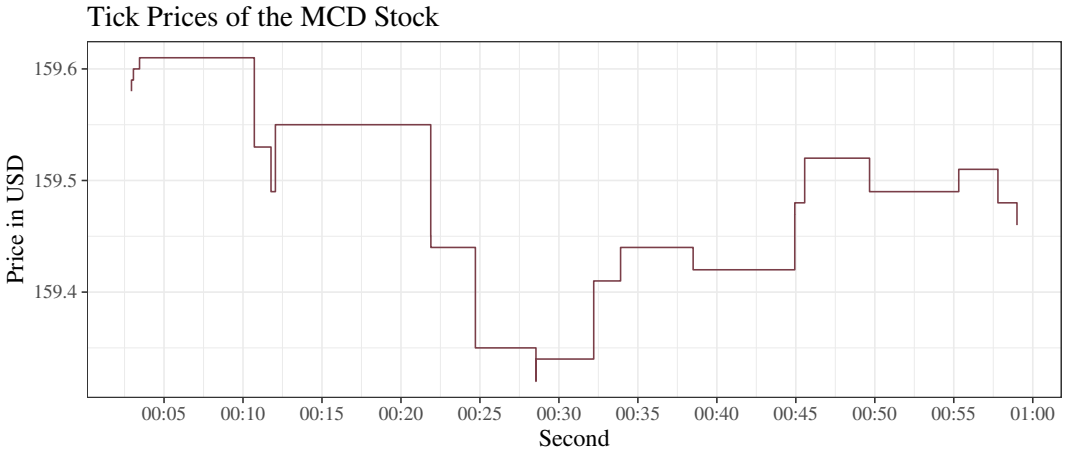

Figure 1.4: Tick prices of the MCD stock during the first minute at 10 a.m. on February 22, 2018.

1.1.2 Market Microstructure

The *market microstructure theory* studies trading process and formation of prices and volumes. The term *market microstructure* was coined by Garman (1976). For a survey of this topic, see Madhavan (2000) and Biais et al. (2005). We focus on implications of the market microstructure theory on the characteristics of the price process. Delbaen and Schachermayer (1994) show that under the assumption of no arbitrage, the *efficient price* process must follow a semimartingale. This efficient price is, however, unobservable due to various frictions in the trading process such as discretness of price values and bid-ask spread. Much of the high-frequency literature is devoted to uncover properties of the efficient price, especially its variance. The efficient price is further discussed in Section 2.2.2.

The most common approach is to model the observed price as the sum of the efficient price and the so-called *market microstructure noise* capturing all trading frictions and informational effects. For more details about this additive model, see e.g. Aït-Sahalia and Jacod (2014). The market microstructure noise has a significant influence on volatility estimation. Usually, the impact of the noise is assessed by the *volatility signature plot* of Andersen et al. (2000). Rosenbaum (2011) introduced the *microstructure noise index* allowing for more comprehensive assessment. Formal tests for the presence of the market microstructure noise were proposed by Awartani et al. (2009) and Aït-Sahalia and Xiu (2016). Statistical properties of the market microstructure noise were analyzed by Bandi and Russell (2006), Hansen and Lunde (2006), Aït-Sahalia and Yu (2009), Ubukata and Oya (2009), Diebold and Strasser (2013), Mancini (2013), Jacod et al. (2017), Taylor (2016) and Dong and Tse (2017b). Diebold and Strasser (2013) analyzed the impact of behavior of economic agents on cross-dependency of the market microstructure noise. Hendershott and Menkveld (2014) studied deviations from the efficient price caused by price pressures. Tsai and Lyuu (2017) estimated the efficient price contaminated by the noise using a robust Kalman filter. The market microstructure noise is further discussed in Section 2.2.3.

Discretness of price values is closely related to the issue of rounding. Jacod (1996), Delattre and Jacod (1997), Rosenbaum (2009) and Li and Mykland (2015) analyze effects of rounding on continuous stochastic processes, especially on estimation of their volatility. The rounding model is briefly visited in Section 2.2.4.

Another approach capturing market microstructure specifics is the *uncertainty zones* model of Robert and Rosenbaum (2011, 2012).

1.2 Analysis of High-Frequency Data

The high-frequency literature offers many methods for analysis of the price process. Most notably, duration analysis deals with modeling times between financial events such as transactions while volatility analysis deals with estimating and forecasting quadratic variation and integrated variance.

We illustrate the benefits of the proper use of high-frequency data in the following experiment. We consider three approaches in volatility estimation based on the used frequency and method.

- *Use low frequency and ignore the market microstructure noise.* This is the simplest approach without any major consequences as the market microstructure noise is quite negligible at lower frequencies. However, these estimates are not very precise as most of the information contained in the price process is discarded.

- *Use high frequency and ignore the market microstructure noise.* When we use the same methods for high-frequency data as for lower frequencies, we begin to face some issues caused by the market microstructure noise. These estimates are significantly biased because we estimate volatility of the noise rather than the price.

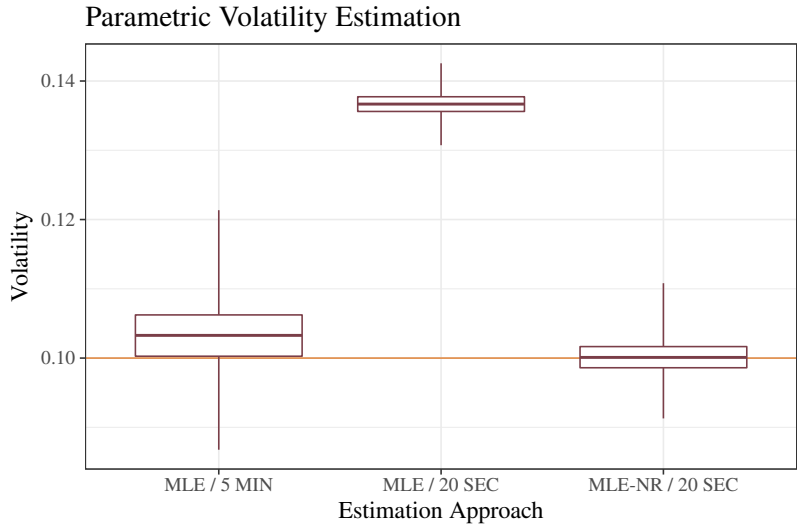

Figure 1.5: Illustrative box plot of volatility estimates by parametric methods.

- *Use high frequency and take the market microstructure noise into account.* The best option is to use the highest frequency possible and methods capable of separating the price and the noise. These estimates are unbiased and more accurate than the ones using lower frequencies.

To confirm these propositions, we simulate the Ornstein–Uhlenbeck process and then estimate its volatility by parametric methods. We compare the maximum likelihood estimator (MLE) using 5-minute data and 20-second data with the noise-robust specification of the maximum likelihood estimator (MLE-NR) of Holý and Tomanová (2018) using 20-second data. Both of these maximum likelihood methods are described in Section 5.1.2. Figure 1.5 shows that the use of higher frequencies and noise-robust methods gives the most accurate estimates. A very similar result can be obtained for non-parametric estimation of integrated variance.

In the rest of this section, we review various methods used in duration and volatility analysis as well as other high-frequency topics.

1.2.1 Duration Analysis

Duration analysis focuses on times between some financial events. There are three commonly analyzed events that can be utilized for various purposes. Durations between successive transactions are known as *trade durations* and can be used as a proxy for trading intensity. Durations until the price changes by a given value are known as *price durations* and can be used as a proxy for volatility. Finally, durations until the trading volume reaches a given amount are known as *volume durations* and can be used as a proxy for liquidity. Engle and Russell (1998) noticed a clustering pattern in durations and proposed to model them by the *autoregressive conditional duration (ACD) model*. For the literature review of duration analysis, see Pacurar (2008), Bauwens and Hautsch (2009) and Saranjeet and Ramanathan (2019).

Many extensions of the original ACD model have been proposed in the literature. Bauwens and Giot (2000) introduced the *logarithmic ACD model* utilizing the logarithmic transformation and exogenous variables. Logarithmic model with a slightly different dynamic was considered by Lunde (1999). Other proposed models include the *fractionally integrated ACD model* of Jasiak (1998), *threshold ACD model* of Zhang et al. (2001), *Box-Cox ACD model* of Hautsch (2001, 2003), *asymmetric ACD model* of Bauwens and Giot (2003), *additive and multiplicative ACD model* of Hautsch (2011), *directional ACD model* of Jeyasreedharan et al. (2014) and *zero-inflated ACD model* of Blasques, Holý and Tomanová (2018). Time-

varying and non-stationary ACD models were studied by Bortoluzzo et al. (2010) and Mishra and Ramanathan (2017). Joint models for durations and prices were proposed by Engle (2000), Grammig and Wellner (2002), Russell and Engle (2005) and Herrera and Schipp (2013). Duration models are described in more detail in Chapter 3.

A different approach for financial events also exists. Russell (1999) modeled transaction arrivals in terms of intensities rather than durations. The main motivation behind this is multivariate analysis. Unlike durations, intensities are defined in continuous time and are therefore suitable for multivariate generalization. Russell (1999) proposed the *autoregressive conditional intensity (ACI)* model based on similar autoregressive structure as the ACD model. For the literature review of various intensity models, see Bauwens and Hautsch (2009).

1.2.2 Volatility Analysis

Volatility is the key object in financial analysis. For the literature review of high-frequency volatility analysis, see Barndorff-Nielsen and Shephard (2007) and McAleer and Medeiros (2008a). Typically, volatility over a given time frame is measured by the *quadratic variation* or *integrated variance*. A natural estimator of quadratic variation is the *realized variance*. It is simple and consistent estimator in the absence of the market microstructure noise. When the noise is present, however, it is biased and inconsistent. Properties of the realized variance were studied by Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002a,b, 2004), Bandi and Russell (2005), Gonçalves and Meddahi (2009) and Fukasawa (2010a,b). It is possible to reduce the bias induced by the market microstructure noise by sampling at lower frequency for the cost of data loss. The optimal sampling frequency was studied by Aït-Sahalia et al. (2005), Zhang et al. (2005), Bandi and Russell (2006, 2008) and De Pooter et al. (2008).

There are many alternative estimators of quadratic variation and integrated variance. The first non-parametric estimator dealing with the market microstructure noise was the *bias-corrected estimator* of Zhou (1996). It was further studied by Hansen and Lunde (2006). Aït-Sahalia et al. (2005) took a parametric approach assuming the Wiener process and proposed the *maximum likelihood estimator*. Xiu (2010) and Aït-Sahalia et al. (2010) further studied this estimator in the context of quasi-maximum likelihood. Parametric approach was also adopted by Holý and Tomanová (2018) for the Ornstein–Uhlenbeck process. The first consistent noise-robust non-parametric estimator was the *two-scale estimator* of Zhang et al. (2005). It was later extended by Zhang (2006) to the *multi-scale estimator* and was further studied by Aït-Sahalia et al. (2011). Barndorff-Nielsen et al. (2008) proposed the *realized kernel estimator*. It was further studied and extended by Barndorff-Nielsen et al. (2009), Bandi and Russell (2011), Barndorff-Nielsen et al. (2011) and Ikeda (2015). Jacod et al. (2009) proposed the *pre-averaging estimator*. It was further studied and extended by Christensen et al. (2010), Hautsch and Podolskij (2013), Jacod and Mykland (2015) and Liu et al. (2017). Other estimators include the *Fourier series estimator* of Malliavin and Mancino (2002), *wavelet estimator* of Høg and Lunde (2003), *Hayashi–Yoshida covariance estimator* of Hayashi and Yoshida (2005), *alternation estimator* of Large (2011), *discrete sine transform estimator* of Curci and Corsi (2012), *least squares estimator* of Nolte and Voev (2012), *uncertainty zones estimator* of Robert and Rosenbaum (2012), *maximum overlap discrete wavelet estimator* of Baruník and Vácha (2015) and *state space estimator* of Nagakura and Watanabe (2015). Sun (2006) and Andersen et al. (2011) established the class of quadratic form estimators to which many of these estimators belong. Some of the estimators were compared by Brownlees and Gallo (2010), Gatheral and Oomen (2010), Sanfelici and Uboldi (2014) and Liu et al. (2015). The estimators of quadratic variation and integrated variance are described in more detail in Section 4.2.

Another topic is volatility modeling and forecasting. Several models were specifically designed to utilize high-frequency volatility. Ghysels et al. (2004, 2006) proposed the *mixed-frequency data sampling (MIDAS) model* considering higher frequencies in explanatory variables. It was further studied and extended by Ghysels et al. (2007), Andreou et al. (2010), Marcellino and Schumacher (2010), Ghysels and Sinko (2011) and Foroni et al. (2015). Corsi (2009) proposed the *heterogeneous autoregressive (HAR)*

model utilizing realized measures over different time horizons. It was further studied and extended by McAleer and Medeiros (2008b), Busch et al. (2011), Patton and Sheppard (2015) and Čech and Baruník (2017). Shephard and Sheppard (2010) proposed the *HEAVY model* relating realized measure to returns. It was further studied and extended by Noureldin et al. (2012). Hansen et al. (2012) proposed the *realized GARCH model* augmenting the regular GARCH model by a realized measure. It was further studied and extended by Watanabe (2012), Hansen et al. (2014), Baruník et al. (2016), Huang et al. (2016) and Jiang et al. (2018). Other studies dealing with volatility forecasting and comparing volatility models include Andersen et al. (2003), Koopman et al. (2005), Aït-Sahalia and Mancini (2008), Chiriac and Voev (2008), Andersen et al. (2011), Çelik and Ergin (2014) and Taylor (2017). Volatility models are described in more detail in Section 4.3.

Volatility at a given time point is known as the *spot volatility* or *instantaneous volatility*. It was estimated by Fan and Wang (2008), Lahalle et al. (2008), Ngo and Ogawa (2009), Kristensen (2010), Ogawa and Sanfelici (2011), Alvarez et al. (2012), Bos et al. (2012), Dahlhaus and Neddermeyer (2014), Zu and Boswijk (2014), Mancini et al. (2015), Bandi and Renò (2018), Liu et al. (2018a) and Liu et al. (2018b).

1.2.3 Higher Moments Analysis

Integrated power variation is a generalization of integrated variance allowing for arbitrary integrated powers of volatility. A special case is the fourth power known as the *integrated quarticity*. The integrated power variation can be estimated by the *realized multipower variation*, which contains the realized variance, bi-power variation and realized quarticity as special cases. It was studied by Barndorff-Nielsen (2004) and Andersen et al. (2012). It is, however, sensitive to the market microstructure noise. For this reason, Podolskij and Vetter (2009) proposed the *modulated multipower variation* which is robust to the market microstructure noise. Jacod and Rosenbaum (2013) further generalized integrated variance and estimated arbitrary functional of volatility in the absence of the noise. Mancino and Sanfelici (2012) focused on the spot volatility and estimated its fourth power known as the *spot quarticity*. They utilized Fourier analysis and considered the market microstructure noise.

1.2.4 Jump Analysis

The price process often contains jumps or can even be solely formed by jumps. Huang and Tauchen (2005), Barndorff-Nielsen and Shephard (2006), Jiang and Oomen (2008) and Christensen et al. (2014) tested whether jumps are present in the price process. Xue et al. (2014) detected jumps using wavelets. Aït-Sahalia and Jacod (2009) proposed the *jump activity index* measuring the degree of the activity of jumps in the price process. It was further studied by Jing et al. (2012b) and Kong (2012). Aït-Sahalia and Jacod (2010) and Jing et al. (2012a) also tested whether the price process can be modeled purely by the jump process. Pure jump processes were further studied and utilized by Oomen (2005, 2006), Large (2011) and Li et al. (2017).

1.2.5 Liquidity Analysis

We also address liquidity and its measurement. A natural way to measure liquidity is the bid-ask spread. Goyenko et al. (2009) advocated the use of modified bid-ask spreads – the *effective spread* and *realized spread*. Liquidity can also be measured using the ACD model for volume durations. This was adopted by Hautsch (2001) and Hautsch (2003). Similarly, Russell (1999) utilized the ACI model based on volume intensity to determine liquidity. Engle and Lange (2001) propose to estimate the so-called *VNET* measuring volume over a price duration. Another approach lies in modeling depth of the order book as followed by Hautsch and Huang (2012) and Härdle et al. (2012).

1.3 Impact of High-Frequency Data

The ultimate goal of any financial analysis is to generate profit or prevent loss. Incorporating high-frequency data into financial analysis helps to achieve such goals. We focus on four crucial aspects of quantitative finance – derivative valuation, risk management, portfolio optimization and trading strategies. We review the literature examining the economic value of high-frequency data.

1.3.1 Derivative Valuation

The first application of high-frequency data analysis we review is derivative valuation. Bollerslev and Zhang (2003) improved the multi-factor asset pricing model by incorporating high-frequency information. Their empirical study shows that high-frequency-based factor loadings yield better returns than the conventional monthly rolling regression-based estimates. Bandi and Russell (2008) and Bandi et al. (2008a) focused on finite sample performance of several quadratic variation estimators with application to option pricing. Another comparison of quadratic variance estimators in the context of option pricing was performed by Sanfelici and Uboldi (2014). Corsi et al. (2013) proposed an option pricing model with HAR forecasts of quadratic variation. In an empirical analysis of S&P 500 index options, they show that their model outperforms competing time-varying and stochastic volatility option pricing models. Stentoft (2008) incorporated realized variance into option pricing model and concluded that the proposed model explains some of the mispricings found when using traditional option pricing models based on daily data. Christoffersen et al. (2014) developed a class of option pricing models utilizing daily returns and realized variance. Their analysis of S&P 500 index showed that realized variance reduces the pricing errors of the benchmark model significantly across moneyness, maturity, and volatility levels. Kenmoe and Sanfelici (2014) utilized high-frequency spot volatility in derivative pricing model and compared several spot volatility estimators. Empirical results showed that using intraday data rather than daily data provides smaller pricing errors. Audrino and Fengler (2015) compared observed realized variance with realized variances implied by the Black-Scholes model, the Heston model and the Bates model. They found that there are significant deviations between the two approaches. Singh and Vipul (2015) tested the performance of Black-Scholes model with the two-scaled realized volatility. Even with the use of high-frequency information, they found that this model is inadequate due to a negative pricing bias. Jeon et al. (2016) evaluated the option market using GARCH-M model and its high-frequency extension in the Bayesian framework. In an empirical study, they found that their model explains a behavior of option prices close to the expiry. Li et al. (2017) focused on the analysis of the VIX index. They captured VIX dynamics as a pure jump semimartingale with infinite jump activity and infinite variation.

1.3.2 Risk Management

High-frequency data can be utilized in evaluating systematic risk. Popular risk measures are the *value-at-risk* and *expected shortfall* also known as *conditional value-at-risk*. One of the earliest uses of high-frequency data in value-at-risk forecasting was in article of Beltratti and Morana (1999). Giot and Laurent (2004) compared the realized variance ARFIMAX model with the daily ARCH model for daily value-at-risk forecasts. They conclude that there is no significant improvement when using realized variance. Kruse (2006) incorporated the realized variance in value-at-risk estimation by extreme value theory and filtered historical simulation. They found that the best performing forecasting models are hybrid specifications based on realized variance and the filtered historical simulation. Value-at-risk estimates by extreme value theory and filtered historical simulation with realized variance were also analyzed by Louzis et al. (2011). Clements et al. (2008) compared several models for volatility and quantile forecasts using high-frequency data and found that the HAR model with empirical distribution provides the most accurate forecasts. McMillan et al. (2008) analyzed intraday periodicity, the presence of short horizon as well as long horizon dependencies and daily realized measures in the context of value-at-risk forecasting. Brownlees and Gallo (2010) compared different volatility measures in the context of value-at-risk forecasting. They found that the realized kernel estimator is superior to the realized variance, bi-power variation,

two-scale realized variance and realized range in terms of value-at-risk predictive ability. Realized range was also used by Shao et al. (2009) in value-at-risk forecasting. Herrera and Schipp (2013) estimated value-at-risk by extreme value theory in combination with autoregressive conditional duration model. Huang and Lee (2013) incorporated high-frequency information in value-at-risk forecasting models by combining forecasts based on different intraday intervals and by combining high-frequency information into a single model. Žikeš and Baruník (2015) modeled value-at-risk by quantile regression with HAR specification of several quadratic variation estimators.

Value-at-risk can also be modeled at intraday level. Several intraday high-frequency risk measures based on quantiles were proposed by Giot (2005), Giot and Grammig (2006), Dionne et al. (2009), So and Xu (2013) and Banulescu et al. (2016).

Although the value-at-risk is dominant in the high-frequency literature, several articles deal with the expected shortfall as well. Guo and Zhang (2008) estimated the expected shortfall using weighted realized variances. Watanabe (2012) utilized realized GARCH model in forecasting value-at-risk and expected shortfall. Bee et al. (2016) used realized extreme value theory for value-at-risk and expected shortfall forecasts.

1.3.3 Portfolio Optimization

An important area of quantitative finance is portfolio optimization. In this application, high-frequency data can be utilized for more precise estimation and prediction of daily volatility. Fleming et al. (2003) measured the economic value of high-frequency data in the context of investment decisions. Their results indicate that a risk-averse investor would be willing to pay substantial fees to capture the observed gains in portfolio performance. Bandi et al. (2008b) evaluated the economic benefits of integrated variance estimates in dynamic portfolio choice when the prices are contaminated by the market microstructure noise. De Pooter et al. (2008) found that when forming the mean-variance efficient stock portfolios with daily rebalancing, the optimal sampling frequency for realized variance in the presence of the market microstructure noise ranges between 30 and 65 minutes. Liu (2009) examined the frequency of portfolio rebalancing at which the use of intraday high-frequency data is beneficial. They found that for monthly rebalancing, the use of daily data is sufficient. However, for daily rebalancing, the use of high-frequency data brings substantial improvements. Hautsch et al. (2015) analyzed high-dimensional portfolio allocations. They found that the predictions based on high-frequency data yield a significantly lower portfolio volatility than methods employing daily returns.

1.3.4 Trading Strategies

There are many trading strategies whether they are labeled as high-frequency trading or operate in longer time horizons. Description of all these strategies is beyond the scope of this thesis and we refer to the book of Aldridge (2013).

We focus only on one particular strategy called the *pairs trading*. It is based on taking advantage of two prices exhibiting strong similarity in the long run that are temporarily out of equilibrium. Liu et al. (2017) introduced the doubly mean-reverting processes for capturing the high-frequency price differences and described related intraday trading strategy. Intraday data for the pairs trading were also utilized by Dunis and Lequeux (2000), Bowen et al. (2010), Peters et al. (2011) and Miao (2014). Holý and Tomanová (2018) modeled the price differences as the Ornstein–Uhlenbeck process and estimated its parameters using ultra-high-frequency data contaminated by the market microstructure noise. They found that ignoring the noise leads to much higher estimates of volatility and speed of reversion parameters resulting in suboptimal decision-making. Using the proposed noise-robust estimator brings a significant additional profit over the strategy based on traditional estimators. We show this application in Chapter 5.

- Chapter 2 -

Practical and Theoretical Framework

High-frequency data have several features requiring a special treatment from both practical and theoretical perspective. Before an econometric analysis, high-frequency data should be subject to data preprocessing including data cleaning, data transformation and data aggregation. In the actual analysis, all used statistical methods should account for the specifics inherent to high-frequency data and persistent after the preprocessing procedures. The most distinctive features of high-frequency data are the following.

- As data are collected at high velocities and large volumes, many errors occur during the collection. Such errors include prices recorded as zeros, misplaced decimal points, missing observations and observations outside the trading hours. Careful data cleaning in the preprocessing step is always necessary when working with high-frequency data. Data cleaning procedures are discussed for example by Falkenberry (2002), Brownlees and Gallo (2006) and Barndorff-Nielsen et al. (2009).
- Data observed at the highest possible frequency are denoted as *ultra-high-frequency data* by Engle (2000). Such data are irregularly spaced, i.e. the spacing of observation times is not constant. As many statistical methods are based on regularly spaced data, irregularly spaced time series can simply be aggregated to equidistant time series. However, this can induce a loss of information or even a bias. Some methods, on the other hand, can be directly utilized for irregularly spaced data (e.g. realized variance) or be modified for such case (e.g. maximum likelihood estimation). Duration analysis pioneered by Engle and Russell (1998) aims to model random times between observations by autoregressive processes.
- Trades happen on either bid or ask side. Whether analyzing transaction data, quote data or both, this needs to be taken into account. For transaction data, the *bid-ask bounce effect* caused by transaction price oscillating between bid and ask prices occurs. It significantly distorts volatility and is treated by modeling the price process contaminated by the so-called market microstructure noise. For quote data, bid and ask prices can be transformed to mid prices as their mean. However, mid prices also exhibit the presence of the market microstructure noise.
- Data are always discrete as they are recorded with a given precision. In some cases, the representation error can be negligible, while in other cases, it can cause a bias or even inconsistent estimates. One approach is to consider a model for continuous values that are observed with some sort of rounding error. For example, the model with market microstructure noise can be utilized. The other approach is to directly model values in a discrete framework. Examples of discrete models include Russell and Engle (2005), Koopman et al. (2015) and Blasques, Holý and Tomanová (2018).

Figure 2.1 illustrates some contradictions between the observed price process and the theoretical price process. One of the tasks of the analyst is to find a balance between characteristics of the original data and assumptions of the considered theory.

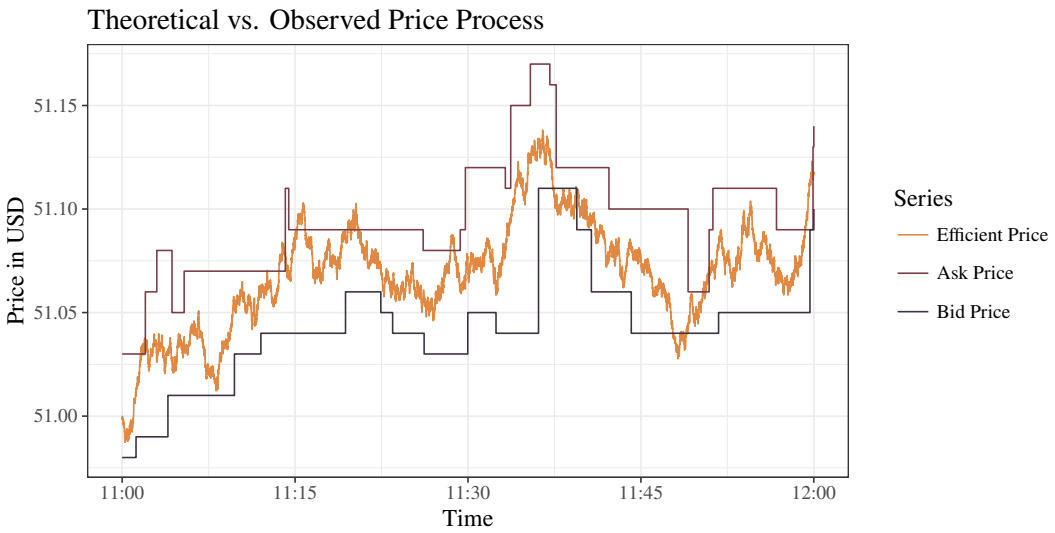

Figure 2.1: Simulated example of the theoretical efficient price and the observed bid and ask prices.

2.1 Preprocessing Procedures

In this section, we discuss some important procedures in the preprocessing of financial high-frequency data. First, we describe necessary steps in the data cleaning procedure for both transaction and quote data. Second, we briefly address the logarithmic returns and mid price interpolation. Third, we present various sampling schemes for temporal data aggregation.

2.1.1 Data Cleaning

In the thesis, we analyze prices of stocks traded on the NYSE and NASDAQ exchanges obtained from the Daily TAQ database of New York Stock Exchange (2019) (or simply NYSE TAQ database). Data cleaning of NYSE TAQ database is performed for example by Brownlees and Gallo (2006) and Barndorff-Nielsen et al. (2009). We describe the standard data cleaning procedure of Barndorff-Nielsen et al. (2009) with some slight modifications.

The cleaning steps can be categorized into three classes. First, *irrelevant entries* are removed. Second, *simultaneous entries* are merged. Third, *erroneous entries* are removed. For the pairs trading strategy in Chapter 5 and the volatility analysis in Chapter 4, we apply steps from all three classes relevant to transaction data. For the duration analysis in Chapter 3, we omit merging simultaneous entries as the goal is to model zero durations between observations.

Irrelevant Entries

Retain entries originating from a single exchange. Delete other entries. This step corresponds to P3 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009). Brownlees and Gallo (2006) stated that they prefer not to discard transaction prices that did not occur on the single exchange. However, in some cases this is not advisable as discussed e.g. by Dufour and Engle (2000). In duration analysis, this cleaning step is often used to reduce the impact of time-delays in the trade updates reporting (exchanges can have different latencies).

Delete all trades and quotes with a timestamp outside the window when the exchange is open. The normal trading hours of the NYSE and NASDAQ exchanges are from 9:30 a.m. to 4:00 p.m. in the eastern time zone. This step corresponds to P1 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries with corrected trades. For the NYSE TAQ database, corrected trades are denoted by the correction indicator 'CORR' other than 0. This step removes trades that were corrected, changed, or signified as cancel or error and corresponds to T1 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries with abnormal trades. For the NYSE TAQ database, abnormal trades are denoted by the sale condition 'COND' having a letter code, except for 'E', 'F' and 'I'. This step rules out data points that the NYSE TAQ database is flagging up as a problem and corresponds to T2 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries which are identified as preferred or warrants. For the NYSE TAQ database, all trades with the non-empty SUFFIX indicator should be deleted.

Simultaneous Entries

Merge entries with the same timestamp. Merging itself can be done using the mean (Aït-Sahalia et al., 2010), the median (Barndorff-Nielsen et al., 2009), the mean weighted by the volume (Christensen et al., 2010), a single random price (Jing et al., 2017) or the last recorded price (Jing et al., 2017). For transaction data using median, this step corresponds to T3 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009). For quote data using median, this step corresponds to Q1 rule of Barndorff-Nielsen et al. (2009). Merging simultaneous entries is quite controversial as it leads to the largest deletion of data and a significant information loss. In volatility analysis, Barndorff-Nielsen et al. (2009) argue that this rule seems inevitable. However, Liu et al. (2018c) estimate integrated variance by the pre-averaging estimator using data with multiple observations at the same time. In duration analysis, simultaneous transactions are also merged in the majority of the literature. The exception is Blasques, Holý and Tomanová (2018) who do not discard zero durations and directly include them in the zero-inflated autoregressive conditional duration model.

Erroneous Entries

Delete entries with the price equal to zero. This step removes obvious errors in the dataset and corresponds to P2 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries for which the spread is negative. This step corresponds to Q2 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries for which the spread is more that 50 times the median spread on that day. This step corresponds to Q3 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries for which the price deviated by more than 10 mean absolute deviations from a rolling centred median of 50 observations. The observation under consideration is excluded in the rolling centered median. The mid price (the mean of the bid and ask prices) can be utilized for the quote data. This step is closely related to the procedure of Brownlees and Gallo (2006) which advocates removing outliers. For quote data, this step corresponds to Q4 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

Delete entries with prices above the ask price plus the bid-ask spread. Delete entries with prices below the bid price minus the bid-ask spread. This step corresponds to T4 rule of the cleaning procedure of Barndorff-Nielsen et al. (2009).

2.1.2 Data Transformation

The goal of this section is to transform observed data into a single price process suitable for further analysis. We establish concepts of the logarithmic price process, logarithmic returns process and mid price process.

Returns and Logarithmic Scale

Let us consider the price process $\tilde{X}_i$ observed at times $i = 0, \dots, n$. This process can represent transaction prices, bid prices, ask prices, or any other kind of prices. In any case, the object of interest of the financial analysis is often the returns process rather than the price process itself. The *raw returns process* is defined as

$$\tilde{Y}_i = \frac{\tilde{X}_i - \tilde{X}_{i-1}}{\tilde{X}_{i-1}}, \qquad i = 1, \dots, n. \tag{2.1}$$

An advantage of the returns process is that it is normalized in the sense that a performance of different assets can be measured by a comparable metric.

Next, we discuss the logarithmic transformation. The *logarithmic price process* is defined as

$$X_i = \log \tilde{X}_i, \qquad i = 0, \dots, n. \tag{2.2}$$

The *logarithmic returns process* is defined as

$$Y_i = \log\left(\frac{\tilde{X}_i}{\tilde{X}_{i-1}}\right) = \log \tilde{X}_i - \log \tilde{X}_{i-1} = X_i - X_{i-1}, \qquad i = 1, \dots, n. \tag{2.3}$$

The logarithmic returns are widely used in finance as they have many desirable properties. First, when the raw return $\tilde{Y}_i$ is small, the logarithmic return Y_i is approximately the same as the raw return $\tilde{Y}_i$. Specifically, we have

$$Y_i = \log\left(1 + \tilde{Y}_i\right) = \tilde{Y}_i + O\left(\tilde{Y}_i^2\right), \qquad \text{for} \quad |\tilde{Y}_i| < 1, \quad i = 1, \dots, n. \tag{2.4}$$

This follows from the sum of infinite series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \tilde{Y}_i^k = \log\left(1 + \tilde{Y}_i\right) \qquad \text{for} \quad |\tilde{Y}_i| < 1, \quad i = 1, \dots, n. \tag{2.5}$$

Second, the logarithmic returns are additive in time. Specifically, the logarithmic return over period $t + s$ is the sum of the logaritmic return over period t and the logaritmic return over period s. Third, the logarithmic returns are more numerically stable. Specifically, the logarithmic transformation reduces a large range of values to a more manageable range. A disadvantage of logarithmic returns lies in the aggregation of assets in a portfolio. Specifically, the raw return of the portfolio is the weighted average of raw returns of the assets in the portfolio. For logarithmic returns, such simple relation does not hold. In the rest of the thesis, we assume that all prices are logarithmic prices.

Mid Price Interpolation

Let us consider the case of quote data. Specifically, we observe the bid price X_i^B and ask price X_i^A at times $i = 0, \dots, n$. Naturally, we have $X_i^B \leq X_i^A$ for all $i = 0, \dots, n$. The efficient price of a financial asset can be approximated by the *mid price* X_i^M interpolated from the bid price X_i^B and ask price X_i^A as

$$X_i^M = \frac{1}{2}\left(X_i^A + X_i^B\right), \qquad i = 0, \dots, n. \tag{2.6}$$

An example of mid price interpolation is shown in Figure 2.2. The mid prices were used in high-frequency data analysis for example by Hansen and Lunde (2006).

2.1.3 Data Aggregation

Let us consider that we observe ultra-high-frequency data at times $T_0 \leq T_1 \leq \cdots \leq T_n$. Without loss of generality, we assume the times of observations fill the interval $[0, 1]$, i.e. $0 = T_0 \leq T_1 \leq \dots \leq T_n = 1$. We present three alternative sampling schemes. Various sampling schemes are studied for example by

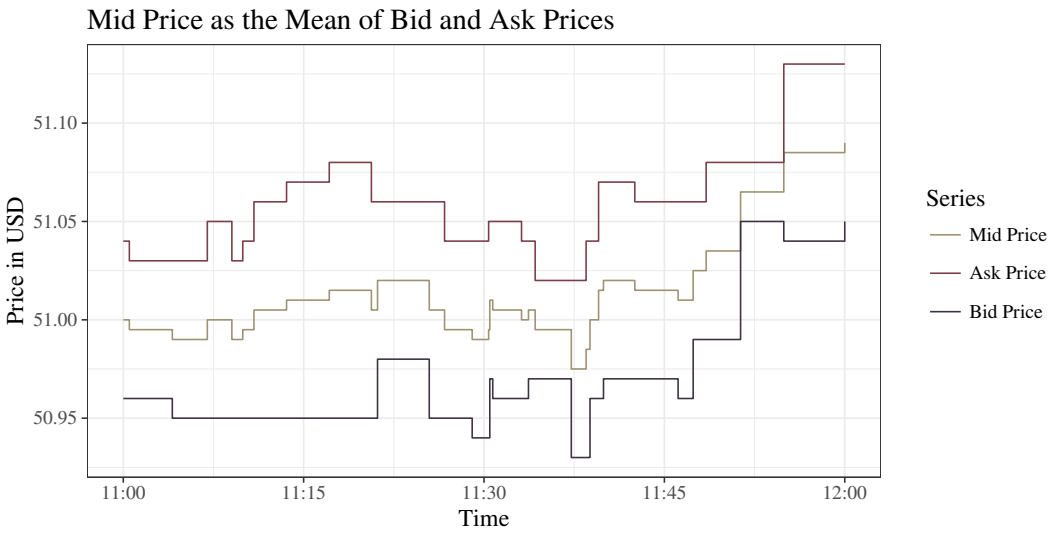

Figure 2.2: Simulated example of the mid price interpolated from the observed bid and ask prices.

Aït-Sahalia and Mykland (2003), Oomen (2005, 2006), Hansen and Lunde (2006), Fukasawa (2010a,b) and Dong and Tse (2017a).

As price values are only available for times T_i, $i = 0, \dots, n$, we must interpolate price value for some other time $S \neq T_i$, $i = 0, \dots, n$. This is usually done by the *last tick method*, which simply takes the last observed value, i.e. the value at time $\max\{i : T_i < S\}$.

Tick Time Sampling

The *tick time sampling* is based on regularly spaced number of ticks. For an initial tick h and sampling frequency s, the tick time sampling is a set of times given by

$$TTS_{h,s} = \left\{ S_j = T_{h+js} : j = 0, \dots, \left\lfloor \frac{n-h}{s} \right\rfloor \right\}. \tag{2.7}$$

The tick time sampling with $h = 0$ and $s = 1$ reduces to the original sampling of ultra-high-frequency data. This sampling is used for example for volatility signature plots and the sparse realized variance defined in Section 4.2.1.

Calendar Time Sampling

The *calendar time sampling* is based on regularly spaced calendar time. For the number of observations m, the calendar time sampling is a set of times given by

$$CTS_m = \left\{ S_j = jm^{-1} : j = 0, \dots m \right\}. \tag{2.8}$$

Business Time Sampling

For the number of observations m and the price process P_t, $t \geq 0$, the *business time sampling* is a set of times given by

$$BTS_m = \left\{ 0 = S_0 < S_1 < \cdots < S_m = 1 : \operatorname{var}\left[P_{S_1} - P_{S_0} \right] = \cdots = \operatorname{var}\left[P_{S_m} - P_{S_{m-1}} \right] \right\}. \tag{2.9}$$

Unlike the tick times and the calendar times, the business times S_j are latent as we do not observe variance of returns. An advantage of the business time sampling is that it yields independent and identically distributed normal returns for a semimartingale price process (see Dong and Tse, 2017a). The returns under business time sampling are also known as *devolatilized returns*.

2.2 Theoretical Models

In this section, we discuss foundations of theoretical models considered in the financial high-frequency data analysis. First, we consider deterministic and stochastic settings for times of observations. Second, we assume the price process to follow semimartingale. Third, we describe the widely use additive model in which the price process is contaminated by the market miscrostructure noise. Fourth, we address the issue of rounding. Fifth, we introduce the model based on interval uncertainty to deal with the bid-ask spread and rounding.

2.2.1 Times of Observations

Let us denote the times of observations as $T_0 \leq T_1 \leq \cdots \leq T_n$. Without loss of generality, we assume the times of observations lie in the interval $[0, 1]$, i.e. $0 \leq T_0 \leq T_1 \leq \cdots \leq T_n \leq 1$. Durations between successive observations are then given by $D_i = T_i - T_{i-1}$, $i = 1, \dots n$. We assume three settings for the times of observations.

In the first setting, observations are equally spaced. Durations are then constant, i.e. $D_i = \Delta = n^{-1}$. In this case, the times of observations are of course deterministic. This is the most elementary assumption in time series analysis. However, it is not very suitable for financial ultra-high-frequency data as calendar time sampling is necessary for this assumption to hold as discussed in Section 2.1.3. We utilize this setting in parts of Chapter 5.

In the second setting, observations are irregularly spaced and their times are deterministic. We further assume that the times of observations are a stricty increasing sequence, i.e. $0 \leq T_0 < T_1 < \cdots < T_n \leq 1$. We utilize this setting in Chapter 4 and parts of Chapter 5.

In the third setting, observations are irregularly spaced and their times are random variables. The times of observations then form a point process on the interval $[0, 1]$. We further assume that times T_i, $i = 0, \dots, n$ are independent from the price process X_i, $i = 0, \dots, n$. We utilize this setting in Chapter 3.

2.2.2 Efficient Price Process

A lot of the high-frequency literature is centered around the concept of the *efficient price* (see e.g. Aït-Sahalia and Jacod, 2014). The efficient price is the latent price of a financial asset with continuous time. However, this is an idealization as prices are observed at discrete transaction times and nothing between transactions actually exists. The efficient price is therefore a theoretical concept of a scaling limit with frequency of observations shrinking to zero. Nevertheless, it is the key subject of financial high-frequency analysis.

Under the assumption of no arbitrage, the efficient price process must follow a semimartingale (see Delbaen and Schachermayer, 1994). Let us denote P_t, $t \geq 0$ as the logarithmic efficient price. *Càdlàg function*[1] is a function defined on the real numbers that is right-continuous everywhere and has left limits everywhere. A *martingale* is a stochastic process for which the conditional expectation of the next value is equal to the present value. A *semimartingale* is defined as the sum of a local martingale and an adapted càdlàg finite-variation process. This decomposition, however, is not unique. A semimartingale can be expressed as

$$P_t = P_0 + \int_0^t D_z \, \mathrm{d}z + \int_0^t V_z \, \mathrm{d}W_z + \sum_{k \colon S_k \leq t} J_k, \tag{2.10}$$

where D_z is a finite variation càdlàg drift process, V_z is an adapted càdlàg volatility process, W_z is a standard Wiener process and J_k are non-zero random variables with random times $0 \leq S_1 < \cdots < S_m \leq 1$. Note that semimartingales are a very general class of processes as D_z and V_z are both time-varying. Semimartingales form the largest class of processes for which the Itô integral can be defined. Examples

[1] From the French "continue à droite, limite à gauche".

of semimartingales include the Wiener process, Itô processes and Lévy processes. A special class of semimartingales are *continuous Itô semimartingales* given by

$$P_t = P_0 + \int_0^t D_z \, \mathrm{d}z + \int_0^t V_z \, \mathrm{d}W_z, \tag{2.11}$$

where D_z is a finite variation càdlàg drift process, V_z is an adapted càdlàg volatility process and W_z is a standard Wiener process.

Literature about Stochastic Calculus

More about semimartingales and related topics can be found in the stochastic calculus literature. For the theory from the financial perspective, see e.g. Steele (2001), Sondermann (2006), Shreve (2004a,b) or Aït-Sahalia and Jacod (2014). For the general theory, see e.g. Chung and Williams (1990), Karatzas and Shreve (1991), Protter (2004) or Klebaner (2005).

2.2.3 Market Microstructure Noise Model

Not surprisingly, there are significant discrepancies between the theoretical efficient price process and the observed price process. Let us assume the latent efficient price process P_t with continuous time $t \geq 0$. Let us consider that we observe price process X_i at discrete times T_i, $i = 0, \ldots, n$. The unobserved price process P_t and the observed price process X_i are then related as

$$X_i = P_{T_i} + E_i, \qquad E_i \sim (0, \omega^2), \qquad i = 0, \ldots, n, \tag{2.12}$$

where E_i is the *market microstructure noise* capturing all the discrepancies. We further denote

$$Y_i = X_i - X_{i-1}, \qquad R_i = P_{T_i} - P_{T_{i-1}}, \qquad F_i = E_i - E_{i-1}, \qquad i = 1, \ldots, n. \tag{2.13}$$

This model is known as the *additive noise model* and it is the most popular model used in the high-frequency literature analyzing volatility of the price process (see e.g. Aït-Sahalia and Jacod, 2014). The noise is a random variable with zero expected value and constant variance ω^2. Generally, it can be dependent in time and dependent on the efficient price. Indeed, Hansen and Lunde (2006) show in an empirical study of DJIA stocks that the market microstructure noise is auto-correlated and cross-correlated. The noise is caused by the following microstructure effects.

- *Discretness of price values.* The efficient price process has continuous values. The observed prices, however, have discrete values. For example, the stocks traded on the NYSE and NASDAQ exchanges are recorded with precision of 2 decimal points (i.e. one cent). This mismatch can be modeled using the rounding error. We further discuss the rounding issue in Section 2.2.4.

- *Sampling issues.* Times of price changes are recorded using discrete values. Additionally, if one of the sampling schemes presented in Section 2.1.3 is adopted, times of observations are modified as well. Any of these sampling alterations contribute to the market microstructure noise.

- *Bid-ask spread.* In transaction data, the bid-ask bounce effect occurs. It is caused by transaction price oscillating between bid and ask prices. This behaviour can be modeled as an error and a part of the market microstructure noise. In quote data, another error occurs as the mid price is interpolated from the bid and ask prices.

- *Informational effects.* In reality, the efficient price process may temporarily deviate from the semimartingale assumption due to various informational effects. Such effects include asymmetric information, partially incorporated information, strategic behavior, trades on different markets, gradual response to a block trade, inventory control effect, difference in trade sizes and price pressure effect.

Source	Data
Frictions in Trading Process	
- Discretness of Price Values	All Data
- Sampling Issues	All Data
- Bid-Ask Bounce	Transaction Data
- Mid Price Interpolation	Quote Data
Informational Effects	
- Asymmetric Information	All Data
- Partially Incorporated Information	All Data
- Strategic Behavior	All Data
- Trades on Different Markets	All Data
- Gradual Response to a Block Trade	All Data
- Inventory Control Effect	All Data
- Difference in Trade Sizes	All Data
- Price Pressure Effect	All Data
Recording Errors	
- Prices Recorded as Zeros	Low Quality Data
- Misplaced Decimal Points	Low Quality Data
- Prices with Wrong Time of Observation	Low Quality Data
- Missing Observations	Low Quality Data

Table 2.1: Overview of causes of the market microstructure noise.

- *Recording errors.* Various recording errors can also be included in the market microstructure noise. However, the treatment of recording errors should be the subject of data cleaning procedure as discussed in Section 2.1.1.

Aït-Sahalia et al. (2011) divide the causes of the noise into three classes: frictions in trading process, informational effects and recording errors. An overview of sources of the market microstructure noise is shown in Table 2.1.

The additive model can also be extended by letting the variance of the market microstructure noise be dependent on the number of observations n (see e.g. Aït-Sahalia and Jacod, 2014). However, in the thesis, we focus only on the case of the constant variance.

2.2.4 Pure Rounding Model

This section follows Holý (2018b) with different notation and application. In the *pure rounding model*, it is assumed that the observed price process X_i is created by rounding down the efficient price process P_{T_i} to d decimal places, i.e.

$$X_i = \lfloor P_{T_i} \rfloor^{[d]}, \qquad i = 0, 1, \dots, n. \tag{2.14}$$

In this model, the rounding is the only source of uncertainty. The rounding can be alternatively defined as rounding up $X_i = \lceil P_{T_i} \rceil^{[d]}$ or rounding to the nearest possible increment $X_i = \lfloor P_{T_i} \rceil^{[d]}$. However, theoretical as well as empirical results would be almost the same and therefore we focus only on rounding down. The observed process can be decomposed into the sum of the efficient price process and the *rounding error* as

$$X_i = P_{T_i} + E_i, \qquad E_i = \lfloor P_{T_i} \rfloor^{[d]} - P_{T_i}. \tag{2.15}$$

This notation corresponds to the additive noise model. Furthermore, we denote the d-th decimal digit of P_{T_i} as $P_{T_i}^{[d]}$. In this setting, the rounding error E_i has the following properties.

1. The distribution of the rounding error E_i is a deterministic transformation of distribution of the efficient price process P_{T_i}.
2. The rounding error E_i is dependent on the efficient price process P_{T_i}.
3. Generally, the rounding error E_i is dependent in time due to time-dependency of the efficient price process P_{T_i}.

Issues related to the rounding errror in financial stochastic processes were studied by Jacod (1996), Delattre and Jacod (1997), Rosenbaum (2009) and Li and Mykland (2015).

Asymptotic Properties of the Rounding Error

The properties of the rounding error are very difficult to work with. However, its asymptotic properties are more pleasant.

1. The rounding error E_i is asymptotically uniformly distributed. Specifically, the d-th decimal digit $P_{T_i}^{[d]}$ has discrete uniform distribution for $d \to \infty$ and the rounding error E_i has continuous uniform distribution for $d \to \infty$. The asymptotic variance of the rounding error is then

$$\text{var}\left[10^d E_i\right] = \frac{1}{12} \qquad \text{for } d \to \infty. \tag{2.16}$$

2. The rounding error E_i is asymptotically uncorrelated with the efficient price process P_{T_i}, i.e.

$$\text{cor}\left[E_i, P_{T_i}\right] = 0 \qquad \text{for } d \to \infty. \tag{2.17}$$

3. The rounding error E_i is asymptotically uncorrelated in time, i.e.

$$\text{cor}\left[E_i, E_j\right] = 0, \qquad \text{for } i \neq j \text{ and } d \to \infty. \tag{2.18}$$

Kosulajeff (1937) showed that the rounding error of a random variable with absolutely continuous distribution function $F(x)$ tends to the continuous uniform distribution. Tukey (1938) showed that the necessary and sufficient condition for the rounding error to have asymptotically continuous uniform distribution is that the Fourier transform of the distribution function $\hat{F}(\xi)$ tends to zero as $|\xi| \to \infty$.

As an illustration, we offer another proof. In the following proposition, we show that the distribution of the d-th decimal digit of a continuous variable P_t tends to the discrete uniform distribution as $d \to \infty$.

Proposition 2.1. *Let $F_{P_t}(x)$ be the distribution function of a continuous variable P_t and $f_{P_t}(x)$ its density function. Let $P_t^{[d]}$ denote the d-th decimal digit of P_t. For all $\varepsilon > 0$, there exists $c \in \mathbb{N}$ such that for all $d \in \mathbb{N}$, $d \geq c$, the distribution of $P_t^{[d]}$ is given by*

$$\text{P}\left[P_t^{[n]} = a\right] \in \left(10^{-1} - \varepsilon, 10^{-1} + \varepsilon\right), \quad a = 0, 1, \ldots, 9. \tag{2.19}$$

Proof. The distribution of the d-th digit follows

$$\begin{aligned} \text{P}\left[P_t^{[d]} = a\right] &= \sum_{b=-\infty}^{\infty} \text{P}\left[(10b + a)10^{-d} \leq P_{T_i} < (10b + a + 1)10^{-d}\right] \\ &= \sum_{b=-\infty}^{\infty} F_{P_t}\left((10b + a + 1)10^{-d}\right) - \sum_{b=-\infty}^{\infty} F_{P_t}\left((10b + a)10^{-d}\right) \end{aligned} \tag{2.20}$$

for $a = 0, 1, \ldots, 9$. From the existence of the right derivative of $F_{P_t}(x)$ in every point x we have

$$\lim_{d_1 \to \infty} \frac{F_{P_t}\left(x + 10^{-d_1}\right) - F_{P_t}(x) - 10^{-d_1} f_{P_t}(x)}{10^{-d_1}} = 0. \tag{2.21}$$

For a countable set of x_b, $b = -\infty, \ldots, \infty$ we have

$$\sum_{b=-\infty}^{\infty} \lim_{d_1 \to \infty} \frac{F_{P_t}\left(x_b + 10^{-d_1}\right) - F_{P_t}\left(x_b\right) - 10^{-d_1} f_{P_t}\left(x_b\right)}{10^{-d_1}} = 0. \tag{2.22}$$

We change the order of limit and sum to get

$$\lim_{d_1 \to \infty} 10^{d_1} \sum_{b=-\infty}^{\infty} \left(F_{P_t}\left(x_b + 10^{-d_1}\right) - F_{P_t}\left(x_b\right) - 10^{-d_1} f_{P_t}\left(x_j\right) \right) = 0. \tag{2.23}$$

We define the set of x_b as $x_b = \lim_{d_0 \to \infty}(10b + a)10^{-d_0}$. For all $\varepsilon_1 > 0$, there exists $\delta_1 > 0$ such that for all d_1 satisfying $10^{-d_1} \leq \delta_1$, we have

$$10^{d_1} \left| \sum_{b=-\infty}^{\infty} F_{P_t}\left(x_b + 10^{-d_1}\right) - \sum_{d=-\infty}^{\infty} F_{P_t}\left(x_b\right) - \sum_{b=-\infty}^{\infty} 10^{-d_1} f_{P_t}\left(x_b\right) \right| < \varepsilon_1, \quad a = 0, 1, \ldots, 9. \tag{2.24}$$

Next, we take a subset of x_b given by $x_b = (10b + a)10^{-d_1}$. For this subset we have weaker inequality

$$\begin{aligned} 10^{d_1} \Bigg| & \sum_{b=-\infty}^{\infty} F_{P_t}\left((10b + a + 1)10^{-d_1}\right) - \sum_{b=-\infty}^{\infty} F_{P_t}\left((10b + a)10^{-d_1}\right) \\ & - \sum_{b=-\infty}^{\infty} 10^{-d_1} f_{P_t}\left((10b + a)10^{-d_1}\right) \Bigg| < \varepsilon_1, \quad a = 0, 1, \ldots, 9. \end{aligned} \tag{2.25}$$

It can be rewritten as

$$10^{d_1} \left| \mathrm{P}\left[P_t^{[d]} = a\right] - 10^{-1} \sum_{b=-\infty}^{\infty} 10^{-d_1+1} f_{P_t}\left((10b + a)10^{-d_1}\right) \right| < \varepsilon_1, \quad a = 0, 1, \ldots, 9. \tag{2.26}$$

This inequality can be further modified to weaker inequality

$$\left| \mathrm{P}\left[P_t^{[d]} = a\right] - 10^{-1} \sum_{b=-\infty}^{\infty} 10^{-d_1+1} f_{P_t}\left((10b + a)10^{-d_1}\right) \right| < \varepsilon_1, \quad i = 0, 1, \ldots, 9. \tag{2.27}$$

Next, let us analyze the second summand in the absolute value. Its limit is the Riemann integral of density function, which equals to 1, i.e.

$$\lim_{d_2 \to \infty} \sum_{b=-\infty}^{\infty} 10^{-d_2+1} f_{P_t}\left((10b + a)10^{-d_2}\right) = \int_{-\infty}^{\infty} f_{P_t}(x)\ \mathrm{d}x = 1, \quad a = 0, 1, \ldots, 9. \tag{2.28}$$

For all $\varepsilon_2 > 0$, there exists $\delta_2 > 0$ such that for all d_2, $10^{-d_2} \leq \delta_2$, we have

$$\left| \sum_{b=-\infty}^{\infty} 10^{-d_2+1} f_{P_t}\left((10b + a)10^{-d_2}\right) - 1 \right| < \varepsilon_2, \quad a = 0, 1, \ldots, 9. \tag{2.29}$$

Finally, we can put together results (2.27) and (2.29). For all ε, we select ε_1 and ε_2 such that $\varepsilon_1 + 10^{-1}\varepsilon_2 \leq \varepsilon$. For ε_1, there exists δ_1 and for ε_2 there exists δ_2 as above. We select c such that $10^{-c} \leq \delta_1$ and $10^{-c} \leq \delta_2$. For all $d \geq c$ we have

$$\left| \mathrm{P}\left[P_t^{[d]} = a\right] - 10^{-1} \right| < \varepsilon, \quad a = 0, 1, \ldots, 9. \tag{2.30}$$

□

Rounding Error in Stock Prices

The question is whether the rounding error present in financial data can be treated according to its asymptotic properties. We analyze the prices of 30 stocks forming Dow Jones Industrial Average (DJIA) index from January to May, 2018. Prices of these stocks have precision of 2 decimal places. If the rounding error is uniformly distributed, its variance is $8.333 \cdot 10^{-6}$ according to (2.16). The knowledge of the variance of the rounding error can be quite useful in quadratic variance estimation and market microstructure noise investigation.

Table 2.2 shows the variances of the error caused by rounding to zero decimals and one decimal. On average, the observed variance is $8.556 \cdot 10^{-2}$ for zero decimals and $8.512 \cdot 10^{-4}$ for one decimal. According to the asymptotic theory, it should be $8.333 \cdot 10^{-2}$ and $8.333 \cdot 10^{-4}$ respectively. We also estimate time-dependence and cross-dependence of the rounding error. The correlation between successive rounding errors is quite strong for the rounding to zero decimals and for one decimal as well suggesting autocorrelation structure. On the contrary, the correlation between rounding errors and observed prices is quite weak.

Figure 2.3 and Figure 2.4 show distribution of the first and second decimal of the CSCO and XOM stock respectively. The first decimal does not quite resemble the uniform distribution. However, there is no evident systematic behaviour common for all stocks. The second decimal is closer to the uniform distribution except for the 0 and 5 digits which are more frequent. Other stocks also exhibit this behavior. This phenomenon is known as the *price clustering* and is observed across different markets (see e.g. Chung et al., 2004; Ahn et al., 2005; Chung et al., 2005; Ohta, 2006; Aşçıoğlu et al., 2007; Brown and Mitchell, 2008; Ikenberry and Weston, 2008; Davis et al., 2014; Blau and Griffith, 2016; Box and Griffith, 2016; Hu et al., 2017; Mishra and Tripathy, 2018). In NYSE and NASDAQ exchanges, price clustering is caused by a portion of traders buying and selling in multiples of dimes (10 cents – the second digit is 0) or nickels (5 cents – the second digit is 0 or 5).

In general, the results are similar to the empirical study of Holý (2018b), which analyzes foreign exchange pairs with precision up to 5 decimals with conclusion that the distribution of the error caused by rounding to three or more decimals is close enough to the uniform distribution from the variance estimation perspective. Our analysis of stock prices shows that although the distribution of the first and second decimal is not exactly uniform and there is significant autocorrelation structure, it is reasonable to approximate the variance of the rounding error using the uniform distribution.

2.2.5 Interval Model

This section follows Holý and Sokol (2018). Both transaction and quote data can be modeled by the *interval model*. For the transaction data, we observe the price of the transaction X_i. We assume the observed price has precision of d digits due to rounding down. The efficient price P_{T_i} is then assumed to lie in the interval

$$P_{T_i} \in [\underline{P_{T_i}}, \overline{P_{T_i}}], \qquad \underline{P_{T_i}} = X_i, \qquad \overline{P_{T_i}} = X_i + 10^{-d}. \tag{2.31}$$

This model can be straightforwardly modified for rounding up or rounding to the nearest possible increment. Similarly to Section 2.2.4, the only source of uncertainty is the rounding. As this model does not deal with the bid-ask bounce effect, it is more suitable to utilize quote data rather than transaction data in the interval model.

For the quote data, we observe the bid price X_i^B (the price for which the trader can sell the financial asset) and the ask price X_i^A (the price for which the trader can buy the financial asset). We have $X_i^B \leq X_i^A$. We assume both bid and ask prices have precision of d digits due to rounding down. The efficient price P_{T_i} is then assumed to lie in the interval

$$P_{T_i} \in [\underline{P_{T_i}}, \overline{P_{T_i}}], \qquad \underline{P_{T_i}} = X_i^B, \qquad \overline{P_{T_i}} = X_i^A + 10^{-d}. \tag{2.32}$$

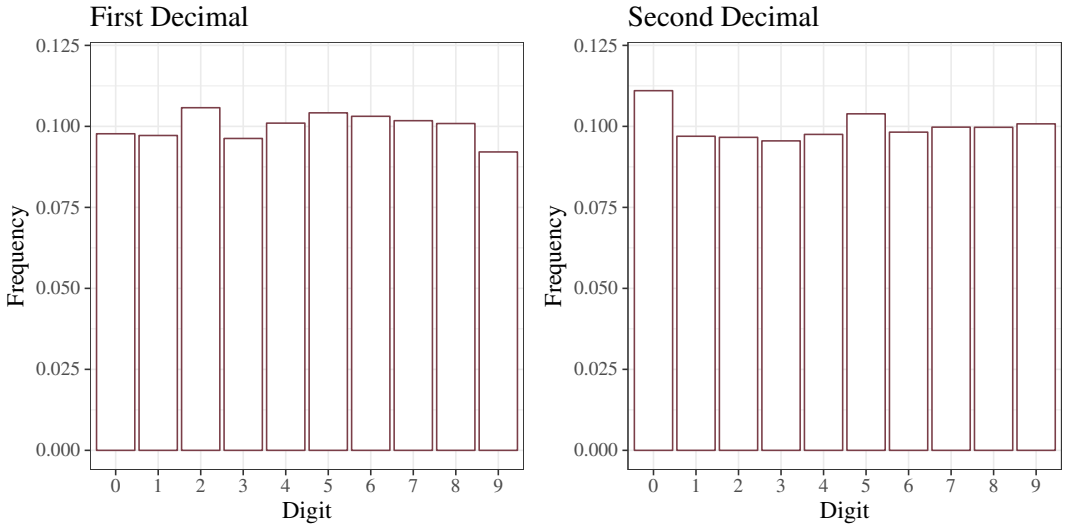

Figure 2.3: Histograms of decimal digits of the CSCO stock.

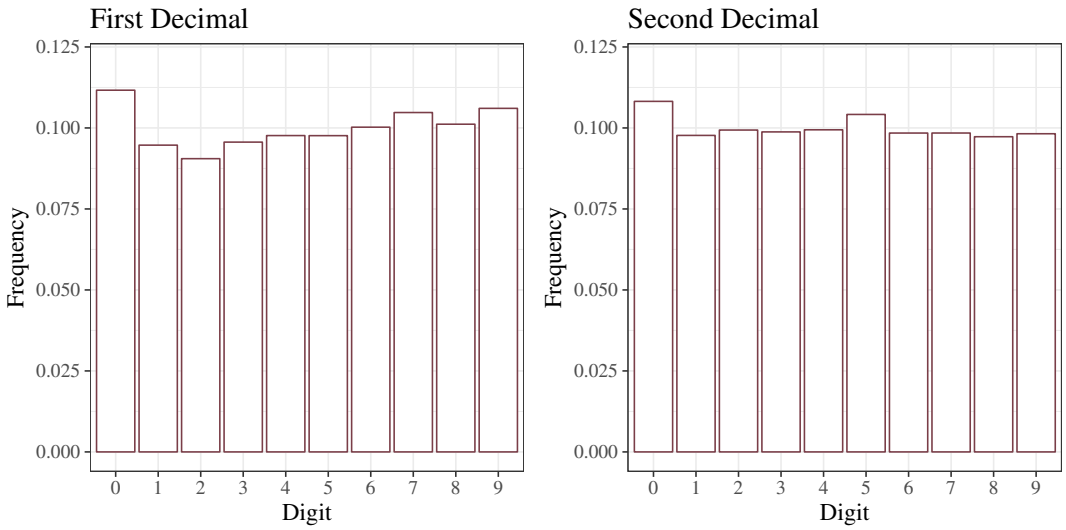

Figure 2.4: Histograms of decimal digits of the XOM stock.

	Zero Decimals			One Decimal		
Stock	Variance	Auto	Cross	Variance	Auto	Cross
AAPL	$9.043 \cdot 10^{-2}$	0.971	-0.013	$8.822 \cdot 10^{-4}$	0.780	0.020
AXP	$8.243 \cdot 10^{-2}$	0.934	0.004	$8.512 \cdot 10^{-4}$	0.561	-0.005
BA	$8.906 \cdot 10^{-2}$	0.816	0.000	$9.079 \cdot 10^{-4}$	0.426	-0.003
CAT	$8.512 \cdot 10^{-2}$	0.895	0.007	$8.615 \cdot 10^{-4}$	0.427	0.001
CSCO	$8.172 \cdot 10^{-2}$	0.991	-0.020	$8.437 \cdot 10^{-4}$	0.922	0.006
CVX	$8.243 \cdot 10^{-2}$	0.936	-0.003	$8.440 \cdot 10^{-4}$	0.564	0.003
DIS	$8.895 \cdot 10^{-2}$	0.953	-0.004	$8.541 \cdot 10^{-4}$	0.652	-0.002
DWDP	$8.581 \cdot 10^{-2}$	0.957	-0.007	$8.518 \cdot 10^{-4}$	0.678	0.005
GE	$8.780 \cdot 10^{-2}$	0.991	-0.038	$8.325 \cdot 10^{-4}$	0.932	0.011
GS	$8.733 \cdot 10^{-2}$	0.857	0.001	$8.701 \cdot 10^{-4}$	0.425	0.008
HD	$8.671 \cdot 10^{-2}$	0.904	-0.004	$8.532 \cdot 10^{-4}$	0.458	0.007
IBM	$8.189 \cdot 10^{-2}$	0.931	-0.037	$8.516 \cdot 10^{-4}$	0.526	0.007
INTC	$8.186 \cdot 10^{-2}$	0.988	-0.005	$8.521 \cdot 10^{-4}$	0.894	0.004
JNJ	$8.730 \cdot 10^{-2}$	0.940	-0.007	$8.458 \cdot 10^{-4}$	0.572	-0.007
JPM	$8.542 \cdot 10^{-2}$	0.959	0.026	$8.449 \cdot 10^{-4}$	0.682	0.005
KO	$8.888 \cdot 10^{-2}$	0.981	0.037	$8.361 \cdot 10^{-4}$	0.844	-0.004
MCD	$8.772 \cdot 10^{-2}$	0.915	0.016	$8.508 \cdot 10^{-4}$	0.469	-0.008
MMM	$8.285 \cdot 10^{-2}$	0.866	-0.005	$8.547 \cdot 10^{-4}$	0.375	0.004
MRK	$8.065 \cdot 10^{-2}$	0.974	0.008	$8.426 \cdot 10^{-4}$	0.786	-0.006
MSFT	$8.856 \cdot 10^{-2}$	0.982	0.035	$8.546 \cdot 10^{-4}$	0.855	0.008
NKE	$8.395 \cdot 10^{-2}$	0.963	-0.004	$8.476 \cdot 10^{-4}$	0.711	0.000
PFE	$8.425 \cdot 10^{-2}$	0.985	0.042	$8.339 \cdot 10^{-4}$	0.865	0.009
PG	$8.615 \cdot 10^{-2}$	0.970	-0.012	$8.405 \cdot 10^{-4}$	0.757	-0.003
TRV	$8.411 \cdot 10^{-2}$	0.895	-0.006	$8.448 \cdot 10^{-4}$	0.435	0.005
UNH	$8.708 \cdot 10^{-2}$	0.868	-0.015	$8.550 \cdot 10^{-4}$	0.403	0.003
UTX	$8.600 \cdot 10^{-2}$	0.919	-0.017	$8.479 \cdot 10^{-4}$	0.517	0.005
V	$8.746 \cdot 10^{-2}$	0.943	0.019	$8.514 \cdot 10^{-4}$	0.604	0.004
VZ	$8.306 \cdot 10^{-2}$	0.976	-0.064	$8.516 \cdot 10^{-4}$	0.805	0.001
WMT	$8.554 \cdot 10^{-2}$	0.960	0.033	$8.468 \cdot 10^{-4}$	0.678	-0.003
XOM	$8.629 \cdot 10^{-2}$	0.972	0.046	$8.322 \cdot 10^{-4}$	0.769	0.016

Table 2.2: Variances of rounding errors, correlations of successive rounding errors (Auto) and correlations of rounding errors with observed prices (Cross) of the 30 DJIA stocks.

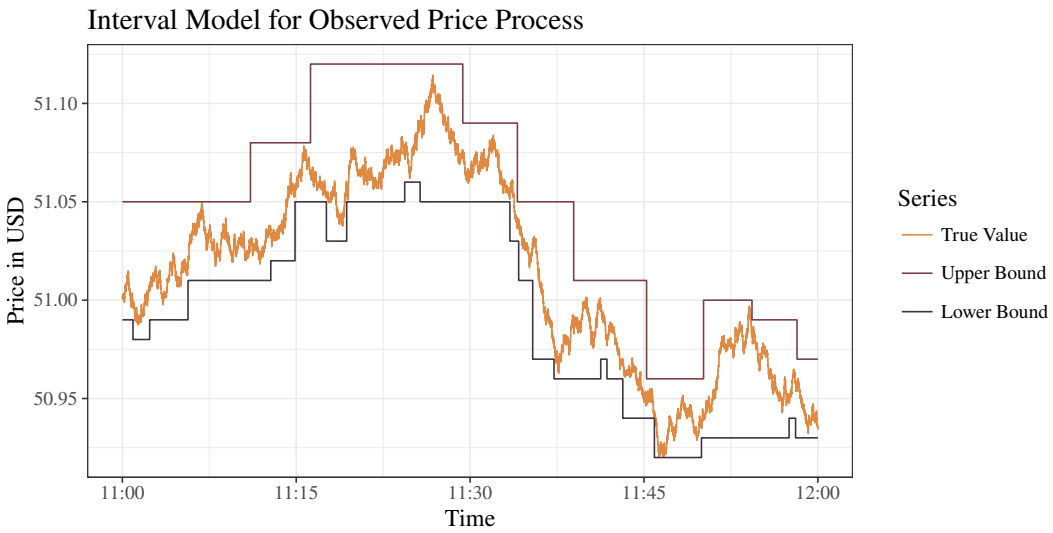

Figure 2.5: Simulated example of the upper and lower bounds in the interval model.

Again, this can be modified for different rounding processes. The interval model for quote data captures both bid-ask spread and discreteness of the prices by interval uncertainty.

Figure 2.5 shows an example of the interval model. Note that we consider the asymptotic case of the infinite number of observations in Figure 2.5. This is a significant difference compared to Figure 2.1 with a finite number of observations for the bid and ask prices.

Literature about Interval Uncertainty

The problem, in which the exact values are not observable but the bounds are, is studied in the theory of partial identification (see e.g. Manski, 2003). As only lower and upper bounds of the price process are available, the goal is to compute lower and upper bounds of certain statistics. However, even some of the basic statistics are not easy to estimate in this setting. The statistics studied in the literature include sample variance (Černý and Sokol, 2015; Ferson et al., 2005; Sokol and Rada, 2016), t-ratio (Černý and Hladík, 2014), entropy (Kreinovich, 1996; Xiang et al., 2007) and higher moments (Kreinovich and Longpre, 2004). An overview of interval uncertainty methods can be found in Ferson et al. (2007) and Nguyen et al. (2012). As we show in the Section 4.1.2 based on Holý and Sokol (2018), the quadratic variation under interval uncertainty is not identifiable and only the information about large jumps in the process can be uncovered.

- Chapter 3 -

Trade Durations

This section follows Blasques, Holý and Tomanová (2018) with some extensions. An important aspect of financial high-frequency data analysis is modeling of durations between events. Among these events belong recording of transactions (denoted as *trade durations*), price changing by a given level (denoted as *price durations*) and volumes reaching a given level (denoted as *volume durations*). We focus on trade durations in this chapter.

Financial durations exhibit strong serial correlation, i.e. long durations are usually followed by long durations and short durations are followed by short durations. To capture this time dependence, Engle and Russell (1998) proposed the *autoregressive conditional duration (ACD) model*. It is analogous to the GARCH volatility model and is similarly popular in the financial durations field. The ACD model has received many extensions over the years. Various continuous distributions with non-negative support and various dynamics of time-varying mean were proposed in the duration literature. Notably, the logarithmic transform for the ACD was utilized by Bauwens and Giot (2000) and Lunde (1999). It allows to omit non-negativity constrains and include exogenous variables. For the survey of duration analysis, see Pacurar (2008), Bauwens and Hautsch (2009), Hautsch (2011) and Saranjeet and Ramanathan (2019).

Financial durations can also be modeled by the class of *generalized autoregressive score (GAS) models* (Creal et al., 2008, 2013), also known as *dynamic conditional score* models (Harvey, 2013). This general class includes many widely used models such as the GARCH model based on the normal distribution of Bollerslev (1986) and ACD model based on the exponential distribution of Engle and Russell (1998). The GAS models capture dynamics of time-varying parameters by the autoregressive term and the term based on the score of the conditional probability density function (or the conditional probability mass function for the case of discrete models). The GAS specification therefore utilizes the entire shape of the underlying distribution. This allows us to formulate novel duration models based on various distributions within a single framework.

We propose a new model for trade durations named the *zero-inflated autoregressive conditional duration (ZIACD) model*. This model is based on the zero-inflated negative binomial distribution and the GAS specification for the time varying parameter. We have two motivations for the proposed model. First, it is suitable for discrete durations as it utilizes a discrete distribution. Second, it allows to capture excessive zero durations caused by split transactions as it utilizes a zero-inflated distribution. In the theoretical part of this chapter, we discuss the consistency and asymptotic normality of the maximum likelihood estimator for the proposed ZIACD model. In the empirical part of this chapter, we show that the proposed discrete ZIACD model is a good fit when duration data is discrete and outperforms traditional continuous models even when duration data is virtually continuous due to its correct treatment of zero values.

Article	Distribution	Parameters	
		Time-Varying	Constant
Engle and Russell (1998)	Exponential	Mean	0
Engle and Russell (1998)	Weibull	Mean	1
Lunde (1999)	Generalized Gamma	Mean	2
Grammig and Maurer (2000)	Burr	Mean	2
Hautsch (2001)	Generalized F	Mean	3
Bhatti (2010)	Birnbaum–Saunders	Median	1
Xu (2013)	Log-Normal	Mean	1
Leiva et al. (2014)	Power-Exponential Birnbaum–Saunders	Median	2
Leiva et al. (2014)	Student's t Birnbaum–Saunders	Median	2
Zheng et al. (2016)	Fréchet	Mean	1

Table 3.1: The use of continuous distributions in ACD models.

3.1 Distributions of Durations

Let $T_0 \leq T_1 \leq \cdots \leq T_n$ be random variables denoting times of transactions. Trade durations are then defined as $X_i = T_i - T_{i-1}$, $i = 1, \ldots, n$. Note that we use slightly different notation than in Section 2 as we denote durations as X_i instead of D_i.

In this section, we focus on the distribution of X_i. We consider both continuous and discrete distributions. In any case, we assume that the distribution of X_i, $i = 1, \ldots, n$ has non-negative support and is dependent on some time-varying parameters $f_i = (f_{i,1} \ldots, f_{i,k})'$, $i = 1, \ldots, n$ and some static parameters $g = (g_1, \ldots, g_l)'$. In the following text, we utilize the *conditional probability density* function $p(x_i|f_i, g)$ for continuous random variable X_i and the *conditional probability mass* function $P[X_i = x_i|f_i, g]$ for discrete random variable X_i. The observed durations are denoted as x_i. We also utilize the score and the Fisher information for time-varying parameters. In the following text, let $p(x_i|f_i, g)$ also denote the conditional probability mass function. The *score* for time-varying parameters f_i is then defined as

$$\nabla(x_i, f_i, g) = \frac{\partial \log p(x_i|f_i, g)}{\partial f_i}. \tag{3.1}$$

The *Fisher information* for time-varying parameters f_i is then defined as

$$\mathcal{I}(f_i, g) = \mathrm{E}\Big[\nabla(x_i, f_i, g)\nabla(x_i, f_i, g)'\Big|f_i, g\Big] = -\mathrm{E}\left[\frac{\partial^2 \log p(x_i|f_i, g)}{\partial f_i \partial f_i'}\middle| f_i, g\right]. \tag{3.2}$$

Note, that the latter equality requires some regularity conditions (see Lehmann and Casella, 1998).

Next, we present various continuous and discrete distributions. We also suggest which parameters should be time-varying and which static with regard to the duration models.

3.1.1 Continuous Distributions

Traditionally, duration models are based on continuous distributions. Table 3.1 reviews continuous distributions used in the autoregressive conditional duration literature. The ACD specification is usually based on the time-varying mean with some additional constant shape parameters. As an illustration, we focus on the generalized gamma distribution and its special cases – the exponential, Weibull and gamma distributions.

Exponential Distribution

The *exponential distribution* is a continuous distribution with non-negative support and one parameter. We consider the scale parameter $\beta_i > 0$ to be time-varying, i.e. $f_i = \beta_i$. The probability density function is

$$p(x_i|\beta_i) = \frac{1}{\beta_i} e^{-\frac{x_i}{\beta_i}} \quad \text{for } x_i \in [0, \infty). \tag{3.3}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \beta_i, \\ \mathrm{var}[X_i] &= \beta_i^2. \end{aligned} \tag{3.4}$$

The score for the parameter β_i is

$$\nabla(x_i, \beta_i) = \beta_i^{-1} \left(x_i \beta_i^{-1} - 1 \right) \quad \text{for } x_i \in [0, \infty). \tag{3.5}$$

The Fisher information for the parameter β_i is

$$\mathcal{I}(\beta_i) = \beta_i^{-2}. \tag{3.6}$$

Weibull Distribution

The *Weibull distribution* is a continuous distribution with strictly positive support and two parameters. We consider the scale parameter $\beta_i > 0$ to be time-varying, while the shape parameter $\varphi > 0$ is static, i.e. $f_i = \beta_i$ and $g = \varphi$. The probability density function is

$$p(x_i|\beta_i, \varphi) = \frac{\varphi}{\beta_i} \left(\frac{x_i}{\beta_i} \right)^{\varphi - 1} e^{-\left(\frac{x_i}{\beta_i} \right)^{\varphi}} \quad \text{for } x_i \in (0, \infty). \tag{3.7}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \beta_i \Gamma \left(1 + \varphi^{-1} \right), \\ \mathrm{var}[X_i] &= \beta_i^2 \Gamma \left(1 + 2\varphi^{-1} \right) - \left(\beta_i \Gamma \left(1 + \varphi^{-1} \right) \right)^2. \end{aligned} \tag{3.8}$$

The score for the parameter β_i is

$$\nabla(x_i, \beta_i, \varphi) = \varphi \beta_i^{-1} \left(x_i^{\varphi} \beta_i^{-\varphi} - 1 \right) \quad \text{for } x_i \in (0, \infty). \tag{3.9}$$

The Fisher information for the parameter β_i is

$$\mathcal{I}(\beta_i, \varphi) = \beta_i^{-2} \varphi^2. \tag{3.10}$$

A special case of the Weibull distribution is the exponential distribution for $\varphi = 1$.

Gamma Distribution

The *gamma distribution* is a continuous distribution with strictly positive support and two parameters. We consider the scale parameter $\beta_i > 0$ to be time-varying, while the shape parameter $\psi > 0$ is static, i.e. $f_i = \beta_i$ and $g = \psi$. The probability density function is

$$p(x_i|\beta_i, \psi) = \frac{1}{\Gamma(\psi)} \frac{1}{\beta_i} \left(\frac{x_i}{\beta_i} \right)^{\psi - 1} e^{-\frac{x_i}{\beta_i}} \quad \text{for } x_i \in (0, \infty). \tag{3.11}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \beta_i \psi, \\ \mathrm{var}[X_i] &= \beta_i^2 \psi. \end{aligned} \tag{3.12}$$

The score for the parameter β_i is

$$\nabla(x_i, \beta_i, \psi) = \beta_i^{-1}\left(x_i\beta_i^{-1} - \psi\right) \quad \text{for } x_i \in (0, \infty). \tag{3.13}$$

The Fisher information for the parameter β_i is

$$\mathcal{I}(\beta_i, \psi) = \beta_i^{-2}\psi. \tag{3.14}$$

A special case of the gamma distribution is the exponential distribution for $\psi = 1$.

Generalized Gamma Distribution

The *generalized gamma distribution* is a continuous probability distribution with strictly positive support and a three-parameter generalization of the two-parameter gamma distribution (Stacy, 1962). It also contains the exponential distribution and the Weibull distribution as special cases. We consider the scale parameter $\beta_i > 0$ to be time-varying, while the shape parameters $\psi > 0$ and $\varphi > 0$ are static, i.e. $f_i = \beta_i$ and $g = (\psi, \varphi)'$. The probability density function is

$$p(x_i|\beta_i, \psi, \varphi) = \frac{1}{\Gamma(\psi)}\frac{\varphi}{\beta_i}\left(\frac{x_i}{\beta_i}\right)^{\psi\varphi-1} e^{-\left(\frac{x_i}{\beta_i}\right)^{\varphi}} \quad \text{for } x_i \in (0, \infty). \tag{3.15}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \beta_i\frac{\Gamma\left(\psi + \varphi^{-1}\right)}{\Gamma(\psi)}, \\ \mathrm{var}[X_i] &= \beta_i^2\frac{\Gamma\left(\psi + 2\varphi^{-1}\right)}{\Gamma(\psi)} - \left(\beta_i\frac{\Gamma\left(\psi + \varphi^{-1}\right)}{\Gamma(\psi)}\right)^2. \end{aligned} \tag{3.16}$$

The score for the parameter β_i is

$$\nabla(x_i, \beta_i, \psi, \varphi) = \varphi\beta_i^{-1}\left(x_i^{\varphi}\beta_i^{-\varphi} - \psi\right) \quad \text{for } x_i \in (0, \infty). \tag{3.17}$$

The Fisher information for the parameter β_i is

$$\mathcal{I}(\beta_i, \psi, \varphi) = \beta_i^{-2}\psi\varphi^2. \tag{3.18}$$

Special cases of the generalized gamma distribution include the gamma distribution for $\varphi = 1$, the Weibull distribution for $\psi = 1$ and the exponential distribution for $\psi = 1$ and $\varphi = 1$.

3.1.2 Discrete Distributions

Traditionally, durations are not considered to be discrete variables. However, as we discuss in Section 3.2.3, discrete distributions for durations are suitable in many cases. We present the Poisson, geometric and negative binomial distributions alongside their zero-inflated modifications.

Non-negative integer variables are commonly analyzed using count data models based on specific underlying distribution, most notably the Poisson distribution and the negative binomial distribution (see Cameron and Trivedi, 2013). A distinctive feature of the Poisson distribution is that its expected value is equal to its variance. This characteristic is too strict in many applications as count data often exhibit overdispersion, a higher variance than the expected value. A generalization of the Poisson distribution overcoming this limitation is the negative binomial distribution with one parameter determining its expected value and another parameter determining its excess dispersion.

The zero-inflated distribution is an extension of a discrete distribution allowing the probability of zero values to be higher than the probability given by the original distribution. In the zero-inflated distribution, values are generated by two components – one component generates only zero values while the other component generates integer values (including zero values) according to the original distribution. Lambert (1992) proposed the zero-inflated Poisson model and Greene (1994) used zero-inflated model for the negative binomial distribution.

Poisson Distribution

The *Poisson distribution* is a discrete distribution with one parameter. We consider the location parameter $\mu_i > 0$ to be time-varying, i.e. $f_i = \mu_i$. The probability mass function is

$$\mathrm{P}[X_i = x_i|\mu_i] = \frac{1}{\Gamma(x_i+1)}\mu_i^{x_i}e^{-\mu_i} \quad \text{for } x_i = 0, 1, 2, \dots . \tag{3.19}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i, \\ \mathrm{var}[X_i] &= \mu_i. \end{aligned} \tag{3.20}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i) = \mu_i^{-1}x_i - 1 \quad \text{for } x_i = 0, 1, 2, \dots . \tag{3.21}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i) = \mu_i^{-1}. \tag{3.22}$$

Zero-Inflated Poisson Distribution

Lambert (1992) proposed the zero-inflated modification of the Poisson distribution. The *zero-inflated Poisson distribution* is a discrete distribution with two parameters. We consider the location parameter $\mu_i > 0$ to be time-varying, while the probability of excessive zero values $\pi \in [0, 1)$ is static, i.e. $f_i = \mu_i$ and $g = \pi$. The variable X_i follows the zero-inflated Poisson distribution if

$$\begin{aligned} X_i &\sim 0 && \text{with probability } \pi, \\ X_i &\sim \mathrm{Poiss}(\mu_i) && \text{with probability } 1-\pi. \end{aligned} \tag{3.23}$$

The first process generates only zeros, while the second process generates values from the Poisson distribution. The probability mass function is

$$\mathrm{P}[X_i = x_i|\mu_i, \pi] = \begin{cases} \pi + (1-\pi)e^{-\mu_i} & \text{for } x_i = 0, \\ (1-\pi)\frac{1}{\Gamma(x_i+1)}\mu_i^{x_i}e^{-\mu_i} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.24}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i(1-\pi), \\ \mathrm{var}[X_i] &= \mu_i(1-\pi)(1+\pi\mu_i). \end{aligned} \tag{3.25}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i, \pi) = \begin{cases} (\pi-1)(1+\pi e^{-\mu_i} - \pi)^{-1} & \text{for } x_i = 0, \\ \mu_i^{-1}x_i - 1 & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.26}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i, \pi) = \pi(\pi-1)(\pi e^{\mu_i} - \pi + 1)^{-1} - (\pi-1)\mu_i^{-1}. \tag{3.27}$$

Geometric Distribution

The *geometric distribution* is a discrete distribution with one parameter. We consider the location parameter $\mu_i > 0$ to be time-varying, i.e. $f_i = \mu_i$. The probability mass function is

$$\mathrm{P}[X_i = x_i|\mu_i] = \frac{1}{1+\mu_i}\left(\frac{\mu_i}{1+\mu_i}\right)^{x_i} \quad \text{for } x_i = 0, 1, 2, \dots . \tag{3.28}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i, \\ \mathrm{var}[X_i] &= \mu_i(1+\mu_i). \end{aligned} \tag{3.29}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i) = \mu_i^{-1}(x_i - \mu_i)(\mu_i + 1)^{-1} \quad \text{for } x_i = 0, 1, 2, \dots . \tag{3.30}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i) = \mu_i^{-1}(\mu_i + 1)^{-1}. \tag{3.31}$$

Zero-Inflated Geometric Distribution

Similarly to the Poisson distribution, the geometric distribution can also be modified to capture excessive zeros. The *zero-inflated geometric distribution* is a discrete distribution with two parameters. We consider the location parameter $\mu_i > 0$ to be time-varying, while the probability of excessive zero values $\pi \in [0, 1)$ is static, i.e. $f_i = \mu_i$ and $g = \pi$. The variable X_i follows zero-inflated negative binomial distribution if

$$\begin{aligned} X_i &\sim 0 && \text{with probability } \pi, \\ X_i &\sim \mathrm{Geom}(\mu_i) && \text{with probability } 1 - \pi. \end{aligned} \tag{3.32}$$

The first process generates only zeros, while the second process generates values from the geometric distribution. The probability mass function is

$$\mathrm{P}[X_i = x_i | \mu_i, \pi] = \begin{cases} \pi + (1-\pi)\frac{1}{1+\mu_i} & \text{for } x_i = 0, \\ (1-\pi)\frac{1}{1+\mu_i}\left(\frac{\mu_i}{1+\mu_i}\right)^{x_i} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.33}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i(1-\pi), \\ \mathrm{var}[X_i] &= \mu_i(1-\pi)(1+\pi\mu_i+\mu_i). \end{aligned} \tag{3.34}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i, \pi) = \begin{cases} (\pi - 1)(\mu_i + 1)^{-1}(\pi\mu_i + 1)^{-1} & \text{for } x_i = 0, \\ \mu_i^{-1}(x_i - \mu_i)(\mu_i + 1)^{-1} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.35}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i, \pi) = \frac{\pi(\pi - 1)}{(\mu_i + 1)^2(\pi\mu_i + 1)} + \frac{1 - \pi}{\mu_i(\mu_i + 1)}. \tag{3.36}$$

Negative Binomial Distribution

The *negative binomial distribution* is a generalization of the Possion distribution and the geometric distribution. It has the location parameter and the dispersion parameter allowing for the variance to be greater than the mean. The negative binomial distribution can be derived in many ways (see Boswell and Patil, 1970). We use the NB2 parameterization of Cameron and Trivedi (1986) derived from the Poisson-gamma mixture distribution. It is the most common parametrization used in the negative binomial regression according to Hilbe (2011) and Cameron and Trivedi (2013). We consider the location parameter $\mu_i > 0$ to be time-varying, while the dispersion parameter $\alpha \geq 0$ is static, i.e. $f_i = \mu_i$ and $g = \alpha$. The probability mass function is

$$\mathrm{P}[X_i = x_i | \mu_i, \alpha] = \frac{\Gamma(x_i + \alpha^{-1})}{\Gamma(x_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{x_i} \quad \text{for } x_i = 0, 1, 2, \dots . \tag{3.37}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i, \\ \mathrm{var}[X_i] &= \mu_i(1 + \alpha\mu_i). \end{aligned} \tag{3.38}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i, \alpha) = \mu_i^{-1}(x_i - \mu_i)(\alpha\mu_i + 1)^{-1} \quad \text{for } x_i = 0, 1, 2, \ldots . \tag{3.39}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i, \alpha) = \mu_i^{-1}(\alpha\mu_i + 1)^{-1}. \tag{3.40}$$

Special cases of the negative binomial distribution include the Poisson distribution for $\alpha = 0$ and the geometric distribution for $\alpha = 1$.

Zero-Inflated Negative Binomial Distribution

Greene (1994) used the zero-inflated model for the negative binomial distribution. The *zero-inflated negative binomial distribution* is a discrete distribution with three parameters. We consider the location parameter $\mu_i > 0$ to be time-varying, while the dispersion parameter $\alpha \geq 0$ and the probability of excessive zero values $\pi \in [0, 1)$ are static, i.e. $f_i = \mu_i$ and $g = (\alpha, \pi)'$. The variable X_i follows the zero-inflated negative binomial distribution if

$$\begin{aligned} X_i &\sim 0 && \text{with probability } \pi, \\ X_i &\sim \mathrm{NB}(\mu_i, \alpha) && \text{with probability } 1 - \pi. \end{aligned} \tag{3.41}$$

The first process generates only zeros, while the second process generates values from the negative binomial distribution. The probability mass function is

$$\mathrm{P}[X_i = x_i | \mu_i, \alpha, \pi] = \begin{cases} \pi + (1 - \pi)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} & \text{for } x_i = 0, \\ (1 - \pi)\frac{\Gamma(x_i + \alpha^{-1})}{\Gamma(x_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{x_i} & \text{for } x_i = 1, 2, \ldots . \end{cases} \tag{3.42}$$

The expected value and variance is

$$\begin{aligned} \mathrm{E}[X_i] &= \mu_i(1 - \pi), \\ \mathrm{var}[X_i] &= \mu_i(1 - \pi)(1 + \pi\mu_i + \alpha\mu_i). \end{aligned} \tag{3.43}$$

The score for the parameter μ_i is

$$\nabla(x_i, \mu_i, \alpha, \pi) = \begin{cases} (\pi - 1)(\alpha\mu_i + 1)^{-1}\Big(1 + \pi(\alpha\mu_i + 1)^{\alpha^{-1}} - \pi\Big)^{-1} & \text{for } x_i = 0, \\ \mu_i^{-1}(x_i - \mu_i)(\alpha\mu_i + 1)^{-1} & \text{for } x_i = 1, 2, \ldots . \end{cases} \tag{3.44}$$

The Fisher information for the parameter μ_i is

$$\mathcal{I}(\mu_i, \alpha, \pi) = \frac{\pi(\pi - 1)}{(\alpha\mu_i + 1)^2\Big(\pi(\alpha\mu_i + 1)^{\alpha^{-1}} - \pi + 1\Big)} + \frac{1 - \pi}{\mu_i(\alpha\mu_i + 1)}. \tag{3.45}$$

Special cases of the zero-inflated negative binomial distribution include the negative binomial distribution for $\pi = 0$, the zero-inflated Poisson distribution for $\alpha = 0$ and the zero-inflated geometric distribution for $\alpha = 1$.

3.2 Models of Durations

In this section, we focus on dynamics of time-varying parameter f_i, $i = 1, \ldots_n$ of the underlying distributions. We present the original autoregressive conditional duration (ACD) model with some of its extensions and modifications. We also present the generalized autoregressive score (GAS) model, which can be utilized for duration modeling as well. The highlight of this section is the proposed zero-inflated autoregressive conditional duration (ZIACD) model based on the zero-inflated negative binomial distribution with the time-varying parameter following the GAS recursion. We derive the model specification and formulate asymptotic properties of the maximum likelihood estimator.

3.2.1 ACD Model and Its Extensions

Engle and Russell (1998) proposed to model durations X_i as

$$X_i = \beta_i E_i, \qquad \mathrm{E}[E_i] = 1, \qquad i = 1, \ldots, n, \tag{3.46}$$

where $\beta_i \geq 0$ is the time-varying mean and E_i are independent and identically distributed random variables with non-negative support and unit mean. This is in line with our (slightly different) framework in which we directly model distribution of X_i with both time-varying and static parameters instead of E_i with only static parameters. Originally, Engle and Russell (1998) considered the exponential and Weibull distributions with unit mean for E_i. In their ACD(p, q) model, the time-varying mean follows recursion

$$\beta_{i+1} = c + \sum_{j=1}^{q} b_j \beta_{i-j+1} + \sum_{j=1}^{p} a_j x_{i-j+1}, \tag{3.47}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i. This specification, however, has some drawbacks. Dufour and Engle (2000) point out that this recursion requires constraints on the parameters to ensure non-negativity of durations. Dufour and Engle (2000) and Fernandes and Grammig (2006) also show that non-linear functional forms of the time-varying mean β_i are more appropriate.

Model (3.46) is utilized in many studies following Engle and Russell (1998) with various underlying distributions for E_i. The recursion for the time-varying mean β_i (3.47) is analogous to the popular GARCH model and similarly to the GARCH model, it has received many extensions and modifications. Next, we list some of the specifications for the mean dynamics proposed in the duration literature. For a more comprehensive overview, see Hautsch (2011).

Specification of Logarithmic Duration Models

The *logarithmic autoregressive conditional duration model* was proposed by Bauwens and Giot (2000). The LACD1(p, q) model is based on recursion

$$\log \beta_{i+1} = c + \sum_{j=1}^{q} b_j \log \beta_{i-j+1} + \sum_{j=1}^{p} a_j \log \left(x_{i-j+1} \beta_{i-j+1}^{-1} \right), \tag{3.48}$$

where c, b_i and a_i are the parameters and x_i are the observed values of X_i. Terms $x_{i-j+1}\beta_{i-j+1}^{-1}$ are residuals, i.e. the observed values of E_i. The logarithmic model allows to include additional variables to the model without the need of sign restrictions on their coefficients. This is the main motivation behind this model.

Another logarithmic autoregressive conditional duration model was proposed by Lunde (1999). The LACD2(p, q) model is based on recursion

$$\log \beta_{i+1} = c + \sum_{j=1}^{q} b_j \log \beta_{i-j+1} + \sum_{j=1}^{p} a_j x_{i-j+1} \beta_{i-j+1}^{-1}, \tag{3.49}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i.

Specification of Additive and Multiplicative Duration Model

The *additive and multiplicative autoregressive conditional duration model* was described by Hautsch (2011). The AMACD(p, r, q) model is based on recursion

$$\beta_{i+1} = c + \sum_{j=1}^{q} b_j \beta_{i-j+1} + \sum_{j=1}^{p} a_j x_{i-j+1} + \sum_{j=1}^{r} d_j x_{i-j+1} \beta_{i-j+1}^{-1}, \tag{3.50}$$

where c, b_j, a_j and d_j are the parameters and x_i are the observed values of X_i. The ACD(p, q) model is a special case of the AMACD(p, r, q) model for $r = 0$.

Specification of Box-Cox Duration Models

Engle (2000) proposed a duration model using the *Box-Cox transformation* of the past innovations. The BCACD1(p, q) model is based on recursion

$$\log \beta_{i+1} = c + \sum_{j=1}^{q} b_j \log \beta_{i-j+1} + \sum_{j=1}^{p} a_j \frac{x_{i-j+1}^{\delta} \beta_{i-j+1}^{-\delta} - 1}{\delta}, \tag{3.51}$$

where c, b_j, a_j and δ are the parameters and x_i are the observed values of X_i. Special cases of the BCACD1(p, q) model include the LACD1(p, q) model for $\delta \to 0$ and LACD2(p, q) model for $\delta = 1$.

Hautsch (2001) and Hautsch (2003) further generalize this model using the Box-Cox transformation for both past innovations and the time-varying mean. The BCACD2(p, q) model is based on recursion

$$\frac{\beta_{i+1}^{\gamma} - 1}{\gamma} = c + \sum_{j=1}^{q} b_j \frac{\beta_{i-j+1}^{\gamma} - 1}{\gamma} + \sum_{j=1}^{p} a_j \frac{x_{i-j+1}^{\delta} \beta_{i-j+1}^{-\delta} - 1}{\delta}, \tag{3.52}$$

where c, b_j, a_j, δ and γ are the parameters and x_i are the observed values of X_i. Special cases of the BCACD2(p, q) model include the ACD(p, q) model for $\delta = 1$ and $\gamma = 1$, LACD1(p, q) model for $\delta \to 0$ and $\gamma \to 0$, LACD2(p, q) model for $\delta = 1$ and $\gamma \to 0$ and BCACD1(p, q) model for $\gamma \to 0$.

3.2.2 GAS Model

Generalized autoregressive score (GAS) models (Creal et al., 2008, 2013), also known as *dynamic conditional score models* (Harvey, 2013), provide a general framework for modeling of time-varying parameters. They capture dynamics of time-varying parameters $f_i = (f_{i,1}, \dots, f_{i,k})'$ by the autoregressive term and the scaled score of the conditional observation density (or the conditional observation probability mass function in the case of discrete distribution). The time-varying parameters f_i in the GAS$(1, 1)$ model follow the recursion

$$f_{i+1} = C + B f_i + A S(f_i, g) \nabla(x_i, f_i, g), \tag{3.53}$$

where $C = (c_1, \dots, c_k)'$ are the constant parameters, $B = \mathrm{diag}(b_1, \dots, b_k)$ are the autoregressive parameters, $A = \mathrm{diag}(a_1, \dots, a_k)$ are the score parameters, $S(f_i, g)$ is a scaling function for the score and $\nabla(x_i, f_i, g)$ is the score defined in (3.1). The score for the time-varying vector f_i is the gradient of the log-likelihood with respect to f_i. It indicates how sensitive the log-likelihood is to parameter f_i. In model (3.53), the score drives the time variation in the parameter f_i and links the shape of the density function (or the probability mass function) directly to the dynamics of f_i. As the scaling function, we consider

- the *unit scaling* $S(f_i, g) = I$,
- the *square root of the inverse of the Fisher information scaling* $S(f_i, g) = \mathcal{I}(f_i, g)^{-\frac{1}{2}}$,

- the *inverse of the Fisher information scaling* $S(f_i, g) = \mathcal{I}(f_i, g)^{-1}$.

Note that each scaling function results in a different GAS model. The long-term mean and unconditional value of the time-varying parameters is $\bar{f} = (I - B)^{-1}C$.

In general, Cox (1981) classifies time series to observation-driven models and parameter-driven models. The GAS models belong to the class of observation-driven models. Koopman et al. (2016) find that observation-driven models based on the score perform comparably to parameter-driven models in terms of predictive accuracy. Observation-driven models (including the GAS model) can be estimated in a straightforward manner by the maximum likelihood method.

Reparametrization

The parameters f_i in (3.53) are assumed to be unbounded. However, some distributions require bounded parameters (e.g. variance greater than zero). The standard solution in the GAS framework is to use an unbounded parametrization $\tilde{f}_i = H(f_i)$, which follows the GAS recursion instead of the original parametrization f_i, i.e.

$$\tilde{f}_{i+1} = \tilde{C} + \tilde{B}\tilde{f}_i + \tilde{A}\tilde{S}(\tilde{f}_i, g)\tilde{\nabla}(x_i, \tilde{f}_i, g), \tag{3.54}$$

where $\tilde{C} = (\tilde{c}_1, \ldots, \tilde{c}_k)'$ are the constant parameters, $\tilde{B} = \mathrm{diag}(\tilde{b}_1, \ldots, \tilde{b}_k)$ are the autoregressive parameters, $\tilde{A} = \mathrm{diag}(\tilde{a}_1, \ldots, \tilde{a}_k)$ are the score parameters, $\tilde{S}(\tilde{f}_i, g)$ is the reparametrized scaling function for the score and $\tilde{\nabla}(x_i, \tilde{f}_i, g)$ is the reparametrized score. The reparametrized score equals to

$$\tilde{\nabla}(x_i, \tilde{f}_i, g) = \dot{H}^{-1}(f_i)\nabla(x_i, f_i, g), \tag{3.55}$$

while the Fisher information of the reparametrized model equals to

$$\tilde{\mathcal{I}}(\tilde{f}_i, g) = \dot{H}'^{-1}(f_i)\mathcal{I}(f_i, g)\dot{H}^{-1}(f_i), \tag{3.56}$$

where $\dot{H}(f_i) = \partial H(f_i)/\partial f_i'$ is the Jacobian matrix of $H(f_i)$.

Higher-Order Generalization

Naturally, the GAS$(1, 1)$ model can be extend to the GAS(p, q) model. The time-varying parameters f_i in the GAS(p, q) model follow the recursion

$$f_{i+1} = C + \sum_{j=1}^{q} B_j f_{i-j+1} + \sum_{j=1}^{p} A_j S(f_{i-j+1}, g)\nabla(x_{i-j+1}, f_{i-j+1}, g), \tag{3.57}$$

where $C = (c_1, \ldots, c_k)'$ are the constant parameters, $B_j = \mathrm{diag}(b_{j,1}, \ldots, b_{j,k})$ are the autoregressive parameters, $A_j = \mathrm{diag}(a_{j,1}, \ldots, a_{j,k})$ are the score parameters, $S(f_i, g)$ is the scaling function for the score and $\nabla(x_i, f_i, g)$ is the score.

Specification of GAS Models Based on the Exponential Distribution

We can obtain various duration models using the GAS specification with various parametrizations and scaling functions. We formulate the models for the case of the exponential distribution (3.3) with the score (3.5). First, by considering the regular parametrization and the unit scaling, we obtain the model

$$\beta_{i+1} = c + \sum_{j=1}^{q} b_j \beta_{i-j+1} + \sum_{j=1}^{p} a_j \left(x_{i-j+1}\beta_{i-j+1}^{-2} - \beta_{i-j+1}^{-1} \right), \tag{3.58}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i.

Second, by considering the regular parametrization and the square root of the inverse of the Fisher information scaling, we obtain the model

$$\beta_{i+1} = c + \sum_{j=1}^{q} b_j \beta_{i-j+1} + \sum_{j=1}^{p} a_j \left(x_{i-j+1} \beta_{i-j+1}^{-1} - 1 \right), \tag{3.59}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i. This specification is equivalent to the AMACD$(p, q, 0)$ model.

Third, by considering the regular parametrization and the inverse of the Fisher information scaling, we obtain the model

$$\beta_{i+1} = c + \sum_{j=1}^{q} b_j \beta_{i-j+1} + \sum_{j=1}^{p} a_j \left(x_{i-j+1} - \beta_{i-j+1} \right), \tag{3.60}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i. This specification is equivalent to the ACD(p, q) model.

Fourth, by considering the logarithmic parametrization and the unit scaling, we obtain the model

$$\log \beta_{i+1} = c + \sum_{j=1}^{q} b_j \log \beta_{i-j+1} + \sum_{j=1}^{p} a_j \left(x_{i-j+1} \beta_{i-j+1}^{-1} - 1 \right), \tag{3.61}$$

where c, b_j and a_j are the parameters and x_i are the observed values of X_i. This specification is equivalent to the LACD2(p, q) model. Note that the square root of the inverse of the Fisher information scaling and the inverse of the Fisher information scaling results in exactly the same model as the Fisher information is equal to 1 for this parametrization.

We can see that within the GAS framework, we can formulate some traditional durations models and construct some brand new models as well. For other distributions with more parameters, the score and the Fisher information are more complex and the expressions for time-varying mean tend to differ from the traditional specifications.

Other Notable GAS Models

The GAS specification includes many commonly used econometric models. Most notably, the GAS model with the normal distribution, the inverse of the Fisher information scaling and time-varying variance results in the GARCH model of Bollerslev (1986). Other continuous models which can be formulated within the GAS framework include the autoregressive conditional intensity model of Russell (1999), dynamic conditional correlation model of Engle (2002), time-varying quantile model of Engle and Manganelli (2004) and dynamic copula model of Patton (2006).

The GAS framework can be utilized for discrete models as well. Koopman et al. (2015) and Koopman et al. (2018) used discrete copulas based on the Skellam distribution for high-frequency stock price changes. Koopman and Lit (2017) used the bivariate Poisson distribution for a number of goals in football matches and the Skellam distribution for a score difference. In a similar fashion, Pikhart and Holý (2018) used the logistic regression for e-sport matches. Gorgi (2018) used the Poisson distribution as well as the negative binomial distribution for offensive conduct reports. Blazsek and Escribano (2016) used the Poisson count panel model for the number of successful patent applications. The Poisson count model of Davis et al. (2003) can be formulated within the GAS framework as well.

For other continuous and discrete GAS models, see the list of papers on the GAS model website of Lucas (2019). A total of 169 published articles, articles in press, working papers and doctoral theses related to the GAS model were listed as of December 31, 2018.

3.2.3 ZIACD Model

Traditional duration models are based on continuous distributions as discussed in Section 3.1.1. All data are, however, inherently discrete. This is also the case of financial durations, whether they are recorded with a second or millisecond precision. Discretness of real data is the first motivation of our approach. Generally, there are three ways to deal with discrete values of observed variables.

- The first approach is to consider random variables to follow a continuous distribution and simply ignore discretness of data. This is a valid and often the best solution when data are recorded with a high precision (e.g. durations with millisecond precision). However, if the precision is lower (e.g. durations with second precision), a bias in estimators increases and the significance level in hypothesis tests is altered (see Schneeweiss et al., 2010). Tricker (1984) and Taraldsen (2011) explore the effects of rounding on the exponential distribution while Tricker (1992) deals with the gamma distribution. In autoregressive processes, the rounding errors can further accumulate making continuous models unreliable (see Zhang et al., 2010 and Li and Bai, 2011).

- The second approach is to consider random variables to follow a continuous distribution and take into account partial identification and interval uncertainty of the observations caused by rounding or grouping (see e.g. Manski, 2003). In financial volatility analysis, discrete values of prices are often (among other effects) captured by the market microstructure noise (see e.g. Hansen and Lunde, 2006). To our knowledge, Grimshaw et al. (2005) is the only paper addressing the issue of rounding in financial durations analysis. They found that ignoring the discretness of data leads to a distortion of time-dependence tests in financial durations.

- The third approach is to consider random variables to follow a discrete distribution. In financial analysis, prices were directly modeled by discrete distributions e.g. by Russell and Engle (2005) and Koopman et al. (2018). Kabasinskas et al. (2012) use discrete distributions to count zero changes in prices. In this section, we follow the discrete approach for financial durations and utilize count time series models (see e.g. Cameron and Trivedi, 2013).

There are many trade durations that are exactly zero or very close to zero. Zero durations can be caused by *split transcations*, i.e. large trades broken into two or more smaller trades. Veredas et al. (2002) offer another explanation as they notice that many simultaneous transactions occur at round prices suggesting many traders post limit orders to be executed at round prices. Zero durations can as well just be independent transactions executed at very similar times and originating from different sources. Whatever the reason for zero durations, ignoring them can cause problems in estimation as many widely used distributions have strictly positive support and zero values have therefore zero density. Liu et al. (2018c) examine the effect of zero durations on integrated variance estimation. The presence of zero durations is the second motivation of our approach. There are several ways how to deal with zero durations.

- The most common approach dating back to Engle and Russell (1998) is to discard zero durations. Specifically, observations with the same timestamp are merged together with the resulting price calculated as an average of prices weighted by volumes. This helps with estimation but the distribution of durations is distorted as zero-durations that are just independent transactions executed at similar times should be kept in the dataset.

- Instead of discarding, Bauwens (2006) sets zero durations to a small given value. This helps with estimation but the distribution of durations is distorted as zero-durations that correspond to split transactions should be omitted from the dataset.

- The information of zero durations can be utilized in a model. Zhang et al. (2001) include indicator of multiple transactions as an explanatory variable in their regression model.

- Another way of incorporating zero durations in a model is to directly include excessive zero values in the underlying distribution. For continuous distributions, zero-augmented models proposed by Hautsch et al. (2014) can be utilized [1]. However, in high-precision data, there are no exact zero values but rather very small positive values, many of which should be considered as zeros. Grammig and Wellner (2002) suggest to treat successive trades with either non-increasing or non-decreasing prices within one second as one large trade (i.e. as zero durations). The issue with this approach is that these successive trades can as well be independent and originate from different sources. Therefore, it is an uneasy task to identify whether close-to-zero durations indicate actual split transactions.

- It is more convenient to model zero durations in a discrete framework. When the values are grouped, zero durations corresponding to split transactions manifest themselves as an excessive probability of the group containg zero values. For discrete distributions, the zero-inflated extension of Lambert (1992) can be utilized. In this section, this is the approach we suggest.

For these two reasons, we propose the zero-inflated autoregressive conditional duration (ZIACD) model. We directly take into account a discreteness of durations and utilize the negative binomial distribution to accommodate for overdispersion in durations (see Boswell and Patil, 1970; Cameron and Trivedi, 1986; Christou and Fokianos, 2014). The excessive zero durations caused by split transactions are captured by the zero-inflated modification of the negative binomial distribution (see Greene, 1994). The time-varying location parameter follows the specification of general autoregressive score (GAS) models (see Creal et al., 2008, 2013; Harvey, 2013).

Specification of Zero-Inflated Duration Model

We operate within a discrete framework and assume trade durations to have discrete values $X_i \in \mathbb{N}_0$, $i = 1, \dots, n$. In the proposed *zero-inflated autoregressive conditional duration (ZIACD) model*, we consider observations to have zero-inflated negative binomial distribution with the time-varying parameter μ_i and static parameters $g = (\alpha, \pi)'$ specified in (3.42). We consider the time-varying parameter to follow the GAS recursion. We follow the theory outlined in Section 3.1.2 and Section 3.2.2.

We use a reparametrization with the exponential link for the location parameter $f_i = H(\mu_i) = \log(\mu_i)$. Parameter $\log(\mu_i)$ then follow recursion

$$f_{i+1} = c + b f_i + a s(x_i, f_i, g), \tag{3.62}$$

where c is the constant parameter, b is the autoregressive parameter, a is the score parameter and $s(x_i, f_i, g) = \tilde{S}(f_i, g)\tilde{\nabla}(x_i, f_i, g)$ is the reparametrized scaled score. Note that both the scaling function $\tilde{S}(f_i, g)$ and the score $\tilde{\nabla}(x_i, f_i, g)$ are with respect to the reparametrization $H(\mu_i)$, which can be obtained from (3.55) and (3.56). The long-term mean and unconditional value of f_i is then $\bar{f} = (1-b)^{-1}c$ and $\bar{\mu} = e^{(1-b)^{-1}c}$ in the original restricted parametrization.

Next, we present the exact specifications of the scaled score for all three considered scaling functions. The reparametrized scaled score for the unit scaling is equal to

$$s(x_i, f_i, g) = \begin{cases} \dfrac{\exp(f_i)(\pi-1)}{(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)} & \text{for } x_i = 0, \\ \dfrac{x_i - \exp(f_i)}{\alpha\exp(f_i)+1} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.63}$$

[1] The use of zero-augmented models was also suggested by T. V. Ramanathan during the 3rd Conference and Workshop on Statistical Methods in Finance, Chennai, December 16–19, 2017.

The reparametrized scaled score for the square root of the Fisher information scaling is equal to

$$s(x_i, f_i, g) = \begin{cases} \dfrac{\sqrt{\exp(f_i)}\sqrt{\pi-1}}{\sqrt{\pi\exp(f_i)-(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)}\sqrt{1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi}} & \text{for } x_i = 0, \\ \dfrac{(x_i-\exp(f_i))\sqrt{1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi}}{\sqrt{\exp(f_i)}\sqrt{\pi-1}\sqrt{\pi\exp(f_i)-(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)}} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.64}$$

The reparametrized scaled score for the inverse of the Fisher information scaling is equal to

$$s(x_i, f_i, g) = \begin{cases} \dfrac{\alpha\exp(f_i)+1}{\pi\exp(f_i)-(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)} & \text{for } x_i = 0, \\ \dfrac{(x_i-\exp(f_i))(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)}{\exp(f_i)(\pi-1)\left(\pi\exp(f_i)-(\alpha\exp(f_i)+1)\left(1+\pi(\alpha\exp(f_i)+1)^{\alpha^{-1}}-\pi\right)\right)} & \text{for } x_i = 1, 2, \dots . \end{cases} \tag{3.65}$$

As the score and the Fisher information for the zero-inflated negative binomial distribution is rather complicated, the expressions (3.63), (3.64) and (3.65) are also quite complex. However, their interpretation as the scaled score remains simple and straightforward.

Maximum Likelihood Estimation

Let us denote $\theta = (\alpha, \pi, c, b, a)'$ the static parameter vector which defines the dynamics of the GAS model proposed in (3.62). The static parameter vector θ is estimated by the method of maximum likelihood

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \hat{L}_n(\theta), \tag{3.66}$$

where $\hat{L}_n(\theta)$ denotes the log likelihood function. The log likelihood is obtained from a sequence of n observations $x_1, \dots, x_n$, which depends on the filtered time-varying parameter $\hat{f}_1(\theta), ..., \hat{f}_n(\theta)$, and is given by

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_i(x_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \mathrm{P}\left[X_i = x_i | \hat{f}_i(\theta), \theta\right]. \tag{3.67}$$

We approach the asymptotic theory of the maximum likelihood estimation within the traditional asymptotic framework described e.g. in Gallant and White (1988), Bougerol (1993), White (1994), Pötscher and Prucha (1997), Straumann and Mikosch (2006) and Blasques (2017).

Below, we show that the maximum likelihood estimator of the ZIACD model is consistent and asymptotically normal. First, we present general asymptotic theorems as formulated in Blasques (2017). Next, we derive conditions for the asymptotic properties of the proposed ZIACD model with general scaling. Finally, we discuss how these conditions can be verified for the case of the unit scaling. We present results for the ZIACD model without proofs and refer to Blasques, Holý and Tomanová (2018) for all proofs.

Filter Invertibility

Filter invertibility is crucial for statistical inference in the context of observation-driven time-varying parameter models (see e.g. Straumann and Mikosch, 2006; Wintenberger, 2013; Blasques et al., 2014).

First, let us define some basic concepts. A random sequence $\{z_i\}_{i\in\mathbb{N}}$ is said to be *strictly stationary* if the distribution of every finite sub-vector is invariant in time. A random sequence $\{z_i\}_{i\in\mathbb{N}}$ is said to be *ergodic* if and only if, every event occurs with probability 0 or 1 over an infinite amount of time. The filter $\{\hat{f}_i(\theta)\}_{i\in\mathbb{N}}$ initialized at some point $\hat{f}_1 \in \mathbb{R}$ is said to be *invertible* if $\hat{f}_i(\theta)$ converges almost surely exponentially fast to a unique limit strictly stationary and ergodic sequence $\{f_i(\theta)\}_{i\in\mathbb{Z}}$, i.e.

$$\gamma^i \left|\hat{f}_i(\theta) - f_i(\theta)\right| \xrightarrow{a.s.} 0 \quad \text{as } i \to \infty \quad \text{for some } \gamma > 1. \tag{3.68}$$

Let $L_n(\theta)$ denote the log likelihood which depends on the limit time-varying parameter $f_1(\theta), ..., f_n(\theta)$

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(x_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \mathrm{P}\left[X_i = x_i | f_i(\theta), \theta\right]. \tag{3.69}$$

Let $L_\infty(\theta)$ denote the limit log likelihood function

$$L_\infty(\theta) = \mathrm{E}\left[\ell_i(\theta)\right] = \mathrm{E}\left[\log \mathrm{P}\left[X_i = x_i | f_i(\theta), \theta\right]\right]. \tag{3.70}$$

Let $\mathrm{E}_{x_i>0}$ denote the conditional expectation $\mathrm{E}_{x_i>0}[\cdot] = \mathrm{E}[\cdot|x_i > 0]$. Finally, let $\log^+$ denote the positive part of the natural logarithm.

In Proposition 3.1, we establish the filter invertibility of the ZIACD model. In Example 3.1, we show how the conditions of Proposition 3.1 can be verified for the case of the unit scaling. Proposition 3.1 is based on Theorem 3.1.

Theorem 3.1. *For some $\theta \in \Theta$, let $\{\hat{f}_i(\theta, \hat{f}_1)\}_{i\in\mathbb{N}}$ be a random sequence initialized at $i = 1$ with value $\hat{f}_1 \in \mathcal{X} \subseteq \mathbb{R}$ and generated by the Markov dynamic system $\hat{f}_{i+1} = \phi(\hat{f}_i, \varepsilon_i, \theta)\ \forall i \in \mathbb{N}$ with differentiable function $\phi : \mathcal{X} \times \mathbb{R}^{n_\varepsilon} \times \Theta \to \mathcal{X}$ and elements $\hat{f}_i(\theta, \hat{f}_1)$ taking values in $\mathcal{X} \subseteq \mathbb{R}$ for every $i \in \mathbb{N}$. Suppose further that the following conditions hold:*

(i) $\{\varepsilon_i\}_{i\in\mathbb{N}}$ *is an exogenous n_ε-variate strictly stationary and ergodic sequence.*

(ii) There exists $\hat{f}_1 \in \mathcal{X}$ such that $\mathrm{E}\left[\log^+ |\phi(\hat{f}_1, \varepsilon_i, \theta)|\right] < \infty$.

(iii) The dynamical system is contracting on average

$$\mathrm{E}\left[\log \sup_{f\in\mathcal{X}} \left|\frac{\partial \phi(f, \varepsilon_i, \theta)}{\partial f}\right|\right] < 0. \tag{3.71}$$

Then $\{\hat{f}_i(\theta, \hat{f}_1)\}_{i\in\mathbb{N}}$ converges exponentially almost surely fast to a unique strictly stationary and ergodic sequence $\{f_i(\theta)\}_{i\in\mathbb{N}}$ as $i \to \infty$, i.e. $|\hat{f}_i(\theta, \hat{f}_1) - f_i(\theta)| \overset{e.a.s.}{\to} 0$ as $i \to \infty$.

Proof. See Theorem 3.1 of Bougerol (1993). □

Proposition 3.1. *Let the observed data $\{x_i\}_{i\in\mathbb{N}}$ be strictly stationary and ergodic and let Θ be a compact set which ensures that*

(i) $\log^+ \sup_{\theta\in\Theta} \left|s(0, \hat{f}_1(\theta), \theta)\right| < \infty$,

(ii) $\mathrm{E}_{x_i>0}\left[\log^+ \sup_{\theta\in\Theta} \left|s(x_i, \hat{f}_1(\theta), \theta)\right|\right] < \infty$,

(iii) $\mathrm{P}[x_i = 0] \log \sup_f \sup_{\theta\in\Theta} \left|a\frac{\partial s(0,f,\theta)}{\partial f} + b\right| + \mathrm{P}[x_i > 0]\mathrm{E}_{x_i>0}\left[\log \sup_f \sup_{\theta\in\Theta} \left|a\frac{\partial s(x_i,f,\theta)}{\partial f} + b\right|\right] < 0.$

Then the filter $\{\hat{f}_i(\theta)\}_{i\in\mathbb{N}}$ defined in (3.62) *is invertible, uniformly in $\theta \in \Theta$.*

Proof. See Proposition 1 of Blasques, Holý and Tomanová (2018). □

Example 3.1. *Consider the case of the score model for the zero-inflated negative binomial distribution with the unit scaling. We note that the conditions of Proposition 3.1 are satisfied for strictly stationary data* $\{x_i\}_{t\in\mathbb{Z}}$ *with finite logarithmic moment* $\mathrm{E}[\log^+|x_i|] < \infty$*, and for a compact parameter space* $\Theta = [\pi^-, \pi^+] \cdot [\alpha^-, \alpha^+] \cdot [c^-, c^+] \cdot [a^-, a^+] \cdot [\beta^-, \beta^+]$ *satisfying restrictions*

$$\begin{aligned} \frac{a^+(\pi^- - 1)^2}{2\alpha^-} + \frac{a^+|\pi^- - 1|}{(\alpha^-)^2} + b^+ < 1, \\ \mathrm{E}_{x_i>0}\left[\log\left(\frac{a^+(\alpha^+ x_i + 1)}{4\alpha^-} + b^+\right)\right] < 0. \end{aligned} \tag{3.72}$$

Proof. See Example 1 of Blasques, Holý and Tomanová (2018). □

Consistency of the Estimator

Theorem 3.4 below establishes the strong consistency of the maximum likelihood estimator $\hat{\theta}_n$ as the sample size n diverges to infinity. It uses the invertibility properties established in Proposition 3.1 for the ZIACD model and obtains the consistency of the maximum likelihood estimator by imposing some additional moment conditions. The moment conditions in Theorem 3.4 are written as high-level conditions that apply to most maximum likelihood settings. The high-level formulation of these assumptions gives us flexibility in applying these results to a wide range of designs of our score model. However, it can also be unfortunately abstract. Luckily, for the ZIACD model with the unit scaling, both moment conditions can be substituted for a simple moment bound directly on the data $\mathrm{E}[x_i] < \infty$ as shown in Example 3.2. Theorem 3.4 is based on Theorem 3.2 and Theorem 3.3.

Theorem 3.2. *Let* Θ *be a compact subset of* $\mathbb{R}^n$*, for some* $n \in \mathbb{N}$ *and* $Q_n : \mathbb{R}^n \times \Theta \to \mathbb{R}$ *be criterion function such that:*

(i) $Q_n(x_n, \cdot) : \Theta \to \mathbb{R}$ *is continuous on* Θ *for each* $x_n \in \mathbb{R}^n$*.*

(ii) $Q_n(\cdot, \theta) : \mathbb{R}^n \to \mathbb{R}$ *is continuous on* $\mathbb{R}^n$ *for each* $\theta \in \Theta$*.*

Then, there exists a measurable map $\hat{\theta}_n : \Omega \to \Theta$ *satisfying*

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} Q_n(x_n, \theta). \tag{3.73}$$

Proof. See Theorem 2.11 of White (1994). □

Theorem 3.3. *Let* $\hat{\theta}_n$ *be an estimator satisfying the conditions of Theorem 3.2. Further suppose that:*

(i) The criterion function Q_n *converges uniformly almost surely over* Θ *to the limit deterministic function* Q_∞ *as* $n \to \infty$

$$\sup_{\theta\in\Theta} \left|Q_n(x_n, \theta) - Q_\infty(\theta)\right| \xrightarrow{a.s.} 0 \qquad \text{as } n \to \infty. \tag{3.74}$$

(ii) The parameter $\theta_0 \in \Theta$ *is the identifiably unique maximizer of the limit criterion function* Q_∞

$$\sup_{\theta\in S^c(\theta_0,\gamma)} Q_\infty(\theta) < Q_\infty(\theta_0). \tag{3.75}$$

Then the estimator $\hat{\theta}_n$ *is strongly consistent for* θ_0*, i.e.* $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ *as* $n \to \infty$*.*

Proof. See Theorem 3.4 of White (1994). □

Theorem 3.4. *Let the conditions of Proposition 3.1 hold, the likelihood have one finite moment and the score have finite logarithmic moment,*

$$\mathrm{E}\left[\ell_i(x_i,\theta)\right] < \infty \quad \textit{and} \quad \mathrm{E}\left[\log^+ \sup_f |\nabla(x_i,f)|\right] < \infty. \tag{3.76}$$

Finally, suppose θ_0 be the unique maximizer of the limit log likelihood function $\mathrm{E}\left[\ell_i(x_i,\cdot)\right] : \Theta \to \mathbb{R}$ over the parameter space Θ, i.e. $\mathrm{E}\left[\ell_i(x_i,\theta_0)\right] > \mathrm{E}\left[\ell_i(x_i,\theta)\right]\ \forall\ \theta \in \Theta : \theta \neq \theta_0$. Then $\hat{\theta}_n$ is strongly consistent for θ_0, i.e. $\hat{\theta}_n \xrightarrow{a.s.} \theta_0 \in \Theta$ as $n \to \infty$.

Proof. See Theorem 1 of Blasques, Holý and Tomanová (2018). □

Example 3.2. *Consider again the score model for the zero-inflated negative binomial distribution with the unit scaling.*

(i) The finite moment for the log likelihood $\mathrm{E}[\ell_i(x_i,\theta)] < \infty$ stated in Theorem 3.4 holds trivially if the data has one finite moment $\mathrm{E}[x_i] < \infty$.

(ii) Additionally, the finite logarithmic moment $\mathrm{E}[\log^+ \sup_f |\nabla(x_i,f)|] < \infty$ stated in Theorem 3.4 also holds under $\mathrm{E}[x_i] < \infty$.

Proof. See Example 2 of Blasques, Holý and Tomanová (2018). □

Asymptotic Normality of the Estimator

Theorem 3.6 establishes the $\sqrt{n}$-consistency rate of $\hat{\theta}_n$ and the asymptotic normality of the standardized estimator $\sqrt{n}(\hat{\theta}_n - \theta_0)$ as $n \to \infty$. Again, the theorem is formulated using some high-level assumptions. Example 3.3 shows how these assumptions can be verified for the ZIACD model with the unit scaling. Theorem 3.6 is based on Theorem 3.5.

Theorem 3.5. *Let $\hat{\theta}_n$ be a consistent extremum estimator for a parameter θ_0 that lies in the interior of a compact parameter space Θ. Suppose further that*

(i) The scaled criterion derivative is asymptotically normal at θ_0

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} N(0,\Sigma) \qquad \textit{as } n \to \infty, \tag{3.77}$$

(ii) The second derivative of the criterion converges uniformly

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^2 Q_n(x_i,\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 Q_\infty(\theta)}{\partial\theta\partial\theta'}\right\| \xrightarrow{p} 0 \qquad \textit{as } n \to \infty. \tag{3.78}$$

(iii) The second derivative of the limit criterion $Q''_\infty(\theta_0)$ is invertible.

Then we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \Omega\Sigma\Omega'\right)$ as $n \to \infty$, where

$$\Omega = \left(\frac{\partial^2 Q_\infty(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}. \tag{3.79}$$

Proof. See Theorem 6.2 of White (1994). □

Theorem 3.6. *Let the conditions of Theorem 3.4 hold. Furthermore, let the zero-inflated negative binomial score model be correctly specified and* $\theta_0 \in \text{int}(\Theta)$*. Additionally, assume that*

(i) the first-order derivatives of the log likelihood have four finite moments at θ_0*,*

$$\mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial f_i}\right\|^4\right] < \infty \quad \textit{and} \quad \mathrm{E}\left[\left\|\frac{\partial \ell_i(x_i,\theta_0)}{\partial \theta}\right\|^4\right] < \infty, \tag{3.80}$$

(ii) the second-order derivatives of the log likelihood have one uniform finite moment,

$$\begin{gathered}\mathrm{E}\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^2 \ell_i(x_i,\theta)}{\partial f_i \partial \theta'}\right\|\right] < \infty, \quad \mathrm{E}\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^2 \ell_i(x_i,\theta)}{\partial f_i^2}\right\|\right] < \infty, \\ \mathrm{E}\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^2 \ell_i(x_i,\theta)}{\partial \theta \partial \theta'}\right\|\right] < \infty,\end{gathered} \tag{3.81}$$

(iii) the third-order derivatives of the log likelihood have uniform finite logarithmic moment,

$$\begin{gathered}\mathrm{E}\left[\log^+ \sup_{\theta\in\Theta}\left\|\frac{\partial^3 \ell_i(x_i,\theta_0)}{\partial f_i^2 \partial \theta'}\right\|\right] < \infty, \quad \mathrm{E}\left[\log^+ \sup_{\theta\in\Theta}\left\|\frac{\partial^3 \ell_i(x_i,\theta_0)}{\partial f_i^3}\right\|\right] < \infty, \\ \mathrm{E}\left[\log^+ \sup_{\theta\in\Theta}\left\|\frac{\partial^3 \ell_i(x_i,\theta_0)}{\partial \theta \partial \theta' \partial f}\right\|\right] < \infty,\end{gathered} \tag{3.82}$$

(iv) the first and second derivatives of the filtering process converge almost surely, exponentially fast, to a limit stationary and ergodic sequence,

$$\left\|\frac{\partial \hat{f}_i(\theta_0)}{\partial \theta} - \frac{\partial f_i(\theta_0)}{\partial \theta}\right\| \overset{e.a.s.}{\to} 0 \quad \textit{and} \quad \sup_{\theta\in\Theta}\left\|\frac{\partial^2 \hat{f}_i(\theta)}{\partial \theta \partial \theta'} - \frac{\partial^2 f_i(\theta)}{\partial \theta \partial \theta'}\right\| \overset{e.a.s.}{\to} 0 \quad \textit{as} \quad i \to \infty, \tag{3.83}$$

with four finite moments

$$\mathrm{E}\left[\left\|\frac{\partial f_i(\theta_0)}{\partial \theta}\right\|^4\right] < \infty \quad \textit{and} \quad \mathrm{E}\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^2 f_i(\theta)}{\partial \theta \partial \theta'}\right\|^4\right] < \infty. \tag{3.84}$$

Then the estimator is asymptotically Gaussian

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, \mathcal{I}(\theta_0)^{-1}) \quad \textit{as} \quad n \to \infty, \tag{3.85}$$

where $\mathcal{I}(\theta_0)^{-1}$ *denotes the inverse of the Fisher information.*

Proof. See Theorem 2 of Blasques, Holý and Tomanová (2018). □

Example 3.3. *Let us revisit once again the score model for the zero-inflated negative binomial distribution with the unit scaling.*

(i) The finite moments imposed in conditions (i), (ii) and (iii) of Theorem 3.6 can be verified by taking the appropriate derivatives of the log likelihood and applying standard moment inequalities. For example, it is easy to see that the four finite moments for score term $\partial \ell_i(x_i,\theta_0)/\partial f_i$ *can be obtained if the data has four finite moments.*

(ii) Similarly, the invertibility conditions stated in condition (iv) of Theorem 3.6 can be verified by applying Theorem 3.1 to the derivative filters.

Proof. See Example 3 of Blasques, Holý and Tomanová (2018). □

Pseudo-True Parameters and Kullback-Leibler Divergence

Finally, we discuss pseudo-true parameters given by the maximum likelihood estimator. We follow Blasques (2017) in this section. The *pseudo-true parameter* θ_0 is defined as the unique maximizer of the limit criterion function L_∞

$$\theta_0 \in \arg\max_{\theta\in\Theta} L_\infty(\theta). \tag{3.86}$$

In the context of maximum likelihood estimation, the parameter θ_0 is the most likely parameter value given an infinitely large sample. For the maximum likelihood estimator based on discrete random variable X_i, the pseudo-true parameter θ_0 is given by

$$\theta_0 \in \arg\max_{\theta\in\Theta} \mathrm{E}\left[\log \mathrm{P}[X_i = x_i|\theta]\right]. \tag{3.87}$$

This means that it is also the unique minimizer of

$$\theta_0 \in \arg\min_{\theta\in\Theta} \left(\mathrm{E}\left[\log \mathrm{P}_0[X_i = x_i]\right] - \mathrm{E}\left[\log \mathrm{P}[X_i = x_i|\theta]\right]\right), \tag{3.88}$$

where $\mathrm{P}_0[X_i = x_i]$ is the true unknown probability mass of X_i. This quantity is known as the *Kullback-Leibler divergence* between the conditional probability mass $\mathrm{P}[X_i = x_i|\theta]$ implied by the model and the true probability mass $\mathrm{P}_0[X_i = x_i]$. Kullback-Leibler divergence is a measure of how one probability distribution is different from another distribution. It was introduced by Kullback and Leibler (1951).

When the model is correctly specified, the parameter θ_0 then corresponds to the true parameter as the Kullback-Leibler divergence is minimized at the point θ_0 where $\mathrm{P}[X_i = x_i|\theta] = \mathrm{P}_0[X_i = x_i]$. When the model is mis-specified, θ_0 is the parameter value that provides the best approximation to the data generating process in Kullback-Leibler divergence.

3.3 Application to Discrete Trade Durations

In an empirical study, we analyze 30 stocks that form Dow Jones Industrial Average (DJIA) index. For more details about the analyzed stocks, see Appendix A. The data are taken from April to May, 2018. We clean data according to the procedure described in Section 2.1.1 omitting the step merging simultaneous transactions.

Basic statistical characteristics after data cleaning are presented in Table 3.2. We give a special attention to the IBM stock as many other studies including Engle and Russell (1998). Figure 3.1 shows trading intensity during trading hours for several trading days of the IBM stock. We can see that there is clear autocorrelation, although each day has a different course. Generally, more trades occur both at the beginning and at the end of a day while the lunch-time is a quiet period with less trades. This behavior is well captured by the ACD models.

3.3.1 Models Performance

We compare models based on the Poisson, geometric and negative binomial distribution together with their zero-inflated versions. First, we evaluate in-sample performance of the discrete models with unit scaling. Second, we evaluate their out-of-sample performance. In both cases, the zero-inflated negative binomial distribution is the best choice. Third, we compare the unit scaling with the square root of the inverse of the Fisher information scaling and the inverse of the Fisher information scaling. We argue that there are not significant differences among the three considered scaling functions as the results are very similar in our application.

	April 2018				May 2018			
Stock	Mean	Var.	n	n_0/n	Mean	Var.	n	n_0/n
AAPL	0.3325	1.2398	1 033 149	0.8553	0.4801	2.4949	816 757	0.8328
AXP	4.8469	107.9742	95 618	0.5306	6.4839	162.4193	75 848	0.4912
BA	2.6756	50.5394	170 401	0.6826	3.1278	59.5706	153 908	0.6605
CAT	3.1267	51.0194	145 195	0.6020	4.0331	72.0750	119 633	0.5611
CSCO	1.1183	16.7652	399 797	0.8270	1.2048	24.6283	394 975	0.8467
CVX	3.0356	33.2949	147 191	0.5309	3.1524	42.1781	150 090	0.5713
DIS	2.6397	24.0900	166 417	0.5275	2.3905	22.6665	191 029	0.5553
DWDP	2.9671	38.7329	151 087	0.5548	3.7913	67.6563	126 805	0.5735
GE	2.1086	29.1117	206 774	0.6332	2.5770	39.3522	179 714	0.5870
GS	2.8363	48.6788	159 905	0.6408	3.9585	80.4010	122 094	0.5854
HD	3.2071	45.2574	140 834	0.5647	3.5334	49.5467	134 424	0.5400
IBM	3.1991	47.0185	141 173	0.5665	4.5602	70.0383	105 697	0.4835
INTC	0.6562	5.6652	630 689	0.8367	1.0788	14.0168	427 737	0.8143
JNJ	2.6894	29.1461	164 944	0.5564	3.6135	47.0778	131 119	0.5035
JPM	1.0586	4.9859	368 021	0.6508	1.5880	10.0435	274 251	0.5938
KO	3.7592	73.1405	121 132	0.5439	4.8639	108.7622	99 634	0.5195
MCD	4.6241	93.0448	99 920	0.5275	5.6159	116.5453	86 826	0.4791
MMM	3.9806	76.2675	115 427	0.5751	6.0409	146.2070	81 086	0.5102
MRK	2.2550	24.0990	191 893	0.5724	2.7490	35.0742	169 067	0.5683
MSFT	0.4358	2.2432	860 371	0.8452	0.6300	4.8465	676 757	0.8313
NKE	3.3765	45.2257	133 242	0.5118	3.9377	63.5151	121 253	0.5070
PFE	2.9778	41.5950	149 534	0.5527	3.3798	65.5733	140 374	0.5881
PG	2.5414	28.9706	172 395	0.5565	3.3726	43.9835	139 142	0.5067
TRV	9.4651	389.7584	50 208	0.4810	10.6297	423.1976	46 946	0.4385
UNH	3.8590	74.8488	119 276	0.5849	5.3471	130.9015	91 672	0.5573
UTX	4.2728	78.2773	107 689	0.5386	5.8129	133.8689	8 3 967	0.4886
V	2.4218	26.5329	182 851	0.5991	3.3445	44.5483	142 026	0.5543
VZ	2.9317	42.0215	152 214	0.5574	3.7413	69.2588	127 330	0.5403
WMT	3.1127	32.1728	143 003	0.4904	2.7936	28.5493	165 904	0.5227
XOM	1.8195	1 4.6438	232 388	0.5896	1.8329	16.4846	243 875	0.6141

Table 3.2: The sample mean of durations, sample variance of durations, number of observations n and ratio n_0/n of durations shorter than 1 second .

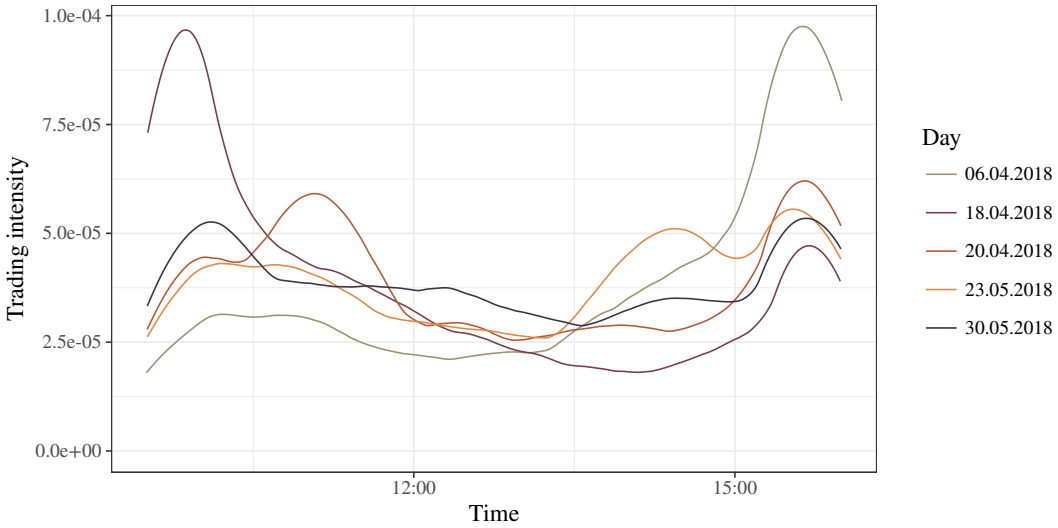

Figure 3.1: Daily trading intensity estimated by the Epanechnikov kernel density for the IBM stock.

In-Sample Results

We fit durations rounded down to seconds of the 30 DJIA stocks using data from April, 2018. To evaluate in-sample fit of the models, we use the *Akaike information criterion (AIC)* (Akaike, 1974) defined as

$$AIC = 2q - 2n\hat{L}_n(\hat{\theta}), \tag{3.89}$$

where $q = 3k + l$ is the number of parameters.

We find that the model based on the zero-inflated negative binomial distribution is the best fit. Estimated parameters are reported in Table 3.3. There is clear evidence of overdispersion, i.e. the variance higher than expected value. Table 3.2 shows that sample variance is much higher than sample mean. According to Table 3.3, the estimated value of dispersion parameter α in the zero-inflated negative binomial model ranges between 1.37 and 2.78 depending on the stock. This favors the negative binomial distribution over Poisson distribution with fixed $\alpha = 0$ and geometric distribution with fixed $\alpha = 1$. Overdispersion is also supported by AIC of the models reported in Table 3.4. The Poisson distribution has the highest AIC for all stocks followed by the geometric distribution. One possible reason for overdispersion could just be the presence of excessive zeros. Zero-inflated Poisson and geometric distributions perform better than the original distributions. However, they are inferior to the zero-inflated negative binomial distribution suggesting there is overdispersion present in non-zero values as well.

Our analysis also reveals the presence of excessive zeros suggesting the existence of the process generating only zero values (i.e. split transactions) alongside the process generating regular durations. According to Table 3.3, the estimated probability of excessive zeros π in the zero-inflated negative binomial model ranges between 0.21 and 0.75 depending on the stock. This corresponds to the ratio of excessive zeros to all zeros ranging between 0.37 and 0.90. Again, the presence of excessive zeros is supported by a decrease in AIC in the zero-inflated distributions as reported in Table 3.4. Table 3.5, Figure 3.2 and Figure 3.3 illustrate shortcomings of the regular negative binomial distribution. In this model, the probability of zero values is underestimated while probabilities of values equal to 1 and 2 are overestimated. The zero-inflated negative binomial distribution better captures probabilities of zero as well as positive values.

Out-of-Sample Results

We forecast durations during May, 2018 for 30 DJIA stocks. We use the models estimated using April, 2018 durations and perform one-step-ahead forecasts. Again, we compare models based on the Poisson,

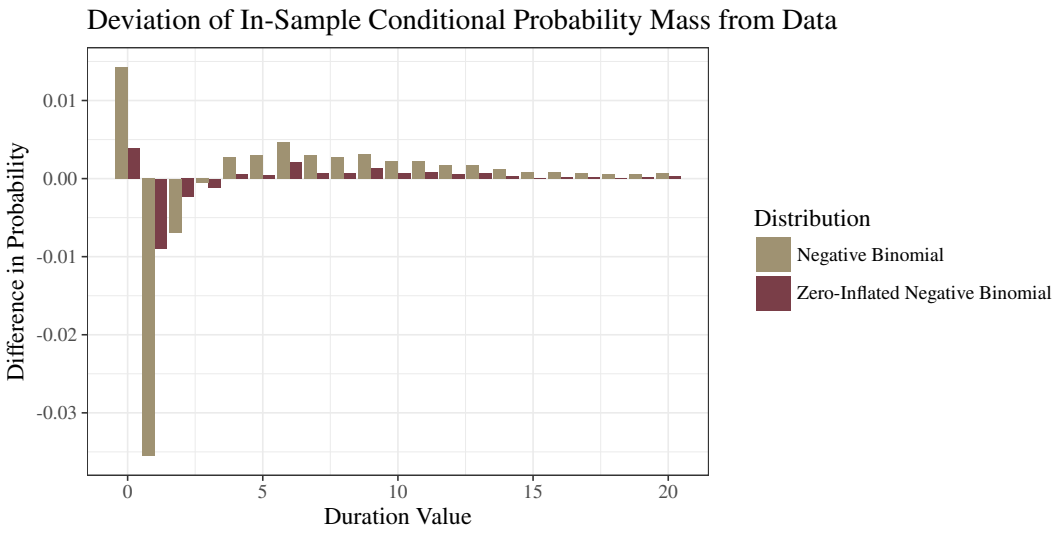

Figure 3.2: Deviation of average in-sample conditional probability mass of duration models based on the negative binomial and zero-inflated negative binomial distributions from data for the IBM stock.

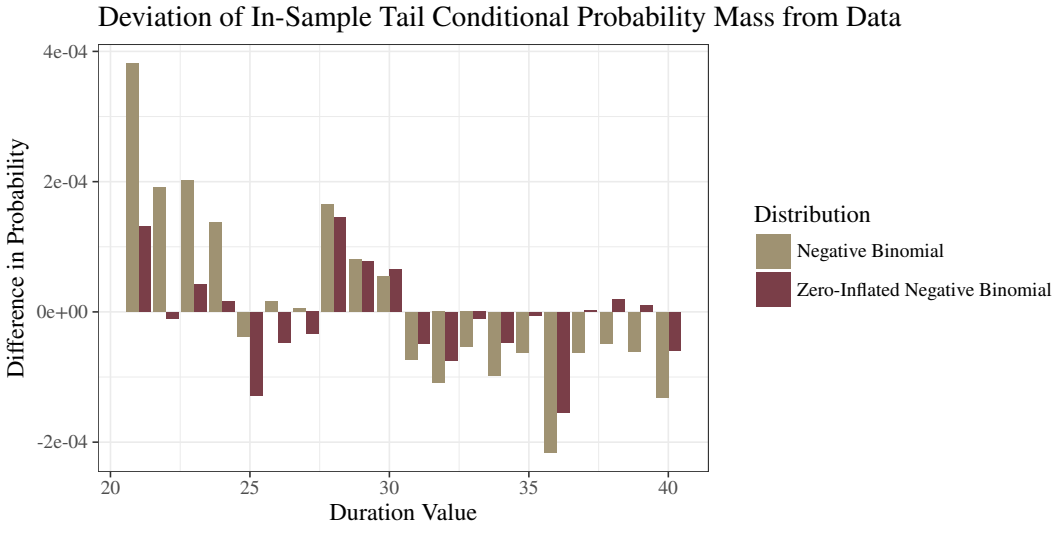

Figure 3.3: Deviation of average in-sample tail conditional probability mass of duration models based on the negative binomial and zero-inflated negative binomial distributions from data for the IBM stock.

Stock	c	b	a	μ	α	π	$\pi n/n_0$
AAPL	-0.0005	0.9992	0.1365	0.5536	2.1981	0.5423	0.6341
AXP	0.0023	0.9986	0.1043	5.2194	1.5091	0.3406	0.6421
BA	0.0015	0.9989	0.0746	4.3259	1.5896	0.5531	0.8105
CAT	0.0013	0.9989	0.0864	3.2437	1.6095	0.3946	0.6555
CSCO	0.0009	0.9992	0.0909	3.0562	1.6550	0.7468	0.9031
CVX	0.0033	0.9973	0.0683	3.4714	1.4882	0.3240	0.6106
DIS	0.0025	0.9979	0.0443	3.2370	1.4352	0.3174	0.6019
DWDP	0.0027	0.9978	0.0758	3.4297	1.5376	0.3544	0.6388
GE	0.0011	0.9986	0.1234	2.1912	2.3913	0.2963	0.4681
GS	0.0019	0.9986	0.0958	4.0308	1.6823	0.4753	0.7419
HD	0.0025	0.9981	0.0575	3.7249	1.5584	0.3783	0.6701
IBM	0.0013	0.9991	0.0781	3.8894	1.4851	0.3619	0.6389
INTC	-0.0000	0.9997	0.1022	0.9973	1.8114	0.6785	0.8109
JNJ	0.0016	0.9986	0.0923	3.0145	1.5536	0.3168	0.5695
JPM	0.0003	0.9976	0.0950	1.1269	1.6171	0.2746	0.4220
KO	0.0023	0.9983	0.0938	3.8186	1.7727	0.3350	0.6161
CD	0.0029	0.9981	0.0974	4.6377	1.6524	0.3277	0.6214
MMM	0.0019	0.9988	0.0466	4.9913	1.5057	0.4231	0.7358
MRK	0.0019	0.9975	0.0629	2.1716	2.2495	0.2106	0.3680
MSFT	-0.0006	0.9986	0.1834	0.6531	2.7830	0.5261	0.6225
NKE	0.0043	0.9967	0.0921	3.6153	1.5318	0.3029	0.5920
PFE	0.0014	0.9988	0.0361	3.1641	1.9766	0.3014	0.5455
PG	0.0015	0.9984	0.0420	2.5336	1.8560	0.2657	0.4776
TRV	0.0135	0.9945	0.0768	11.8486	1.5690	0.3745	0.7789
UNH	0.0021	0.9986	0.0866	4.7575	1.6127	0.4194	0.7172
UTX	0.0044	0.9972	0.0787	4.9029	1.6319	0.3650	0.6779
V	0.0013	0.9988	0.0746	2.8667	1.3962	0.4026	0.6721
VZ	0.0011	0.9987	0.0718	2.2735	1.7328	0.3017	0.5413
WMT	0.0014	0.9987	0.0673	2.9697	1.3694	0.2639	0.5383
XOM	0.0018	0.9973	0.0577	1.9454	1.8167	0.2764	0.4688

Table 3.3: Estimated parameters of duration model based on the zero-inflated negative binomial distribution.

geometric and negative binomial distributions together with their zero-inflated versions and we restrict ourselves to the unit scaling. Let n denote the number of in-sample observations and m the number of out-of-sample observations. We evaluate forecasting accuracy of the models using a score rule based on the out-of-sample likelihood. For a single prediction at time i, we use the *logarithmic score (LS)* (see e.g. Amisano and Giacomini, 2007; Bao et al., 2007; Diks et al., 2011) defined as

$$LS_i = \log \mathrm{P}[X_i = x_i | \hat{f}_i, \hat{g}], \quad i = n+1, \ldots, n+m, \tag{3.90}$$

where $\mathrm{P}[X_i = x_i | \hat{f}_i, \hat{g}]$ is the forecasted probability of the actual value x_i at time i. Higher values of LS indicate higher prediction accuracy. For a comparison of models A and B, we adopt the *Diebold-Mariano test* (Diebold and Mariano, 1995). Let LS_i^A denote the logarithmic score for the model A and LS_i^B for the model B at time i. Let us define difference between logarithmic scores of the two models as

$$D_i^{A,B} = LS_i^A - LS_i^B, \quad i = n+1, \ldots, n+m. \tag{3.91}$$

Stock	P	G	NB	ZIP	ZIG	ZINB
AAPL	1 752 835	1 436 127	1 274 892	1 329 073	1 324 993	1 273 668
AXP	1 116 440	453 481	405 616	692 445	427 167	402 954
BA	1 714 949	671 432	530 018	793 071	545 287	525 603
CAT	1 251 216	589 332	516 469	764 941	541 448	513 647
CSCO	2 242 102	1 108 899	741 638	970 148	753 770	735 754
CVX	1 292 853	626 788	574 269	776 359	588 582	571 962
DIS	1 276 179	681 546	631 054	807 139	643 547	628 753
DWDP	1 380 739	631 931	568 598	794 463	586 941	566 040
GE	1 585 084	729 731	640 220	892 066	674 969	639 487
GS	1 599 302	646 101	535 237	797 104	553 623	531 869
HD	1 391 154	605 618	536 811	773 215	551 925	534 074
IBM	1 409 646	587 759	527 272	762 254	551 672	524 298
INTC	2 156 336	1 275 969	988 674	1 134 875	1 025 972	984 691
JNJ	1 352 421	655 228	600 622	800 594	623 946	598 487
JPM	1 425 972	983 194	935 525	1 020 591	970 479	934 605
KO	1 240 075	547 311	483 538	776 119	503 712	481 649
MCD	1 138 054	475 858	424 350	710 930	443 140	422 252
MMM	1 241 189	528 115	455 307	721 713	469 907	451 916
MRK	1 413 460	717 368	656 316	876 315	682 954	655 864
MSFT	2 007 496	1 398 027	1 183 376	1 274 651	1 229 614	1 182 340
NKE	1 294 371	589 112	539 781	766 888	556 507	537 530
PFE	1 399 374	634 262	565 807	824 666	583 032	564 604
PG	1 378 713	673 862	619 087	834 969	643 021	618 046
TRV	1 125 310	308 609	263 795	643 669	271 665	261 338
UNH	1 271 795	538 710	460 092	736 672	476 276	456 995
UTX	1 209 229	514 266	450 870	733 816	463 881	448 394
V	1 434 679	694 257	622 431	809 263	643 856	618 864
VZ	1 211 910	617 784	561 035	805 736	587 971	559 527
WMT	1 222 539	608 654	574 724	762 183	595 135	572 621
XOM	1 400 264	802 635	743 016	905 705	766 784	742 016

Table 3.4: In-sample Akaike information criterion of duration models based on the Poisson (P), geometric (G), negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated geometric (ZIG) and zero-inflated negative binomial (ZINB) distributions.

	Duration Value					
Distribution	0	1	2	3	4	5
Observed Data	0.5664	0.0865	0.0595	0.0437	0.0350	0.0279
Negative Binomial	0.5521	0.1220	0.0664	0.0442	0.0322	0.0248
Zero-Inflated Negative Binomial	0.5625	0.0954	0.0618	0.0448	0.0344	0.0274

Table 3.5: Average in-sample conditional probability mass for the IBM stock.

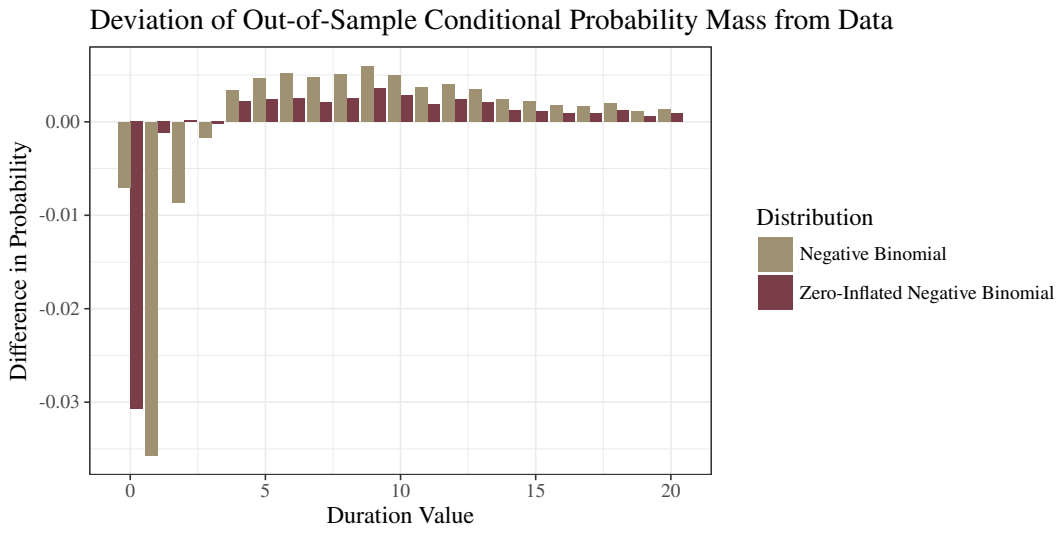

Figure 3.4: Deviation of average out-of-sample conditional probability mass of duration models based on the negative binomial distribution and zero-inflated negative binomial distributions from data for the IBM stock.

The mean and standard deviation are

$$\overline{D}^{A,B} = \frac{1}{m}\sum_{n+1}^{n+m} D_i^{A,B}, \qquad \sigma_D^{A,B} = \sqrt{\frac{1}{m-1}\sum_{n+1}^{n+m}\left(D_i^{A,B} - \overline{D}^{A,B}\right)^2}. \tag{3.92}$$

Diebold-Mariano test statistic is then defined as

$$DM^{A,B} = \sqrt{m}\frac{\overline{D}^{A,B}}{\sigma_D^{A,B}}. \tag{3.93}$$

Under the null hypothesis of equal performance of both models, the statistic has asymptotically standard normal distribution.

We compare the zero-inflated negative binomial distribution with the other considered distributions. Diebold-Mariano test statistics are reported in Table 3.6. All values are positive, which means that the zero-inflated negative binomial distribution outperforms all the other distributions. The values are also quite high, which means that the zero-inflated negative binomial distribution is significantly better at any reasonable significance level. These out-of-sample results together with in-sample results clearly show that the duration model based on the zero-inflated negative binomial distribution is the most suitable model among the considered candidates.

However, there are some shortcomings in the predictive ability of our models. Table 3.7 and Figure 3.4 illustrate forecasted probability mass of the negative binomial and zero-inflated negative binomial distributions. We can see that the zero-inflated negative binomial distribution is a very good fit for positive values but overestimates zero value for the IBM stock. This could be explained by a decrease in probability of excessive zeros in May, 2018. Indeed, we can see in Table 3.2 that the ratio of all zero values decreased from 57% to 48% from April to May for the IBM stock.

We leave the analysis of long-term dynamics of excessive zero probability as a topic for future research. In the context of financial duration modeling, non-stationary ACD models were studied by Bortoluzzo et al. (2010) and Mishra and Ramanathan (2017).

Stock	ZINB/P	ZINB/G	ZINB/NB	ZINB/ZIP	ZINB/ZIG
AAPL	182.6095	182.7811	22.9323	78.5214	110.7903
AXP	116.7725	97.0725	19.2860	78.2483	31.5584
BA	122.6371	164.7156	29.1190	80.7895	31.5362
CAT	137.0735	127.0594	27.7917	83.5686	47.1496
CSCO	121.9222	230.0715	36.7724	71.1828	47.6283
CVX	138.1669	123.9111	16.7692	84.9286	24.4014
DIS	147.2722	103.5195	14.4460	75.0423	40.0487
DWDP	111.6772	139.3729	28.7878	78.4778	37.4006
GE	102.2640	101.3909	9.2927	74.7982	73.1739
GS	104.2343	128.0335	26.5007	79.8998	50.4813
HD	125.3915	85.2530	14.8932	77.9376	23.4781
IBM	111.4394	91.4346	23.4384	88.2387	49.8152
INTC	126.7347	233.6059	46.4575	76.5541	54.6382
JNJ	115.4911	91.8415	17.0701	85.7032	46.5001
JPM	133.5494	107.1435	21.8151	84.5418	80.2947
KO	116.0503	110.0412	19.3464	75.4929	49.2069
MCD	119.3942	93.0403	21.0761	79.8904	41.5543
MMM	115.9969	80.4366	12.1684	78.3872	26.0296
MRK	113.5194	100.2440	15.2540	68.4814	28.4557
MSFT	143.1199	211.3403	21.9804	78.4489	73.6277
NKE	109.3419	111.8271	24.1253	79.8502	51.2108
PFE	107.0852	115.4612	12.7901	75.2372	44.2632
PG	105.6457	72.6883	16.7427	69.5813	34.0538
TRV	90.0728	62.7401	19.8402	64.8239	30.7718
UNH	118.2081	117.8648	26.2378	74.8625	39.2313
UTX	108.5396	79.4963	11.1877	73.6593	29.6606
V	122.8782	119.7647	23.7352	88.0254	55.6638
VZ	115.9128	110.8392	15.2525	72.5115	43.6104
WMT	147.7816	105.3120	17.0288	85.8218	75.0586
XOM	137.5378	105.4905	9.3978	72.0103	37.6252

Table 3.6: Out-of-sample Diebold-Mariano test statistic comparing duration model based on the zero-inflated negative binomial distribution (ZINB) with duration models based on the Poisson (P), geometric (G), negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated geometric (ZIG) distributions.

	Duration Value					
Distribution	0	1	2	3	4	5
Observed Data	0.4833	0.0824	0.0595	0.0456	0.0392	0.0330
Negative Binomial	0.4904	0.1181	0.0681	0.0473	0.0357	0.0283
Zero-Inflated Negative Binomial	0.5141	0.0836	0.0593	0.0459	0.0370	0.0306
Generalized Gamma with Discarding	0.4349	0.1139	0.0748	0.0555	0.0436	0.0354
Generalized Gamma with Truncating	0.5462	0.0908	0.0575	0.0421	0.0329	0.0267

Table 3.7: Average out-of-sample conditional probability mass for the IBM stock.

Scaling Function

So far, we have used only the unit scaling. In this section, we compare the unit scaling $S(f_i, g) = I$ with the square root of the inverse of the Fisher information scaling $S(f_i, g) = \mathcal{I}(f_i, g)^{-\frac{1}{2}}$ and the inverse of the Fisher information scaling $S(f_i, g) = \mathcal{I}(f_i, g)^{-1}$. The results of both in-sample and out-of-sample analysis are reported in Table 3.8. It is evident that there is no universally best scaling. Each of the three considered scalings leads to the lowest AIC for some stocks and the highest AIC for other stocks. Out-of-sample analysis is also inconclusive. For some stocks (e.g. AXP and KO), Diebold-Mariano test shows no significant differences between the models. For some stocks (e.g. BA, CVX), a single model is significantly preferred. However, this may be inconsistent with the in-sample preference as in the case of CVX suggesting the choice of scaling may change in time. Overall, differences between estimated coefficients are quite negligible. For these reasons, we use only the unit scaling throughout the section.

3.3.2 Discrete vs. Continuous Approach

We assess both motivations for the discrete approach by comparing discrete distributions with the exponential, Weibull, gamma and generalized gamma distributions within the GAS framework. The exponential distribution and the Weibull distribution were proposed to model financial durations by Engle and Russell (1998), while the generalized gamma distribution was proposed by Lunde (1999). Both Bauwens et al. (2004) and Fernandes and Grammig (2005) found than the generalized gamma distribution is more adequate than the exponential, Weibull and Burr distributions. The study Xu (2013) shows that the log-normal distribution does not outperform the generalized gamma distribution either. For these reasons, the generalized gamma distribution is our main candidate for the competing continuous distribution. In our comparison, we do not consider the generalized F distribution as it has four parameters and in most cases of financial durations reduces to the generalized gamma distribution as discussed by Hautsch (2003) and Hautsch (2011). We also do not consider Birnbaum-Saunders distribution as it models median instead of mean and therefore does not strictly belong to the traditional ACD class.

First, in a simulation study, we study discretness of data and show how various degrees of rounding affect discrete and continuous models. Second, in an empirical study, we study zero durations and show how various treatments of zero values induce loss of information. We find that the proposed discrete approach is superior from both perspectives.

Simulation Study

In a simulation study, we explore the influence of rounding on estimation of a GAS model based on discrete and continuous distributions. For this purpose we restrict ourselves to a comparison of the exponential distribution (a special case of the generalized gamma distribution) with the geometric distribution (a special case of the negative binomial distribution) as the geometric distribution is the discrete analogue of the exponential distribution. Specifically, if a random variable X_i follows the exponential distribution with the scaling parameter β_i, the variable rounded down to the nearest integer $\lfloor X_i \rfloor$ follows the geometric distribution with the parameter μ_i. The parameters β_i and μ_i are then related by

$$\mu_i = \frac{1}{\exp\left(\beta_i^{-1}\right) - 1}, \qquad \beta_i = \frac{1}{\log\left(\mu_i^{-1} + 1\right)}. \tag{3.94}$$

We use the geometric distribution reparametrized according to (3.94) so both GAS specifications model the same parameter.

We simulate 1000 observations following the GAS specification based on the exponential distribution with true parameters $c = 0$, $b = 0.9$, $a = 0.1$ and unconditional value of the scale parameter equal to 1. Then, we round down the observations to a given number of decimal places. Finally, we estimate the GAS model using the rounded observations. The simulation is performed 1000 times.

	In-Sample AIC			Out-of-Sample DM	
Stock	I	$\mathcal{I}^{-\frac{1}{2}}$	$\mathcal{I}^{-1}$	$I/\mathcal{I}^{-\frac{1}{2}}$	$I/\mathcal{I}^{-1}$
AAPL	1 273 668	1 274 470	1 273 668	15.3206	1.3221
AXP	402 954	402 740	402 866	0.6763	-1.3888
BA	525 603	525 655	525 603	5.6978	3.6410
CAT	513 647	513 687	513 647	-2.5933	4.0843
CSCO	735 754	735 724	735 754	-1.1811	-13.9296
CVX	571 962	572 371	571 958	12.5461	2.1536
DIS	628 753	629 133	628 762	4.3508	0.4103
DWDP	566 040	566 272	566 033	2.2668	-3.5294
GE	639 487	639 493	639 487	-3.8839	6.6827
GS	531 869	531 811	531 869	-2.8932	-2.2265
HD	534 074	534 448	534 080	8.8983	0.1088
IBM	524 298	524 168	524 248	-6.7754	-15.5891
INTC	984 691	984 638	984 691	-12.4427	-3.5159
JNJ	598 487	598 264	598 427	2.0586	-6.7557
JPM	934 605	934 760	934 623	-3.3771	-8.4668
KO	481 649	481 611	481 649	0.9593	-0.1152
MCD	422 252	422 307	422 219	-5.2202	-10.3754
MMM	451 916	452 175	451 915	7.2350	5.3239
MRK	655 864	656 471	655 866	4.9090	-0.9545
MSFT	1 182 340	1 182 245	1 182 306	-8.7383	-5.2511
NKE	537 530	537 322	537 401	-0.3547	-2.9612
PFE	564 604	564 857	564 604	9.4416	-1.9040
PG	618 046	619 369	618 049	15.0810	-1.8765
TRV	261 338	261 409	261 342	3.6553	2.1572
UNH	456 995	456 830	456 995	-1.5005	2.8147
UTX	448 394	448 594	448 399	7.0043	5.7113
V	618 864	618 927	618 854	-6.0081	-6.4800
VZ	559 527	559 931	559 546	4.3381	-2.3639
WMT	572 621	572 373	572 418	-7.1384	-8.6033
XOM	742 016	742 818	742 028	12.1894	-2.5186

Table 3.8: In-sample Akaike information criterion and out-of-sample Diebold-Mariano test statistic for duration models based on the zero-inflated negative binomial distribution with the unit scaling I, the square root of the inverse of the Fisher information scaling $\mathcal{I}^{-\frac{1}{2}}$ and the inverse of the Fisher information scaling $\mathcal{I}^{-1}$.

Estimate	G(0)	G(1)	G(2)	E(0)	E(1)	E(2)	E(∞)
c	0.005	0.005	0.005	0.055	0.007	0.005	0.005
b	0.039	0.037	0.037	0.039	0.037	0.037	0.037
a	0.029	0.027	0.027	0.031	0.027	0.026	0.026
β	0.051	0.051	0.051	0.424	0.067	0.051	0.051

Table 3.9: Mean absolute errors of the parameters estimated from a simulated GAS model based on the geometric (G) and exponential (E) distributions with data rounded down to a given precision as denoted in parentheses.

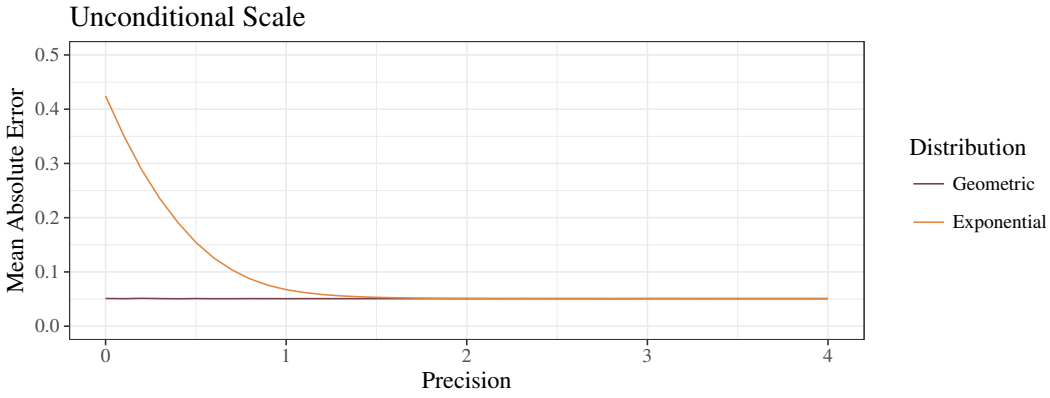

Figure 3.5: Mean absolute error of the unconditional scale estimated from a simulated GAS model based on the geometric and exponential distributions with data rounded down to a given precision.

In Figure 3.5 and Table 3.9, we see the results of the simulation experiment. Both exponential distribution and geometric distribution identify the autoregressive parameter b and the score parameter a under any degree of rounding. The model with geometric distribution also estimates the constant parameter c and the unconditional scale with a minimal error under any degree of rounding. The model with the exponential distribution, however, gives a biased estimate of the constant parameter c and therefore the biased unconditional scale when the rounding is significant. The results show that it is more appropriate to use correctly specified discrete distribution when the continuous process has rounded values.

Out-of-Sample Comparison

We resume the empirical analysis with the continuous approach. For this purpose, we use the original unrounded durations. As they have a precision of 6 decimal places or more for some stocks, it is quite suitable to model them using continuous distributions. However, a numerical problem with close-to-zero values arises. There are two ways how to deal with close-to-zero durations. The first option is to *discard* close-to-zero values. This is a very common approach dating back to Engle and Russell (1998). The second option is to *truncate* close-to-zero values. This is a less used approach proposed by Bauwens (2006). We compare proposed discrete approach with the continuous approach that discards close-to-zero values and the approach that truncates close-to-zero values. In all cases, the original data are modified. All three approaches alter observations and discarding close-to-zero values also reduces the number of observations. For this reason, we focus on the out-of-sample forecasts, in which we do not discard observations.

In the estimation process, we face some numerical issues. We consider close-to-zero values lower than 0.001. This is an empirically selected treshold that leads to convergence of the estimator for most stocks. When the close-to-zero values are present, the likelihood function increases far above a reasonable limit for the Weibull, gamma and generalized gamma distributions. This is more significant for frequently traded stocks such as AAPL, CSCO, INTC and MSFT. Note that these are the four stocks in the DJIA index traded on NASDAQ while the rest is traded on NYSE. The estimation of the exponential distribution is unaffected by close-to-zero values as it contains zero in its support. As the estimation procedure, we use a combination of the Nelder–Mead algorithm (NM) (Nelder and Mead, 1965) and the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) (Broyden, 1970a,b; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) implemented in the open-source NLopt library (Johnson, 2019). In the case of the four most traded stocks and truncating close-to-zero values, neither algorithm does converge. This is because of a huge number of close-to-zero values. Specifically, 54% for AAPL, 70% for CSCO, 65% for INTC and 59% for MSFT. JPM is also a frequently traded stock but has only 22% of close-to-zero values and its convergence is therefore unaffected.

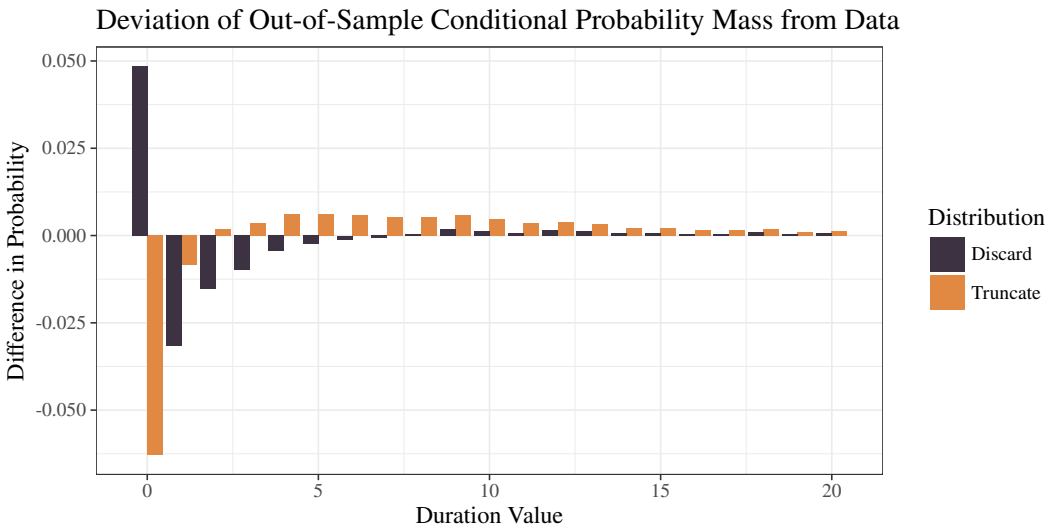

Figure 3.6: Deviation of average out-of-sample conditional probability mass of duration model based on the generalized gamma distribution from data with close-to-zero values either discarded or truncated for the IBM stock.

For evaluation, we use the logarithmic score with Diebold-Mariano test statistic as in Section 3.3.1. To be able to compare the discrete ZINB model with continuous models, we evaluate all models on the same discrete grid. For continuous distributions, we modify the logarithmic score (3.90) to

$$LS_i = \log \mathrm{P}[\underline{x}_i < X_i \leq \overline{x}_i | \hat{f}_i, \hat{g}], \quad i = n+1, \ldots, n+m, \tag{3.95}$$

where $\underline{x}_i$ is the value of the actual observation rounded down to the nearest integer while $\overline{x}_i$ is its value rounded up to the nearest integer.

Table 3.10 and Table 3.11 report the Diebold-Mariano test statistic which compares the ZIACD model with models based on continuous distributions. For most stocks, the values are positive and quite high indicating the ZIACD model produces more precise forecasts. For the GE stock, the test statistic indicates similar performance of the ZIACD model with models discarding close-to-zero values based on the gamma and generalized gamma distributions. For the DWDP, MRK and PFE stocks, the test statistic indicates similar performance of the ZIACD model with models truncating close-to-zero values based on the gamma and generalized gamma distributions. Overall, the results imply that the loss of decimal places in the discrete approach is of less importance than the loss of close-to-zero values in the continuous approach. With regard to continuous distributions, the results do not clearly show which zero treatment is the best in terms of predictive accuracy. When truncating close-to-zero values in frequently traded stocks, however, the estimation does not converge as previously discussed. Table 3.7 and Figure 3.6 show us the shortcomings of both zero treatments. Discarding close-to-zero values leads to underestimation of zero values while truncating them results in overestimation. In both cases, the distributions are significantly distorted.

3.3.3 Discussion

In an empirical study, we analyze 30 stocks that form Dow Jones Industrial Average index with values of trade durations rounded down to seconds. We compare the Poisson, geometric and negative binomial distributions together with their zero-inflated modifications. We find that the proposed ZIACD model is a good fit as it captures both overdispersion and excessive zero values. We argue that zero or close-to-zero durations should not be removed from data as they contain important information and their removal distorts the estimated distribution. This is because only part of them is actually caused by split transactions

	Close-to-Zero Values Discarded			
Stock	ZINB/E	ZINB/W	ZINB/G	ZINB/GG
AAPL	106.9908	67.8638	52.8650	60.0127
AXP	73.2643	17.9373	15.4072	16.6935
BA	140.3922	39.4326	50.5946	50.5621
CAT	114.6549	32.9050	24.4430	23.5386
CSCO	129.4904	81.3149	107.4502	118.5973
CVX	109.4995	14.2293	22.3204	35.2534
DIS	96.5620	12.1016	11.9358	17.1803
DWDP	111.6567	29.2596	38.8707	41.7382
GE	66.1663	10.9305	-1.3031	-1.4056
GS	102.1125	31.3903	22.3592	22.1285
HD	93.6815	15.4376	17.9416	17.7019
IBM	64.6579	26.8612	5.5353	5.7665
INTC	117.5481	74.0000	90.9294	105.2006
JNJ	66.9729	18.5150	15.9069	13.7316
JPM	79.1881	22.9344	1.4320	12.6533
KO	101.5418	21.3555	22.6800	20.5254
MCD	75.8810	22.3171	7.9754	7.8753
MMM	79.3406	17.1030	13.7454	13.1427
MRK	103.9809	10.3362	21.5666	15.5205
MSFT	112.6267	69.5370	78.2352	86.7501
NKE	80.1893	24.5685	19.0817	23.7879
PFE	117.6369	5.2013	22.2919	18.7775
PG	90.6864	10.4339	8.5639	8.2226
TRV	60.9655	17.5976	9.2966	9.5009
UNH	92.3429	28.6549	20.5646	23.0692
UTX	70.2034	11.6773	7.9131	9.1428
V	86.0509	29.1858	14.2810	21.9443
VZ	102.1814	10.9463	18.2628	17.8147
WMT	71.2637	16.7298	4.0405	12.2646
XOM	117.5449	6.0814	15.2313	11.3587

Table 3.10: Out-of-sample Diebold-Mariano test statistic comparing duration model based on the zero-inflated negative binomial distribution (ZINB) with duration models based on the exponential (E), Weibull (W), gamma (G) and generalized gamma (GG) distributions with close-to-zero values discarded.

	Close-to-Zero Values Truncated			
Stock	ZINB/E	ZINB/W	ZINB/G	ZINB/GG
AAPL	-	-	-	-
AXP	73.1534	27.3467	9.4732	8.1273
BA	133.6607	114.4500	59.6291	38.0899
CAT	109.7037	56.2401	14.2042	8.8725
CSCO	-	-	-	-
CVX	108.1999	21.3970	7.1023	19.5490
DIS	94.7389	42.1165	17.4469	15.8197
DWDP	112.3021	27.3929	1.6377	1.7461
GE	68.3601	50.8589	16.9671	54.1558
GS	100.4230	59.2803	11.0896	83.9239
HD	92.7702	31.8002	31.1084	30.3689
IBM	65.1237	71.4568	29.2430	20.5622
INTC	-	-	-	-
JNJ	66.4611	65.0253	24.9040	23.5725
JPM	80.5696	84.5780	37.5159	30.4522
KO	102.3197	44.7889	10.8268	8.4080
MCD	71.3971	48.3043	18.5807	10.0202
MMM	81.3998	28.1204	13.8402	12.8514
MRK	103.4981	21.2388	-1.4653	-1.3740
MSFT	-	-	-	-
NKE	88.6302	52.3342	12.6283	16.2392
PFE	114.1245	14.8362	-1.1426	0.4174
PG	89.8912	31.9824	18.6263	18.9892
TRV	60.9610	33.5640	18.0302	12.4438
UNH	86.6440	40.1384	13.9099	8.8888
UTX	68.0445	34.1515	17.9190	18.3269
V	82.1849	65.5784	22.2183	18.2378
VZ	100.3199	27.4085	4.9086	5.0960
WMT	75.9355	61.6785	19.6004	12.4074
XOM	111.3729	28.1351	8.6675	9.2472

Table 3.11: Out-of-sample Diebold-Mariano test statistic comparing duration model based on the zero-inflated negative binomial distribution (ZINB) with duration models based on the exponential (E), Weibull (W), gamma (G) and generalized gamma (GG) distributions with close-to-zero values truncated.

while the rest is due to execution of independent transactions at similar times. The portion of zeros caused by split transactions ranges from 37% up to 90% depending on the stock with the average of 63%.

We also compare the proposed ZIACD model with the commonly used continuous models based on the exponential, Weibull, gamma and generalized gamma distributions. In a simulation study, we find that when data are rounded, the estimates of the continuous model are biased while the proper use of the discrete model identifies true parameters. Further, we resume with the empirical study. Our original duration data have very high precision and as we round them to seconds for the discrete model, we lose some information. Continuous approach, however, also causes a loss of information as close-to-zero durations need to be removed or set to a given threshold value for estimation purposes. We find that the loss of decimals is less severe than the loss of zeros and the proposed ZIACD model outperforms considered continuous models in terms of predictive accuracy.

Our proposed model can be utilized in a joint modeling of prices and durations. It also allows to study the trading process from the market microstructure perspective.

- Chapter 4 -

Quadratic Variation

The focus of financial time series analysis is the price process. Cont (2001) presents several stylized empirical facts regarding prices in various financial markets. One of these facts is the absence of linear autocorrelations in returns. Another fact is clustering of volatility. For these two reasons, financial analysis focuses more on the second moment than the first moment.

Quadratic variation is a measure of volatility over a given time interval. It is a common tool in analysis of stochastic processes, especially suitable for high-frequency data (Aït-Sahalia and Jacod, 2014). In financial econometrics, there are two topics regarding quadratic variation – ex-post estimation and ex-ante forecasts. In this chapter, we present the highlights of both worlds. We also compare presented methods in simulation and empirical studies.

First, we deal with ex-post estimation of quadratic variation. Quadratic variation is an unobserved quantity because of two reasons. We do not observe the price process in continuous time, but only in a finite number of times. We also observe the price process contaminated by the market microstructure noise. When the noise is not present, quadratic variation can be straightforwadly estimated by the realized variance (Barndorff-Nielsen and Shephard, 2002b). When the noise is present, however, the realized variance is significantly biased and inconsistent. In the case of the white noise, it linearly diverges to infinity, while in the case of time-dependent and cross-dependent noise, it can have more complex bias (Hansen and Lunde, 2006). Luckily, many alternative estimators of quadratic variation have been proposed in the literature. These include the two-scale estimator of Zhang et al. (2005), realized kernel estimator of Barndorff-Nielsen et al. (2008), pre-averaging estimator of Jacod et al. (2009) and least squares estimator of Nolte and Voev (2012). All of these estimators can be conveniently expressed as a quadratic form (Sun, 2006; Andersen et al., 2011).

Second, we deal with ex-ante forecasts of quadratic variation. Once we know how to estimate historical quadratic variation, the next question is how to forecast its future values. Traditional time series models can be utilized, see e.g. Andersen et al. (2003) and Aït-Sahalia and Mancini (2008). Models specifically designed for high-frequency data include the HAR model of Corsi (2009) and realized GARCH model of Hansen et al. (2012).

We also approach quadratic variation from a perspective of interval uncertainty. It is natural to consider prices to be observed as intervals due to discretness of price values and bid-ask spread. However, we show that the lack of any further assumptions makes quadratic variation under this setting unidentifiable. Only some information about volatility of large jumps can be uncovered. Nevertheless, these results are important as they show necessity of more strict assumptions.

4.1 Theory of Quadratic Variation

Our main focus is the quadratic variation of a process. Without loss of generality, we limit ourselves to the time interval $[0, 1]$. Let us consider a sampling of the process P_t at discrete times $0 = T_0 < T_1 < \ldots < T_n = 1$. The *quadratic variation* of the process P_t is then given by

$$QV = \operatorname*{plim}_{\Delta_n \to 0} \sum_{i=1}^{n} \left(P_{T_i} - P_{T_{i-1}} \right)^2, \tag{4.1}$$

where plim denotes the limit in probability and $\Delta_n = \max\{T_1 - T_0, T_2 - T_1, \ldots, T_n - T_{n-1}\}$ is the maximal lag between the observations.

Quadratic variation is not the only variation of the process. The *absolute variation* of the process P_t is defined as

$$V = \operatorname*{plim}_{\Delta_n \to 0} \sum_{i=1}^{n} \left| P_{T_i} - P_{T_{i-1}} \right|, \tag{4.2}$$

where $\Delta_n = \max\{T_1 - T_0, T_2 - T_1, \ldots, T_n - T_{n-1}\}$ is the maximal lag between the observations. It can be shown that for the process with finite absolute variation, the quadratic variation exists and is equal to zero. It can also be shown that the process with positive quadratic variation has infinite absolute variation. See the literature listed in Section 2.2.2 for more details.

We present two frameworks for the quadratic variation. First, we briefly present the widespread approach of the semimartingale theory and stochastic calculus. Second, we introduce the interval approach based on the interval uncertainty.

4.1.1 Stochastic Calculus Approach

We build on the efficient price model presented in Section 2.2.2. Let us remind the expression for semimartingale (2.10) with the following decomposition

$$P_t = P_0 + \underbrace{\int_0^t D_z \, \mathrm{d}z}_{P_t^D} + \underbrace{\int_0^t V_z \, \mathrm{d}W_z}_{P_t^V} + \underbrace{\sum_{k : S_k \leq t} J_k}_{P_t^J}, \tag{4.3}$$

where D_z is a finite variation càdlàg drift process, V_z is an adapted càdlàg volatility process, W_z is a standard Wiener process and J_k are non-zero random variables with random times S_k.

An important result of stochastic calculus is that quadratic variation exists for every semimartingale. Quadratic variation plays a major part in stochastic calculus as it appears in the integration by parts formula and the stochastic change of variables formula known as Ito's lemma. See the literature listed in Section 2.2.2 for more details.

Integrated Variance and Jump Variance

Let us define two following volatility measures based on the semimartingale (4.3). The *integrated variance* is given by

$$IV = \int_0^t V_z^2 \, \mathrm{d}z. \tag{4.4}$$

The *jump variance* is given by

$$JV = \sum_{k : S_k \leq t} J_k^2. \tag{4.5}$$

Both of these measures are related to quadratic variation as we show bellow.

Let us consider decomposition (4.3). Quadratic variation for the drift component P_t^D is zero. For the volatility component P_t^V, it is equal to the integrated variance IV. Finally, for the jump component P_t^J, it is equal to the jump variance JV. This means that, for a continuous semimartingale without jumps (2.11), quadratic variation is equal to the integrated variance, i.e. $QV = IV$. For a general semimartingale (4.3), we have $QV = IV + JV$. Again, see the literature listed in Section 2.2.2 for more details

A portion of the high-frequency literature is devoted to estimation of integrated variance rather then quadratic variation as it may be suitable to filter out jumps in prices. In this chapter, however, we mostly focus on the case of continuous semimartingale (2.11), in which the quadratic variation is equal to the integrated variance.

Integrated Power Variation

The integrated variance can be generalized to higher powers. The *integrated power variation* of order p is defined as

$$IPV^{(p)} = \int_0^t V_z^p \, \mathrm{d}z. \tag{4.6}$$

Notably, the fourth power of volatility is called the *integrated quarticity* and is given by

$$IQ = IPV^{(4)} = \int_0^t V_z^4 \, \mathrm{d}z. \tag{4.7}$$

4.1.2 Interval Approach

This section follows Holý and Sokol (2018) with a slightly different notation. We build on the interval model presented in Section 2.2.5. First, we decompose the quadratic variation into the continuous part and the finite-jump part. Next, we introduce the interval quadratic variation and show some of its properties. Finally, we briefly illustrate one possible use of the interval quadratic variation in a simulation study.

Decomposition of Quadratic Variation

Let us consider the proccess P_t with continuous time $t \geq 0$. We assume this process can be decomposed into the *continuous component* P_t^C with continuous path and the *jump component* P_t^J with a finite number of discrete jumps, i.e. $P_t = P_t^C + P_t^J$. Formally, the jump component is defined as

$$P_t^J = \sum_{i\,:\,0 \leq S_i \leq t} J_i, \qquad t \geq 0, \tag{4.8}$$

where J_i are non-zero random variables with non-equal random times S_i. We define the *continuous variance* CV and the *jump variance* JV as the quadratic variation for the continuous component P_t^C and jump component P_t^J respectively.

Proposition 4.1. *Quadratic variation can be decomposed into* $QV = CV + JV$.

Proof. We decompose quadratic variation as

$$\begin{aligned} QV &= \operatorname*{plim}_{\Delta^n \to 0} \sum_{i=1}^{n} \left(P_{T_i}^C - P_{T_{i-1}}^C + P_{T_i}^J - P_{T_{i-1}}^J \right)^2 \\ &= \operatorname*{plim}_{\Delta^n \to 0} \left(\sum_{i=1}^{n} \left(P_{T_i}^C - P_{T_{i-1}}^C \right)^2 + \sum_{i=1}^{n} \left(P_{T_i}^J + P_{T_{i-1}}^J \right)^2 + 2 \sum_{i=1}^{n} \left(P_{T_i}^C - P_{T_{i-1}}^C \right) \left(P_{T_i}^J - P_{T_{i-1}}^J \right) \right) \\ &= CV + JV + 2 \operatorname*{plim}_{\Delta^n \to 0} \sum_{i=1}^{n} \left(P_{T_i}^C - P_{T_{i-1}}^C \right) \left(P_{T_i}^J - P_{T_{i-1}}^J \right). \end{aligned} \tag{4.9}$$

As $P^J_{T_i} - P^J_{T_{i-1}}$ is non-zero only in a finite number of cases, we have

$$\operatorname*{plim}_{\Delta^n\to 0} \sum_{i=1}^{n} \left(P^C_{T_i} - P^C_{T_{i-1}}\right)\left(P^J_{T_i} - P^J_{T_{i-1}}\right) = \sum_{i:0\le S_i\le 1} \operatorname*{plim}_{\Delta^n\to 0}\left(P^C_{S_i} - P^C_{S_i-\Delta^n}\right)\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right). \tag{4.10}$$

Because

$$\operatorname*{plim}_{\Delta^n\to 0}\left(P^C_{S_i} - P^C_{S_i-\Delta^n}\right) = 0 \quad \text{and} \quad \operatorname*{plim}_{\Delta^n\to 0}\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right) = J_i, \tag{4.11}$$

we have

$$\operatorname*{plim}_{\Delta^n\to 0}\left(P^C_{S_i} - P^C_{S_i-\Delta^n}\right)\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right) = 0, \tag{4.12}$$

and therefore,

$$\sum_{i:0\le S_i\le 1} \operatorname*{plim}_{\Delta^n\to 0}\left(P^C_{S_i} - P^C_{S_i-\Delta^n}\right)\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right) = 0. \tag{4.13}$$

□

Proposition 4.2. *Quadratic variation for the jump component is* $JV = \sum_{i:0\le S_i\le 1} J_i^2$.

Proof. As $P^J_{T_i} - P^J_{T_{i-1}}$ is non-zero only in a finite number of cases, we have

$$JV = \operatorname*{plim}_{\Delta^n\to 0} \sum_{i=1}^{n} \left(P^J_{T_i} - P^J_{T_{i-1}}\right)^2 = \sum_{i:0\le S_i\le 1} \operatorname*{plim}_{\Delta^n\to 0}\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right)^2. \tag{4.14}$$

Because

$$\operatorname*{plim}_{\Delta^n\to 0}\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right) = J_i, \tag{4.15}$$

we have

$$\operatorname*{plim}_{\Delta^n\to 0}\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right)^2 = J_i^2, \tag{4.16}$$

and therefore,

$$\sum_{i:0\le S_i\le 1} \operatorname*{plim}_{\Delta^n\to 0}\left(P^J_{S_i} - P^J_{S_i-\Delta^n}\right)^2 = \sum_{i:0\le S_i\le 1} J_i^2. \tag{4.17}$$

□

Interval Setup

Let us consider we do not observe the process P_t, $t \ge 0$ but rather a collection of intervals $[\underline{P}_t, \overline{P}_t]$ guaranteeing that $P_t \in [\underline{P}_t, \overline{P}_t]$, $t \ge 0$. This setup, in which the true values are not observable but the bounds are available, is studied in the area of partial identification. Due to our weak assumptions, the only information we can infer about any statistic from the observable intervals $[\underline{P}_t, \overline{P}_t]$, $t \ge 0$ is its lower and upper bound. In the case of quadratic variation, $QV \in [\underline{QV}, \overline{QV}]$, where $\underline{QV}$ and $\overline{QV}$ is the lower and upper bound of the form

$$\begin{aligned}
\underline{QV} &= \operatorname*{plim}_{\Delta^n\to 0} \min\left\{ \sum_{i=1}^{n} \left(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\right)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \right\}, \\
\overline{QV} &= \operatorname*{plim}_{\Delta^n\to 0} \max\left\{ \sum_{i=1}^{n} \left(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\right)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \right\}.
\end{aligned} \tag{4.18}$$

We let $\tilde{P}_t$ denote any possible process satisfying $\underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i}$ for $i = 0,\dots,n$. Bellow, we work with the set of all possible processes $\tilde{P}_t$ rather than the true process P_t as it is unobservable.

We assume a process $\tilde{P}_t$ can be decomposed into the continuous component $\tilde{P}_t^C$ and the jump component $\tilde{P}_t^J$ just like the original process P_t. We further decompose the jump component into the *small-jump component* $\tilde{P}_t^{SJ}$ containing only jumps smaller or equal to a given threshold κ in absolute value and the *large-jump component* $\tilde{P}_t^{LJ}$ containing only jumps larger than $\kappa > 0$ in absolute value, i.e. $\tilde{P}_t^J = \tilde{P}_t^{SJ} + \tilde{P}_t^{LJ}$. Formally, the small-jump component and the large-jump component are defined as

$$\tilde{P}_t^{SJ} = \sum_{i:0\le S_i\le t, |J_i|\le\kappa} \tilde{J}_i, \qquad \tilde{P}_t^{LJ} = \sum_{i:0\le S_i\le t, |\tilde{J}_i|>\kappa} J_i, \qquad t \ge 0, \tag{4.19}$$

where $\tilde{J}_i$ are non-zero random variables with non-equal random times S_i. For a given $\kappa > 0$, the price component can then be decomposed into $\tilde{P}_t = \tilde{P}_t^C + \tilde{P}_t^{SJ} + \tilde{P}_t^{LJ}$.

We define $\underline{CV}$, $\overline{CV}$, $\underline{SJV}$, $\overline{SJV}$, $\underline{LJV}$, $\overline{LJV}$ as the minimal/maximal quadratic variation for processes $\tilde{P}_t^C/\tilde{P}_t^{SJ}/\tilde{P}_t^{LJ}$ over all possible processes $\tilde{P}_t \in [\underline{P}_t, \overline{P}_t]$. For simplification, we assume all intervals $[\underline{P}_t, \overline{P}_t]$ have constant width $\omega = \overline{P}_t - \underline{P}_t$ for all $t \ge 0$. The threshold separating small and large jumps is then set to $\kappa = 2\omega$.

Properties of Quadratic Variation Under Interval Uncertainty

In the following propositions, we investigate properties of quadratic variation under interval uncertainty. We find that quadratic variation is unbounded from above. Quadratic variation for both the continuous component and the small-jump component is also unbounded from above. Quadratic variation for the large-jump component is, however, bounded from above and therefore partially identifiable.

Proposition 4.3. *Quadratic variation is unbounded from above, i.e.* $\overline{QV} = \infty$.

Proof. The upper bound of quadratic variation is given by

$$\overline{QV} = \operatorname*{plim}_{\Delta^n\to 0} \max\Big\{ \sum_{i=1}^n \Big(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\Big)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \Big\}. \tag{4.20}$$

Let us consider process $\hat{P}_{T_i}$ defined for $i = 0$ as $\hat{P}_{T_0} = \underline{P}_{T_0}$ and for $i = 1,\dots,n$ as $\hat{P}_{T_i} = \underline{P}_{T_i}$ if $|\hat{P}_{T_{i-1}} - \underline{P}_{T_i}| > |\hat{P}_{T_{i-1}} - \overline{P}_{T_i}|$ or $\hat{P}_{T_i} = \overline{P}_{T_i}$ else. For any two consecutive values, we have $|\hat{P}_{T_i} - \hat{P}_{T_{i-1}}| \ge \frac{\omega}{2}$. Therefore, we have

$$\frac{\omega^2}{4} n \le \sum_{i=1}^n \Big(\hat{P}_{T_i} - \hat{P}_{T_{i-1}}\Big)^2 \le \max\Big\{ \sum_{i=1}^n \Big(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\Big)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \Big\}. \tag{4.21}$$

For $\Delta^n \to 0$, we have $n \to \infty$ and quadratic variation diverges to infinity. □

Proposition 4.4. *Quadratic variation for the continuous component as well as the small-jump component is unbounded from above, i.e.* $\overline{CV} = \infty$ *and* $\overline{SJV} = \infty$ *respectively.*

Proof. The upper bound of quadratic variation for the continuous component is given by

$$\overline{CV} = \operatorname*{plim}_{\Delta^n\to 0} \max\Big\{ \sum_{i=1}^n \Big(\tilde{P}^C_{T_i} - \tilde{P}^C_{T_{i-1}}\Big)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \Big\}. \tag{4.22}$$

while the upper bound of quadratic variation for the small-jump component is given by

$$\overline{SJV} = \operatorname*{plim}_{\Delta^n\to 0} \max\Big\{ \sum_{i=1}^n \Big(\tilde{P}^{SJ}_{T_i} - \tilde{P}^{SJ}_{T_{i-1}}\Big)^2 : \ \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i = 0,\dots,n \Big\}, \tag{4.23}$$

where $\tilde{P}_t^{SJ} = \sum_{j:0\le S_j^S\le t,|J_j^S|>2\omega} J_j^S$, $t \ge 0$ is the small-jump component with jumps $\tilde{J}_j^S$ and jump times $\tilde{S}_j^S$ corresponding to a process $\tilde{P}_t$. As there are only finite number of values that can be attributed to jumps, a non-trivial time interval $[a,b] \subset [0,1]$ with $\min\{\operatorname*{plim}_{\Delta^n\to 0} |\overline{P}_t - \underline{P}_{t-\Delta^n}|, \operatorname*{plim}_{\Delta^n\to 0} |\underline{P}_t - \overline{P}_{t-\Delta^n}|\} \le 2\omega$ for each $t \in (a,b]$ exists. Let $\overline{QV}_{[a,b]}$ denote the upper bound of quadratic variation on interval $[a,b]$. This interval cannot contain any large jumps, i.e. $\overline{QV}^{LJ}_{[a,b]} = 0$, and the upper bound of quadratic variation for continuous component is indistinguishable from the small-jump component, i.e. $\overline{QV}_{[a,b]} = \overline{QV}^{C}_{[a,b]} = \overline{QV}^{SJ}_{[a,b]}$. From Theorem 4.3, we have $\overline{QV}_{[a,b]} = \infty$. Finally, we have $\overline{CV} \ge \overline{QV}^{C}_{[a,b]} = \infty$ and $\overline{SJV} \ge \overline{QV}^{SJ}_{[a,b]} = \infty$. □

Proposition 4.5. *The lower and upper bounds of quadratic variation for the large-jump component are finite and respectively given by*

$$\begin{aligned}\underline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \sum_{i=1}^{n} \underline{\chi}^2_{T_i,T_{i-1}} \mathbb{I}_{\{\underline{\chi}_{T_i,T_{i-1}}>2\omega\}},\\ \overline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \sum_{i=1}^{n} \overline{\chi}^2_{T_i,T_{i-1}} \mathbb{I}_{\{\overline{\chi}_{T_i,T_{i-1}}>2\omega\}},\end{aligned} \tag{4.24}$$

where $\mathbb{I}$ *denotes the indicator function and*

$$\begin{aligned}\underline{\chi}_{T_i,T_{i-1}} &= \min\{|\overline{P}_{T_i} - \underline{P}_{T_{i-1}}|, |\underline{P}_{T_i} - \overline{P}_{T_{i-1}}|\},\\ \overline{\chi}_{T_i,T_{i-1}} &= \max\{|\overline{P}_{T_i} - \underline{P}_{T_{i-1}}|, |\underline{P}_{T_i} - \overline{P}_{T_{i-1}}|\}.\end{aligned} \tag{4.25}$$

Proof. The lower and upper bounds of quadratic variation for the large-jump component are defined as

$$\begin{aligned}\underline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \min\Big\{ \sum_{i=1}^{n} \Big(\tilde{P}^{LJ}_{T_i} - \tilde{P}^{LJ}_{T_{i-1}}\Big)^2 \,:\, \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i=0,\dots,n\Big\},\\ \overline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \max\Big\{ \sum_{i=1}^{n} \Big(\tilde{P}^{LJ}_{T_i} - \tilde{P}^{LJ}_{T_{i-1}}\Big)^2 \,:\, \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i=0,\dots,n\Big\},\end{aligned} \tag{4.26}$$

where $\tilde{P}_t^{LJ} = \sum_{j:0\le S_j^L\le t,|J_j^L|>2\omega} J_j^L$, $t \ge 0$ is the large-jump component with jumps $\tilde{J}_j^L$ and jump times $\tilde{S}_j^L$ corresponding to a process $\tilde{P}_t$. An absolute difference between two consecutive values $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}^{LJ}_{T_i} - \tilde{P}^{LJ}_{T_i-\Delta^n}|$ can either be $|\tilde{J}_j^L| > 2\omega$ or zero from the definition of the large-jump component. If a large jump $\tilde{J}_j^L$ occurs at time T_i, we have $2\omega < \operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_i-\Delta^n}| = \operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}^{LJ}_{T_i} - \tilde{P}^{LJ}_{T_i-\Delta^n}| = |\tilde{J}_j^L|$ as the absolute difference between two consecutive values in limit is zero for both the continuous and small-jump component. Therefore, we can ignore $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_i-\Delta^n}| \le 2\omega$ as they do not correspond to a large jump. On the other hand, $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_i-\Delta^n}| > 2\omega$ may correspond to a large jump and must be included. We can replace $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}^{LJ}_{T_i} - \tilde{P}^{LJ}_{T_i-\Delta^n}|$ with $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_i-\Delta^n}|$ as we have a finite number of $\operatorname*{plim}_{\Delta^n\to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_i-\Delta^n}| > 2\omega$. The lower and upper bounds are then given by

$$\begin{aligned}\underline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \min\Big\{ \sum_{i=1}^{n} \Big(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\Big)^2 \mathbb{I}_{\{|\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}|>2\omega\}} \,:\, \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i=0,\dots,n\Big\},\\ \overline{LJV} &= \operatorname*{plim}_{\Delta^n\to 0} \max\Big\{ \sum_{i=1}^{n} \Big(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}\Big)^2 \mathbb{I}_{\{|\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}|>2\omega\}} \,:\, \underline{P}_{T_i} \le \tilde{P}_{T_i} \le \overline{P}_{T_i},\ i=0,\dots,n\Big\}.\end{aligned} \tag{4.27}$$

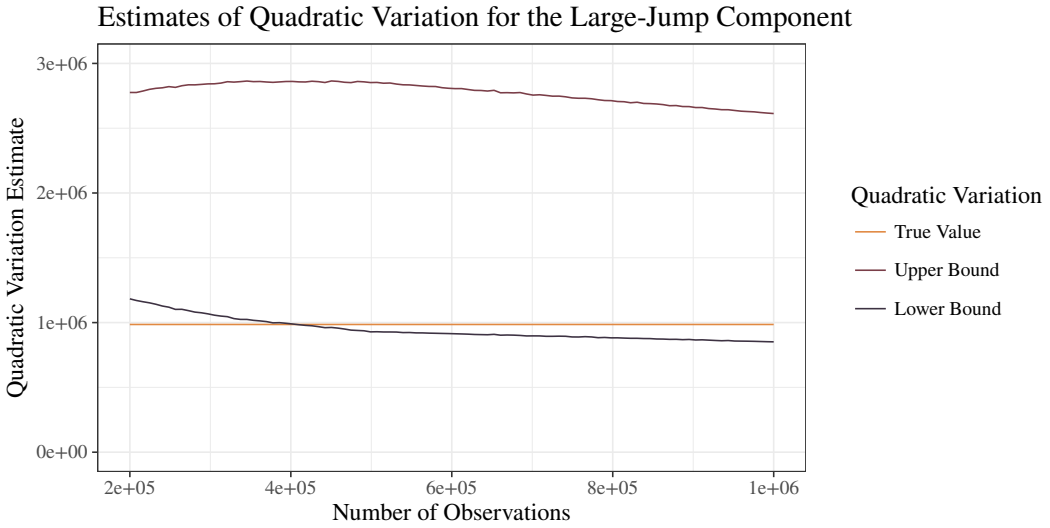

Figure 4.1: A simulation of lower and upper bounds of quadratic variation for the large-jump component.

As two jumps cannot occur at the same time, we have

$$\begin{aligned}\underline{LJV} &= \operatorname*{plim}_{\Delta^n \to 0} \sum_{i=1}^{n} \min\left\{ \left(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}} \right)^2 \mathbb{I}_{\{|\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}| > 2\omega\}} \ : \ \underline{P}_{T_i} \leq \tilde{P}_{T_i} \leq \overline{P}_{T_i} \right\}, \\ \overline{LJV} &= \operatorname*{plim}_{\Delta^n \to 0} \sum_{i=1}^{n} \max\left\{ \left(\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}} \right)^2 \mathbb{I}_{\{|\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}| > 2\omega\}} \ : \ \underline{P}_{T_i} \leq \tilde{P}_{T_i} \leq \overline{P}_{T_i} \right\}.\end{aligned} \tag{4.28}$$

Finally, we can restrict the minimization and maximization to extreme points and get the desired expressions with $\underline{\chi}_{T_i,T_{i-1}}$ and $\overline{\chi}_{T_i,T_{i-1}}$ respectively. Both bounds are finite because $\operatorname*{plim}_{\Delta^n \to 0} |\tilde{P}_{T_i} - \tilde{P}_{T_{i-1}}| > 2\omega$ can occur only at finite number of times. □

Simulation of Interval Quadratic Variation

In a simulation study, we illustrate the finite-sample properties of estimation of quadratic variation for the large-jump component. We simulate 1 000 000 observations as the sum of the continuous and jump components. The continuous component is simulated as the Wiener process with zero mean and unit standard deviation. The jump component contains 10 000 jumps with values generated from the normal distribution with zero mean and standard deviation equal to 10 and with times generated according to the exponential distribution. The lower and upper bounds of observed intervals are given by rounding down to the nearest integer and rounding up to the nearest integer respectively.

The finite-sample counterpart of quadratic variation is called the realized variance. Simulated interval estimates of quadratic variation (i.e. interval realized variances) for various numbers of observations are presented in Figure 4.1. We can see that for a smaller number of observations, the interval estimate is not precise as it omits some jumps. For a larger number of observations, the interval estimate converges and the bounds contain the true value of quadratic variation for the large-jump component.

4.2 Estimators of Quadratic Variation

We follow the framework presented in Section 2.2.3 and consider the price process to be given by the additive model $X_i = P_{T_i} + E_i$, $i = 0, \ldots, n$ observed at times $0 = T_0 < T_1 < \cdots < T_n = 1$. Unless otherwise stated, we assume the price process P_{T_i} to follow a continuous semimartingale (2.11) through

this section. In this case, the quadratic variation QV is equal to the integrated variance IV. We also assume that we measure volatility of the price process on interval $[0, 1]$.

We estimate quadratic variation QV by non-parametric methods within a unified framework based on a quadratic form. The class of *qudratic estimators* was introduced by Sun (2006) and independently by Andersen et al. (2011). Estimators in this class can be formulated as a *quadratic form*

$$QE = Y'WY = \sum_{i=1}^{n}\sum_{j=1}^{n} Y_i w_{i,j} Y_j, \tag{4.29}$$

where $Y = (Y_1, \dots, Y_n)' = (X_1 - X_0, \dots, X_n - X_{n-1})'$ is a vector of returns and $W = (w_{i,j})_{i=1,j=1}^{n,n}$ is a matrix of weights for returns determining an estimator. It can also be rewritten using the actual prices $X = (X_0, X_1, \dots X_n)'$ instead of returns as

$$QE = Y'WY = X'U'WUX = X'VX = \sum_{i=0}^{n}\sum_{j=0}^{n} X_i v_{i,j} X_j, \tag{4.30}$$

where $V = (v_{i,j})_{i=0,j=0}^{n,n}$ is a matrix of weights for prices determining an estimator based on the matrix W and matrix $U = (u_{i,j})_{i=1,j=0}^{n,n}$ with elements

$$u_{i,j} = \begin{cases} 1 & \text{for } j = i+1, \\ -1 & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.31}$$

The class of quadratic estimators includes the realized variance as well as two-scale estimator of Zhang et al. (2005), realized kernel estimator of Barndorff-Nielsen et al. (2008), pre-averaging estimator of Jacod et al. (2009) and least squares estimator of Nolte and Voev (2012).

4.2.1 Realized Variance

A natural estimator of quadratic variation is the *realized variance* defined as

$$RV_n = \sum_{i=1}^{n} \left(X_i - X_{i-1}\right)^2 = \sum_{i=1}^{n} Y_i^2. \tag{4.32}$$

It is simply the sum of squared returns and finite-sample version of quadratic variation. As the returns are random variables, realized variance is also a random variable.

Under the very general assumption of semimartingale (2.10), the realized variance converges to quadratic variation in probability

$$\lim_{n\to\infty} \mathrm{P}\left[|RV_n - QV| > \varepsilon\right] = 0 \qquad \forall\, \varepsilon > 0. \tag{4.33}$$

However, as Barndorff-Nielsen and Shephard (2002b) note, this result lacks a theory of measurement error. They also argue that for a stronger result, additional assumptions are needed.

Next, consider a continuous semimartingale (2.11). Recall that quadratic variation QV is equal to integrated variance IV in this case. Barndorff-Nielsen and Shephard (2002a) show that realized variance converges to integrated variance at rate $\sqrt{n}$. They also derive asymptotic distribution of the estimator

$$\frac{RV_n - IV}{\sqrt{\frac{2}{3}\sum_{i=1}^{n} Y_i^4}} \to \mathrm{N}(0, 1). \tag{4.34}$$

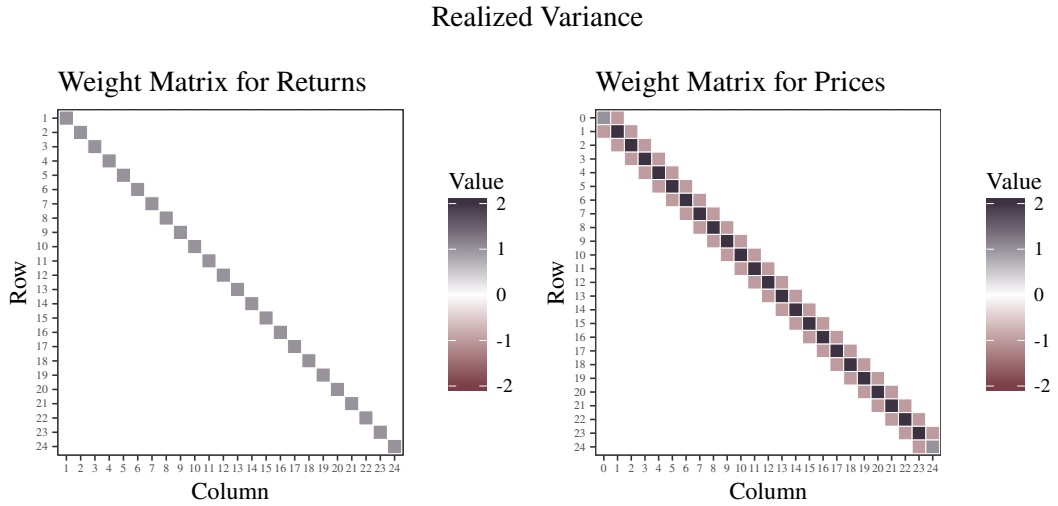

Figure 4.2: Quadratic form of realized variance with $n = 24$.

Bias of Realized Variance

The situation complicates as we contaminate the price process by the market microstructure noise. We consider the additive noise model discussed in Section 2.2.3. Specifically, let the observed price follow the process $X_i = P_{T_i} + E_i$, $i = 0, \dots, n$, where P_{T_i} is the semimartingale efficient price and E_i is the noise with zero mean and variance ω^2.

In this setting, realized variance is biased and inconsistent estimator of quadratic variation. Let us decompose realized variance as

$$\begin{aligned} RV_n &= \sum_{i=1}^{n} \left(X_i - X_{i-1}\right)^2 \\ &= \sum_{i=1}^{n} \left(P_{T_i} - P_{T_{i-1}} + E_i - E_{i-1}\right)^2 \\ &= \sum_{i=1}^{n} \left(P_{T_i} - P_{T_{i-1}}\right)^2 + 2\sum_{i=1}^{n} \left(P_{T_i} - P_{T_{i-1}}\right)\left(E_i - E_{i-1}\right) + \sum_{i=1}^{n} \left(E_i - E_{i-1}\right)^2 \\ &= \sum_{i=1}^{n} R_i^2 + 2\sum_{i=1}^{n} R_i F_i + \sum_{i=1}^{n} F_i^2. \end{aligned} \tag{4.35}$$

Following Hansen and Lunde (2006), we investigate the bias of realized variance under various noise settings. First, let us assume that the market microstructure noise E_i is weakly stationary with autocovariance function $\pi(s)$ and variance $\omega^2 = \pi(0)$. The bias of realized variance is then

$$\mathrm{E}[RV_n - QV] = 2\sum_{i=1}^{n} \mathrm{E}[R_i F_i] + 2n\left(\omega^2 - \pi\left(T_i - T_{i-1}\right)\right). \tag{4.36}$$

The second bias term $2n\left(\omega^2 - \pi\left(T_i - T_{i-1}\right)\right)$ is always non-negative, while the first term $2\sum_{i=1}^{n} \mathrm{E}[R_i F_i]$ can have any value. This means that the bias can be either positive or negative. As Hansen and Lunde (2006) argue, the negative bias is possible only if the innovations in the noise process F_i are negatively correlated with the returns R_i, i.e. $\mathrm{E}[R_i F_i] < 0$.

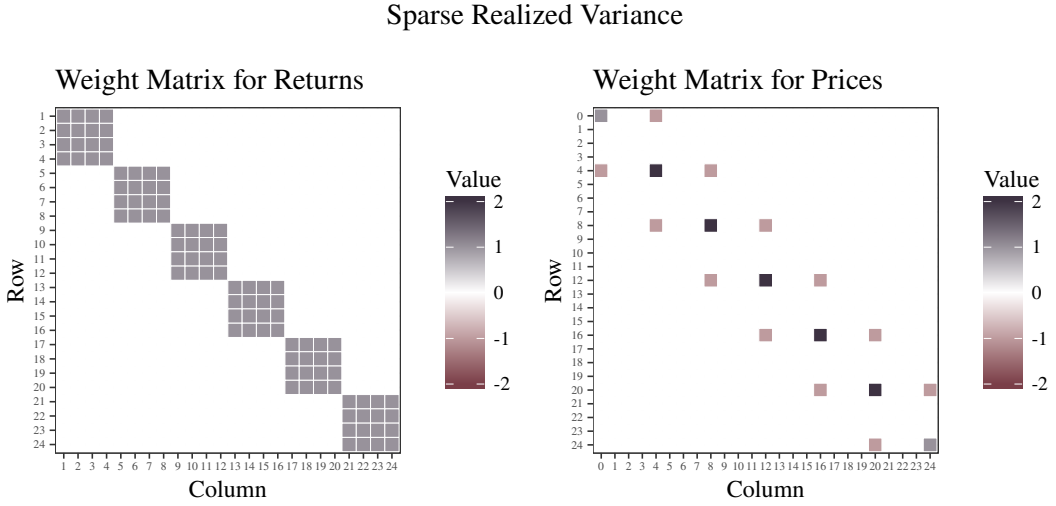

Figure 4.3: Quadratic form of sparse realized variance with $n = 24$, $h = 1$ and $s = 4$.

Next, let us consider the market microstructure noise to be the white noise with variance ω^2 independent of the price process. The bias of realized variance is then

$$\mathrm{E}[RV_n - QV] = 2n\omega^2. \tag{4.37}$$

In other words, the bias is positive and realized variance linearly diverges to infinity with increasing number of observations n.

Sparse Realized Variance

The bias of realized variance can be reduced by sampling at lower frequencies. However, this is at cost of data loss. This approach is called the *sparse realized variance*. Let h denote the initial observation and s denote the sampling interval for ticks. For example $h = 2$ and $s = 3$ would correspond to observations at times $\{T_2, T_5, T_8, T_{11}, \dots\}$. The number of used observations is then

$$m(n, h, s) = \left\lfloor \frac{n - h}{s} \right\rfloor, \tag{4.38}$$

where $\lfloor \cdot \rfloor$ denotes rounding down. The sparse realized variance is then defined as

$$SRV_{n,h,s} = \sum_{i=1}^{m(n,h,s)} \left(X_{is+h} - X_{(i-1)s+h} \right)^2 = \sum_{i=1}^{m(n,h,s)} Y^2_{(i-1)s+h,is+h}. \tag{4.39}$$

The optimal sampling frequency of the realized variance was studied by Aït-Sahalia et al. (2005), Zhang et al. (2005), Bandi and Russell (2006, 2008) and De Pooter et al. (2008).

Average Realized Variance

Sparse realized variance uses only a fraction of available observations. To fully utilize all data, the *average realized variance* can be adopted (Zhang et al., 2005). It averages sparse realized variances over subgrids given by different initial observations h. For a given sampling interval s, it is defined as

$$ARV_{n,s} = \frac{1}{s} \sum_{h=1}^{s} SRV_{n,h,s}, \tag{4.40}$$

where $SRV_{n,s}$ is given by (4.39). Although this approach reduces the impact of the noise, the average realized variance is still a biased estimator of the quadratic variation.

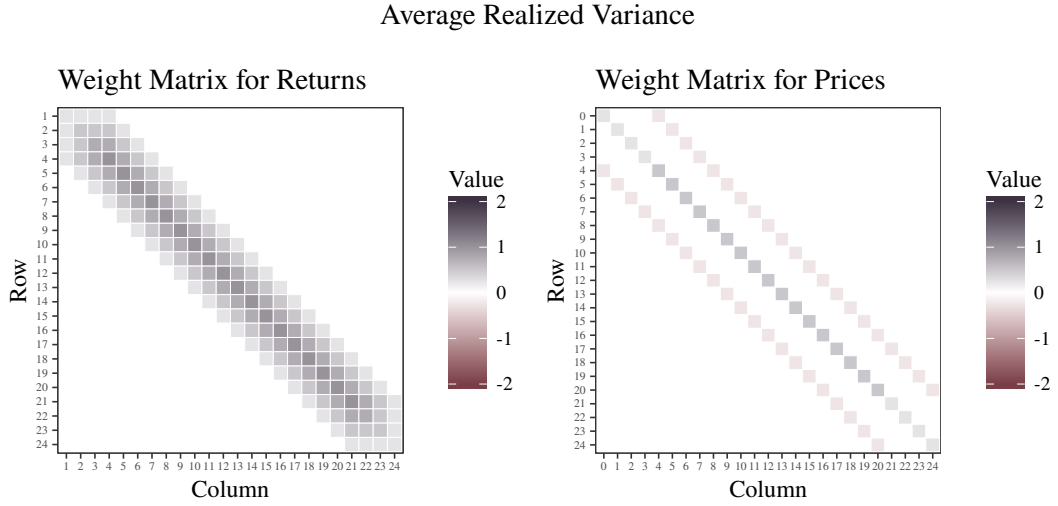

Figure 4.4: Quadratic form of average realized variance with $n = 24$ and $s = 4$.

Quadratic Form

This section follows Holý (2017e). The realized variance (4.32) can easily be expressed as a quadratic estimator using the weight matrix W_n^{RV} given by elements

$$w_{i,j}^{RV} = \begin{cases} 1 & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.41}$$

An example of this weight matrix is shown in Figure 4.2.

The sparse realized variance (4.39) with initial observation h and sampling interval s can be expressed as a quadratic estimator using the weight matrix $W_{n,h,s}^{SRV}$ given by elements

$$w_{i,j}^{SRV} = \begin{cases} 1 & \text{for } (k-1)s + h \le i, \quad j \le ks + h - 1, \quad k = 1, \dots, m(n,h,s), \\ 0 & \text{otherwise.} \end{cases} \tag{4.42}$$

It is visualized in Figure 4.3.

The average realized variance (4.40) with sampling interval s can be expressed as a quadratic estimator using the weight matrix

$$W_{n,s}^{ARV} = \frac{1}{s} \sum_{h=1}^{s} W_{n,h,s}^{SRV}, \tag{4.43}$$

where $W_{n,h,s}^{SRV}$ is given by (4.42). This weight matrix is shown in Figure 4.4.

Simulation of the Impact of Market Microstructure Noise

This section loosely follows Holý (2016). We illustrate the bias of realized variance under various noise settings using simulations. We consider the following model for simulations. The number of observations is set to $n = 23\,400$. The times of observations T_i, $i = 0, \dots, n$ are generated by the Poisson point process and normalized to interval $[0, 1]$. The observed price is given by the additive model $X_i = P_{T_i} + E_i$, $i = 0, \dots, n$. The efficient price P_t follows the Wiener process

$$\mathrm{d}P_t = \sigma \mathrm{d}W_t, \qquad t \in [0, 1]. \tag{4.44}$$

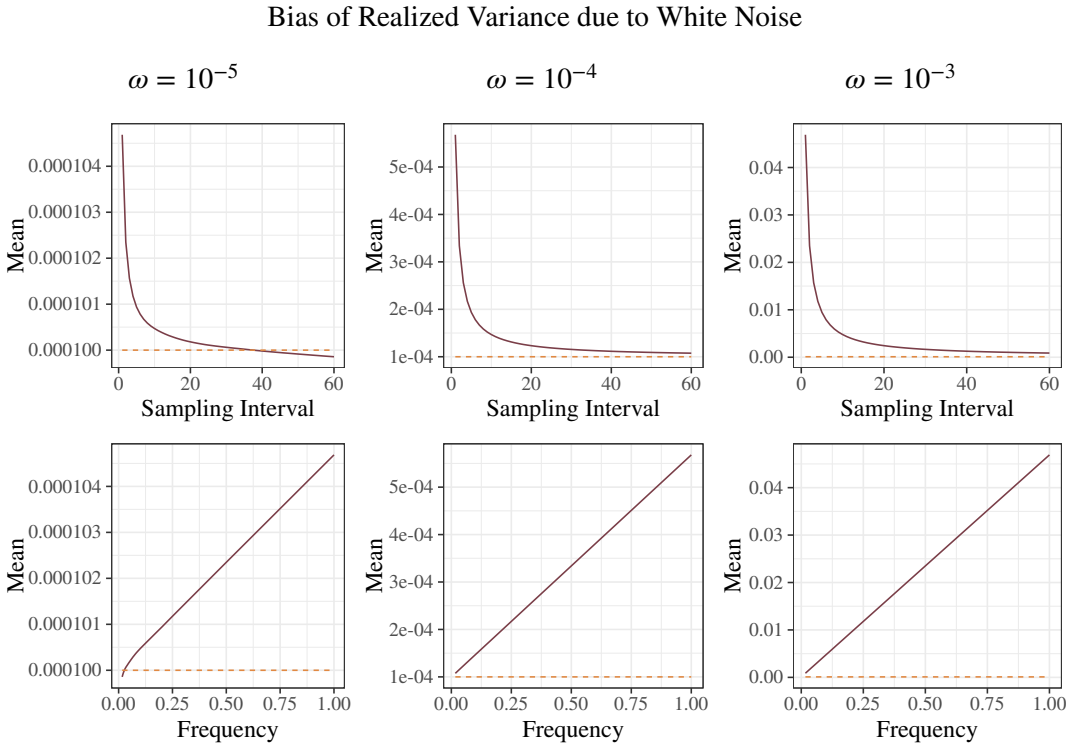

Figure 4.5: Simulated means of realized variance (solid lines) with true values (dotted lines) for the W-CV-WN-1, W-CV-WN-2 and W-CV-WN-3 models.

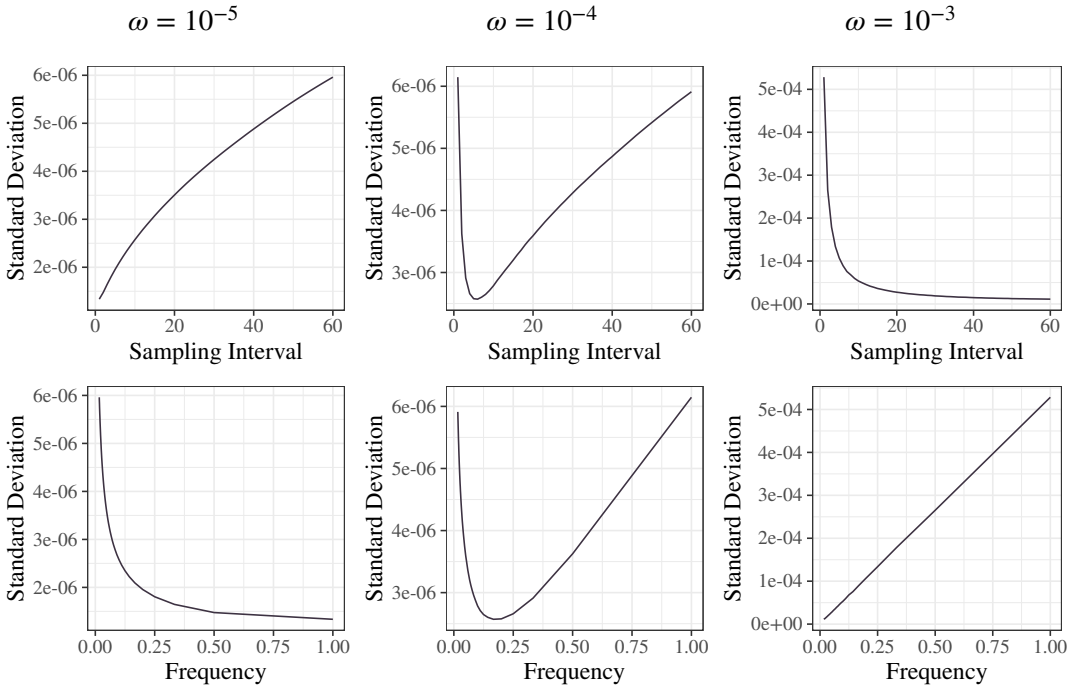

Figure 4.6: Simulated standard deviations of realized variance for the W-CV-WN-1, W-CV-WN-2 and W-CV-WN-3 models.

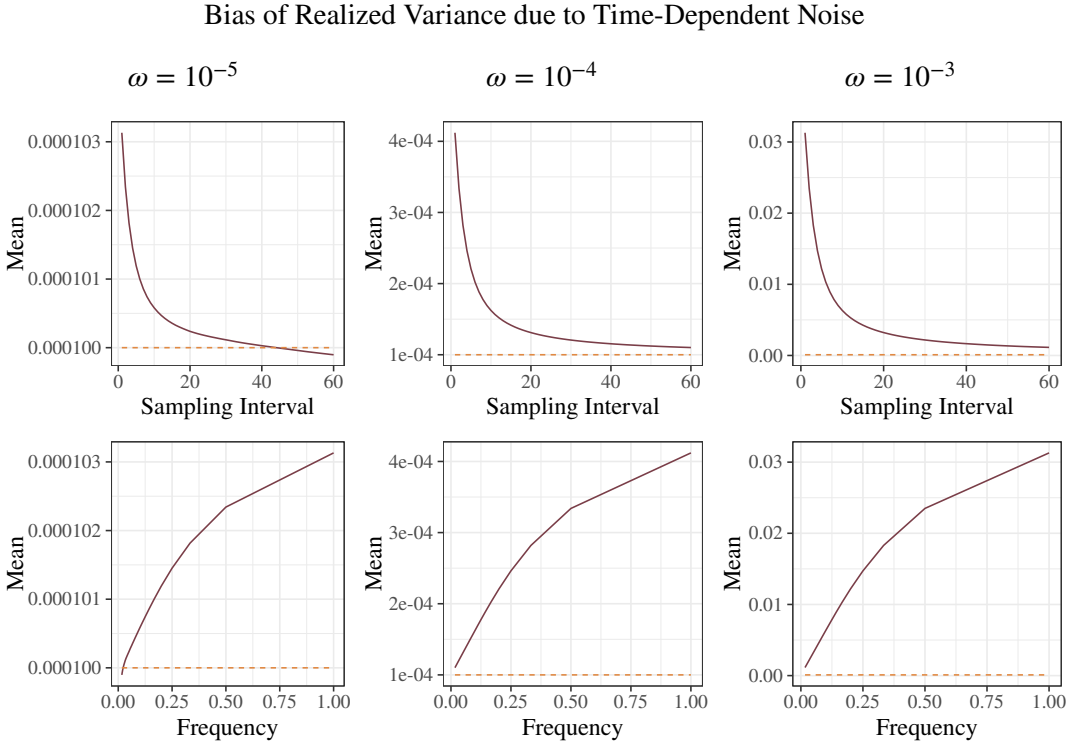

Figure 4.7: Simulated means of realized variance (solid lines) with true values (dotted lines) for the W-CV-TDN-1, W-CV-TDN-2 and W-CV-TDN-3 models.

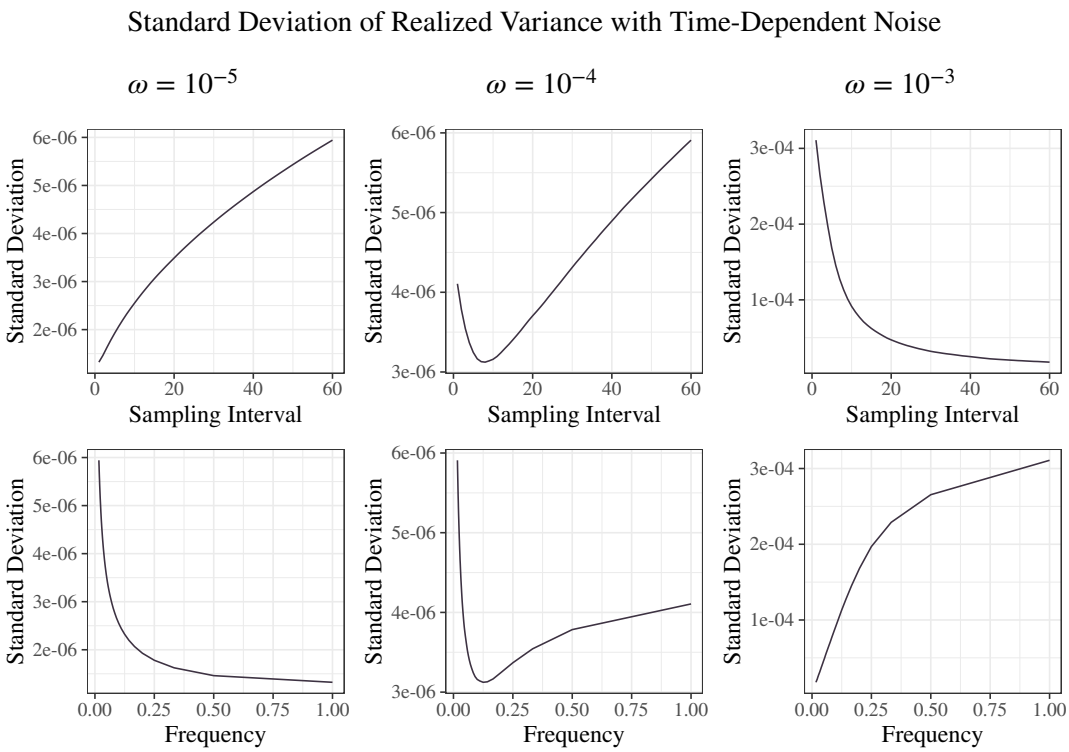

Figure 4.8: Simulated standard deviations of realized variance for the W-CV-TDN-1, W-CV-TDN-2 and W-CV-TDN-3 models.

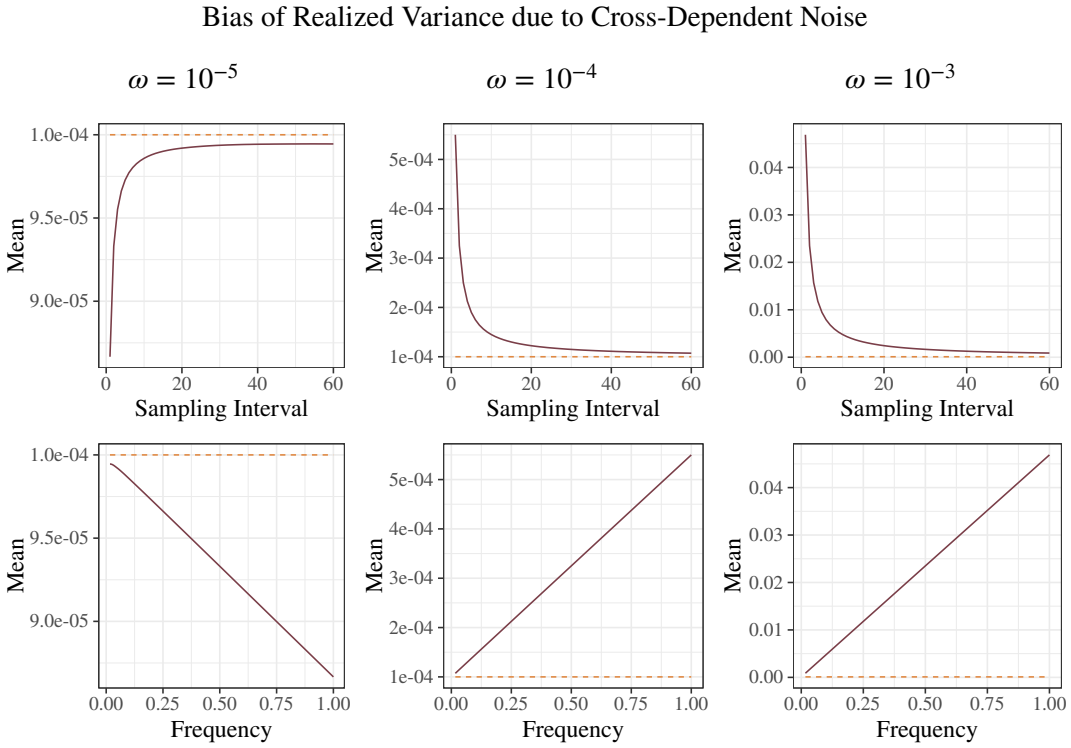

Figure 4.9: Simulated means of realized variance (solid lines) with true values (dotted lines) for the W-CV-CDN-1, W-CV-CDN-2 and W-CV-CDN-3 models.

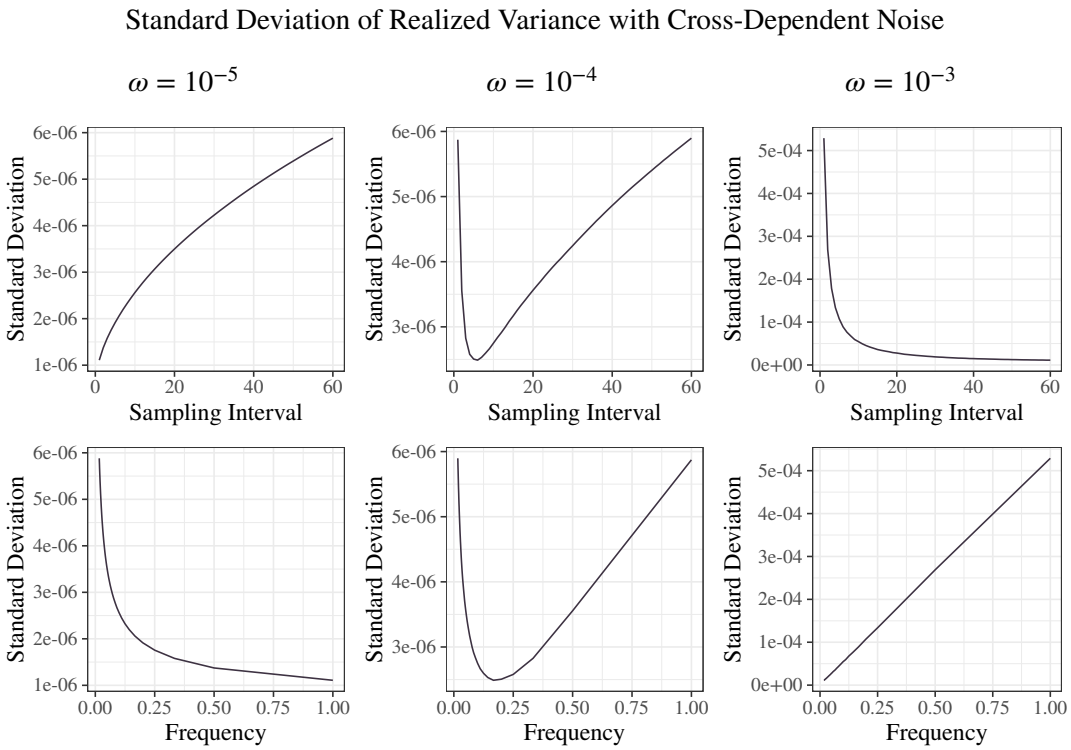

Figure 4.10: Simulated standard deviations of realized variance for the W-CV-CDN-1, W-CV-CDN-2 and W-CV-CDN-3 models.

Model	σ	ω	φ	θ
W-CV-WN-1	0.01	10^{-5}		
W-CV-WN-2	0.01	10^{-4}		
W-CV-WN-3	0.01	10^{-3}		
W-CV-TDN-1	0.01	10^{-5}	0.5	
W-CV-TDN-2	0.01	10^{-4}	0.5	
W-CV-TDN-3	0.01	10^{-3}	0.5	
W-CV-CDN-1	0.01	10^{-5}		-0.1
W-CV-CDN-2	0.01	10^{-4}		-0.1
W-CV-CDN-3	0.01	10^{-3}		-0.1

Table 4.1: Parameter values for the simulation models based on the Wiener process (W) with constant volatility (CV) and either white noise (WN), time-dependent noise (TDN) or cross-dependent noise (CDN).

The market microstructure noise follows the process

$$E_i = \varphi E_{i-1} + \theta \left(P_{T_i} - P_{T_{i-1}} \right) + U_i, \qquad U_i \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \omega^2\right), \qquad i = 1, \ldots, n. \tag{4.45}$$

We consider 9 simulation settings in total with different values of parameters σ, ω, φ and θ. These scenarios are listed in Table 4.1. Each simulation is performed 10 000 times.

A visual tool for investigating the impact of the market microstructure noise is the *volatility signature plot*. This technique was introduced by Andersen et al. (2000). It shows the dependency of bias of estimated volatility on sampling frequency. More specifically, our volatility signature plots show the average realized variance $ARV_{n,s}$ with different values of sampling intervals s. We also present dependency on frequency $f = 1/s$. Volatility signature plots are shown in figures 4.5, 4.7 and 4.9 while their standard deviations are shown in figures 4.6, 4.8 and 4.10.

The market microstructure noise following the white noise (scenarios W-CV-WN-1, W-CV-WN-2 and W-CV-WN-3) is investigated in figures 4.5 and 4.6. We can see that the bias is linear in the frequency for all considered variances of the noise. This is in line with (4.37). The behavior of standard deviation of the estimator, however, varies. Asymptotically, the variance of realized variance diverges to infinity. We see this behavior for the noise with large variance $\omega = 10^{-3}$. Specifically, standard deviation diverges as $O(f)$ for $f \to \infty$, i.e. variance diverges as $O(f^2)$ for $f \to \infty$. For the noise with small variance $\omega = 10^{-5}$, standard deviation can be approximated by $O(f^{-1/2})$ for $f \to 0$ and variance by $O(f^{-1})$ for $f \to 0$. Plots with medium variance $\omega = 10^{-4}$ illustrate how one behaviour transits into the other. The analytic expression of the variance of the realized variance under the independent white noise setting can be found e.g. in Zhang et al. (2005), Hansen and Lunde (2006) and Bandi and Russell (2008).

The market microstructure noise following the time-dependent noise (scenarios W-CV-TDN-1, W-CV-TDN-2 and W-CV-TDN-3) is investigated in figures 4.7 and 4.8. As suggested by (4.36), realized variance also diverges to infinity under this noise setting. The bias is however no longer linear.

The market microstructure noise following the cross-dependent noise (scenarios W-CV-CDN-1, W-CV-CDN-2 and W-CV-CDN-3) is investigated in figures 4.9 and 4.10. When the noise is negatively correlated with the efficient price, realized variance can diverge to minus infinity as discussed by Hansen and Lunde (2006). We observe this behaviour for the noise with small variance $\omega = 10^{-5}$.

4.2.2 Two-Scale Estimator

The first unbiased and consistent non-parametric estimator of the integrated variance proposed in the literature is the *two-scale estimator* of Zhang et al. (2005). It combines average realized variance at lower frequency as a biased estimate of quadratic variation with realized variance at highest possible frequency as an estimator of the noise variance (and therefore the bias under white noise assumption). For a given sampling interval s, it is defined as

$$TSE_{n,s} = \frac{n}{n - \tilde{m}(n,s)} ARV_{n,s} - \frac{\tilde{m}(n,s)}{n - \tilde{m}(n,s)} RV_n, \tag{4.46}$$

where the realized variance RV_n is given by (4.32), the average realized variance $ARV_{n,s}$ is given by (4.40) and

$$\tilde{m}(n,s) = \frac{1}{s} \sum_{h=1}^{s} \left\lfloor \frac{n-h}{s} \right\rfloor. \tag{4.47}$$

The estimator is consistent assuming the market microstructure noise follows the white noise. Zhang et al. (2005) show that the optimal choice for the number of subgrids s^* is

$$s^* = c^* n^{2/3}, \qquad c^* = \left(\frac{12\omega^4}{IQ} \right)^{1/3}, \tag{4.48}$$

where ω^2 is the variance of the noise and IQ is the integrated quarticity given by (4.7). The two-scale estimator then converges at rate $n^{1/6}$.

Multi-Scale Estimator

Zhang (2006) generalizes the two-scale estimator to the *multi-scale estimator*. It utilizes multiple average realized variances to cancel out the noise. The multi-scale estimator converges at rate $n^{1/4}$, which is the best achievable convergence rate for estimators of integrated variance (see Zhang, 2006). The estimator was further studied by Aït-Sahalia et al. (2011).

Quadratic Form

This section follows Holý (2017e). The two-scale estimator is a quadratic estimator with weight matrix

$$W_{n,s}^{TSE} = \frac{n}{n - \tilde{m}(n,s)} W_{n,s}^{ARV} - \frac{\tilde{m}(n,s)}{n - \tilde{m}(n,s)} W_n^{RV}, \tag{4.49}$$

where $W_{n,s}^{ARV}$ is given by (4.43) and W_n^{RV} is given by (4.41). As we can see in Figure 4.11, the structure is similar to the average realized variance. However, unlike the average realized variance it is a consistent estimator under the white noise assumption.

4.2.3 Realized Kernel Estimator

A popular estimator of integrated variance is the *realized kernel estimator* of Barndorff-Nielsen et al. (2008). It is consistent even for the time-dependent and cross-dependent noise. It is defined as

$$RK_{n,k} = RV_n + \sum_{l=1}^{k} K\left(\frac{l-1}{k} \right) \left(RA_{n,l} + RA_{n,-l} \right), \tag{4.50}$$

where $K(\cdot)$ is a kernel function and $RA_{n,l}$ is the *realized autocovariance* defined as

$$RA_{n,l} = \sum_{i=1}^{n} \left(X_i - X_{i-1} \right) \left(X_{i-l} - X_{i-l-1} \right) = \sum_{i=1}^{n} Y_i Y_{i-l}. \tag{4.51}$$

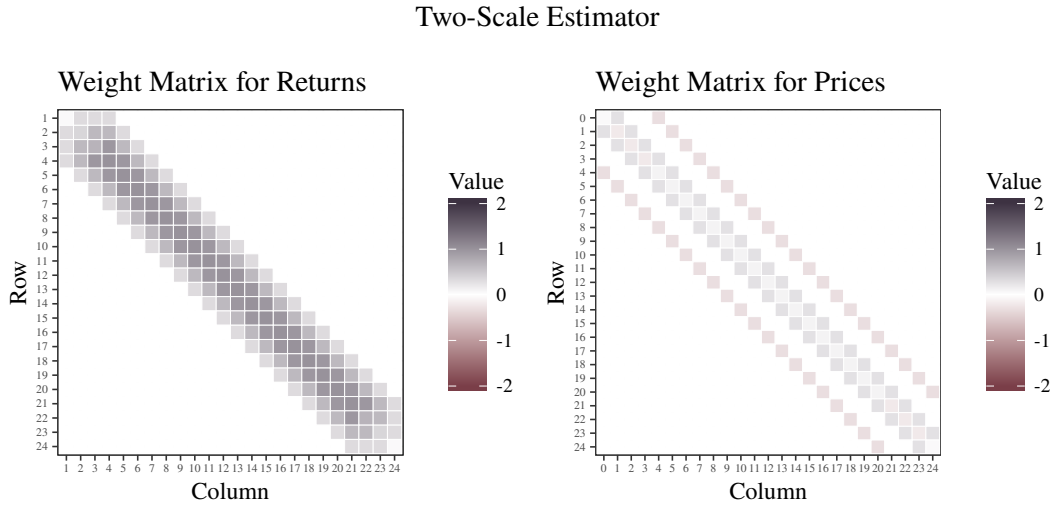

Figure 4.11: Quadratic form of two-scale estimator with $n = 24$ and $s = 4$.

Further, let us define ζ^2 as the *noise-to-signal ratio*

$$\zeta^2 = \frac{\omega^2}{\sqrt{IQ}}, \tag{4.52}$$

where ω^2 is the variance of the noise and IQ is the integrated quarticity given by (4.7).

First, let us consider kernel functions satisfying $K(0) = 1$ and $K(1) = 0$ with the bandwidth k selected as

$$k = \left\lfloor c\zeta^{4/3} n^{\frac{2}{3}} \right\rfloor, \tag{4.53}$$

where c is a constant depending on the kernel function. The realized kernel estimator is then asymptotically mixed Gaussian and converges at rate $n^{1/6}$.

Second, let us consider kernel functions satisfying $K(0) = 1$, $K(1) = 0$, $K'(0) = 0$ and $K'(1) = 0$ with the bandwidth k selected as

$$k = \left\lfloor c\zeta n^{1/2} \right\rfloor, \tag{4.54}$$

where c is a constant depending on the kernel function. The realized kernel estimator is then asymptotically mixed Gaussian and converges at rate $n^{1/4}$. We list some appropriate kernel functions with their optimal values c^* in the next section.

The realized kernel estimator was further studied and extended by Barndorff-Nielsen et al. (2009), Bandi and Russell (2011), Barndorff-Nielsen et al. (2011) and Ikeda (2015).

Kernel Functions

A simple kernel function satisfying $K(0) = 1$ and $K(1) = 0$ is the *Bartlett kernel* given by

$$K(x) = 1 - x, \qquad 0 \leq x \leq 1. \tag{4.55}$$

For the optimal bandwidth k^* in (4.53), we have $c^* = 2.28$. Interestingly, the realized kernel estimator with the Bartlett kernel function has asymptotically the same distribution as the two-scale estimator (see Barndorff-Nielsen et al., 2008).

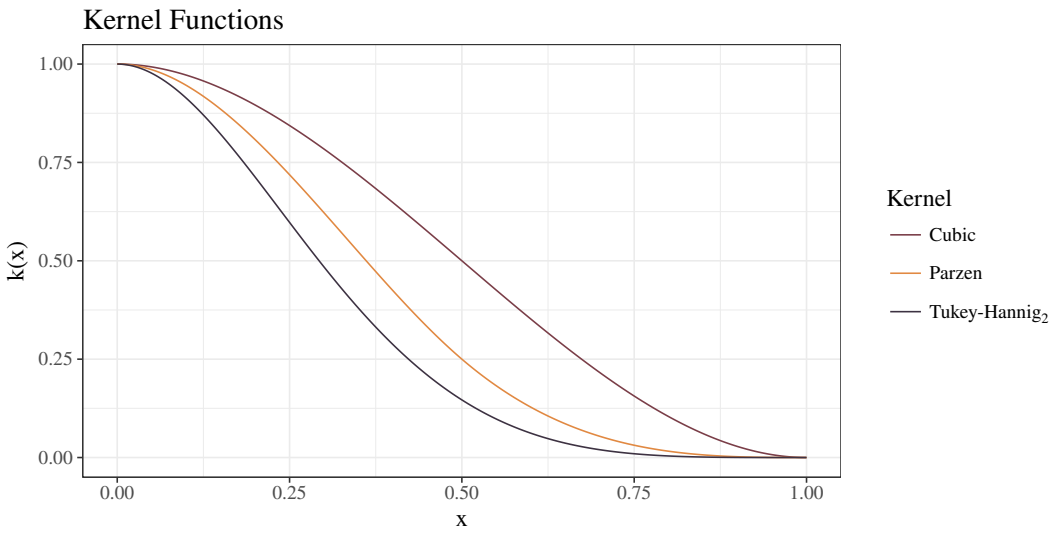

Figure 4.12: Various kernel functions.

Additionally, we consider three kernel functions satisfying $K(0) = 1$, $K(1) = 0$, $K'(0) = 0$ and $K'(1) = 0$. First, the *cubic kernel* is given by

$$K(x) = 1 - 3x^2 + 2x^3, \qquad 0 \leq x \leq 1. \tag{4.56}$$

For the optimal bandwidth k^* in (4.54), we have $c^* = 3.68$. The realized kernel estimator with the cubic kernel function has asymptotically the same distribution as the multi-scale estimator (see Barndorff-Nielsen et al., 2008).

Second, the *Parzen kernel* is given by

$$K(x) = \begin{cases} 1 - 6x^2 + 6x^3, & 0 \leq x \leq \frac{1}{2}, \\ 2(1-x)^3, & \frac{1}{2} < x \leq 1. \end{cases} \tag{4.57}$$

For the optimal bandwidth k^* in (4.54), we have $c^* = 4.77$.

Third, the *Tukey-Hanning kernel of order two* is given by

$$K(x) = \sin^2\left(\frac{\pi}{2}(1-x)^2\right), \qquad 0 \leq x \leq 1. \tag{4.58}$$

For the optimal bandwidth k^* in (4.54), we have $c^* = 5.74$.

The last three defined kernel functions are illustrated in Figure 4.12. For more alternative kernel functions, see Barndorff-Nielsen et al. (2008).

Quadratic Form

This section follows Holý (2017e). The realized kernel estimator can be expressed as a quadratic form with weight matrix $W_{n,k}^{RK}$ given by elements

$$w_{i,j}^{RK} = \begin{cases} 1 & \text{for } j = i, \\ K\left(\frac{l-1}{k}\right) & \text{for } |i-j| = l,\ l = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{4.59}$$

An example of the structure of the weight matrix for the realized kernel with Tukey-Hanning kernel of order two is shown in Figure 4.13.

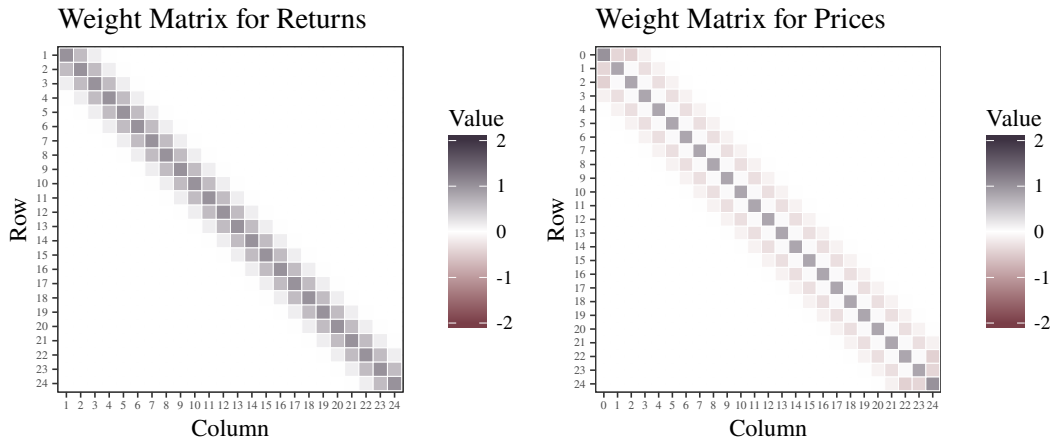

Figure 4.13: Quadratic form of realized kernel estimator with $n = 24$ and $k = 4$.

4.2.4 Pre-Averaging Estimator

The market microstructure noise can be removed by locally averaging returns. The *pre-averaging estimator* of Jacod et al. (2009) is based on this idea. The estimator is consistent for the time-dependent and cross-dependent noise. First, let us define the *averaged returns* as

$$Z_i = \sum_{l=1}^{k} G\left(\frac{l}{k}\right)\left(X_{i+l-1} - X_{i+l-2}\right) = \sum_{l=1}^{k} G\left(\frac{l}{k}\right) Y_{i+l-1}, \tag{4.60}$$

where $G(\cdot)$ is a function given by

$$G(x) = \min\left(x, 1 - x\right). \tag{4.61}$$

A direct analogue of realized variance with averaged returns is then given by

$$PAV_{n,k} = \sum_{i=1}^{n-k+1} Z_i^2. \tag{4.62}$$

Hautsch and Podolskij (2013) suggest to select the window size k as

$$k^* = \left\lfloor \theta\sqrt{n} \right\rfloor, \qquad \theta = 0.8. \tag{4.63}$$

This value of k^* leads to the optimal convergence rate of $n^{1/4}$ for the estimator. Finally, the pre-averaging estimator is defined as

$$PAE_{n,k} = \left(1 - \frac{\psi_{1,k} n^{-1}}{2\theta^2 \psi_{2,k}}\right)^{-1} \left(\frac{\sqrt{n}}{(n-k+2)\theta\psi_{2,k}} PAV_{n,k} - \frac{\psi_{1,k} n^{-1}}{2\theta^2 \psi_{2,k}} RV_n\right), \tag{4.64}$$

where

$$\begin{aligned} \psi_{1,k} &= k \sum_{l=1}^{k} \left(G\left(\frac{l+1}{k}\right) - G\left(\frac{l}{k}\right)\right)^2, \\ \psi_{2,k} &= \frac{1}{k} \sum_{l=1}^{k-1} G^2\left(\frac{l}{k}\right). \end{aligned} \tag{4.65}$$

The pre-averaging estimator was further studied and extended by Christensen et al. (2010), Hautsch and Podolskij (2013), Jacod and Mykland (2015) and Liu et al. (2017).

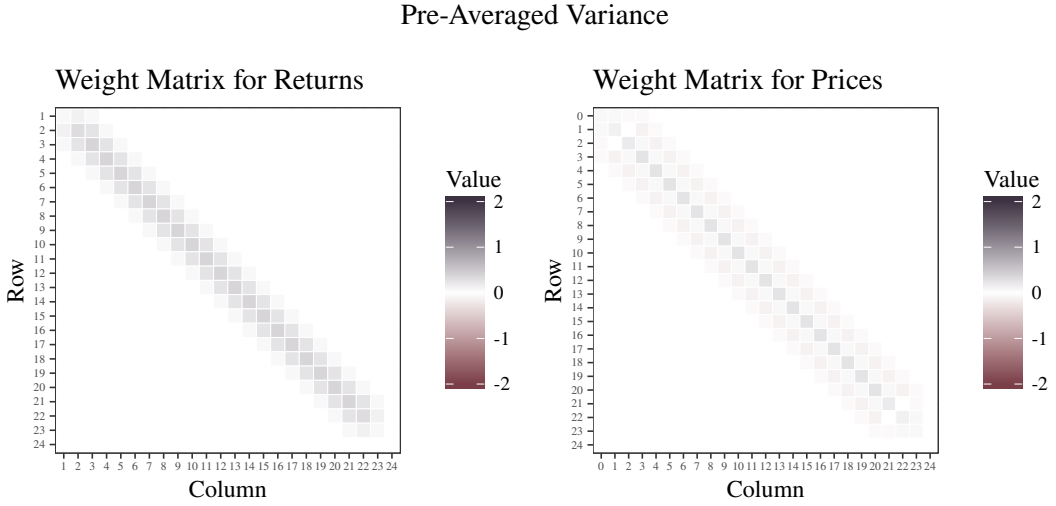

Figure 4.14: Quadratic form of pre-averaged variance with $n = 24$ and $k = 4$.

Quadratic Form

This section follows Holý (2017e). The pre-averaged variance can be formulated as a quadratic form with weight matrix $W^{PAV}_{n,k} = S'S$, where S is a matrix with $n-k$ rows and $n-1$ columns given by elements

$$s_{i,j} = \begin{cases} G\left(\frac{l}{k}\right) & \text{for } j = i + l - 1,\ l = 1, \dots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{4.66}$$

The structure of the weight matrix is shown in Figure 4.14. The pre-averaging estimator can then be formulated as a quadratic form with weight matrix

$$W^{PAE}_{n,k} = \frac{2\theta\sqrt{n}}{(2\theta^2 - \psi_{1,k}n^{-1})(n-k+2)} W^{PAV}_{n,k} - \frac{\psi_{1,k}}{2\theta^2\psi_{2,k}n - \psi_{1,k}} W^{RV}_{n}. \tag{4.67}$$

where $W^{PAV}_{n,k}$ is given by (4.66) and W^{RV}_{n} is given by (4.41). The weight matrix of the pre-averaging estimator is visualized in Figure 4.15.

4.2.5 Least Squares Estimator

This section loosely follows Holý (2017b). The idea behind the *least squares estimator* of Nolte and Voev (2012) is quite simple. Several realized variances are estimated using different numbers of observations n_i (i.e. different data subsamples). Assuming white noise, the bias of these estimates should be linearly dependent on the number of subsampled observations. The expected value of sparse realized variance (4.39) with the initial observation h and the sampling interval s is

$$\mathrm{E}[SRV_{n,h,s}] = IV + 2m(n,h,s)\omega^2, \tag{4.68}$$

where $m(n,h,s)$ is the number of observations utilized by the sparse realized variance given by (4.38). We can model this behaviour using linear regression of the form

$$SRV_{n,h,s} = \alpha + \beta m(n,h,s) + \varepsilon_{h,s}, \quad \varepsilon_{h,s} \overset{i.i.d.}{\sim} \mathrm{N}(0,\eta^2), \quad s = 1,\dots,k, \quad h = 1,\dots,s. \tag{4.69}$$

In other words we fit a line in signature volatility plot in Figure 4.5. Coefficient α then represents the integrated variance IV and coefficient β represents double the variance of the noise $2\omega^2$.

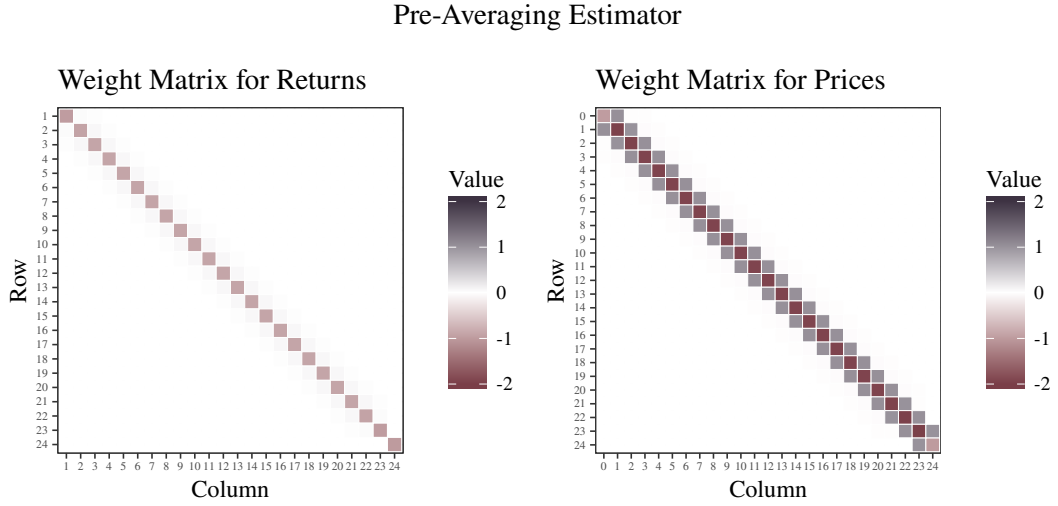

Figure 4.15: Quadratic form of pre-averaging estimator with $n = 24$ and $k = 4$.

The least squares estimator is given by

$$LSE_{n,k} = \frac{\bar{N}_2 \sum_{h=1}^{s} SRV_{n,h,s} - \bar{N}_1 \sum_{h=1}^{s} N_{h,s} SRV_{n,h,s}}{\bar{N}_0 \bar{N}_2 - \bar{N}_1^2}, \tag{4.70}$$

where

$$\bar{N}_d = \sum_{s=1}^{k} \sum_{h=1}^{s} |m(n,h,s)|^d \,. \tag{4.71}$$

Nolte and Voev (2012) find the optimal k^* as

$$k^* = \left\lfloor an^b \right\rfloor, \tag{4.72}$$

where

$$\begin{aligned} a &= \left(\frac{33.75\omega^4 \left(\pi^2 - 4\left(\gamma_0^2 + 2\gamma_1\right)\right)}{IQ} \right)^{1/3}, \\ b &= \frac{2}{3}\left(1 - \frac{\log(\log(n))}{\log(n)}\right), \end{aligned} \tag{4.73}$$

where π is the Archimedes' constant (approximately 3.14159), γ_0 is the Euler–Mascheroni constant (approximately 0.57722), γ_1 is the first Stieltjes constant (approximately -0.07282), ω^2 is the variance of the noise and IQ is the integrated quarticity given by (4.7).

Heteroskedasticity

Model (4.69) assumes that the variance of the error term $\varepsilon_{h,s}$ is constant. However, the variance of the sparse realized variance is not constant as illustrated in Figure 4.6 resulting in heteroskedasticity. The model can be improved by accounting for this variance. It is, however, quite complex as it is dependent on the integrated quarticity and the fourth moment of the noise. Holý (2017b) suggests to approximate standard deviation of the sparse realized variance by $O\left(m(n,h,s)^{-1/2}\right)$. The linear regression model then takes the form

$$m(n,h,s)^{\frac{1}{2}} SRV_{n,h,s} = \alpha m(n,h,s)^{\frac{1}{2}} + \beta m(n,h,s)^{\frac{3}{2}} + m(n,h,s)^{\frac{1}{2}} \varepsilon_{h,s}, \quad \varepsilon_{h,s} \overset{i.i.d.}{\sim} \mathrm{N}(0,\eta^2). \tag{4.74}$$

This approximation is valid only for a finite sample and noise with relatively small variance as suggested by Figure 4.6.

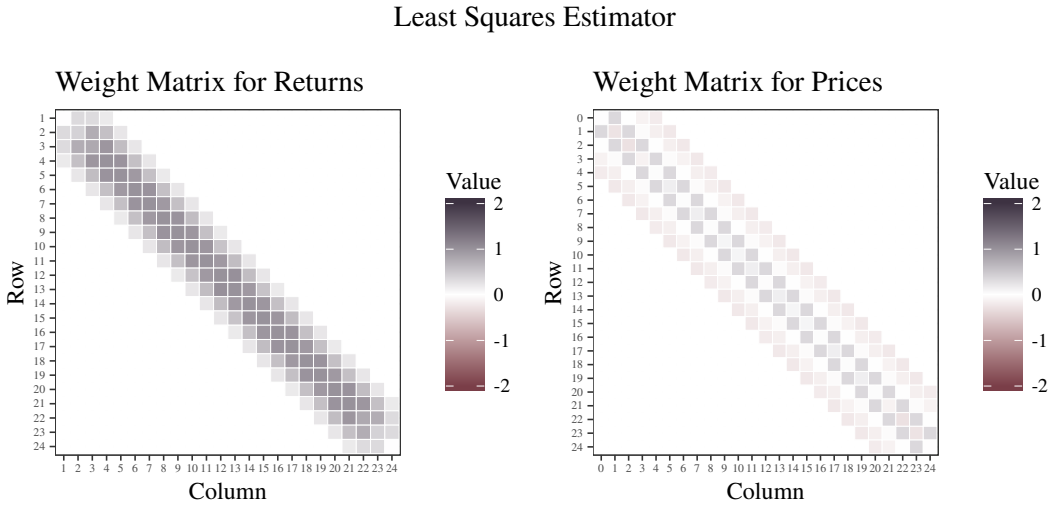

Figure 4.16: Quadratic form of least squares estimator with $n = 25$ and $k = 4$.

Test for the Presence of the Noise

Modeling the bias of realized variance by the ordinary least squares allows us to test for the presence of the market microstructure noise quite easily within the framework of linear regression. As suggested by Holý (2017b), the t-test for zero value of coefficient β in model (4.69) or (4.74) is actually the test for the presence of the noise.

Quadratic Form

This section follows Holý (2017e). The least squares estimator is a quadratic estimator with weight matrix

$$W_{n,k}^{LSE} = \sum_{s=1}^{k} \sum_{h=1}^{s} \frac{\bar{N}_2 - \bar{N}_1 N_{h,s}}{\bar{N}_0 \bar{N}_2 - \bar{N}_1^2} W_{n,h,s}^{SRV}, \tag{4.75}$$

where $W_{n,h,s}^{SRV}$ is given by (4.42). The structure of the weight matrix of the least squares estimator is similar to the average realized variance in Figure 4.4 and two-scale estimator in Figure 4.11. This is because all these methods are based on sparse realized variances.

4.3 Models of Quadratic Variation

To forecast daily volatility, various parametric models relating daily returns with latent volatility and intraday realized measure are utilized. The realized measure RM_i for day $i = 1, \ldots, n$ can be the realized variance or an estimator of integrated variance robust to the market miscrostructure noise. In this section, we present three models – the traditional time series model ARIMA, the HAR model of Corsi (2009) utilizing realized measures over various time horizons and the realized GARCH model of Hansen et al. (2012) based on the GARCH model augmented by realized measure.

4.3.1 ARIMA Model

Autoregressive models for realized measure forecasting were utilized e.g. by Andersen et al. (2003) and Aït-Sahalia and Mancini (2008). We consider the *autoregressive integrated moving average (ARIMA)* model (see e.g. Shumway and Stoffer, 2011). The ARIMA(p, d, q) model for realized measure RM_i is

given by

$$\Delta^d RM_i = \alpha + \sum_{j=1}^{p} \varphi_j \Delta^d RM_{i-j} + \sum_{j=1}^{p} \theta_j \varepsilon_{i-j} + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4.76}$$

where operator Δ^d denotes the d-th difference, ε_i is the uncorrelated white noise with variance σ^2 and $\alpha, \varphi_j, \theta_j$ are the parameters. Without some additional constraints on parameters, this model does not ensure the non-negativity of realized measure.

Choosing the Order

The question is how to choose the autoregressive order p, differencing order d and moving-average order q. As we need to estimate a large number of models in our empirical analysis in Section 4.4.2, we adopt the automatic framework of Hyndman and Khandakar (2008). In this procedure, the differencing order d is selected according to successive KPSS unit-root tests (Kwiatkowski et al., 1992)[1]. The autoregressive order p and the moving-average order q are then selected by minimizing the Akaike information criterion (Akaike, 1974). This approach was utilized e.g. by Holý (2017c) and Holý (2017d).

Logarithmic Transformation

Andersen et al. (2003) argue in their empirical study that the distribution of logarithm of realized variance is closer to the Gaussian distribution than the distribution of realized variance. For this reason, we also consider the logarithm of realized measure to follow the ARIMA process. The logarithmic ARIMA(p, d, q) model for realized measure RM_i is given by

$$\Delta^d \log RM_i = \alpha + \sum_{j=1}^{p} \varphi_j \Delta^d \log RM_{i-j} + \sum_{j=1}^{p} \theta_j \varepsilon_{i-j} + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4.77}$$

where operator Δ^d denotes the d-th difference, ε_i is the uncorrelated white noise with variance σ^2 and $\alpha, \varphi_j, \theta_j$ are the parameters. An advantage of the logarithmic model over the regular model is that it does not require constrains ensuring non-negativity of realized measure.

Gonçalves and Meddahi (2011) and Taylor (2017) generalized the logarithm transformation in the context of volatility forecasting by considering Box-Cox transformations for realized measures.

Relating Realized Measure to Returns

Models (4.76) and (4.77) concern only with dynamics of realized measure. To jointly model realized measure RM_i and daily returns Y_i, Aït-Sahalia and Mancini (2008) suggest extending the model (4.76) or (4.77) by equation

$$Y_i = M_i + \sqrt{RM_i}\chi_i, \qquad i = 1, \dots, n, \tag{4.78}$$

where M_i is a process for mean and χ_i are i.i.d. N(0, 1).

Daily returns Y_i can be defined in several ways. *Daytime return (open-to-close return)* measures the return generated by a stock during trading hours. It is based on the difference between the opening price of a given day and the closing price of that day. *Overnight return (close-to-open return)* measures the return generated by a stock when the market is closed. It is based on the difference between the closing price of a given day and the opening price of the next trading day. Overnight return and daytime return together form the *total daily return (close-to-close return)*, which is based on the difference between the closing price of a given day and the closing price of the next trading day. Differences between daytime, overnight and total daily returns were studied for example by Wang et al. (2009), Kelly and Clark (2011), Tsai et al. (2012) and Ochiai and Nacher (2019).

[1] When considering seasonality in the model, the procedure additionaly performs the extended Canova-Hansen test to determine seasonal differencing (Canova and Hansen, 1995)

4.3.2 HAR Model

Corsi (2009) proposes to model daily realized measures by the *heterogeneous autoregressive (HAR) model*. It features realized measures over different time horizons motivated by market agents operating at different frequencies. A standard HAR model explains daily realized measure by realized measure of the last day, week (last 5 days) and month (last 22 days). The HAR model for realized measure RM_i is given by

$$RM_i = \alpha + \beta_1 RM_{i-1} + \beta_2 \frac{1}{5} \sum_{j=1}^{5} RM_{i-j} + \beta_3 \frac{1}{22} \sum_{j=1}^{22} RM_{i-j} + \varepsilon_i, \qquad i = 1, \dots, n \tag{4.79}$$

where ε_i is an independent and identically distributed innovation process with zero mean and α, β_1, β_2, β_3 are the parameters.

The HAR model was further studied and extended by McAleer and Medeiros (2008b), Busch et al. (2011), Patton and Sheppard (2015) and Čech and Baruník (2017).

Logarithmic Transformation

Simalarly to Section 4.3.1, we also consider the HAR model for the logarithm of realized measure. The logarithmic HAR model for realized measure RM_i is given by

$$\log RM_i = \alpha + \beta_1 \log RM_{i-1} + \beta_2 \frac{1}{5} \sum_{j=1}^{5} \log RM_{i-j} + \beta_3 \frac{1}{22} \sum_{j=1}^{22} \log RM_{i-j} + \varepsilon_i, \qquad i = 1, \dots, n \tag{4.80}$$

where ε_i is an independent and identically distributed innovation process with zero mean and α, β_1, β_2, β_3 are the parameters.

4.3.3 Realized GARCH Model

Hansen et al. (2012) propose the *realized GARCH* model to jointly model observed returns, latent volatility and realized measure. It is a modification of the *generalized autoregressive conditional heteroskedasticity (GARCH)* model of Engle (1982) and Bollerslev (1986). In the realized GARCH model, lagged realized variances are used instead of lagged errors and the measurement equation of realized measure is added. The realized GARCH(p, q) model for realized measure RM_i is given by

$$\begin{aligned} Y_i &= M_i + \sqrt{H_i}\chi_i \\ H_i &= \omega + \sum_{j=1}^{p} \beta_j H_{i-j} + \sum_{j=1}^{q} \gamma_j RM_{i-j} \\ RM_i &= \xi + \varphi H_i + \tau(\chi_i) + \varepsilon_i \end{aligned} \tag{4.81}$$

with the leverage function $\tau(\cdot)$ given by

$$\tau(x) = \eta_1 x + \eta_2 \left(x^2 - 1\right), \tag{4.82}$$

where M_i is a process for mean, χ_i are i.i.d. N(0, 1), ε_i are i.i.d. N$(0, \sigma^2)$ and ω, β_j, γ_j, ξ, φ, η_1, η_2, σ are the parameters. Note that Y_i and RM_i are observable while H_i is latent. Daily returns Y_i can be either open-to-close returns or close-to-close returns similarly to Section 4.3.1. The last equation relates realized measure to the latent volatility with the leverage function $\tau(\cdot)$ allowing for asymmetric response in volatility to return shocks.

The realized GARCH model was further studied and extended by Watanabe (2012), Hansen et al. (2014), Baruník et al. (2016), Huang et al. (2016) and Jiang et al. (2018).

Logarithmic Transformation

Hansen et al. (2012) also suggest to use the realized GARCH model with logarithmic volatility and realized measure. The logarithmic realized GARCH(p, q) model for realized measure RM_i is given by

$$\begin{aligned} Y_i &= M_i + \sqrt{H_i}\chi_i \\ \log H_i &= \omega + \sum_{j=1}^{p} \beta_j \log H_{i-j} + \sum_{j=1}^{q} \gamma_j \log RM_{i-j} \\ \log RM_i &= \xi + \varphi \log H_i + \tau(\chi_i) + \varepsilon_i, \end{aligned} \tag{4.83}$$

where M_i is a process for mean, χ_i are i.i.d. N(0, 1), ε_i are i.i.d. N(0, σ^2) and ω, β_j, γ_j, ξ, φ, η_1, η_2, σ are the parameters.

4.4 Application to Daily Volatility

We assess suitability of quadratic variation estimators and models for daily volatility of various stocks. We study 30 stocks forming Dow Jones Industrial Average (DJIA) index as of March 19, 2015. We have data from January, 2015 until June, 2018. We clean data according to the procedure described in Section 2.1.1. For more details about analyzed stocks, see Appendix A. Our goal is to find an estimator with the best finite sample performance and a model with the most precise forecasting ability.

4.4.1 Estimators Performance

We compare the realized variance with the noise-robust estimators described in Section 4.2. For this purpose, we do not aggregate observations and utilize tick data. We denote the realized variance (4.32) as RV, two-scale estimator 4.46 as TSE, realized kernel estimator (4.50) with cubic kernel (4.56) as RK-C, realized kernel estimator (4.50) with Parzen kernel (4.57) as RK-P, realized kernel estimator (4.50) with Tukey-Hanning kernel of order two (4.58) as RK-TH2, pre-averaging estimator (4.64) as PAE and least squares estimator (4.70) as LSE.

Simulation Study

This section follows Holý and Černý (2017). To compare finite-sample performance of estimators, we conduct a simulation study. We consider the following model for simulations. The number of observations is set to $n = 23\,400$. The times of observations T_i, $i = 0, \ldots, n$ are generated by the Poisson point process and normalized to interval $[0, 1]$. The observed price is given by the additive model $X_i = P_{T_i} + E_i$, $i = 0, \ldots, n$. The efficient price P_{T_i} follows the process

$$\begin{aligned} \mathrm{d}P_t &= \tau(\mu - P_t)\mathrm{d}t + \sigma e^{G_t}\mathrm{d}W_t, \\ \mathrm{d}G_t &= -\kappa G_t \mathrm{d}t + \eta \mathrm{d}V_t, \end{aligned} \tag{4.84}$$

where W_t and V_t are Wiener processes correlated with coefficient ρ. This is the *one-factor stochastic volatility (SV1F) model* commonly used in simulations (see e.g. Barndorff-Nielsen and Shephard, 2004; Huang and Tauchen, 2005). The market microstructure noise follows the process

$$E_i = \varphi E_{i-1} + \theta\left(P_{T_i} - P_{T_{i-1}}\right) + U_i, \qquad U_i \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \omega^2\right), \qquad i = 1, \ldots, n. \tag{4.85}$$

We consider 8 simulation settings in total with different values of parameters μ, τ, σ, κ, η, ρ, ω, φ and θ. These scenarios are listed in Table 4.2. We denote constant volatility as CV, stochastic volatility as SV, no noise as P, white noise as WN, time-dependent noise as TDN and cross-dependent noise as CDN. Each simulation is performed 10 000 times.

Model	μ	τ	σ	κ	η	ρ	ω	φ	θ
OU-CV-P	1	10	0.01						
OU-CV-WN	1	10	0.01				0.0001		
OU-CV-TDN	1	10	0.01				0.0001	0.5	
OU-CV-CDN	1	10	0.01				0.0001		-0.1
OU-SV-P	1	10	0.01	0.1	0.1	-0.5			
OU-SV-WN	1	10	0.01	0.1	0.1	-0.5	0.0001		
OU-SV-TDN	1	10	0.01	0.1	0.1	-0.5	0.0001	0.5	
OU-SV-CDN	1	10	0.01	0.1	0.1	-0.5	0.0001		-0.1

Table 4.2: Parameter values for the simulation models based on the Ornstein–Uhlenbeck process (OU) with either constant volatility (CV) or stochastic volatility (SV) and either no noise (P), white noise (WN), time-dependent noise (TDN) or cross-dependent noise (CDN).

Model	RV	TSE	RK-C	RK-P	RK-TH2	PAE	LSE
OU-CV-P	0.0088	0.1037	0.0163	0.0161	0.0162	0.0544	0.0091
OU-CV-WN	4.6793	0.1053	0.1436	0.1561	0.1635	0.0564	0.0217
OU-CV-TN	3.1192	0.0991	0.5752	0.5875	0.5927	0.0535	1.2179
OU-CV-CN	4.4996	0.1039	0.1261	0.1370	0.1444	0.0551	0.0216
OU-SV-P	0.1485	0.1846	0.1496	0.1504	0.1500	0.1603	0.1488
OU-SV-WN	4.6559	0.1849	0.1773	0.1826	0.1850	0.1631	0.1531
OU-SV-TN	3.0941	0.1823	0.5513	0.5687	0.5712	0.1614	1.1752
OU-SV-CN	4.4734	0.1901	0.1739	0.1764	0.1783	0.1636	0.1562

Table 4.3: Mean absolute errors of quadratic variation estimated by various non-parametric methods using simulations of several price models.

The results of simulations are reported in Table 4.3. For the case of process without the noise OU-CV-P and OU-SV-P, the realized variance performs the best due to its simplicity. However, in other scenarios, it is clearly biased. Generally, quadratic variation in scenarios with stochastic volatility OU-SV-P, OU-SV-WN, OU-SV-TDN and OU-SV-CDN is much harder to estimate than in the case of constant volatility OU-CV-P, OU-CV-WN, OU-CV-TDN and OU-CV-CDN. The realized kernel estimator and least squares estimator are strongly affected by time dependence in the noise as seen in scenarios OU-CV-TDN and OU-SV-TDN. The pre-averaging estimator is, on the other hand, stable for various noise settings and overall performs the best. This is consistent with the study of Holý and Černý (2017) which uses a different simulation setup.

Evidence in Stock Prices

We estimate quadratic variation for the DJIA stocks from January, 2015 to June, 2018. First, we investigate structure of the market microstructure noise. We utilize the volatility signature plot of Andersen et al. (2000). Figure 4.17 shows the noise behavior on four examplary days. In some cases, we can see decreasing value of estimated quadratic variation with increasing number of observations. This was also observed by Hansen and Lunde (2006) in a similar dataset. The only reason for this kind of bias is a negative correlation between the noise and the efficient price. Overall, we can see that the noise is relatively small but has complex structure. This is also in line with Hansen and Lunde (2006).

Median quadratic variations estimated by various methods are reported in Table (4.4). We resort to median values as volatility reaches extreme values on a few days as illustrated in Figure 4.18. The estimated

Stock	RV	TSE	RK-C	RK-P	RK-TH2	PAE	LSE
AAPL	0.7939	0.7443	0.8493	0.8493	0.8481	0.8349	0.8122
AXP	0.6819	0.5252	0.6920	0.6905	0.6917	0.6616	0.6937
BA	0.8729	0.6079	0.8700	0.8654	0.8686	0.7718	0.8750
CAT	1.1546	0.8859	1.1584	1.1584	1.1584	1.1045	1.1727
CSCO	0.7195	0.6507	0.7219	0.7224	0.7261	0.7926	0.7342
CVX	0.8950	0.7327	0.8889	0.8906	0.8895	0.8412	0.9048
DD	1.0498	0.7832	1.0444	1.0478	1.0474	0.7919	0.9998
DIS	0.6196	0.4863	0.6289	0.6298	0.6301	0.6253	0.6293
GE	0.7436	0.5988	0.7541	0.7552	0.7547	0.7418	0.7486
GS	1.1380	0.8336	1.1285	1.1260	1.1282	1.0082	1.1378
HD	0.6933	0.5118	0.6912	0.6912	0.6912	0.6490	0.6963
IBM	0.6348	0.4512	0.6396	0.6387	0.6381	0.5709	0.6347
INTC	0.9306	0.8571	0.9416	0.9403	0.9420	1.0471	0.9464
JNJ	0.4982	0.3569	0.4941	0.4941	0.4942	0.4461	0.5030
JPM	0.9249	0.9071	0.9597	0.9631	0.9614	0.6399	0.9326
KO	0.4219	0.3278	0.4254	0.4259	0.4255	0.4017	0.4290
MCD	0.5459	0.3696	0.5500	0.5515	0.5513	0.4853	0.5535
MMM	0.5282	0.3311	0.5258	0.5261	0.5267	0.4071	0.5142
MRK	0.6930	0.5833	0.6994	0.7001	0.6997	0.7175	0.7091
MSFT	0.7373	0.7233	0.7751	0.7748	0.7751	0.8133	0.7516
NKE	0.8627	0.6791	0.8795	0.8799	0.8799	0.8613	0.8810
PFE	0.6823	0.5315	0.6848	0.6863	0.6856	0.6903	0.6966
PG	0.4492	0.3454	0.4633	0.4643	0.4636	0.4438	0.4575
TRV	0.5960	0.3747	0.5856	0.5849	0.5856	0.4585	0.5756
UNH	0.9228	0.6144	0.9153	0.9149	0.9150	0.7778	0.9080
UTX	0.6400	0.4267	0.6391	0.6383	0.6392	0.5160	0.6334
V	0.5778	0.4297	0.5897	0.5896	0.5898	0.5575	0.5848
VZ	0.5938	0.5113	0.6055	0.6054	0.6073	0.6355	0.6021
WMT	0.6135	0.4594	0.6184	0.6187	0.6190	0.5869	0.6182
XOM	0.6459	0.5569	0.6677	0.6684	0.6688	0.6489	0.6514

Table 4.4: Medians of daily quadratic variation estimated by the various non-parametric methods.

values do not distinctly differ as the noise induces sometimes increasing and sometimes decreasing bias in our dataset. Based on our simulation study, we consider the values estimated by the pre-averaging estimator as the most reliable.

4.4.2 Models Performance

We compare forecasting ability of models for realized measures described in Section 4.3. As realized measure we adopt the pre-averaging estimator (4.64). We denote the ARIMA model (4.76) as ARIMA, logarithmic ARIMA model (4.77) as ARIMA-LN, HAR model (4.79) as HAR, logarithmic HAR model (4.80) as HAR-LN, logarithmic realized GARCH model (4.83) with open-to-close returns as RGARCH-OC and logarithmic realized GARCH model (4.83) with close-to-close returns as RGARCH-CC. We also consider some elementary naive models. The forecast equal to the value of previous day is denoted as PREV, the forecast equal to the mean of past historical values is denoted as MEAN and the forecast equal to the median of past historical values is denoted as MED.

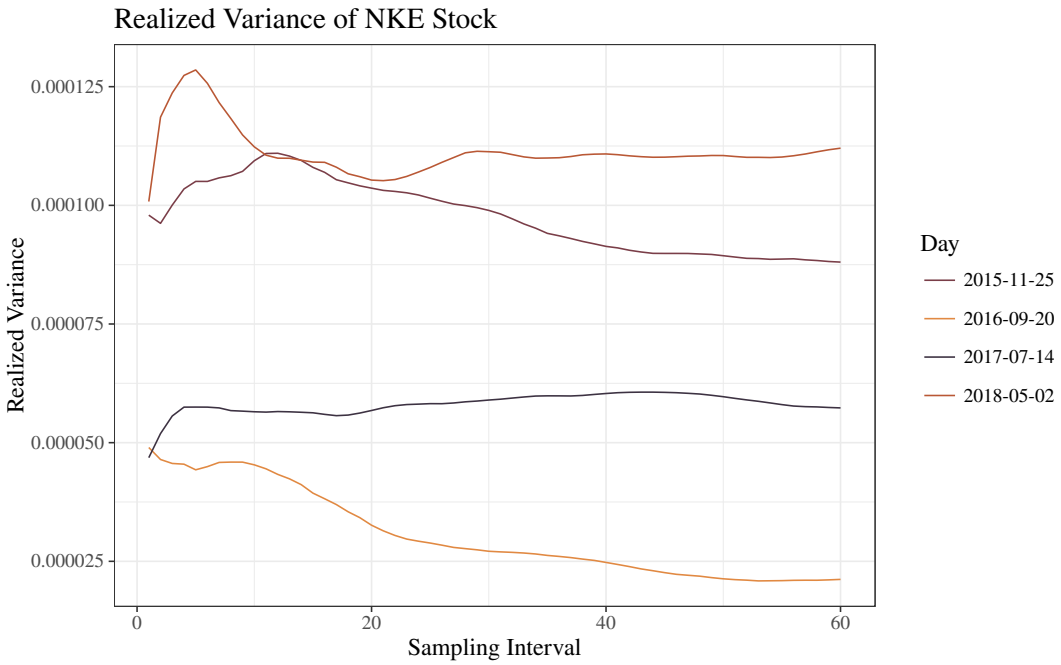

Figure 4.17: Volatility signature plots of NKE stock on various days.

Evidence in Stock Prices

Figure 4.18 shows examples of time series we forecast. There is some degree of autocorrelation present but also some observations with extreme values. We focus on one-step-ahead forecasts of realized measure. Table 4.5 and Table 4.6 report median absolute errors of forecasts. Generally, models with logarithmic transformation perform much better. Interestingly, the naive PREV model outperforms the ARIMA and HAR models, which are sensitive to extreme observations. The HAR-LN, however, has the lowest median absolute error among all considered models. In this case, the logarithmic scale reduces the impact of large values of volatility. Realized GARCH model with logarithmic transform also performs adequately. We find that the use of the returns given by the difference between the closing and opening price of the same day (the RGARCH-OC model) is more suitable in this application than the returns given by the difference between closing prices of successive days (the RGARCH-CC model). We see that models HAR-LN, RGARCH-OC and RGARCH-CC specifically designed for high-frequency data have more accurate forecasts than the traditional ARIMA-LN model.

4.4.3 Discussion

First, we compare estimators of quadratic variation in the presence of the market microstructure noise. In a simulation study, we find that the pre-averaging estimator is the best alternative for finite-sample estimation of quadratic variation. It exhibits relatively small errors even for the time-dependent and cross-dependent noise. Although it is outperformed by other methods in some scenarios, it has rather stable performance and does not cause extreme errors in any scenario. For this reason, we recommend to use the pre-averaging estimator for quadratic variation estimation.

Second, we compare models for realized measure forecasting. Following the comparison of ex-post estimators, we select the pre-averaging estimator as realized measure. The empirical study shows that it is necessary to use the logarithmic transformation for realized measure. If the goal is to solely forecast realized measure, the logarithmic HAR model is the best choice. If the goal is to jointly model returns, volatility and realized measure, the logarithmic realized GARCH model performs adequately.

Stock	PREV	MEAN	MED	ARIMA	HAR
AAPL	2.4758	4.8201	3.3948	3.0910	3.0659
AXP	2.2267	3.2660	2.2909	2.5527	2.6609
BA	2.7419	4.1360	3.4876	3.0323	3.1444
CAT	3.6633	5.6677	5.2383	4.3074	4.2637
CSCO	2.3029	3.6396	2.6921	2.5117	2.5477
CVX	2.5448	3.5907	3.5389	2.6197	2.5545
DD	3.7109	4.0291	3.2170	2.9344	3.2416
DIS	1.9707	3.1009	2.3421	2.3182	2.1531
GE	2.8384	3.8303	3.6604	2.7167	2.8193
GS	3.5255	4.9245	3.9007	3.6179	3.7763
HD	2.1249	3.6600	2.2731	2.5155	2.4113
IBM	1.7954	2.7260	2.3295	2.0096	2.0655
INTC	2.8695	4.4331	3.4076	3.1777	3.0712
JNJ	1.3633	2.2745	1.7236	1.6964	1.6971
JPM	3.7248	6.4754	3.6022	4.7961	6.0715
KO	1.1205	1.8137	1.2731	1.3014	1.3422
MCD	1.5160	2.0884	1.7430	1.6954	1.8347
MMM	1.3189	2.0605	1.5252	1.4372	1.5226
MRK	2.1655	4.0791	2.9284	2.7302	2.9306
MSFT	2.2494	3.9312	3.0617	2.6711	2.5744
NKE	3.1729	4.1483	3.4244	3.0669	3.1101
PFE	2.0662	4.0580	2.9021	2.5046	2.6306
PG	1.3232	2.0572	1.5838	1.4780	1.4745
TRV	1.6313	2.3427	1.6163	1.7352	1.8324
UNH	2.5298	4.1908	3.1899	2.9564	3.0985
UTX	1.8267	2.4309	2.1382	1.7816	2.0530
V	1.5682	2.6779	1.7819	1.5799	1.7273
VZ	2.1086	3.4119	2.3123	2.6512	2.5718
WMT	2.0958	3.1225	2.2289	2.3191	2.3879
XOM	1.8118	2.8572	2.4249	1.8313	1.9118
Average	2.2794	3.5281	2.7078	2.5212	2.6182

Table 4.5: Median absolute errors of one-step-ahead forecasts of daily quadratic variation estimated by pre-averaging estimator and forecasted by various models.

Stock	ARIMA-LN	HAR-LN	RGARCH-OC	RGARCH-CC
AAPL	2.2378	2.3462	2.2626	2.2687
AXP	1.9345	1.9302	1.9438	1.9634
BA	2.4374	2.3241	2.5698	2.4730
CAT	3.4316	3.4671	3.4081	3.3660
CSCO	2.0163	1.9260	1.9852	1.9455
CVX	2.2778	2.1953	2.2525	2.2287
DD	3.4348	3.5842	3.2550	3.5192
DIS	1.7278	1.6612	1.7039	1.6843
GE	2.4083	2.3728	2.4204	2.4286
GS	3.1508	2.9903	3.1162	3.0968
HD	1.7844	1.6996	1.7256	1.7676
IBM	1.6292	1.6186	1.6015	1.6069
INTC	2.5469	2.4392	2.4392	2.4826
JNJ	1.2440	1.2047	1.2419	1.2468
JPM	3.9702	3.9288	3.5737	3.9726
KO	1.0423	1.0305	0.9936	0.9989
MCD	1.3366	1.3377	1.2722	1.3436
MMM	1.1328	1.1839	1.1396	1.1080
MRK	1.9819	1.9992	2.0442	2.0039
MSFT	2.0891	2.0948	2.1460	2.1849
NKE	2.7593	2.6539	2.7646	2.6978
PFE	1.8920	1.9375	1.9826	1.9320
PG	1.1477	1.1395	1.1265	1.1652
TRV	1.3386	1.2985	1.2981	1.3158
UNH	2.1366	2.1240	2.2049	2.2066
UTX	1.5626	1.5421	1.6019	1.5320
V	1.3699	1.3891	1.3779	1.4104
VZ	1.8120	1.6991	1.7323	1.7572
WMT	1.8722	1.7712	1.7557	1.7442
XOM	1.6851	1.6300	1.7383	1.6688
Average	2.0463	2.0173	2.0226	2.0373

Table 4.6: Median absolute errors of one-step-ahead forecasts of daily quadratic variation estimated by the pre-averaging estimator and forecasted by various logarithmic models.

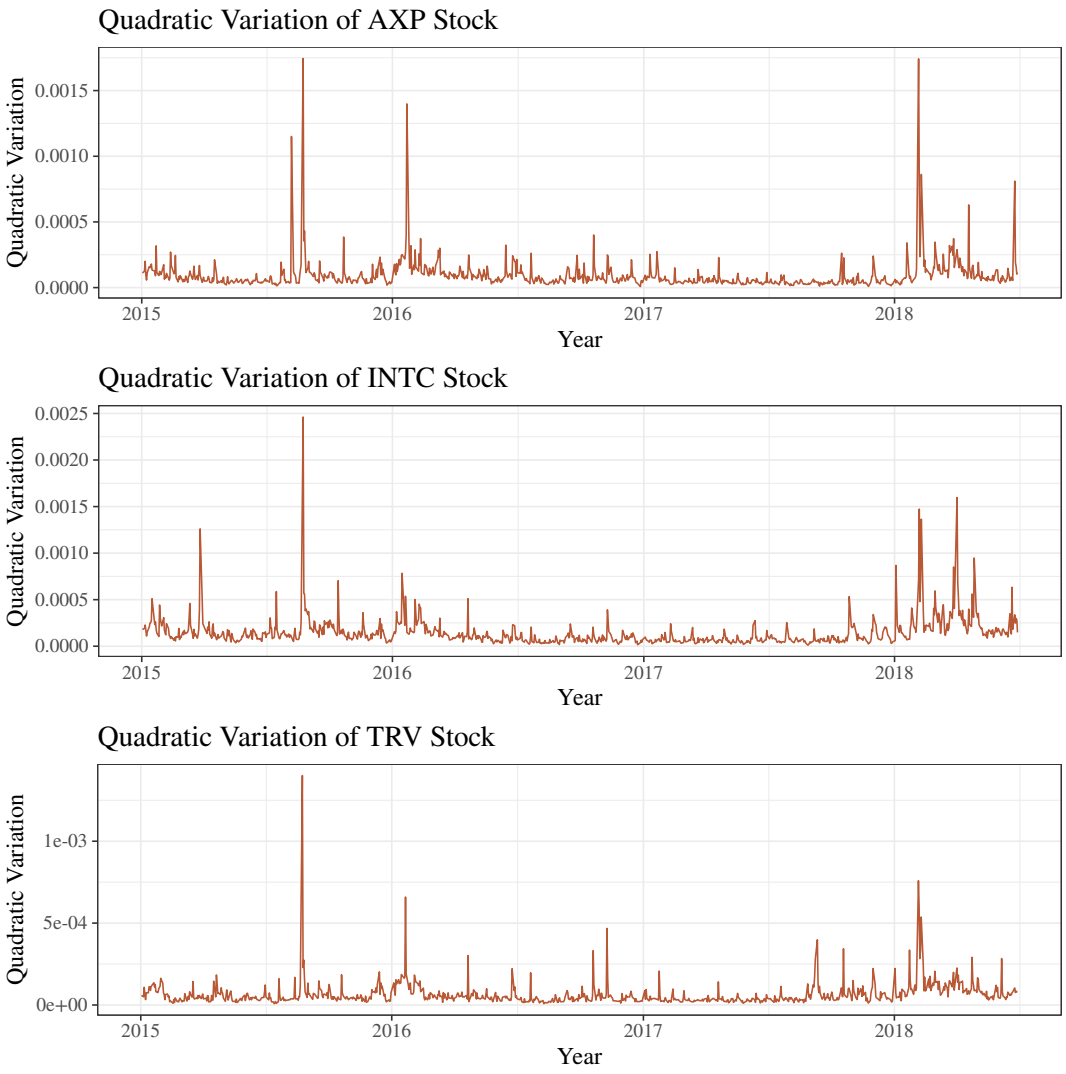

Figure 4.18: Daily quadratic variation of AXP, INTC and TRV stocks estimated by the pre-averaging method.

- Chapter 5 -

Ornstein–Uhlenbeck Process

This section follows Holý and Tomanová (2018). In finance, many different time series tend to move to their mean values over time. This behaviour is known as the mean reversion and is often captured by the Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930). It can be used to model currency exchange rates (Ball and Roma, 1994; Gil-Alana, 2000) and commodity prices (Schwartz, 1997). A major application of the Ornstein–Uhlenbeck process is the modeling of interest rates by the so-called Vasicek model (Vasicek, 1977; Hull and White, 1990; Babbs and Nowman, 1999; Andresen et al., 2014). The Ornstein–Uhlenbeck process can also be utilized to model stochastic volatility of financial assets (Barndorff-Nielsen and Shephard, 2001; Griffin and Steel, 2006; Hofmann and Schulz, 2016; Peng et al., 2016; Benth et al., 2018). Another application is the trading strategy called the pairs trading. It is based on a tendency of the spread between highly correlated time series to return to its long-term mean value making the movement of the prices predictable and profitable (Elliott et al., 2005; Bertram, 2010; Cummins and Bucca, 2012; Zeng and Lee, 2014; Liu et al., 2017). An example of Ornstein–Uhlenbeck process path is shown in Figure 5.1.

The Ornstein–Uhlenbeck process can be utilized when analyzing financial high-frequency data. In general, high-frequency time series exhibit specific characteristics such as heavy tailed distribution, the presence of jumps and market microstructure noise. In the Ornstein–Uhlenbeck model, the first two characteristics are often captured by generalizing the background driving process to the Lévy process (Barndorff-Nielsen and Shephard, 2001). The Ornstein–Uhlenbeck process driven by the Lévy process was further studied by Masuda (2004), Lindner and Maller (2005), Brockwell et al. (2007), Borovkov and Novikov (2008), Behme and Lindner (2012), Fasen (2013), Pakkanen et al. (2017) and Kevei (2018).

We focus on challanges surrounding the market microstructure noise. The majority of the literature concerning the market microstructure noise is focused on non-parametric volatility estimation. However, a parametric modeling is also important as it can be directly utilized in forecasting and decision-making. In this section, we estimate parameters of the Gaussian Ornstein–Uhlenbeck process in the presence of the independent Gaussian noise. The noise-robust approach we propose has several important implications and advantages. We show that Ornstein–Uhlenbeck parameters estimated by methods ignoring the noise are biased and inconsistent. In addition, we demonstrate that even when the variance of the noise is relatively small and one would simply decide to ignore it (which is unfortunately quite common in practice), it has a great impact on estimated parameters. The reliance of market participants on this biased estimates can lead to wrong decisions and have harmful consequences as we illustrate in an application to the pairs trading strategy. The pitfall of this lies in the fact that estimated parameters might appear as reliable values at the first sight but they are actually multiple times higher than their true values. As we argue this is caused by the fact that the Ornstein–Uhlenbeck process contaminated by the independent Gaussian white noise and observed at discrete equidistant times follows ARMA(1,1) process instead of AR(1) process. We make use of this finding and propose a noise-robust estimator based on the ARMA(1,1) reparametrization. We also deal with the situation when the observations are not equidistant and propose noise-robust estimator based on the maximum likelihood.

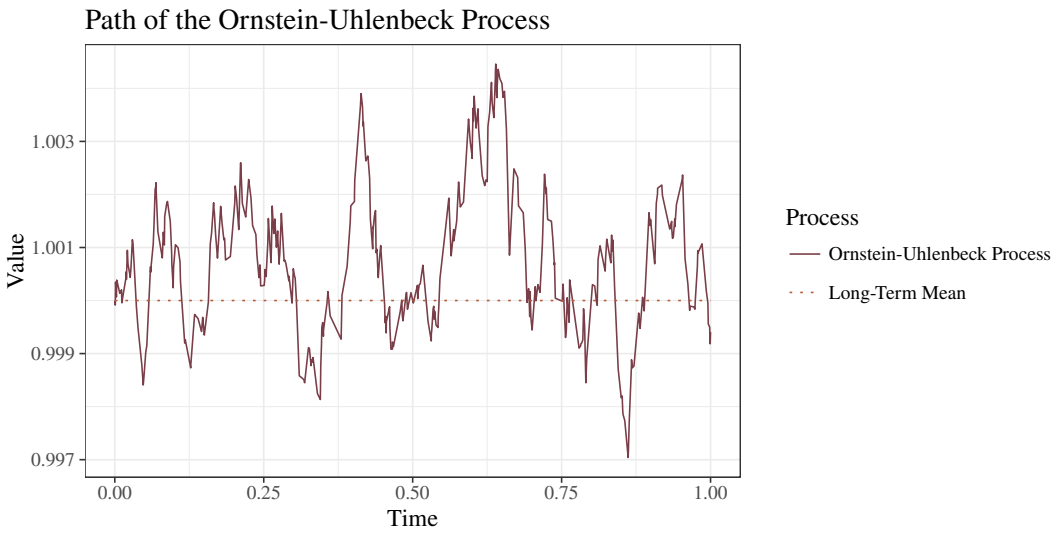

Figure 5.1: Simulated path of the Ornstein–Uhlenbeck process with parameters $\mu = 1$, $\tau = 10$ and $\sigma^2 = 10^{-4}$.

5.1 Estimators of Ornstein–Uhlenbeck Process

The *Ornstein–Uhlenbeck process* $P_t, t \geq 0$ is a process satisfying stochastic differential equation

$$\mathrm{d}P_t = \tau(\mu - P_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{5.1}$$

where W_t is a Wiener process, μ is a parameter representing long-term mean, $\tau > 0$ is a parameter representing speed of reversion and $\sigma > 0$ is a parameter representing instantaneous volatility. This stochastic differential equation has solution

$$P_t = P_0 e^{-\tau t} + \mu(1 - e^{-\tau t}) + \sigma \int_0^t e^{-\tau(t-s)} \mathrm{d}W_s. \tag{5.2}$$

When assuming $P_0 \sim \mathrm{N}(\mu, \sigma^2/2\tau)$ and $P_0 \perp W_t, t \geq 0$, the Ornstein–Uhlenbeck process P_t is a stationary process with normally distributed increments and unconditional moments

$$\begin{aligned} \mathrm{E}[P_t] &= \mu, \\ \mathrm{var}[P_t] &= \frac{\sigma^2}{2\tau}, \\ \mathrm{cov}[P_t, P_s] &= \frac{\sigma^2}{2\tau} e^{-\tau|t-s|}, \quad t \neq s. \end{aligned} \tag{5.3}$$

For a given initial value p_0, the Ornstein–Uhlenbeck process P_t is a non-stationary process with normally distributed increments and conditional moments

$$\begin{aligned} \mathrm{E}[P_t|P_0 = p_0] &= p_0 e^{-\tau t} + \mu\Big(1 - e^{-\tau t}\Big), \\ \mathrm{var}[P_t|P_0 = p_0] &= \frac{\sigma^2}{2\tau}\Big(1 - e^{-2\tau t}\Big), \\ \mathrm{cov}[P_t, P_s|P_0 = p_0] &= \frac{\sigma^2}{2\tau}\Big(e^{-\tau|t-s|} - e^{-\tau(t+s)}\Big), \quad t \neq s. \end{aligned} \tag{5.4}$$

In practice, we do not observe continuous paths of the process. Instead, we only observe the process P_{T_i} at a finite number of discrete times $0 = T_0 < T_1 < \ldots < T_n = 1$, where T_i are deterministic times

of observations. Without loss of generality, we restrict ourselves to the time interval $[0, 1]$. We further assume that the observed process is contaminated by independent white noise $E_i \sim \mathrm{N}(0, \omega^2)$. For the observed discrete process X_i, we utilize the additive noise model

$$X_i = P_{T_i} + E_i, \qquad i = 0, \ldots, n. \tag{5.5}$$

When assuming $P_0 \sim \mathrm{N}(\mu, \sigma^2/2\tau)$ and P_0 independent of W_{T_i}, $i \geq 0$, the observed process X_i is a stationary process with normally distributed increments and unconditional moments

$$\begin{aligned} \mathrm{E}[X_i] &= \mu, \\ \mathrm{var}[X_i] &= \frac{\sigma^2}{2\tau} + \omega^2, \\ \mathrm{cov}[X_i, X_j] &= \frac{\sigma^2}{2\tau} e^{-\tau|T_i - T_j|}, \quad i \neq j. \end{aligned} \tag{5.6}$$

For a given x_0 the observed process X_i is a non-stationary process with normally distributed increments and conditional moments

$$\begin{aligned} \mathrm{E}[X_i|X_0 = x_0] &= \mathrm{E}[P_0|X_0 = x_0]e^{-\tau T_i} + \mu\Big(1 - e^{-\tau T_i}\Big), \\ \mathrm{var}[X_i|X_0 = x_0] &= \mathrm{var}[P_0|X_0 = x_0]e^{-2\tau T_i} + \frac{\sigma^2}{2\tau}\Big(1 - e^{-2\tau T_i}\Big) + \omega^2, \\ \mathrm{cov}[X_i, X_j|X_0 = x_0] &= \mathrm{var}[P_0|X_0 = x_0]e^{-\tau(T_i + T_j)} + \frac{\sigma^2}{2\tau}\Big(e^{-\tau|T_i - T_j|} - e^{-\tau(T_i + T_j)}\Big), \quad i \neq j, \end{aligned} \tag{5.7}$$

where

$$\begin{aligned} \mathrm{E}[P_0|X_0 = x_0] &= \frac{x_0\sigma^2 + 2\tau\mu\omega^2}{\sigma^2 + 2\tau\omega^2}, \\ \mathrm{var}[P_0|X_0 = x_0] &= \frac{\sigma^2\omega^2}{\sigma^2 + 2\tau\omega^2}. \end{aligned} \tag{5.8}$$

The above conditional distribution is derived using the following proposition with $P = P_0$, $\mu_P = \mu$, $\sigma_P^2 = \sigma^2/(2\tau)$, $E = E_0$, $\mu_E = 0$, $\sigma_E^2 = \omega^2$ and $X = X_0$.

Proposition 5.1. *Let $P \sim N(\mu_P, \sigma_P^2)$, $E \sim N(\mu_E, \sigma_E^2)$ and $P \perp E$. Let $X = P + E$. The conditional probability density function is then*

$$f_P(p|X = x) = \frac{1}{\sqrt{2\pi\sigma_C^2(x)}} \exp\left\{ -\frac{\big(p - \mu_C(x)\big)^2}{2\sigma_C^2(x)} \right\}, \tag{5.9}$$

where

$$\begin{aligned} \mu_C(x) &= \frac{\mu_P\sigma_E^2 - \mu_E\sigma_P^2 + x\sigma_P^2}{\sigma_P^2 + \sigma_E^2}, \\ \sigma_C^2(x) &= \frac{\sigma_P^2\sigma_E^2}{\sigma_P^2 + \sigma_E^2}. \end{aligned} \tag{5.10}$$

Proof. The joint probability density function of P and X is given by

$$\begin{aligned} g_{P,X}(p, x) &= \frac{1}{\sqrt{2\pi\sigma_P^2}} \exp\left\{ -\frac{(p - \mu_P)^2}{2\sigma_P^2} \right\} \frac{1}{\sqrt{2\pi\sigma_E^2}} \exp\left\{ -\frac{(x - p - \mu_E)^2}{2\sigma_E^2} \right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_P^2}\sqrt{2\pi\sigma_E^2}} \exp\left\{ -\frac{\sigma_P^2 + \sigma_E^2}{2\sigma_P^2\sigma_E^2} p^2 + \frac{\mu_P\sigma_E^2 + x\sigma_P^2 - \mu_E\sigma_P^2}{\sigma_P^2\sigma_E^2} p \right. \\ &\quad \left. + \frac{2x\mu_E\sigma_P^2 - \mu_P^2\sigma_E^2 - x^2\sigma_P^2 - \mu_E^2\sigma_P^2}{2\sigma_P^2\sigma_E^2} \right\}. \end{aligned} \tag{5.11}$$

Using the property of Gaussian function integral

$$\int_{-\infty}^{\infty} \exp\left\{-ap^2 + bp + c\right\} \mathrm{d}p = \sqrt{\frac{\pi}{a}} \exp\left\{\frac{b^2}{4a} + c\right\}, \tag{5.12}$$

we get the marginal probability density function

$$\begin{aligned} h_X(x) &= \int_{-\infty}^{\infty} g_{P,X}(p,x)\mathrm{d}p \\ &= \frac{1}{\sqrt{2\pi\sigma_P^2}\sqrt{2\pi\sigma_E^2}} \sqrt{\frac{\pi}{\frac{\sigma_P^2+\sigma_E^2}{2\sigma_P^2\sigma_E^2}}} \exp\left\{ \frac{\left(\frac{\mu_P\sigma_E^2 + x\sigma_P^2 - \mu_E\sigma_P^2}{\sigma_P^2\sigma_E^2}\right)^2}{4\left(\frac{\sigma_P^2+\sigma_E^2}{2\sigma_P^2\sigma_E^2}\right)} + \frac{2x\mu_E\sigma_P^2 - \mu_P^2\sigma_E^2 - x^2\sigma_P^2 - \mu_E^2\sigma_P^2}{2\sigma_P^2\sigma_E^2} \right\} \\ &= \frac{1}{\sqrt{2\pi\left(\sigma_P^2+\sigma_E^2\right)}} \exp\left\{-\frac{(\mu_P - x + \mu_E)^2}{2(\sigma_P^2+\sigma_E^2)}\right\}. \end{aligned} \tag{5.13}$$

The conditional probability density function is then derived as

$$\begin{aligned} f_P(p|X=x) &= \frac{g_{P,X}(p,x)}{h_X(x)} \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma_P^2\sigma_E^2}{\sigma_P^2+\sigma_E^2}}} \exp\left\{-\frac{(p-\mu_P)^2}{2\sigma_P^2} - \frac{(x-p-\mu_E)^2}{2\sigma_E^2} + \frac{(\mu_P - x + \mu_E)^2}{2(\sigma_P^2+\sigma_E^2)}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_C^2(x)}} \exp\left\{-\frac{\left(p-\mu_C(x)\right)^2}{2\sigma_C^2(x)}\right\}. \end{aligned} \tag{5.14}$$

□

Let us analyze the situation in which we assume observations to follow the Ornstein–Uhlenbeck process P_{T_i} but they actually follow the noisy process X_i. From (5.3) and (5.6) we have unconditional moments

$$\begin{aligned} \mathrm{E}[X_i] &= \mathrm{E}[P_{T_i}], \\ \mathrm{var}[X_i] &= \mathrm{var}[P_{T_i}] + \omega^2, \\ \mathrm{cov}[X_i, X_j] &= \mathrm{cov}[P_{T_i}, P_{T_j}], \quad i \neq j. \end{aligned} \tag{5.15}$$

This means that an unbiased estimate of the expected value of X_i is also an unbiased estimate of the expected value of P_{T_i}. The same applies for the autocovariance function of X_i and the autocovariance function of P_{T_i}. An unbiased estimate of the variance of X_i, on the contrary, is a positively biased estimator of the variance of P_{T_i}. Because of this, the autocorrelation function

$$\mathrm{cor}[X_i, X_j] = \mathrm{cor}[P_{T_i}, P_{T_j}] - \frac{2\tau\omega^2}{\sigma^2 + 2\tau\omega^2} e^{-\tau|T_i - T_j|}, \quad i \neq j \tag{5.16}$$

also differes from the autocorrelation function of P_{T_i}. Figure 5.2 shows the autocorrelation function of the process with and without the noise. To sum up, the misspecification of the process does not affect unconditional expected value and autocovariance estimation, but does affect unconditional variance and autocorrelation estimation.

Our goal is to estimate the parameters μ, τ, σ of the Ornstein–Uhlenbeck process P_{T_i} and the parameter ω of the market microstructure noise E_i from the observed process X_i. For this purpose, we propose the method of moments estimator, maximum likelihood estimator and the estimator reparametrizing discretized Ornstein–Uhlenbeck process with the noise as an ARMA(1,1) process.

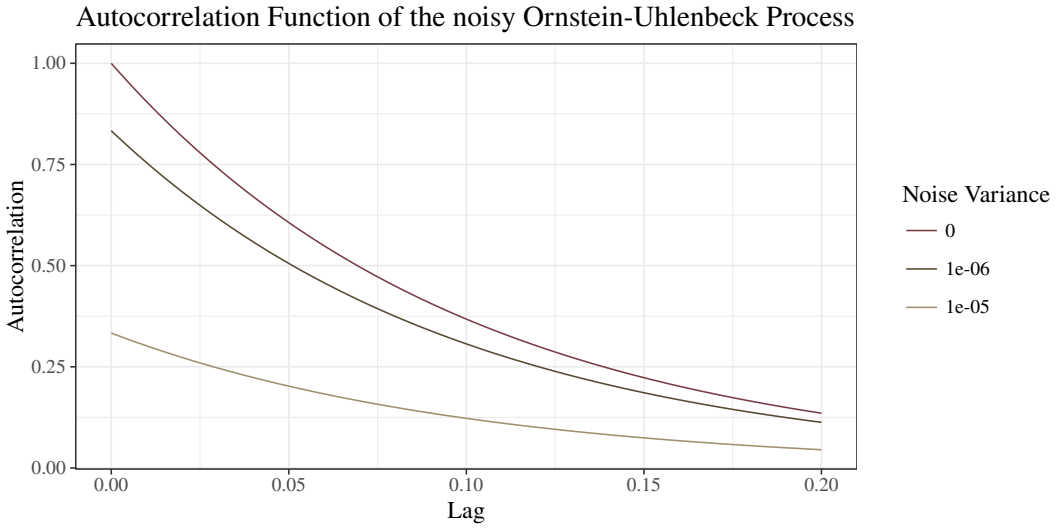

Figure 5.2: The autocorrelation function of the process X_i with parameters $\mu = 1$, $\tau = 10$, $\sigma^2 = 10^{-4}$ and various values of ω^2.

5.1.1 Method of Moments

The *method of moments* is based on relating theoretical values of random variable moments to their finite-sample estimates. The advantage of the method of moments lies in its simplicity and closed-form solution. It is often used as an initial solution for more sofisticated methods such as the maximum likelihood estimator. In this section, we assume that the times of observations T_i are equally spaced and $T_i - T_{i-1} = n^{-1}$.

Noise-Sensitive Estimator

First, we derive the method of moments for the case of the equidistantly sampled Ornstein–Uhlenbeck process with no noise. As we need to estimate parameters μ, τ and σ, we utilize three unconditional moments

$$\begin{aligned} \mathrm{E}[P_{T_i}] &= \mu, \\ \mathrm{var}[P_{T_i}] &= \frac{\sigma^2}{2\tau}, \\ \mathrm{cov}[P_{T_i}, P_{T_{i-1}}] &= \frac{\sigma^2}{2\tau} e^{-\tau n^{-1}}, \end{aligned} \tag{5.17}$$

We can estimate these moments using observed values $p_{T_0}, p_{T_1}, \ldots, pT_n$ as

$$\begin{aligned} M_{1,n} &= \frac{1}{n+1} \sum_{i=0}^{n} p_{T_i}, \\ M_{2,n} &= \frac{1}{n} \sum_{i=0}^{n} (p_{T_i} - M_{1,n})^2, \\ M_{3,n} &= \frac{1}{n-1} \sum_{i=1}^{n} (p_{T_i} - M_{1,n})(p_{T_{i-1}} - M_{1,n}), \end{aligned} \tag{5.18}$$

By solving equations

$$\mathrm{E}[P_{T_i}] = M_{1,n}, \quad \mathrm{var}[P_{T_i}] = M_{2,n}, \quad \mathrm{cov}[P_{T_i}, P_{T_{i-1}}] = M_{3,n}, \tag{5.19}$$

we get estimates

$$\begin{aligned}\hat{\mu} &= M_{1,n},\\ \hat{\tau} &= n \log \frac{M_{2,n}}{M_{3,n}},\\ \hat{\sigma}^2 &= 2nM_{2,n} \log \frac{M_{2,n}}{M_{3,n}}.\end{aligned} \tag{5.20}$$

Illustration of Bias

We illustrate the bias of the method of moments when the Ornstein–Uhlenbeck process is contaminated by the white noise with standard deviation ω. Parameter μ can be consistently estimated by sample mean. For the other two parameters, the situation is more difficult. Parameter τ can be estimated using equation

$$\begin{aligned}\tau_{P,n} &= n \log \frac{\mathrm{var}[P_{T_{i-1}}]}{\mathrm{cov}[P_{T_i}, X_{T_{i-1}}]}\\ &= -n \log \mathrm{cor}[P_{T_i}, P_{T_{i-1}}].\end{aligned} \tag{5.21}$$

The method of moments replaces the theoretical correlation in this equation by the sample correlation to estimate τ. However, if the actual process follows X_i, the equality (5.21) does not hold and instead we have

$$\begin{aligned}\tau_{X,n} &= n \log \frac{\mathrm{var}[X_{i-1}]}{\mathrm{cov}[X_i, X_{i-1}]}\\ &= -n \log \mathrm{cor}[X_i, X_{i-1}]\\ &= -n \log \left(\frac{\sigma^2}{\sigma^2 + 2\tau\omega^2} e^{-\tau(T_i - T_{i-1})} \right)\\ &= \tau_{P,n} - n \log \frac{\sigma^2}{\sigma^2 + 2\tau\omega^2}.\end{aligned} \tag{5.22}$$

The estimate $\tau_{X,n}$ is a function of the number of observations, which for $n \to \infty$ linearly diverges to infinity. Similarly, parameter σ can be estimated using equation

$$\begin{aligned}\sigma^2_{P,n} &= 2n\mathrm{var}[P_{T_i}] \log \frac{\mathrm{var}[P_{T_{i-1}}]}{\mathrm{cov}[P_{T_i}, X_{T_{i-1}}]}\\ &= -2n\mathrm{var}[P_{T_i}] \log \mathrm{cor}[P_{T_i}, P_{T_{i-1}}].\end{aligned} \tag{5.23}$$

When the process is noisy, we have

$$\begin{aligned}\sigma^2_{X,n} &= 2n\mathrm{var}[X_i] \log \frac{\mathrm{var}[X_{i-1}]}{\mathrm{cov}[X_i, X_{i-1}]}\\ &= -2n\mathrm{var}[X_i] \log \mathrm{cor}[X_i, X_{i-1}]\\ &= -2n \left(\frac{\sigma^2}{2\tau} + \omega^2 \right) \log \left(\frac{\sigma^2}{\sigma^2 + 2\tau\omega^2} e^{-\tau(T_i - T_{i-1})} \right)\\ &= \sigma^2_{P,n} + 2\tau\omega^2 - 2n\left(\frac{\sigma^2}{2\tau} + \omega^2 \right) \log \frac{\sigma^2}{\sigma^2 + 2\tau\omega^2},\end{aligned} \tag{5.24}$$

which also linearly diverges to infinity for $n \to \infty$. We show the bias of $\tau_{X,n}$ and $\sigma^2_{X,n}$ in Figure 5.3.

Noise-Robust Estimator

In the noise-robust variant of the method of moments estimator, we additionally need to estimate the standard deviation of the noise ω. As we estimate four parameters of the observed process X_i, we utilize

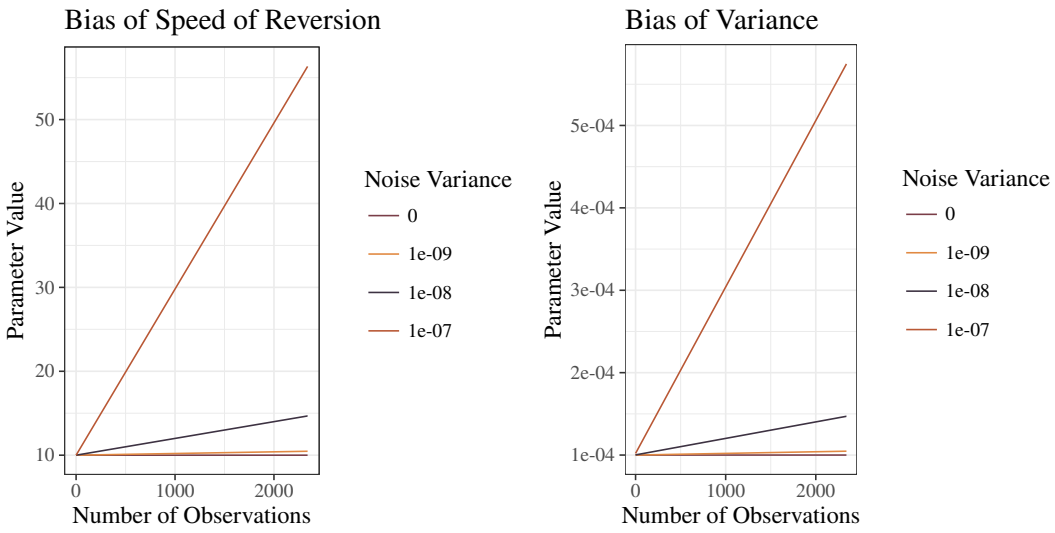

Figure 5.3: The bias of functions $\tau_{X,n}$ and $\sigma^2_{X,n}$ with parameters $\mu = 1$, $\tau = 10$, $\sigma^2 = 10^{-4}$ and various values of ω^2.

four unconditional moments

$$\begin{aligned} \mathrm{E}[X_i] &= \mu, \\ \mathrm{var}[X_i] &= \frac{\sigma^2}{2\tau} + \omega^2, \\ \mathrm{cov}[X_i, X_{i-1}] &= \frac{\sigma^2}{2\tau} e^{-\tau\Delta}, \\ \mathrm{cov}[X_i, X_{i-2}] &= \frac{\sigma^2}{2\tau} e^{-2\tau\Delta}. \end{aligned} \tag{5.25}$$

We can estimate these moments using observed values $x_0, x_1, \ldots, x_n$ as

$$\begin{aligned} M_{1,n} &= \frac{1}{n+1} \sum_{i=0}^{n} x_i, \\ M_{2,n} &= \frac{1}{n} \sum_{i=0}^{n} (x_i - M_{1,n})^2, \\ M_{3,n} &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - M_{1,n})(x_{i-1} - M_{1,n}), \\ M_{4,n} &= \frac{1}{n-2} \sum_{i=2}^{n} (x_i - M_{1,n})(x_{i-2} - M_{1,n}). \end{aligned} \tag{5.26}$$

By solving equations

$$\mathrm{E}[X_i] = M_{1,n}, \quad \mathrm{var}[X_i] = M_{2,n}, \quad \mathrm{cov}[X_i, X_{i-1}] = M_{3,n}, \quad \mathrm{cov}[X_i, X_{i-2}] = M_{4,n}, \tag{5.27}$$

we get estimates

$$\begin{aligned}
\hat{\mu} &= M_{1,n}, \\
\hat{\tau} &= \frac{1}{\Delta} \log \frac{M_{3,n}}{M_{4,n}}, \\
\hat{\sigma}^2 &= 2\frac{1}{\Delta} \frac{M_{3,n}^2}{M_{4,n}} \log \frac{M_{3,n}}{M_{4,n}}, \\
\hat{\omega}^2 &= M_{2,n} - \frac{M_{3,n}^2}{M_{4,n}}.
\end{aligned} \tag{5.28}$$

Higher moments and higher lags of autocovariance function can also be used. However, because we use this method mainly as initial estimates, we do not focus on finding the optimal set of moments.

Online and Streaming Estimation Perspective

It is natural to consider financial high-frequency data X_i as a data stream. A *streaming algorithm* can examine a sequence of inputs in a single pass only. The available memory is limited and cannot store all data. We can store only a constant number of real variables (i.e. not depending on a size of our data stream). An *online algorithm* is based on a similar idea of a single pass. The focus here is more on the updating scheme rather than the memory constraints. With a new observation, a statistic of interest is updated using the previous value of the statistic, the new observation and possibly some auxiliary variables.

We show that the method of moments estimator of the Ornstein–Uhlenbeck process is indeed a streaming and online algorithm. Our sample moments can be recursively computed as

$$\begin{aligned}
M_{1,n} &= \frac{n}{n+1} M_{1,n-1} + \frac{1}{n+1} X_n. \\
M_{2,n} &= \frac{n-1}{n} M_{2,n-1} + \frac{1}{n}(X_n - M_{1,n})^2 + \left(M_{1,n-1} - M_{1,n}\right)^2. \\
M_{3,n} &= \frac{n-2}{n-1} M_{3,n-1} + \frac{1}{n-1}\left(X_n - M_{1,n}\right)\left(X_{n-1} - M_{1,n}\right) + \left(M_{1,n-1} - M_{1,n}\right)^2 \\
&\quad + \frac{1}{n-1}\left(M_{1,n-1} - M_{1,n}\right)\left(X_1 + X_{n-1} - 2M_{1,n}\right), \\
M_{4,n} &= \frac{n-3}{n-2} M_{4,n-1} + \frac{1}{n-2}\left(X_n - M_{1,n}\right)\left(X_{n-2} - M_{1,n}\right) + \left(M_{1,n-1} - M_{1,n}\right)^2 \\
&\quad + \frac{1}{n-2}\left(M_{1,n-1} - M_{1,n}\right)\left(X_1 + X_2 + X_{n-2} + X_{n-1} - 4M_{1,n}\right).
\end{aligned} \tag{5.29}$$

As all four sample moments can be expressed in a recursive form as an update of their previous values, this estimation method is an online algorithm. It is required only to store variables $M_{1,n-1}$, $M_{2,n-1}$, $M_{3,n-1}$, $M_{4,n-1}$, X_1, X_2, X_{n-2}, X_{n-1} at time n and this estimation method is therefore a streaming algorithm.

Other examples of online or streaming algorithms from the statistics and econometrics field include the estimation and diagnostics of linear regression (Černý, 2018), estimation of GARCH process (Aknouche and Guerbyenne, 2006; Hendrych and Cipra, 2018), estimation of unobserved mean-reverting spread (Triantafyllopoulos and Montana, 2011), estimation of spot volatility (Lahalle et al., 2008; Dahlhaus and Neddermeyer, 2014) and detection of changepoints in a data stream (Bodenham and Adams, 2017).

5.1.2 Maximum Likelihood Method

A widely used method for parameter estimation is the *maximum likelihood estimator*. It maximizes the likelihood function (or, equivalently, the logarithmic likelihood function) given the observations. In our case, it utilizes the normal conditional density function for the Ornstein–Uhlenbeck process. In some simple cases, the maximum likelihood estimators are available in a closed form. Tang and Chen (2009)

present the closed-form estimates for the regularly spaced Ornstein–Uhlenbeck process without the noise. We focus on the more general case of the irregularly spaced Ornstein–Uhlenbeck process contaminated by the noise. As its likelihood is more complicated, we present it only as an optimization problem. In this section, we allow for irregularly spaced observations with deterministic times of observations T_i.

Noise-Sensitive Estimator

In the case of the Ornstein–Uhlenbeck process without the noise, the maximum likelihood estimates are obtained by maximizing the logarithmic likelihood function given by

$$L(\mu, \tau, \sigma^2) = \sum_{i=1}^{n} \log f_{P_{T_i}} \left(p_{T_i} | P_{T_{i-1}} = p_{T_{i-1}} \right), \tag{5.30}$$

where $f_{P_{T_i}} \left(p_{T_i} | P_{T_{i-1}} = p_{T_{i-1}} \right)$ is the conditional density function of the observations. According to equation (5.4), it is the conditional density function of the normal distribution

$$f_{P_{T_i}} \left(p_{T_i} | P_{T_{i-1}} = p_{T_{i-1}} \right) = \frac{1}{\sqrt{2\pi \mathrm{var}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}]}} \exp \left\{ - \frac{\left(p_{T_i} - \mathrm{E}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}] \right)^2}{2\mathrm{var}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}]} \right\}, \tag{5.31}$$

with conditional moments

$$\begin{aligned} \mathrm{E}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}] &= p_{T_{i-1}} e^{-\tau(T_i - T_{i-1})} + \mu \left(1 - e^{-\tau(T_i - T_{i-1})} \right), \\ \mathrm{var}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}] &= \frac{\sigma^2}{2\tau} \left(1 - e^{-2\tau(T_i - T_{i-1})} \right). \end{aligned} \tag{5.32}$$

The logarithmic likelihood function can be simplified to

$$L(\mu, \tau, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \log \left(2\pi \mathrm{var}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}] \right) - \frac{1}{2} \sum_{i=1}^{n} \frac{\left(p_{T_i} - \mathrm{E}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}] \right)^2}{\mathrm{var}[P_{T_i} | P_{T_{i-1}} = p_{T_{i-1}}]}. \tag{5.33}$$

The estimates are then given by

$$(\hat{\mu}, \hat{\tau}, \hat{\sigma}^2)' = \arg\max_{\mu, \tau, \sigma^2} L(\mu, \tau, \sigma^2) \qquad \text{s. t.} \qquad \sigma^2 \geq 0. \tag{5.34}$$

Noise-Robust Estimator

In the case of the Ornstein–Uhlenbeck process contaminated by the noise, the maximum likelihood estimates are obtained by maximizing the logarithmic likelihood function given by

$$L(\mu, \tau, \sigma^2, \omega^2) = \sum_{i=1}^{n} \log f_{X_i} \left(x_i | X_{i-1} = x_{i-1} \right), \tag{5.35}$$

where $f_{X_i} \left(x_i | X_{i-1} = x_{i-1} \right)$ is the conditional density function of the observations. According to Proposition 5.1, it is the conditional density function of the normal distribution

$$f_{X_i}(x_i | X_{i-1} = x_{i-1}) = \frac{1}{\sqrt{2\pi \mathrm{var}[X_i | X_{i-1} = x_{i-1}]}} \exp \left\{ - \frac{\left(x_i - \mathrm{E}[X_i | X_{i-1} = x_{i-1}] \right)^2}{2\mathrm{var}[X_i | X_{i-1} = x_{i-1}]} \right\} \tag{5.36}$$

with conditional moments

$$\begin{aligned} \mathrm{E}[X_i | X_{i-1} = x_{i-1}] &= \frac{x_{i-1}\sigma^2 + 2\tau\mu\omega^2}{\sigma^2 + 2\tau\omega^2} e^{-\tau(T_i - T_{i-1})} + \mu \left(1 - e^{-\tau(T_i - T_{i-1})} \right), \\ \mathrm{var}[X_i | X_{i-1} = x_{i-1}] &= \frac{\sigma^2\omega^2}{\sigma^2 + 2\tau\omega^2} e^{-2\tau(T_i - T_{i-1})} + \frac{\sigma^2}{2\tau} \left(1 - e^{-2\tau(T_i - T_{i-1})} \right) + \omega^2. \end{aligned} \tag{5.37}$$

The logarithmic likelihood function can be simplified to

$$L(\mu, \tau, \sigma^2, \omega^2) = -\frac{1}{2}\sum_{i=1}^{n} \log\left(2\pi \mathrm{var}[X_i|X_{i-1} = x_{i-1}]\right) - \frac{1}{2}\sum_{i=1}^{n} \frac{\left(x_t - \mathrm{E}[X_i|X_{i-1} = x_{i-1}]\right)^2}{\mathrm{var}[X_i|X_{i-1} = x_{i-1}]}. \quad (5.38)$$

The estimates are then given by

$$(\hat{\mu}, \hat{\tau}, \hat{\sigma}^2, \hat{\omega}^2)' = \arg\max_{\mu,\tau,\sigma^2,\omega^2} L(\mu, \tau, \sigma^2, \omega^2) \qquad \text{s. t.} \qquad \sigma^2 \geq 0,\ \omega^2 \geq 0. \quad (5.39)$$

5.1.3 Time Series Reparametrization

The time series reparametrization lies in the following three steps. First, we reparametrize the discretized equidistant process to a commonly used and studied time series model. Second, we estimate parameters of the time series model, e.g. by the conditional-sum-of-squares or maximum likelihood estimators. Third, we transform the estimates back to the original parametrization. One possible disadvantage is that the reparametrization does not respect parameter restrictions. In our case, σ^2 and ω^2 parameters should be non-negative, but the reparametrization allows for negative values. In this section, we assume the times of observations T_i are equally spaced and denote $\Delta = T_i - T_{i-1} = n^{-1}$.

It is well known that the discretized Ornstein–Uhlenbeck process corresponds to an AR(1) process. Aït-Sahalia et al. (2005) reparametrized the discretized Wiener process contaminated by the white noise as an ARIMA(0,1,1) process. As the discretized Wiener process without the noise is an ARIMA(0,1,0) process, the noise therefore induces a moving average component of order one. We show that the same happens for the discretized Ornstein–Uhlenbeck process contaminated by the white noise as it corresponds to an ARMA(1,1) process.

Noise-Sensitive Estimator

When the noise is not present, the discrete process P_{T_i} can be reparametrized as an AR(1) process. Using (5.2), the process P_{T_i} can be rewritten as

$$P_{T_i} = P_{T_{i-1}} e^{-\tau\Delta} + \mu(1 - e^{-\tau\Delta}) + \sigma \int_{T_{i-1}}^{T_i} e^{-\tau(\Delta - s)} \mathrm{d}W_s. \quad (5.40)$$

We denote

$$\begin{aligned} \alpha &= \mu(1 - e^{-\tau\Delta}), \\ \varphi &= e^{-\tau\Delta}. \end{aligned} \quad (5.41)$$

We further denote

$$V_i = \sigma \int_{T_{i-1}}^{T_i} e^{-\tau(\Delta - s)} \mathrm{d}W_s. \quad (5.42)$$

From equation (5.4) we have that the random variable V_i is normally distributed with variance

$$\gamma^2 = \mathrm{var}[V_i] = \frac{\sigma^2}{2\tau}\left(1 - e^{-2\tau\Delta}\right). \quad (5.43)$$

The random variable V_i is independent from $P_{T_{i-1}}$. Using (5.41) and (5.43), we can reparametrize the process (5.40) as an AR(1) process

$$P_{T_i} = \alpha + \varphi P_{T_{i-1}} + V_i, \qquad V_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \gamma^2). \quad (5.44)$$

We can estimate parameters α, φ and γ^2 by any suitable method. Finally, by solving equations

$$\begin{aligned} \hat{\alpha} &= \hat{\mu}(1 - e^{-\hat{\tau}\Delta}), \\ \hat{\varphi} &= e^{-\hat{\tau}\Delta}, \\ \hat{\gamma}^2 &= \frac{\hat{\sigma}^2}{2\hat{\tau}}\left(1 - e^{-2\hat{\tau}\Delta}\right), \end{aligned} \quad (5.45)$$

we get estimates

$$\begin{aligned}\hat{\mu} &= \frac{\hat{\alpha}}{1-\hat{\varphi}},\\ \hat{\tau} &= -\frac{1}{\Delta}\log\hat{\varphi},\\ \hat{\sigma}^2 &= -2\frac{1}{\Delta}\frac{\hat{\gamma}^2}{1-\hat{\varphi}^2}\log\hat{\varphi}.\end{aligned} \tag{5.46}$$

Noise-Robust Estimator

When the process P_{T_i} is contamined by the white noise, the discrete process X_i can be reparametrized as an ARMA(1,1) process. Using (5.2) with initial time T_{i-1}, the process X_i can be decomposed as

$$\begin{aligned}X_i &= P_{T_i} + E_i\\ &= \mu(1-e^{-\tau\Delta}) + P_{T_{i-1}}e^{-\tau\Delta} + \sigma\int_0^\Delta e^{-\tau(\Delta-s)}\mathrm{d}W_s + E_i\\ &= \mu(1-e^{-\tau\Delta}) + X_{i-1}e^{-\tau\Delta} + \sigma\int_0^\Delta e^{-\tau(\Delta-s)}\mathrm{d}W_s + E_i - E_{i-1}e^{-\tau\Delta},\end{aligned} \tag{5.47}$$

where the last equality holds because $P_{T_{i-1}} = X_{i-1} - E_{i-1}$. We denote

$$\begin{aligned}\alpha &= \mu(1-e^{-\tau\Delta}),\\ \varphi &= e^{-\tau\Delta}.\end{aligned} \tag{5.48}$$

We further denote

$$U_i = \sigma\int_0^\Delta e^{-\tau(\Delta-s)}\mathrm{d}W_s + E_i - E_{i-1}e^{-\tau\Delta}. \tag{5.49}$$

Using (5.7) we have that the random variable U_i is normally distributed with moments

$$\begin{aligned}\mathrm{E}[U_i] &= 0,\\ \mathrm{var}[U_i] &= \frac{\sigma^2}{2\tau}(1-e^{-2\tau\Delta}) + \omega^2(1+e^{-2\tau\Delta}),\\ \mathrm{cov}[U_i, U_{i-1}] &= -\omega^2 e^{-\tau\Delta},\\ \mathrm{cov}[U_i, U_{i-j}] &= 0, \qquad j > 1.\end{aligned} \tag{5.50}$$

Using substitutions (5.48) and (5.49), we rewrite (5.47) as

$$X_i = \alpha + \varphi X_{i-1} + U_i. \tag{5.51}$$

Let us define a moving average process of order one $\tilde{U}_i$, $i \geq 0$ as

$$\tilde{U}_i = \theta V_{i-1} + V_i, \qquad V_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \gamma^2). \tag{5.52}$$

Variable $\tilde{U}_i$ is then normally distributed with moments

$$\begin{aligned}\mathrm{E}[\tilde{U}_i] &= 0,\\ \mathrm{var}[\tilde{U}_i] &= \gamma^2(1+\theta^2),\\ \mathrm{cov}[\tilde{U}_i, \tilde{U}_{i-1}] &= \theta\gamma^2.\\ \mathrm{cov}[\tilde{U}_i, \tilde{U}_{i-j}] &= 0, \qquad j > 1.\end{aligned} \tag{5.53}$$

We show that the process $\{U_i\}_{i\geq 0}$ is equivalent to the process $\{\tilde{U}_i\}_{i\geq 0}$ for the right choice of γ and θ parameters satisfying

$$\begin{aligned}\mathrm{var}[U_i] &= \mathrm{var}[\tilde{U}_i],\\ \mathrm{cov}[U_i, U_{i-1}] &= \mathrm{cov}[\tilde{U}_i, \tilde{U}_{i-1}].\end{aligned} \tag{5.54}$$

The joint distribution of the process $\{U_i\}_{i\geq 0}$ is identical to the joint distribution of the process $\{\tilde{U}_i\}_{i\geq 0}$ as both processes are normally distributed with zero first moment and the same autocovariation function. We can then rewrite (5.51) as

$$X_i = \alpha + \varphi X_{i-1} + \tilde{U}_i. \tag{5.55}$$

This is an ARMA(1,1) process of the form

$$X_i = \alpha + \varphi X_{i-1} + \theta V_{i-1} + V_i, \qquad V_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \gamma^2). \tag{5.56}$$

We can estimate parameters α, φ, θ and γ^2 by any suitable method. Substitution (5.48) and equivalency (5.54) with (5.50) and (5.53) imply

$$\begin{aligned} \hat{\alpha} &= \hat{\mu}(1 - e^{-\hat{\tau}\Delta}), \\ \hat{\varphi} &= e^{-\hat{\tau}\Delta}, \\ \hat{\gamma}^2(1 + \hat{\theta}^2) &= \frac{\hat{\sigma}^2}{2\hat{\tau}}(1 - e^{-2\hat{\tau}\Delta}) + \hat{\omega}^2(1 + e^{-2\hat{\tau}\Delta}), \\ \hat{\theta}\hat{\gamma}^2 &= -\hat{\omega}^2 e^{-\hat{\tau}\Delta}. \end{aligned} \tag{5.57}$$

Finally, by solving this system of equations, we get estimates

$$\begin{aligned} \hat{\mu} &= \frac{\hat{\alpha}}{1 - \hat{\varphi}}, \\ \hat{\tau} &= -\frac{1}{\Delta}\log\hat{\varphi}, \\ \hat{\sigma}^2 &= -2\frac{1}{\Delta}\frac{\hat{\gamma}^2(\hat{\varphi} + \hat{\theta}^2\hat{\varphi} + \hat{\theta}\hat{\varphi}^2 + \hat{\theta})}{\hat{\varphi}(1 - \hat{\varphi}^2)}\log\hat{\varphi}, \\ \hat{\omega}^2 &= -\frac{\hat{\theta}\hat{\gamma}^2}{\hat{\varphi}}. \end{aligned} \tag{5.58}$$

5.2 Application to Pairs Trading Strategy

As an application of the noise-robust high-frequency estimators of the Ornstein–Uhlenbeck process, we analyze the *pairs trading strategy* based on stochastic spread. This application allows us to evaluate the added value of the noise-robust estimators compared to the noise-sensitive estimators in terms of profit.

The idea behind pairs trading lies in taking an advantage of financial markets that are out of equilibrium. When some pairs of prices exhibit strong similarity in the long run and they are currently far enough from their equilibrium, traders might profit by taking a long position in one security and a short position in the other security in a predetermined ratio. When the price spread reverts back to its mean level, the positions are closed and the profit is made. Typically, two similar commodities (e.g. West Texas Intermediate crude oil and Brent crude oil) or two stocks of companies in the same industry (e.g. Coca-Cola company and Pepsi company) are traded. The pairs trading can be further generalized to trading of groups of securities. For a comprehensive review of the pairs trading literature, see Krauss (2017). There are three commonly used approaches in pairs trading.

- The *distance approach* was introduced by Gatev et al. (2006). In this method, a pair of comoving and potentionally profitable securities is selected according to some distance metric in the so-called formation period. Trading itself is controlled by elementary non-parametric entry and exit rules. For example, a trade is opened when the spread diverges by two standard deviations and closed when the spread returns to its long-term equilibrium. In comparison to other methods, the distance method is rather simple and easy to use. Other studies following the distance approach include Perlin (2009), Bowen et al. (2010), Do and Faff (2010), Mori and Ziobrowski (2011), Broussard and Vaihekoski (2012), Do and Faff (2012), Huck (2013), Huck (2015), Jacobs (2015), Jacobs and Weber (2015), Bowen and Hutchinson (2016) and Rinne and Suominen (2017).

- The *cointegration approach* was described in detail by Vidyamurthy (2004). It relies on cointegration analysis and suitable pairs of securities are identified by cointegration tests. Trading signals are again some elementary rules just as in the case of the distance method. Other studies following the cointegration approach include Wahab et al. (1994), Girma and Paulson (1999), Simon (1999), Dunis and Lequeux (2000), Emery and Liu (2002), Liu and Chou (2003), Lin et al. (2006), Puspaningrum et al. (2010), Cheng et al. (2011), Gutierrez and Tse (2011), Peters et al. (2011), Galenko et al. (2012), Caldeira and Moura (2013), Li et al. (2014), Miao (2014) and Clegg and Krauss (2018).

- The *stochastic spread approach* was suggested by Elliott et al. (2005). The focus of this approach is more on the time series analysis of a given pair of securities rather than the selection of securities. Typically, the spread process is modeled by a mean-reverting autoregressive process with discrete time or the Ornstein–Uhlenbeck process with continuous time. Entry and exit signals are generated in an optimal way (e.g. maximazing mean profit). Other studies following the stochastic spread approach include Bertram (2009), Bertram (2010), Kanamura et al. (2010), Triantafyllopoulos and Montana (2011), Cummins and Bucca (2012), Bogomolov (2013), Song and Zhang (2013), Zeng and Lee (2014), Leung and Li (2015), De Moura et al. (2016), Göncü and Akyildirim (2016) and Liu et al. (2017).

A few of the above mentioned studies deal with intraday trading and high-frequency data. Namely, Bowen et al. (2010) use 60-minute data, Dunis and Lequeux (2000) 30-minute data, Miao (2014) 15-minute data, Peters et al. (2011) 10-minute data and Liu et al. (2017) 5-minute data. However, none of these studies utilizes tick data. Our aim is therefore to bring an insight into the pairs trading strategy in the context of ultra-high-frequency data.

In our study, we focus on stocks of 7 Big Oil companies traded on New York Stock Exchange (NYSE). Stocks of Chevron (CVX), Phillips 66 (PSX) and ExxonMobil (XOM) companies are primarily listed on NYSE while stocks of BP (BP), Eni (E), Royal Dutch Shell (RDS-A) and Total (TOT) companies are primarily listed on some other exchanges and only secondary listed on NYSE. The analyzed stocks are described in Appendix A in more detail. As all 7 companies are in the same industry and they are all influenced by crude oil prices, some degree of comovement of their stock prices can be expected. The 7 considered stocks can form 21 possible pairs in total. Figure 5.4 illustrates daily price movement of BP and RDS-A stocks. Our pairs trading strategy falls into the category of stochastic spread methods and consists of the following steps.

- First, we analyze historical intraday data. We separately estimate the parameters of the Ornstein–Uhlenbeck process for each considered pair on each considered day. Some days exhibit strong mean-reversion suggesting the Ornstein–Uhlenbeck process with high speed of reversion as illustrated in the upper plot of Figure 5.5 while others exhibit random walk behaviour suggesting the Wiener process as illustrated in the lower plot of Figure 5.5. Days with high speed of reversion and high volatility offer more opportunities for profit.

- Second, we utilize time series models to capture time-varying nature of daily parameter values. This allows us to predict future parameter values. In other words, we assume the prices during a single day in future to follow the Ornstein–Uhlenbeck process with forecasted parameters.

- Third, assuming the Ornstein–Uhlenbeck process with specific parameters, we find the optimal entry and exit signals together with the expected profit and the variance of the profit for a given pair on a given day. Based on the values of the mean profit and its variance, we decide whether to trade the given pair on the given day or not. If the decision is positive, the trading is then controlled by the optimal entry and exit signals.

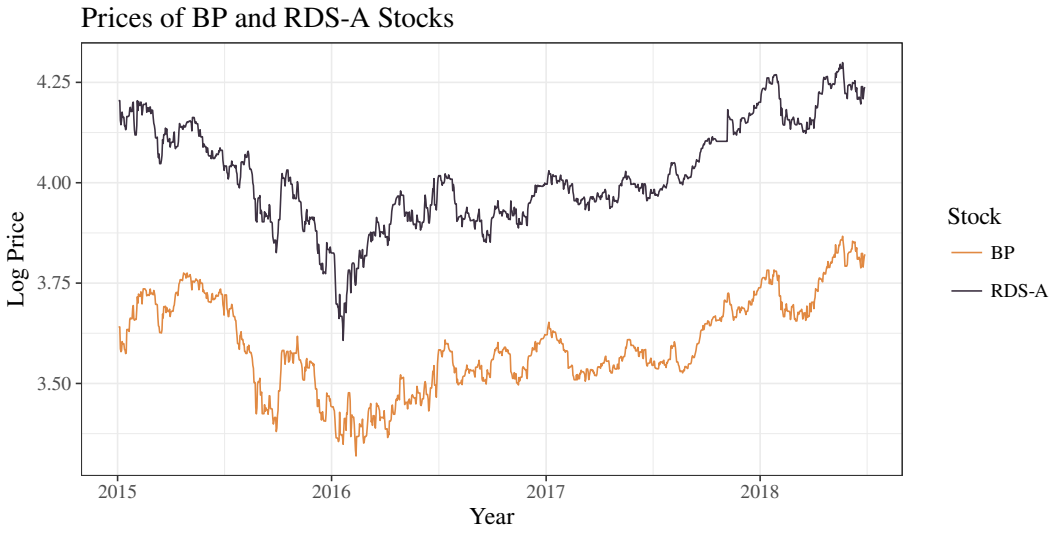

Figure 5.4: Prices of BP and RDS-A stocks.

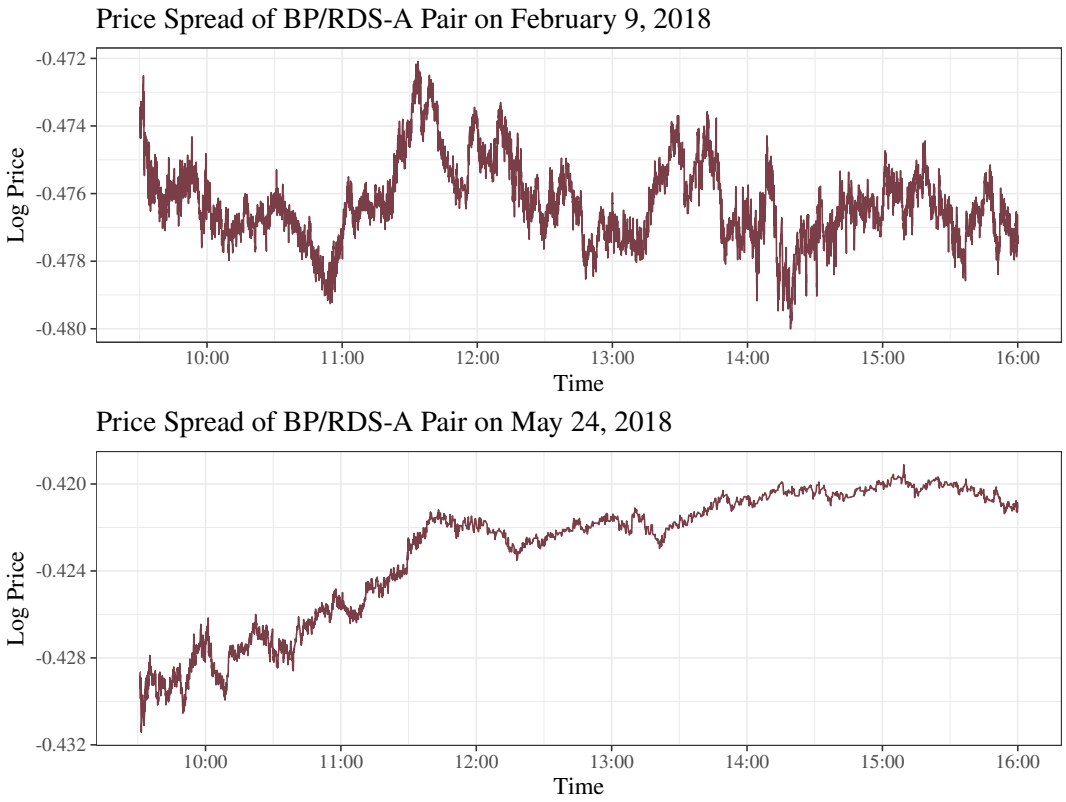

Figure 5.5: Price spread of BP/RDS-A pair resembling Ornstein–Uhlenbeck process on February 9, 2018 and Wiener process on May 24, 2018.

5.2.1 Estimators Performance

We compare the presented noise-sensitive and noise-robust estimators of the Ornstein–Uhlenbeck process together with non-parametric estimators of variance. For the methods requiring equidistant observations, we aggregate tick data to 1-minute data using the *previous tick method*. The noise-sensitive method of moments is denoted as 1MIN-MOM and its noise-robust modification as 1MIN-MOM-NR. The approach based on the reparametrization to time series models estimates parameters by the conditional sum-of-squares and is denoted as 1MIN-AR for the noise-sensitive reparametrization to the AR(1) process and 1-MIN-ARMA-NR for the noise-robust reparametrization to the ARMA(1,1) process. The noise-sensitive and noise-robust maximum likelihood estimators based on 1-minute data are denoted as 1MIN-MLE and 1MIN-MLE-NR respectively while their tick-data counterparts are denoted as TICK-MLE and TICK-MLE-NR respectively.

The variance of the process can also be estimated by non-parametric methods. Since the parameter σ^2 of the Ornstein–Uhlenbeck process is equal to the quadratic variation of the process over time interval $(0, 1)$, we can estimate σ^2 by non-parametric estimators of quadratic variation. The straightforward estimator of quadratic variation is the realized variance. However, as shown for example by Hansen and Lunde (2006), it is biased and inconsistent in the presence of the market microstructure noise. We denote the realized variance based on 1-minute data as 1MIN-RV and TICK-RV for tick data. There are many noise-robust alternatives for the non-parametric quadratic variation estimation in the literature. One of the method is the realized kernel estimator proposed by Barndorff-Nielsen et al. (2008). We utilize the variant with the modified Tukey-Hanning kernel and denote it as 1MIN-RK-TH2 for 1-minute data and TICK-RK-TH2 for tick data. Another noise-robust method is the pre-averaging estimator of Jacod et al. (2009). It is denoted as 1MIN-PAE for 1-minute data and TICK-PAE for tick data. The variance of the noise ω^2 is estimated using biased realized variance RV_n adjusted for the noise-robust estimate RM_n (either the realized kernel or the pre-averaging estimate) $\hat{\omega}^2 = (RV_n - RM_n)/2n$, where n is the number of observations.

Simulation Study

We evaluate the finite-sample performance of the proposed estimators using simulations. We simulate the observed price process as the Ornstein–Uhlenbeck process with parameters $\mu = 10^{-1}$, $\tau = 10$ and $\sigma^2 = 10^{-4}$ contaminated by the independent Gaussian white noise with variance $\omega^2 = 10^{-8}$. We select the values of parameteres to resemble values reported in the empirical study of the 7 Big Oil companies. The simulated observations are irregularly spaced and the times of observations are generated by the Poisson point process. We perform the simulation 10 000 times, each with 23 400 observations. The number of observations corresponds to durations between price changes to be one second on average during 6.5 hours long trading day.

The results of simulations are reported in Table 5.1. We compare the estimators by mean absolute errors of estimated parameters. Generally, the noise-robust estimators based on tick data outperform the noise-robust estimators based on 1-minute data while the noise-sensitive estimators based on tick data are outperformed by the noise-sensitive estimators based on 1-minute data. This is because the noise-robust estimators can utilize the additional information from tick data while the noise-sensitive estimators are more biased with more observations. We further investigate this property in Figure 5.6 in the empirical study. When considering only 1-minute data, the best parametric estimator is the 1MIN-ARMA-NR. However, for the volatility estimation based on 1-minute data, non-parametric estimators 1MIN-RK-TH2 and 1MIN-PAE are superior to parametric estimators. When considering both tick data and 1-minute aggregation, the best parametric estimator is the TICK-MLE-NR. The shortcoming of this estimator is slightly worse estimation of μ, but it is compensated by the lowest mean absolute error of τ and σ^2 parameters. On the other hand, its noise-sensitive variant TICK-MLE performs very poorly due to the misspecification of the process (omitting the noise). The TICK-MLE-NR even outperforms the non-parametric TICK-RK-TH2 and TICK-PAE estimators in the estimation of the variance σ^2. In the rest of the study, we work solely with tick data and focus only on the TICK-MLE and TICK-MLE-NR estimators.

Method	μ	τ	σ	ω
1MIN-MOM	$7.3437 \cdot 10^{-4}$	$0.4932 \cdot 10^{2}$	$0.9855 \cdot 10^{-2}$	-
1MIN-MOM-NR	$7.3437 \cdot 10^{-4}$	$0.2128 \cdot 10^{2}$	$0.4392 \cdot 10^{-2}$	$3.6847 \cdot 10^{-5}$
1MIN-AR	$7.3437 \cdot 10^{-4}$	$0.4893 \cdot 10^{2}$	$0.9905 \cdot 10^{-2}$	-
1MIN-ARMA-NR	$7.3437 \cdot 10^{-4}$	$0.1414 \cdot 10^{2}$	$0.2782 \cdot 10^{-2}$	$2.8072 \cdot 10^{-5}$
1MIN-MLE	$7.4366 \cdot 10^{-4}$	$0.4898 \cdot 10^{2}$	$0.9905 \cdot 10^{-2}$	-
TICK-MLE	$7.3500 \cdot 10^{-4}$	$9.4178 \cdot 10^{2}$	$8.6879 \cdot 10^{-2}$	-
1MIN-MLE-NR	$7.4366 \cdot 10^{-4}$	$0.2124 \cdot 10^{2}$	$0.4464 \cdot 10^{-2}$	$3.6924 \cdot 10^{-5}$
TICK-MLE-NR	$7.4058 \cdot 10^{-4}$	$0.0586 \cdot 10^{2}$	$0.0259 \cdot 10^{-2}$	$0.0652 \cdot 10^{-5}$
1MIN-RV	-	-	$0.9893 \cdot 10^{-2}$	-
TICK-RV	-	-	$1.3830 \cdot 10^{-2}$	-
1MIN-RK-TH2	-	-	$0.1380 \cdot 10^{-2}$	$0.3247 \cdot 10^{-5}$
TICK-RK-TH2	-	-	$0.0790 \cdot 10^{-2}$	$0.1812 \cdot 10^{-5}$
1MIN-PAE	-	-	$0.0319 \cdot 10^{-2}$	$0.0823 \cdot 10^{-5}$
TICK-PAE	-	-	$0.0327 \cdot 10^{-2}$	$0.0835 \cdot 10^{-5}$

Table 5.1: Mean absolute errors of parameters estimated by various methods from the simulated noisy Ornstein–Uhlenbeck process with true parameters $\mu = 1$, $\tau = 10$, $\sigma^2 = 10^{-4}$ and $\omega^2 = 10^{-8}$. Estimators based on 1-minute data are denoted as 1MIN while estimators based on tick data as TICK. The noise-sensitive method of moments is denoted as MOM, the noise-robust method of moments as MOM-NR, the noise-sensitive AR(1) reparametrization as AR, the noise-robust ARMA(1,1) reparametrization as ARMA-NR, the noise-sensitive maximum likelihood as MLE, the noise-robust maximum likelihood as MLE-NR, the realized variance as RV, the realized kernel estimator as RK-TH2 and the pre-averaging estimator as PAE.

Evidence in Stock Prices

We analyze high-frequency data of the 7 Big Oil stocks traded on NYSE from January 2, 2015 to June 29, 2018 consisting of 880 trading days. We perform data cleaning procedure described in Section 2.1.1.

The first question is whether the market microstructure noise is indeed present in the observed prices. As the high-frequency data studies agree that the noise is present (e.g. Hansen and Lunde, 2006), we address the issue only briefly using a graphical analysis. In Figure 5.6, we adopt the so-called volatility signature plot introduced by Andersen et al. (2000). The plot shows the dependence of the average estimated value of variance on the sampling interval. For tick data, the sampling interval k refers to data consisting of each k-th observation. For example, value 1 corresponds to complete tick data while value 2 corresponds to every second observation being dropped. The number of observations for sampling interval k is approximately n/k, where n is the number of observations of complete tick data. We can see in Figure 5.6 that the variance estimated by the noise-sensitive method increases with the number of observations. This is exactly the behaviour caused by the market microstructure noise. Noise-robust estimator, on the other hand, sticks around a constant value. For $k = 1$, the bias of the TICK-MLE method is quite big causing very distorted image of the price volatility.

The second question about the market microstructure noise is whether the independent white noise assumption is met in practice. Hansen and Lunde (2006) analyze stocks traded on the NYSE and NASDAQ exchanges and find that the market microstructure noise present in prices is dependent in time and dependent on efficient prices. Using volatility signature plots, they notice decreasing volatility with increasing number of observations n/k, which can be explained only by the innovations in the noise process negatively correlated with the efficient returns. We further discuss this issue in Section 4.2.1. When we analyze stock prices, we achieve the same results. However, when we analyze spreads between pairs of stocks, the volatility estimated by the noise-sensitive method is distinctly increasing with shorter sam-

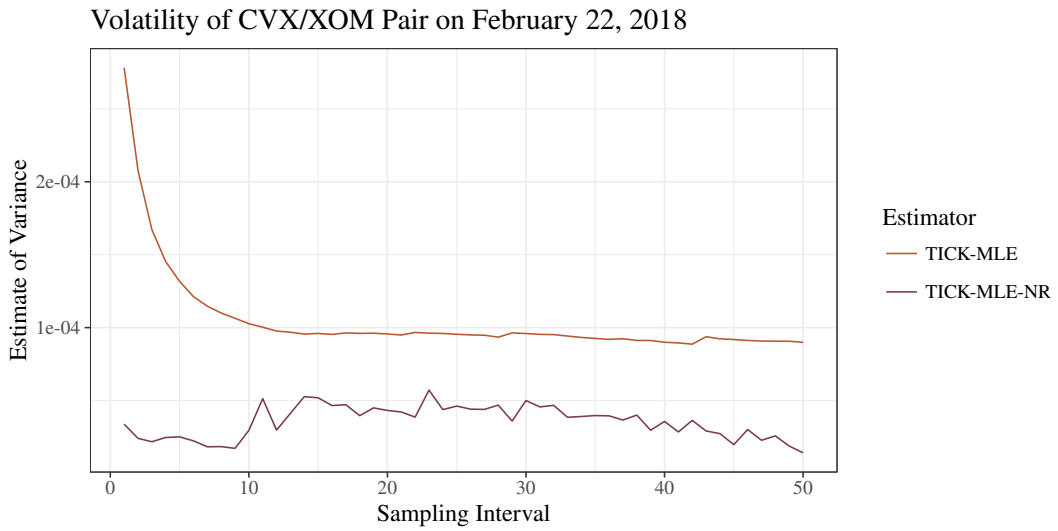

Figure 5.6: Volatility signature plot of CVX/XOM pair on February 22, 2018.

pling interval in the vast majority of days as shown in the example in Figure 5.6. We argue that the noise in the spread process has twice as many sources than the noise in a price process which diminishes dependency of the noise. For this reason, we consider the white noise assumption reasonable for the pair spread process, even when it is not suitable for the price process itself.

The average parameters estimated by the TICK-MLE and TICK-MLE-NR methods for each pair are reported in Table 5.2. The estimated means μ are quite similar for the two methods while parameters τ and σ are much higher for the TICK-MLE method. On average, the speed of reversion τ is 6.36 times higher and the standard error σ is 2.21 higher (the variance σ^2 is 4.78 higher) when estimated by the TICK-MLE method. Note that Table 5.2 reports standard deviation σ and not variance σ^2. Following our theory and Figure 5.6, we argue that the estimates of τ and σ by the TICK-MLE method are significantly biased and this estimator should be avoided. The proposed TICK-MLE-NR method, on the other hand, is not affected by the noise while utilizing all available tick data.

Numerical Estimation

The log likelihood function is maximized by numerical methods. As an initial solution, we use the method of moments estimates. The optimal solution is then found iteratively by the Subplex algorithm (SBPLX) (Rowan, 1990), a variant of the Nelder–Mead algorithm (NM) (Nelder and Mead, 1965), implemented in the open-source NLopt library (Johnson, 2019).

During the estimation, we face the following issue concerning with distribution assumptions. We consider the Ornstein–Uhlenbeck process based on the normal distribution. This is quite restrictive assumption as financial data often exhibit heavy tails and the presence of jumps. Although somewhat rare, large jumps can cause problems for the estimators based on the maximum likelihood. A large jump over short period of time is not consistent with the assumed volatility process which is proportional to the time period and the maximum likelihood estimator attributes this jump to the noise component. This results in zero variance of the Ornstein–Uhlenbeck process σ^2 and overestimation of the noise variance ω^2. To avoid such problems, we consider large jumps to be outliers and remove them from data for the estimation purposes. We remove 1 % of all observations with the lowest log likelihood at initial parameter values. In the subsequent analysis, removed observations are again included. An inclusion of jumps in the model is possible improvement of the method which we leave for the future research.

	TICK-MLE Estimator			TICK-MLE-NR Estimator		
Pair	μ	τ	σ	μ	τ	σ
BP / CVX	-1.0546	29.9279	$1.7401 \cdot 10^{-2}$	-1.0552	3.9524	$0.6676 \cdot 10^{-2}$
BP / E	0.1149	35.7787	$1.7458 \cdot 10^{-2}$	0.1149	6.5204	$0.9428 \cdot 10^{-2}$
BP / PSX	-0.8416	25.5638	$1.9848 \cdot 10^{-2}$	-0.8418	4.1090	$0.8054 \cdot 10^{-2}$
BP / RDS-A	-0.4256	65.3009	$1.7258 \cdot 10^{-2}$	-0.4256	6.4025	$0.6903 \cdot 10^{-2}$
BP / TOT	-0.3324	49.7040	$1.8318 \cdot 10^{-2}$	-0.3326	7.0370	$0.8302 \cdot 10^{-2}$
BP / XOM	-0.8230	26.4986	$1.5536 \cdot 10^{-2}$	-0.8240	3.0105	$0.6908 \cdot 10^{-2}$
CVX / E	1.1697	18.0115	$1.6074 \cdot 10^{-2}$	1.1690	4.4924	$0.8804 \cdot 10^{-2}$
CVX / PSX	0.2129	21.9259	$1.8796 \cdot 10^{-2}$	0.2130	3.9646	$0.8357 \cdot 10^{-2}$
CVX / RDS-A	0.6286	28.4128	$1.7110 \cdot 10^{-2}$	0.6284	3.7525	$0.6507 \cdot 10^{-2}$
CVX / TOT	0.7223	24.6909	$1.7381 \cdot 10^{-2}$	0.7233	4.8954	$0.8190 \cdot 10^{-2}$
CVX / XOM	0.2316	33.6227	$1.5280 \cdot 10^{-2}$	0.2316	3.2192	$0.5324 \cdot 10^{-2}$
E / PSX	-0.9566	19.8339	$1.9473 \cdot 10^{-2}$	-0.9565	5.6958	$1.1065 \cdot 10^{-2}$
E / RDS-A	-0.5393	32.5781	$1.6314 \cdot 10^{-2}$	-0.5396	5.9070	$0.8110 \cdot 10^{-2}$
E / TOT	-0.4473	56.8259	$1.9806 \cdot 10^{-2}$	-0.4472	10.1327	$0.9431 \cdot 10^{-2}$
E / XOM	-0.9389	15.2862	$1.3825 \cdot 10^{-2}$	-0.9388	3.4012	$0.7846 \cdot 10^{-2}$
PSX / RDS-A	0.4160	22.8768	$1.9327 \cdot 10^{-2}$	0.4161	3.6241	$0.8121 \cdot 10^{-2}$
PSX / TOT	0.5094	23.1866	$2.0295 \cdot 10^{-2}$	0.5096	5.5541	$1.0106 \cdot 10^{-2}$
PSX / XOM	0.0186	19.4030	$1.7160 \cdot 10^{-2}$	0.0186	2.8399	$0.7516 \cdot 10^{-2}$
RDS-A / TOT	0.0933	51.7991	$1.7501 \cdot 10^{-2}$	0.0934	7.1039	$0.6976 \cdot 10^{-2}$
RDS-A / XOM	-0.3985	25.8139	$1.5277 \cdot 10^{-2}$	-0.3987	2.7363	$0.5473 \cdot 10^{-2}$
TOT / XOM	-0.4907	21.0923	$1.5298 \cdot 10^{-2}$	-0.4910	3.5759	$0.7001 \cdot 10^{-2}$
Average	-0.1491	30.8635	$1.7368 \cdot 10^{-2}$	-0.1492	4.8537	$0.7862 \cdot 10^{-2}$

Table 5.2: Average values of the Ornstein–Uhlenbeck process parameters estimated by the noise-sensitive estimator TICK-MLE and the noise-robust estimator TICK-MLE-NR.

5.2.2 Models Performance

In this section, we present the time series models used for time-varying parameters of the Ornstein–Uhlenbeck process. We assume values of parameters can change on each day $i = 1, \ldots, h$. In other words, we assume the time-varying parameters to follow piecewise constant process, in which parameters are constant during the whole day. For each parameter, we consider separate model. The main purpose of these models is to forecast future values of the parameters.

Models with time-varying parameters were studied for example by Swamy and Tinsley (1980), Tucci (1995) and Cai (2007) in the context of linear regression and Peiris (1986), Bibi and Francq (2003), Francq and Gautier (2004) and Azrak and Mélard (2006) in the context of ARMA time series.

Daily mean parameter μ_i is modeled as the AR(1) process with the opening price $X_{0,i}$ on day i as an exogenous variable, i.e.

$$\mu_i = a + b\mu_{i-1} + cX_{0,i} + \varepsilon_i, \qquad i = 1, \ldots, h, \tag{5.59}$$

where a, b, c are the coefficients and ε_i is the Gaussian white noise. This is a very similar idea to the doubly mean-reverting process of Liu et al. (2017). In their study, they consider the prices to follow two mean-reverting processes on two frequencies. The low frequency corresponds to daily opening and closing prices while the high frequency corresponds to intraday prices. In our case, the low frequency mean-reverting process is represented by the autoregressive process for the daily mean parameter.

Daily speed of reversion parameter τ_i is modeled only by the mean value, i.e.

$$\tau_i = a' + \varepsilon_i', \qquad i = 1, \dots, h, \tag{5.60}$$

where a' is the coefficient and ε_i' is the Gaussian white noise. The one-step-ahead forecast of τ_i is then simply the average of its past values. We resort to this static model as we find no autocorrelation structure in the empirical study.

For the daily variance parameter σ_i^2, we utilize the HAR model of Corsi (2009). They model volatility by the realized variance over different time periods. Specifically, the daily realized variance is dependent on the realized variance of the previous day, the realized variance of the previous week and the realized variance of the previous month. This model is further discussed in Section 4.3.2. In our case, the logarithm of the parameter σ_i^2 follows the autoregressive process

$$\log \sigma_i^2 = a'' + b'' \log \sigma_{i-1}^2 + c'' \frac{1}{5} \sum_{j=1}^{5} \log \sigma_{i-j}^2 + d'' \frac{1}{22} \sum_{j=1}^{22} \log \sigma_{i-j}^2 + \varepsilon_i'', \qquad i = 1, \dots, h, \tag{5.61}$$

where a'', b'', c'', d'' are the coefficients and ε_i'' is the Gaussian white noise.

Evidence in Stock Prices

We train the models using a rolling window of 132 days (approximately 6 months) and perform one-step-ahead forecasts. The median coefficients of determination and the median absolute errors of one-step-ahead forecasts of the Ornstein–Uhlenbeck process parameters are reported in Table 5.3. We resort to the median statistics because there are several days with extreme volatility as illustrated in Figure 5.7. We find that the model (5.59) for the long-term mean parameter explains 96 % of the variance of μ_i on average while the model (5.61) for the variance parameter explains 25 % of the variance of σ_i^2 on average. By definition, the model (5.60) for speed of reversion parameter explains exactly 0 % of the variance of τ. Overall, we find that the models for μ_i and σ_i^2 parameters are satisfactory while the parameter τ_i is very hard to predict.

5.2.3 Optimal Strategy

For a given pair of stocks A and B, the pairs trading strategy is based on the *logarithmic price spread process*

$$P_t = \log\left(\frac{A_t}{B_t}\right) = \log A_t - \log B_t, \tag{5.62}$$

where A_t is the price of stock A and B_t is the price of stock B. We model the process P_t as the Ornstein–Uhlenbeck process given by (5.1) with a long-term mean μ, speed of reversion τ and instantaneous volatility $\sigma > 0$. The strategy itself consists of the following steps. First, we wait until the logarithmic price spread P_t reaches a given entry level a at time t_1. Without loss of generality, we assume the entry level a is greater than the long-term mean μ, i.e. $a > \mu$. When the entry level is reached, we simultaneously enter short position in stock A and long posistion in stock B. We expect the price of A to go down and price of B to go up, i.e. the spread to revert to its long-term mean. When the logarithmic price spread P_t reaches a given exit level $b < a$ at time t_2, we clear both positions and make profit. The profit from stock A in terms of continuous compound rate of return is $\log A_{t_1} - \log A_{t_2}$ while the profit from stock B is $\log B_{t_2} - \log B_{t_1}$. Adding a transaction cost c for the whole pairs trade, we have the total profit

$$\begin{aligned} r &= \log A_{t_1} - \log A_{t_2} + \log B_{t_2} - \log B_{t_1} - c \\ &= P_{t_1} - P_{t_2} - c \\ &= a - b - c. \end{aligned} \tag{5.63}$$

	MedR^2		MedAE		
Pair	μ	σ^2	μ	τ	σ^2
BP / CVX	0.9675	0.3551	$4.0942 \cdot 10^{-3}$	1.9615	$0.8709 \cdot 10^{-5}$
BP / E	0.9827	0.2072	$3.1735 \cdot 10^{-3}$	3.5419	$1.5074 \cdot 10^{-5}$
CVX / E	0.9487	0.3146	$5.1204 \cdot 10^{-3}$	2.3681	$1.2568 \cdot 10^{-5}$
BP / PSX	0.9586	0.2875	$5.4505 \cdot 10^{-3}$	2.1768	$1.2934 \cdot 10^{-5}$
CVX / PSX	0.9436	0.2286	$5.7493 \cdot 10^{-3}$	2.2265	$1.1816 \cdot 10^{-5}$
E / PSX	0.9726	0.3406	$5.9111 \cdot 10^{-3}$	2.6825	$2.0679 \cdot 10^{-5}$
BP / RDSA	0.9816	0.2018	$2.4141 \cdot 10^{-3}$	3.9130	$0.6701 \cdot 10^{-5}$
CVX / RDSA	0.9660	0.2744	$4.7031 \cdot 10^{-3}$	1.9110	$0.8604 \cdot 10^{-5}$
E / RDSA	0.9768	0.2564	$3.3555 \cdot 10^{-3}$	3.3164	$1.3228 \cdot 10^{-5}$
PSX / RDSA	0.9403	0.3339	$5.6414 \cdot 10^{-3}$	1.8570	$1.1742 \cdot 10^{-5}$
BP / TOT	0.9653	0.2604	$2.9551 \cdot 10^{-3}$	3.7663	$1.1387 \cdot 10^{-5}$
CVX / TOT	0.9555	0.3332	$5.0260 \cdot 10^{-3}$	2.5196	$1.2753 \cdot 10^{-5}$
E / TOT	0.9681	0.2514	$2.6592 \cdot 10^{-3}$	5.1985	$1.5872 \cdot 10^{-5}$
PSX / TOT	0.9361	0.3448	$5.5260 \cdot 10^{-3}$	2.7046	$1.7659 \cdot 10^{-5}$
RDSA / TOT	0.9830	0.2249	$2.5487 \cdot 10^{-3}$	3.8992	$1.0673 \cdot 10^{-5}$
BP / XOM	0.9689	0.1988	$4.7299 \cdot 10^{-3}$	1.6379	$0.6464 \cdot 10^{-5}$
CVX / XOM	0.9523	0.2228	$3.7614 \cdot 10^{-3}$	1.8123	$0.5655 \cdot 10^{-5}$
E / XOM	0.9127	0.1664	$5.0604 \cdot 10^{-3}$	1.9967	$1.1112 \cdot 10^{-5}$
PSX / XOM	0.9607	0.1839	$5.6828 \cdot 10^{-3}$	1.5600	$0.8629 \cdot 10^{-5}$
RDSA / XOM	0.9726	0.1903	$4.0018 \cdot 10^{-3}$	1.3017	$0.5872 \cdot 10^{-5}$
TOT / XOM	0.9624	0.1673	$4.4425 \cdot 10^{-3}$	1.9390	$0.9087 \cdot 10^{-5}$
Average	0.9608	0.2545	$4.3813 \cdot 10^{-3}$	2.5853	$1.1296 \cdot 10^{-5}$

Table 5.3: Median coefficients of determination MedR^2 and median absolute errors MedAE of one-step-ahead forecasts of the Ornstein–Uhlenbeck process parameters estimated by the TICK-MLE-NR method.

After the trade, we again wait for the spread P_t to reach the entry level a and repeat the whole trading cycle. The trading cycle is thus composed of two parts. In the first part, we hold short and long positions in stocks A and B respectively, while in the second part, we wait until the next trading signal. We denote the duration of the trading cycle as

$$\mathcal{T} = \mathcal{T}_{a\to b} + \mathcal{T}_{b\to a}, \tag{5.64}$$

where $\mathcal{T}_{a\to b}$ is the first passage time from a to b and $\mathcal{T}_{b\to a}$ is the first passage time from b to a.

In this strategy, we short stock A and long stock B. The opposite strategy can be adopted as well. In that case, when reaching the entry level $a' < \mu$, we long A and short B. Then, when reaching the exit level $b' > a'$, we make profit $b' - a' - c$. Since the Ornstein–Uhlenbeck process is symmetric around μ, the second strategy for stocks A and B is identical to the first strategy for stocks B and A. For simplicity, we focus only on the first case for stocks A and B with $a > \mu$.

Our goal is to determine the values of entry signal a and exit signal b for a given transaction cost c and static parameters μ, τ and σ. To optimally select signals a and b, we closely follow the framework of Bertram (2009) and Bertram (2010), also adopted by Cummins and Bucca (2012), Zeng and Lee (2014) and Göncü and Akyildirim (2016). All these papers focus on maximazing the expected profit while Bertram (2010) also deals with maximazing the Sharpe ratio. In our work, we adopt the *mean-variance optimization* related to the modern portfolio theory pioneered by Markowitz (1952). We formulate the problem as the maximization of the expected profit for a given level of maximum variance. If the level of maximum variance is large enough, the problem reduces to the maximization of the expected profit.

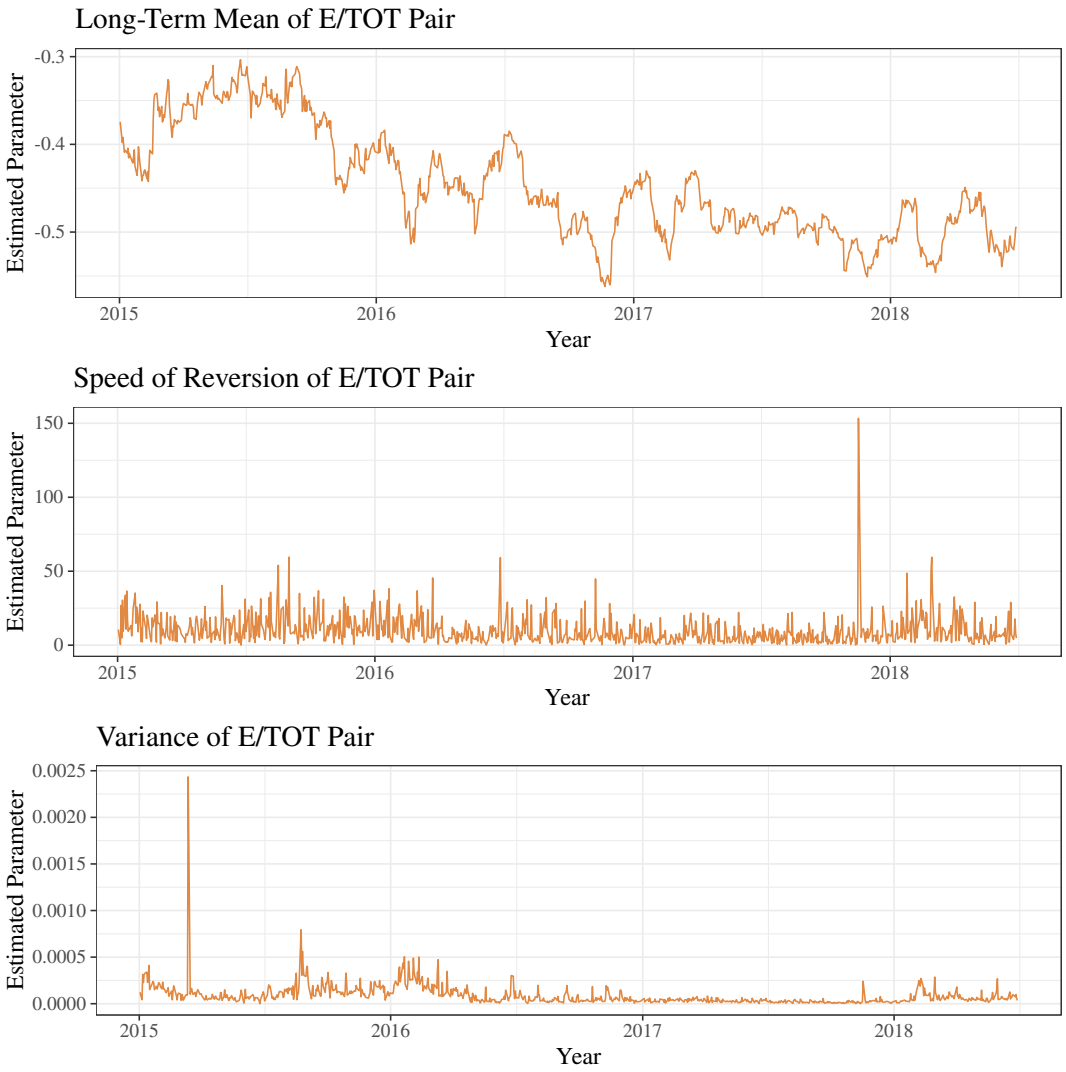

Figure 5.7: Estimated daily parameters of the Ornstein–Uhlenbeck process of E/TOT pair.

Let Z_t be the random profit of the strategy over time t. For a given entry signal a, exit signal b and transaction cost c, it is equal to

$$Z_t = (a - b - c)\, N_t, \tag{5.65}$$

where N_t is the counting process representing the number of trades during time t. Because the profit per trade $a - b - c$ is always constant, the only randomness lies in the process N_t. Next, let us define the *expected profit per unit time* and *variance of profit per unit time* as

$$\begin{aligned} Z_M &= \lim_{t\to\infty} \frac{\mathrm{E}[Z_t]}{t} = \lim_{t\to\infty} \frac{(a-b-c)\,\mathrm{E}N_t}{t}, \\ Z_V &= \lim_{t\to\infty} \frac{\mathrm{var}[Z_t]}{t} = \lim_{t\to\infty} \frac{(a-b-c)^2\,\mathrm{var}\,N_t}{t}. \end{aligned} \tag{5.66}$$

As in Bertram (2010), using the results from the renewal theory for the expected value and variance (see e.g. Cox, 1962; Cox and Miller, 1965), we obtain

$$\begin{aligned} Z_M &= \frac{a-b-c}{\mathrm{E}\mathcal{T}}, \\ Z_V &= \frac{(a-b-c)^2\,\mathrm{var}\mathcal{T}}{(\mathrm{E}\mathcal{T})^3}, \end{aligned} \tag{5.67}$$

where $\mathcal{T}$ is the trading cycle duration given by (5.64). In our mean-variance optimization, we utilize these two moments per unit time.

Dimensionless System

Following Bertram (2010) and Zeng and Lee (2014), we reparametrize the Ornstein–Uhlenbeck process (5.1) to the *dimensionless system*. We transform the process to

$$\tilde{P}_t = \sqrt{\frac{2\tau}{\sigma^2}}\left(P_t - \mu\right), \tag{5.68}$$

and perform the time dilation $\tilde{t} = \tau t$. Using Itô's lemma, we have

$$\mathrm{d}\tilde{P}_{\tilde{t}} = -\tilde{P}_{\tilde{t}}\mathrm{d}\tilde{t} + \sqrt{2}\mathrm{d}W_{\tilde{t}}. \tag{5.69}$$

A major advantage of this reparametrization is that it does not depend on parameters μ, τ and σ^2. For this reason, the subsequent analysis of first passage times and optimal signals is much more simple. The dimensionless system also allows us to study the impact of biased parameters on the pairs trading strategy. The reparametrized entry level, exit level and transaction cost are respectively

$$\begin{aligned}
\tilde{a} &= \sqrt{\frac{2\tau}{\sigma^2}}(a-\mu), \qquad & a &= \sqrt{\frac{\sigma^2}{2\tau}}\tilde{a} + \mu, \\
\tilde{b} &= \sqrt{\frac{2\tau}{\sigma^2}}(b-\mu), \qquad & b &= \sqrt{\frac{\sigma^2}{2\tau}}\tilde{b} + \mu, \\
\tilde{c} &= \sqrt{\frac{2\tau}{\sigma^2}}c, & c &= \sqrt{\frac{\sigma^2}{2\tau}}\tilde{c}.
\end{aligned} \tag{5.70}$$

The reparametrized duration of trading cycle is

$$\tilde{\mathcal{T}} = \tau\mathcal{T}, \qquad \mathcal{T} = \frac{1}{\tau}\tilde{\mathcal{T}}. \tag{5.71}$$

Finally, the reparametrized expected profit per unit time and variance of profit per unit time are respectively

$$\begin{aligned}
\tilde{Z}_M &= \sqrt{\frac{2}{\tau\sigma^2}}Z_M, \qquad & Z_M &= \sqrt{\frac{\tau\sigma^2}{2}}\tilde{Z}_M, \\
\tilde{Z}_V &= \frac{2}{\sigma^2}Z_V, & Z_V &= \frac{\sigma^2}{2}\tilde{Z}_V.
\end{aligned} \tag{5.72}$$

First Passage Times

The key variable in expression for moments per time (5.67) is the duration of trading cycle. In the dimensionless system, it is equal to

$$\tilde{\mathcal{T}} = \tilde{\mathcal{T}}_{\tilde{a}\to\tilde{b}} + \tilde{\mathcal{T}}_{\tilde{b}\to\tilde{a}}. \tag{5.73}$$

When assuming $\tilde{a} > 0$ and $\tilde{b} < \tilde{a}$, it is the sum of the first passage time from $\tilde{a}$ to $\tilde{b}$ and the first passage time from $\tilde{b}$ to $\tilde{a}$ defined as

$$\begin{aligned}
\tilde{\mathcal{T}}_{\tilde{a}\to\tilde{b}} &= \inf\left\{t : \tilde{P}_t < \tilde{b} | \tilde{P}_0 = \tilde{a}\right\}, \\
\tilde{\mathcal{T}}_{\tilde{b}\to\tilde{a}} &= \inf\left\{t : \tilde{P}_t > \tilde{a} | \tilde{P}_0 = \tilde{b}\right\}.
\end{aligned} \tag{5.74}$$

In this section, we present the expected value and variance of the trading cycle duration. These results are based on the explicit expressions of the first-passage-time moments derived by Ricciardi and Sato (1988). We denote the gamma function as $\Gamma(\cdot)$ and digamma function as $\psi(\cdot)$ (see Appendix D for definition).

The expected values of the first passage times from $\tilde{a}$ to $\tilde{b}$ and from $\tilde{b}$ to $\tilde{a}$ are respectively

$$\begin{aligned} \mathrm{E}\tilde{\mathcal{T}}_{\tilde{a}\to\tilde{b}} &= \phi_1(-\tilde{b}) - \phi_1(-\tilde{a}), \\ \mathrm{E}\tilde{\mathcal{T}}_{\tilde{b}\to\tilde{a}} &= \phi_1(\tilde{a}) - \phi_1(\tilde{b}), \end{aligned} \tag{5.75}$$

where

$$\phi_1(z) = \frac{1}{2}\sum_{k=1}^{\infty} \frac{\left(\sqrt{2}z\right)^k}{k!}\Gamma\left(\frac{k}{2}\right). \tag{5.76}$$

The expected value of the trading cycle duration is then

$$\mathrm{E}\tilde{\mathcal{T}} = \sum_{k=1}^{\infty} \frac{\left(\sqrt{2}\tilde{a}\right)^{2k-1} - \left(\sqrt{2}\tilde{b}\right)^{2k-1}}{(2k-1)!}\Gamma\left(\frac{2k-1}{2}\right). \tag{5.77}$$

The variances of the first passage times from $\tilde{a}$ to $\tilde{b}$ and from $\tilde{b}$ to $\tilde{a}$ are respectively

$$\begin{aligned} \mathrm{var}\tilde{\mathcal{T}}_{\tilde{a}\to\tilde{b}} &= \left(\phi_1(-\tilde{b})\right)^2 - \phi_2(-\tilde{b}) + \phi_2(-\tilde{a}) - \left(\phi_1(-\tilde{a})\right)^2, \\ \mathrm{var}\tilde{\mathcal{T}}_{\tilde{b}\to\tilde{a}} &= \left(\phi_1(\tilde{a})\right)^2 - \phi_2(\tilde{a}) + \phi_2(\tilde{b}) - \left(\phi_1(\tilde{b})\right)^2, \end{aligned} \tag{5.78}$$

where $\phi_1(z)$ is given by (5.76) and

$$\phi_2(z) = \frac{1}{2}\sum_{k=1}^{\infty} \frac{\left(\sqrt{2}z\right)^k}{k!}\Gamma\left(\frac{k}{2}\right)\left(\psi\left(\frac{k}{2}\right) - \psi(1)\right). \tag{5.79}$$

The variance of the trading cycle duration is then

$$\mathrm{var}\tilde{\mathcal{T}} = w_1(\tilde{a}) - w_1(\tilde{b}) - w_2(\tilde{a}) + w_2(\tilde{b}), \tag{5.80}$$

where

$$\begin{aligned} w_1(z) &= \left(\frac{1}{2}\sum_{k=1}^{\infty} \frac{\left(\sqrt{2}z\right)^k}{k!}\Gamma\left(\frac{k}{2}\right)\right)^2 - \left(\frac{1}{2}\sum_{k=1}^{\infty} \frac{\left(-\sqrt{2}z\right)^k}{k!}\Gamma\left(\frac{k}{2}\right)\right)^2, \\ w_2(z) &= \sum_{k=1}^{\infty} \frac{\left(\sqrt{2}z\right)^{2k-1}}{(2k-1)!}\Gamma\left(\frac{2k-1}{2}\right)\psi\left(\frac{2k-1}{2}\right). \end{aligned} \tag{5.81}$$

By applying (5.77) and (5.80) to (5.67), we have the explicit formula for the expected profit per unit time and variance of profit per unit time.

Optimization Problem

We continue to operate within the dimensionless system. For a given transaction cost $\tilde{c}$ and maximum allowed variance per unit time $\tilde{\eta}$, we find the optimal entry signal $\tilde{a}$ and exit signal $\tilde{b}$ by the optimization problem

$$\begin{aligned} \max_{\tilde{a},\tilde{b}} \quad & \tilde{Z}_M(\tilde{a},\tilde{b},\tilde{c}) \\ \text{such that} \quad & \tilde{Z}_V(\tilde{a},\tilde{b},\tilde{c}) \le \tilde{\eta}, \\ & \tilde{b} \le \tilde{a}, \\ & \tilde{a} \ge 0, \end{aligned} \tag{5.82}$$

where the expected profit per unit time $\tilde{Z}_M(\tilde{a},\tilde{b},\tilde{c})$ and variance of profit per unit time $\tilde{Z}_V(\tilde{a},\tilde{b},\tilde{c})$ are given by (5.72). This formulation corresponds to the strategy in which we short stock A and long stock

B. The formulation for the opposite positions strategy with signals $\tilde{a}' = -\tilde{a}$ and $\tilde{b}' = -\tilde{b}$ is symmetrical. In any case, it is a non-linear constrained optimization problem which we solve by numerical methods.

Let us denote $\tilde{a}^*$ the optimal entry signal, $\tilde{b}^*$ the optimal exit signal and $\tilde{Z}_M^*$ the optimal mean profit in the dimensionless system. Our numerical results show that the optimal exit signal is $\tilde{b}^* = -\tilde{a}^*$. This is the exactly same behavior as for the optimal exit signal in the case of unrestricted maximization of the expected profit and maximization of the Sharpe ratio as shown by Bertram (2010). This also means that the waiting part of the trading cycle for the strategy allowing for both long/short and short/long positions reduces to zero as the exit level is equal to the entry level for the strategy with opposite positions, i.e. $\tilde{b}^* = -\tilde{a}^* = \tilde{a}'^*$. The optimal strategy suggests to simply switch positions from short to long for stock A and from long to short for stock B at signal $-\tilde{a}^*$ and vice versa at signal $\tilde{a}^*$.

Impact of Biased Estimates

Next, we investigate the impact of biased estimates of τ and σ^2. As the optimization problem (5.82) itself is formulated in the dimensionless system, it is unaffected by the values of the Ornstein–Uhlenbeck process parameters. Reparametrization (5.68) is, however, affected. This means that the inputs to the optimization problems $\tilde{c}$ and $\tilde{\eta}$ based on the values c and η in the original parametrization can be biased. According to (5.70), the transaction cost $\tilde{c}$ is biased when the ratio of τ and σ^2 is biased. The maximum allowed variance $\tilde{\eta}$ is, similarly to the variance in (5.72), reparametrized as $\tilde{\eta} = 2\eta/\sigma^2$ and is therefore biased when σ^2 is biased. A bias can also occur when the resulting optimal signals $\tilde{a}$ and $\tilde{b}$ are transformed back to a and b in the original parametrization. According to (5.70), the entry level a and exit level b are biased when the ratio of τ and σ^2 is biased. The optimal mean profit per unit time Z_M is also biased when either τ or σ^2 is biased according to (5.72). Overall, the biased estimates of τ and σ^2 have impact on the maximum variance constraint, optimal expected profit and optimal entry and exit signals.

We illustrate the bias of the optimal expected profit when σ^2 is correctly specified but τ is considered 10 times higher than the actual value. In this case, the maximum variance constraint is unbiased. Figure 5.8 shows the efficient frontier of the mean-variance model for the optimization problem based on correctly specified as well as biased parameters. We can see that the optimization problem based on incorrectly specified parameter τ overestimates the optimal mean profit. It also finds suboptimal entry and exit signals resulting in much lower actual mean profit in comparison with the optimal mean profit based on the correct parameters.

5.2.4 Strategy Performance

For a set of parameters of the Ornstein–Uhlenbeck process obtained by the forecasting models and a given maximum allowed variance of the profit η, we find the optimal entry and exit signals together with the maximal expected profit. As the forecasted parameter values are uncertain, we trade only if the expected profit is larger than a given threshold ζ.

We use transaction costs $c = 0.0015$ per round-trip pair-trade. In the literature, this is considered as a moderate level of transaction costs. For example, Avellaneda and Lee (2010), Bertram (2010) and Liu et al. (2017) use an optimistic transaction costs level of 0.0010, Bowen et al. (2010) use a moderate level of 0.0015 and Bogomolov (2013) uses a conservative level of 0.0040.

Trading Algorithm

In this section, we summarize the proposed pairs trading strategy. We describe the strategy in general with notes regarding our specific setting. First, we need to select several parameters of the strategy. The initialization of the strategy requires the following steps.

- A set of potentially tradable pairs is selected. The number of pairs is denoted as p. In our case, we consider $p = 21$ pairs created from 7 stocks.

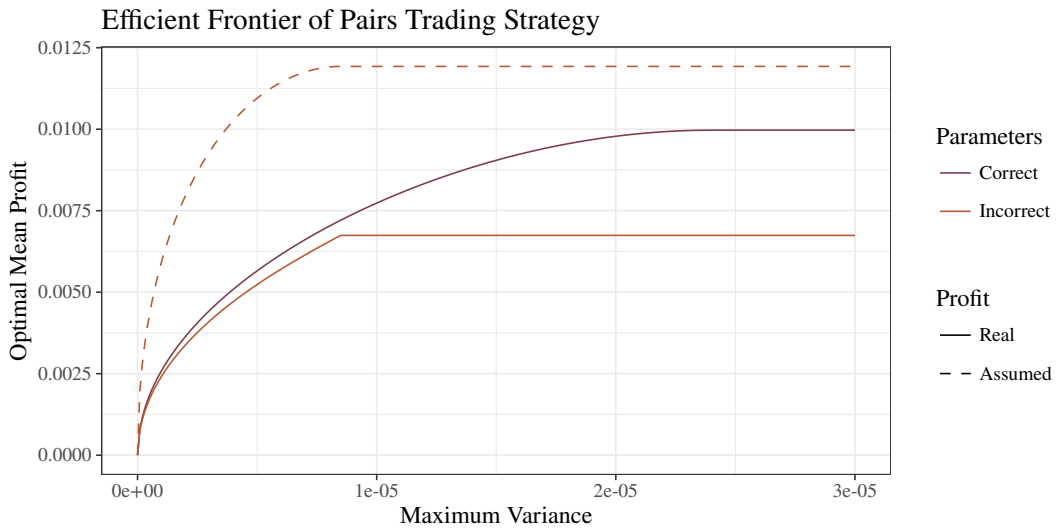

Figure 5.8: Efficient frontier of the mean-variance model for the optimization problem based on correctly specified parameters $\mu = 1$, $\tau = 10$, $\sigma^2 = 10^{-4}$ as well as incorrect parameter $\tau = 100$.

- The length of history h is selected. In our case, we use history of $h = 132$ days corresponding roughly to 6 months.
- The maximum allowed variance η for daily profit is selected. In our case, we consider $\eta = 10^{-5}$, $\eta = 5 \cdot 10^{-5}$, and $\eta = \infty$. The value $\eta = 5 \cdot 10^{-5}$ is found to yield the best results.
- The minimum allowed mean ζ for daily profit is selected. In our case, we consider $\zeta \in (0, 0.7)$. The value 0.009 is found to yield the best results.

Next, we describe the strategy for a single trading day $h + 1$. The execution of the strategy lies in the following steps.

- For each pair $j = 1, \ldots, p$ and each historical day $i = 1, \ldots, h$, the Ornstein–Uhlenbeck parameters $\mu_{j,i}$, $\tau_{j,i}$ and $\sigma^2_{j,i}$ are estimated. In our case, we use the TICK-MLE and TICK-MLE-NR estimators presented in Section 5.1.2.
- For each pair $j = 1, \ldots, p$, the models (5.59), (5.60) and (5.61) for daily Ornstein–Uhlenbeck parameters are estimated using history h. Future parameter values $\mu_{j,h+1}$, $\tau_{j,h+1}$ and $\sigma^2_{j,h+1}$ are then forecasted.
- For each pair $j = 1, \ldots, p$, the optimal entry signal a^*_j, the optimal exit signal b^*_j and the optimal mean profit $Z^*_{M,j}$ are found using (5.82). In this model, the mean profit $Z_{M,j}$ is maximized while the variance of the profit $Z_{V,j}$ is lower than η. For the opposite pairs trade, the optimal entry signal is $a'^*_j = b^*_j$, the optimal exit signal is $b'^*_j = a^*_j$ and the optimal mean profit is $Z'^*_{M,j} = Z^*_{M,j}$.
- For each pair $j = 1, \ldots, p$, it is decided whether this pair will be traded on day $h + 1$ or not. The pair will be traded if its optimal mean is higher than the selected threshold, i.e. $Z^*_{M,j} \geq \zeta$.
- For each tradable pair j, intraday prices are monitored. When the price reaches the entry level a^*_j or a'^*_j, the appropriate pairs trade is entered as described in Section 5.2.3. When the price reaches the exit level $b^*_j = a'^*_j$ or $b'^*_j = a^*_j$, long and short positions are switched. Right before the market closes, both positions are closed regardless the price.

Evidence in Stock Prices

We asses the profitability of the pairs trading strategy for the 21 pairs comprising of the 7 Big Oil companies. As we use 6 months history for the training of the forecasting models, we evaluate the strategy from the second half of the year 2015 to the second half of the year 2018.

We consider $\eta = 10^{-5}$, $\eta = 5 \cdot 10^{-5}$, and $\eta = \infty$ as levels for the maximum allowed variance. Figure 5.9 shows the total daily profit of the strategy based on 21 pairs for various values of the minimum mean profit ζ. We can see that the profit is quite sensitive to the selection of thresholds η and ζ. When the expected mean is not limited, almost all pairs are traded on all days resulting in a huge loss. When the minimum mean profit ζ is set around 0.009, the strategy based on the TICK-MLE-NR estimator performs the best and achieves daily profit up to 0.0069 in terms of the continuous compound rate of return for $\eta = 5 \cdot 10^{-5}$. When we further increase the threshold for minimum mean profit ζ, less trades are carried out and even the profitable trades are cut resulting in decline of the profit. Naturally, the profit converges to zero with increasing minimum mean profit ζ.

Interestingly, the number of trades and the profit are not evenly distributed throughout the years. Figure 5.10 shows the daily number of trades. Most trades are executed during the years 2015, 2016 and 2018 while the year 2017 is quiet period for the strategy based on the TICK-MLE-NR estimator. We attribute this to the lower volatility of the spread prices during 2017 as indicated by Figure 5.7.

Table 5.4 reports daily profit for each pair separately while Table 5.5 reports the number of trades. Generally, pairs with higher estimated values of τ and σ^2 are traded more as their expected profit is also higher. We focus on the TICK-MLE-NR estimator with the most profitable setting of the maximum variance of the profit $5 \cdot 10^{-5}$ and the minimum mean profit 0.009. Table 5.4 indicates that E/PSX, E/TOT and PSX/TOT are the most traded pairs while E/PSX and E/TOT are also the most profitable pairs. Table 5.2 shows that these pairs have the above average estimated values of τ and σ^2.

Finally, we compare the TICK-MLE and TICK-MLE-NR estimators. Figure 5.9 illustrates that both estimators have quite different ideas of the mean profit and its variance. As shown in Section 5.2.3, the values of the moments are quite distorted when the parameter estimates are biased as they are in the case of the TICK-MLE estimator. More important, even when selecting the best thresholds for the minimum mean profit ζ and the maximum variance of the profit η for each method separately, the TICK-MLE-NR estimator significantly outperforms the TICK-MLE estimator. This is because the optimization based on the TICK-MLE estimator finds suboptimal values of entry and exit signals. The TICK-MLE-NR estimator, on the other hand, finds optimal values leading to a much greater profit. This finding is the key result of our pairs trading application.

5.2.5 Discussion

We propose three estimators of the Ornstein–Uhlenbeck process directly taking the market microstructure noise into account. For initial estimates, we propose the closed-form method of moments. For equidistant sampling, we propose an approach based on the reparametrization of the process to the ARMA(1,1) process and subsequent estimation by the maximum likelihood or conditional sum-of-squares methods. For irregularly spaced observations, we propose the method based on the maximum likelihood. We show in a simulation study as well as in an empirical study that the proposed noise-robust estimators outperform the traditional estimators ignoring the noise. The behavior of the estimators is consistent with the high-frequency literature dealing with the market microstructure noise represented for example by Aït-Sahalia et al. (2005) and Hansen and Lunde (2006).

We illustrate the benefits of the proposed estimators in an application to the pairs trading strategy. However, our goal is not to present a ready-to-use pairs trading strategy with a guaranteed profit. That would be quite futile task in the ever-changing financial market. Instead, our study aims to bring an insight to the pairs trading strategy in the context of ultra-high-frequency data. We show that the strategy based

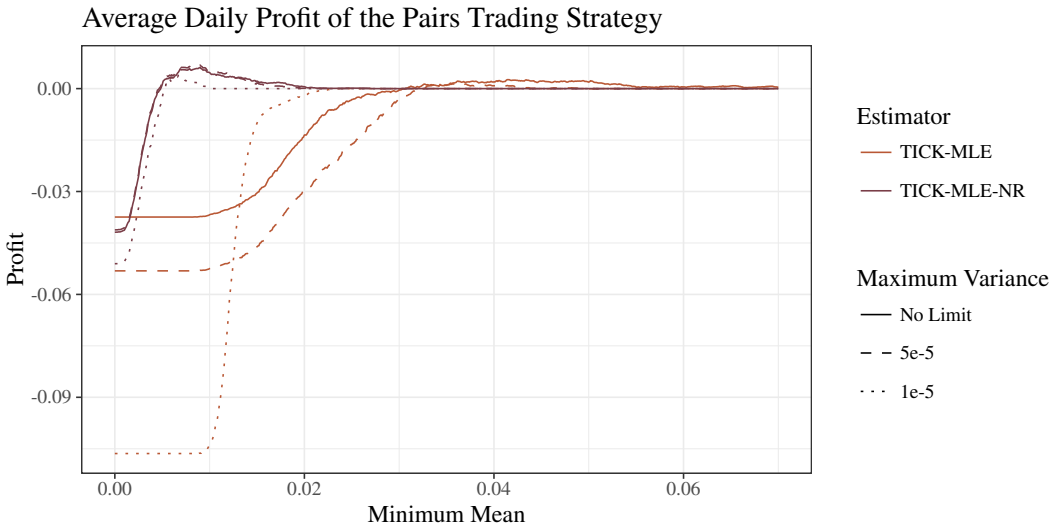

Figure 5.9: Dependence of the daily profit on the maximum variance η and minimum mean ζ for the noise-sensitive and noise-robust estimators.

	TICK-MLE Estimator			TICK-MLE-NR Estimator		
Pair	$\zeta = 0.000$	$\zeta = 0.009$	$\zeta = 0.040$	$\zeta = 0.000$	$\zeta = 0.009$	$\zeta = 0.040$
BP / CVX	-0.003235	-0.003235	0.000000	-0.002608	0.000018	0.000000
BP / E	0.000944	0.000944	0.000002	0.000611	0.001127	0.000000
BP / PSX	-0.004420	-0.004420	0.000000	-0.002965	-0.000011	0.000000
BP / RDSA	-0.002371	-0.002362	-0.000130	-0.002364	-0.000056	0.000000
BP / TOT	-0.002026	-0.002026	0.000151	-0.001873	0.000239	0.000000
BP / XOM	-0.004307	-0.004307	0.000000	-0.003813	-0.000063	0.000000
CVX / E	-0.000925	-0.000946	0.000000	-0.001030	0.000780	0.000000
CVX / PSX	-0.005322	-0.005322	0.000000	-0.004387	-0.000760	0.000000
CVX / RDSA	-0.003398	-0.003417	0.000000	-0.002683	0.000128	0.000000
CVX / TOT	-0.002068	-0.002068	0.000000	-0.001198	0.000607	0.000000
CVX / XOM	-0.004745	-0.004752	0.000000	-0.003945	0.000051	0.000000
E / PSX	-0.001614	-0.001614	0.000000	-0.000254	0.001487	0.000000
E / RDSA	0.000144	0.000144	0.000000	0.000455	0.001219	0.000000
E / TOT	0.001941	0.001941	0.000654	0.001360	0.001823	0.000000
E / XOM	-0.000196	-0.000126	0.000000	-0.000961	-0.000019	0.000000
PSX / RDSA	-0.004710	-0.004710	0.000000	-0.003634	0.000122	0.000000
PSX / TOT	-0.003660	-0.003660	0.000000	-0.001192	-0.000034	0.000000
PSX / XOM	-0.005219	-0.005219	0.000000	-0.003389	0.000031	0.000000
RDSA / TOT	-0.001325	-0.001264	0.000150	-0.001266	0.000128	0.000000
RDSA / XOM	-0.004145	-0.004069	0.000000	-0.003553	0.000000	0.000000
TOT / XOM	-0.002483	-0.002481	0.000000	-0.002504	0.000053	0.000000
Sum	-0.053140	-0.052969	0.000826	-0.041193	0.006868	0.000000

Table 5.4: Average daily profit with $\eta = 5 \cdot 10^{-5}$ and various values of ζ for the noise-sensitive and noise-robust estimators.

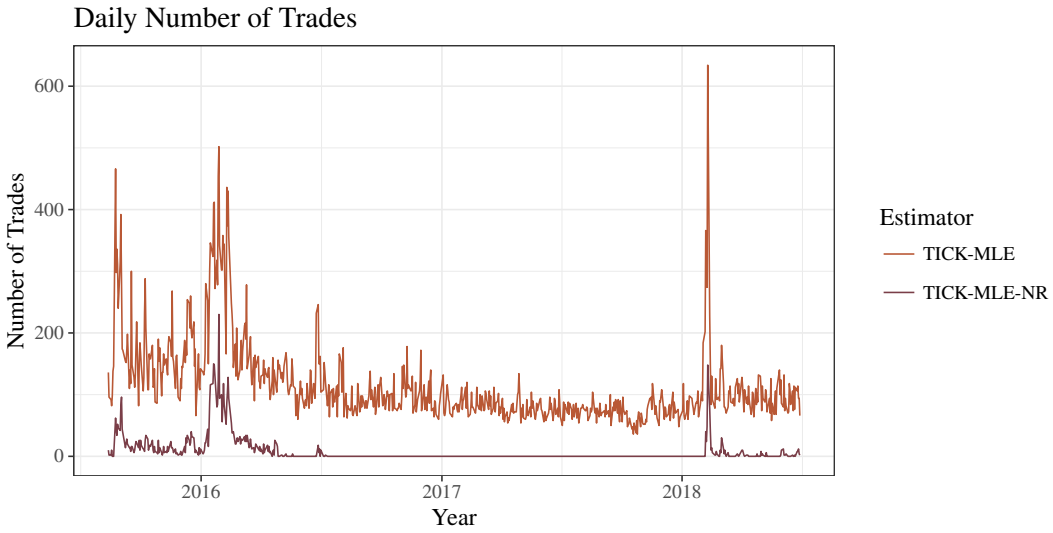

Figure 5.10: Daily number of trades with $\eta = 5 \cdot 10^{-5}$ and $\zeta = 0.009$ for the noise-sensitive and noise-robust estimators.

	TICK-MLE Estimator			TICK-MLE-NR Estimator		
Pair	$\zeta = 0.000$	$\zeta = 0.009$	$\zeta = 0.040$	$\zeta = 0.000$	$\zeta = 0.009$	$\zeta = 0.040$
BP / CVX	5.59	5.59	0.00	4.10	0.14	0.00
BP / E	5.89	5.89	0.03	4.39	0.87	0.00
BP / PSX	5.99	5.99	0.00	4.25	0.25	0.00
BP / RDSA	4.50	4.49	1.73	3.61	0.03	0.00
BP / TOT	5.24	5.24	1.47	3.84	0.25	0.00
BP / XOM	4.83	4.83	0.00	4.04	0.01	0.00
CVX / E	5.94	5.93	0.00	4.37	0.76	0.00
CVX / PSX	6.05	6.05	0.00	4.25	0.34	0.00
CVX / RDSA	5.18	5.17	0.00	3.88	0.12	0.00
CVX / TOT	6.35	6.35	0.00	4.29	0.65	0.00
CVX / XOM	4.62	4.61	0.00	4.06	0.03	0.00
E / PSX	6.79	6.79	0.00	4.88	2.21	0.00
E / RDSA	5.35	5.35	0.00	4.18	0.69	0.00
E / TOT	6.28	6.28	1.50	4.52	1.52	0.00
E / XOM	4.93	4.89	0.00	4.15	0.12	0.00
PSX / RDSA	5.74	5.74	0.00	3.97	0.25	0.00
PSX / TOT	6.42	6.42	0.00	4.61	1.39	0.00
PSX / XOM	5.47	5.47	0.00	4.30	0.10	0.00
RDSA / TOT	5.11	5.10	1.31	3.91	0.22	0.00
RDSA / XOM	4.61	4.59	0.00	4.01	0.00	0.00
TOT / XOM	5.09	5.06	0.00	4.03	0.02	0.00
Sum	115.99	115.84	6.04	87.64	9.96	0.00

Table 5.5: Average daily number of trades with $\eta = 5 \cdot 10^{-5}$ and various values of ζ for the noise-sensitive and noise-robust estimators.

on biased estimates of the Ornstein–Uhlenbeck process parameters caused by ignoring the market microstructure noise leads to a decline in profitability.

It is an inherent characteristic of the pairs trading strategy that it is sensitive to almost all aspects. In the literature, the strategy is found to be sensitive to transaction costs, speed of execution, length of the formation period, changes in model parameters over time, diversity of traded securities and news shocks. These unpleasant properties were studied for example by Bowen et al. (2010), Do and Faff (2012), Huck (2013) and Jacobs and Weber (2015). We add to this long list the sensitivity of the intraday pairs trading strategy to the market microstructure noise.

One possible direction for the future research is an inclusion of the parameter uncertainty in the optimization problem finding the trading signals. In this study as well as many other studies including Bertram (2010), Cummins and Bucca (2012), Zeng and Lee (2014) and Göncü and Akyildirim (2016), the optimization of the trading signals is based on given values of the Ornstein–Uhlenbeck paramaters. In reality, however, the values of parameters are subject to considerable uncertainty. We believe that addressing this issue would increase stability of the profitability and help to remove ambiguity in determining trade opportunities.

- Conclusion -

In the thesis, we deal with financial high-frequency data. We review the high-frequency literature and contribute to the statistical and econometric theory. In the empirical study, we apply the presented methodology to high-frequency prices of the 30 stocks forming Dow Jones Industrial Average index and 7 stocks representing Big Oil companies. All analyzed stocks are traded either on the NYSE exchange or the NASDAQ exchange. Specifically, we investigate three topics in financial high-frequency analysis.

First, we analyze durations between successive transactions. Traditionally, trade durations are modeled by the autoregressive conditional duration (ACD) model. We utilize this model in a discrete framework and particularly focus on zero values of trade durations. Zero values can be caused by multiple transactions occuring at the same time but originating from different sources or by a single order split into multiple transactions for faster execution. Most of the duration literature removes zero values from the analysis. We propose to keep them in dataset and directly model them using the zero-inflated autoregressive conditional duration (ZIACD) model. This model is based on the generalized autoregressive score (GAS) specification with the zero-inflated negative binomial distribution.

Second, we analyze non-parametric volatility of the price process. Intraday volatility is often measured by the quadratic variation. For the class of semimartingale processes, it can be decomposed to the integrated variance for the continuous part and the jump variance for the discontinuous part. We introduce the quadratic variation under interval uncertainty. However, we show that the use of the interval quadratic variation is quite limited in practice as it is not identified. For this reason, we stay within the traditional framework based on the additive model for the market microstructure noise. We illustrate the bias of realized variance due to the noise and compare various noise-robust estimators introduced in the high-frequency literature. In the simulation study, we find that the pre-averaging estimator outperforms the realized variance, two-scale estimator, realized kernel estimator and least squares estimator. We also compare various models used for volatility forecasting. In the empirical study, we find that the HAR model and the realized GARCH model with logarithmic specification outperforms naive models as well as the ARIMA model for the one-step ahead forecasts of the quadratic variation.

Third, we analyze the price process by parametric methods. A popular model in finance is the Ornstein–Uhlenbeck process with continuous time. It is often used to model interest rates, exchange rates, commodity prices, stochastic volatility and spreads between correlated assets. We address the issues in the estimation of the Ornstein–Uhlenbeck process using high-frequency data. The price process is known to be contaminated by the market microstructure noise which causes a significant bias in volatility estimation. We show that the Ornstein–Uhlenbeck process contaminated by the white noise and sampled at discrete equidistant time folows the ARMA(1,1) process instead of the AR(1) process. We also consider irregularly spaced observations and propose a noise-robust estimator based on the maximum likelihood. We illustrate the added value of our proposed approach in an application to the pairs trading strategy. We show that the volatility and speed of reversion parameters estimated by the traditional methods are distinctively biased and many times higher than the parameters estimated by the proposed noise-robust methods. This leads to suboptimal decision-making and significant decrease in profits of the pairs trading strategy.

- Appendix A -

Stock Market

A *stock* is a security representing fractional ownership of a given company. Stocks are commonly traded on *stock exchanges*. Some companies are listed only on a single stock exchange while other companies prefer multiple exchanges. A typical example is a company outside the United States that is primarily listed in its home country and secondarily listed in the United States. Some smaller companies that do not qualify for the listing requirements of the major exchanges can be traded over-the-counter.

Table A.1 and Figure A.1 list the largest stock exchanges in the world measured by the domestic market capitalization. New York City is the capital of the financial world as it is home to the two largest stock exchanges – the NYSE exchange located at 11 Wall Street and the NASDAQ exchange located at 165 Broadway.

To give an idea of the size of the stock market, we report the market capitalization for the exchanges as well as for the individual stocks. The *market capitalization* is the market value of outstanding shares of a publicly traded company. For a given stock i, the market capitalization MC_i is simply the number of shares outstanding N_i times the closing price per share P_i, i.e. $MC_i = N_i P_i$. For a given exchange j, the domestic market capitalization is the sum of market capitalizations of all listed domestic companies as well as foreign companies exclusively listed on the exchange j. The source of the market capitalization of the exchanges is World Federation of Exchanges (2019). For the individual stocks, the source of the numbers of shares outstanding is Nasdaq (2019) while the source of closing prices is Yahoo! (2019).

In the thesis, we analyze two sets of stock. For the duration analysis in Chapter 3 and the volatility analysis in Chapter 4, we use tick data of 30 stocks forming the Dow Jones Industrial Average index. For the pairs trading strategy in Chapter 5, we use tick data of 7 stocks representing the Big Oil companies. The source of our tick data is the Daily TAQ database of New York Stock Exchange (2019).

Dow Jones Industrial Average Index

The *Dow Jones Industrial Average (DJIA) index* is a stock market index consisting of 30 publicly owned American companies. It was founded by Charles Dow on May 26, 1896 and is the second-oldest index

NYSE NASDAQ

Figure A.1: Two largest stock exchanges in the world.

Exchange	Name	Location	Companies	Market Cap
NYSE	New York Stock Exchange	United States	2 292	23 216 420
NASDAQ	Nasdaq Stock Exchange	United States	3 004	10 998 591
JPX	Tokyo Stock Exchange	Japan	3 628	6 059 062
SSE	Shanghai Stock Exchange	China	1 432	4 526 023
EURONEXT	Euronext Stock Exchange	Netherlands	1 239	4 341 984
LSE	London Stock Exchange	United Kingdom	2 491	4 316 300
HKEX	Hong Kong Stock Exchange	Hong Kong	2 215	4 219 596
SZSE	Shenzhen Stock Exchange	China	2 115	3 091 590
TSX	Toronto Stock Exchange	Canada	3 366	2 276 829
BSE	Bombay Stock Exchange	India	5 290	2 121 000
FSX	German Stock Exchange	Germany	509	2 113 779
NSE	National Stock Exchange of India	India	1 951	2 097 899
KRX	Korea Stock Exchange	South Korea	2 151	1 638 023
SIX	Swiss Stock Exchange	Switzerland	268	1 518 557
ASX	Australian Securities Exchange	Australia	2 150	1 450 006
OMX	Stockholm Stock Exchange	Sweden	1 008	1 449 594
TWSE	Taiwan Stock Exchange	Taiwan	934	1 064 851
JSE	Johannesburg Stock Exchange	South Africa	364	1 061 518
BME	Spanish Stock Exchange	Spain	3 046	851 755
BOVESPA	Brazilian Stock Exchange	Brazil	342	783 499

Table A.1: List of 20 stock exchanges with highest domestic market capitalization in millions of USD as of June 29, 2018.

in the United States. Currently, it is owned by S&P Dow Jones Indices company. Figure A.3 shows daily prices as well as daily volume of the index since January, 2000.

Components of the DJIA index changes over time. Table A.2 and Figure A.2 list the composition of the index from September 1, 2017 to June 25, 2018. On September 1, 2017, DuPont (DD) was replaced by DowDuPont (DWDP) due to the merger of Dow Chemical Company with DuPont. On June 26, 2018, General Electric (GE) was replaced by Walgreens Boots Alliance (WBA). For the duration analysis in Chapter 3, we use the 30 stocks that formed the index from September 1, 2017 to June 25, 2018 as listed in Table A.2. For the volatility analysis in Chapter 4, we use the 30 stocks that formed the index from March 19, 2015 to August 31, 2017, i.e. the stocks listed in Table A.2 with the DD stock instead of the DWDP stock.

Big Oil Companies

The term *Big Oil* refers to the 7 largest publicly traded oil and gas companies, also known as the *super-majors*. The national producers and the OPEC oil companies are not considered to be a part of the Big Oil, although they have much greater influence on oil and gas prices.

Table A.3 and Figure A.4 list the 7 Big Oil companies. Some sources exclude Phillips 66 from the Big Oil. Chevron, Phillips 66 and ExxonMobil are American companies listed on the NYSE exchange. BP, Eni, Royal Dutch Shell and Total are European companies primarily listed on European exchanges and secondarily listed on the NYSE exchange. The market capitalization in Table A.3 refers only to stocks traded on the NYSE exchange. The total market capitalization for European companies is therefore much higher as it includes stocks traded on home exchanges. For the pairs trading strategy in Chapter 5, we use only stocks traded on the NYSE exchange as indicated in Table A.3.

Stock	Exchange	Company	Industry	Market Cap
AAPL	NASDAQ	Apple	Consumer Electronics	909 841
AXP	NYSE	American Express	Credit Services	84 383
BA	NYSE	Boeing	Aerospace & Defense	192 753
CAT	NYSE	Caterpillar	Farm & Construction Equipment	81 124
CSCO	NASDAQ	Cisco Systems	Communication Equipment	202 365
CVX	NYSE	Chevron	Oil & Gas	241 602
DIS	NYSE	Walt Disney	Media	156 222
DWDP	NYSE	DowDuPont	Chemicals	152 986
GE	NYSE	General Electric	Conglomerate	118 207
GS	NYSE	Goldman Sachs	Capital Markets	83 313
HD	NYSE	The Home Depot	Home Improvement	225 056
IBM	NYSE	IBM	Computers & Technology	128 240
INTC	NASDAQ	Intel	Semiconductors	231 649
JNJ	NYSE	Johnson & Johnson	Pharmaceuticals	325 452
JPM	NYSE	JPMorgan Chase	Banking	354 778
KO	NYSE	Coca-Cola	Beverages	186 636
MCD	NYSE	McDonald's	Restaurants	123 029
MMM	NYSE	3M	Conglomerate	116 791
MRK	NYSE	Merck & Company	Pharmaceuticals	163 301
MSFT	NASDAQ	Microsoft	Software	757 640
NKE	NYSE	Nike	Apparel	102 029
PFE	NYSE	Pfizer	Pharmaceuticals	212 222
PG	NYSE	Procter & Gamble	Household & Personal Products	196 290
TRV	NYSE	Travelers	Insurance	32 748
UNH	NYSE	UnitedHealth Group	Health Care Plans	235 767
UTX	NYSE	United Technologies	Aerospace & Defense	100 031
V	NYSE	Visa	Credit Services	236 577
VZ	NYSE	Verizon Communications	Telecommunication Services	207 876
WMT	NYSE	Walmart	Retail	255 832
XOM	NYSE	ExxonMobil	Oil & Gas	350 265

Table A.2: List of 30 stocks forming the Dow Jones Industrial Average index from September 1, 2017 to June 25, 2018 with market capitalization in millions of USD as of June 29, 2018.

Stock	Exchange	Company	Listing	Market Cap
BP	NYSE	BP	Secondary	41 574
CVX	NYSE	Chevron	Primary	241 602
E	NYSE	Eni	Secondary	1 361
PSX	NYSE	Phillips 66	Primary	52 318
RDS-A	NYSE	Royal Dutch Shell	Secondary, Class A	32 291
TOT	NYSE	Total	Secondary	12 010
XOM	NYSE	ExxonMobil	Primary	350 265

Table A.3: List of 7 stocks representing the Big Oil companies with market capitalization in millions of USD as of June 29, 2018.

Figure A.2: List of 30 stocks forming the Dow Jones Industrial Average index from September 1, 2017 to June 26, 2018.

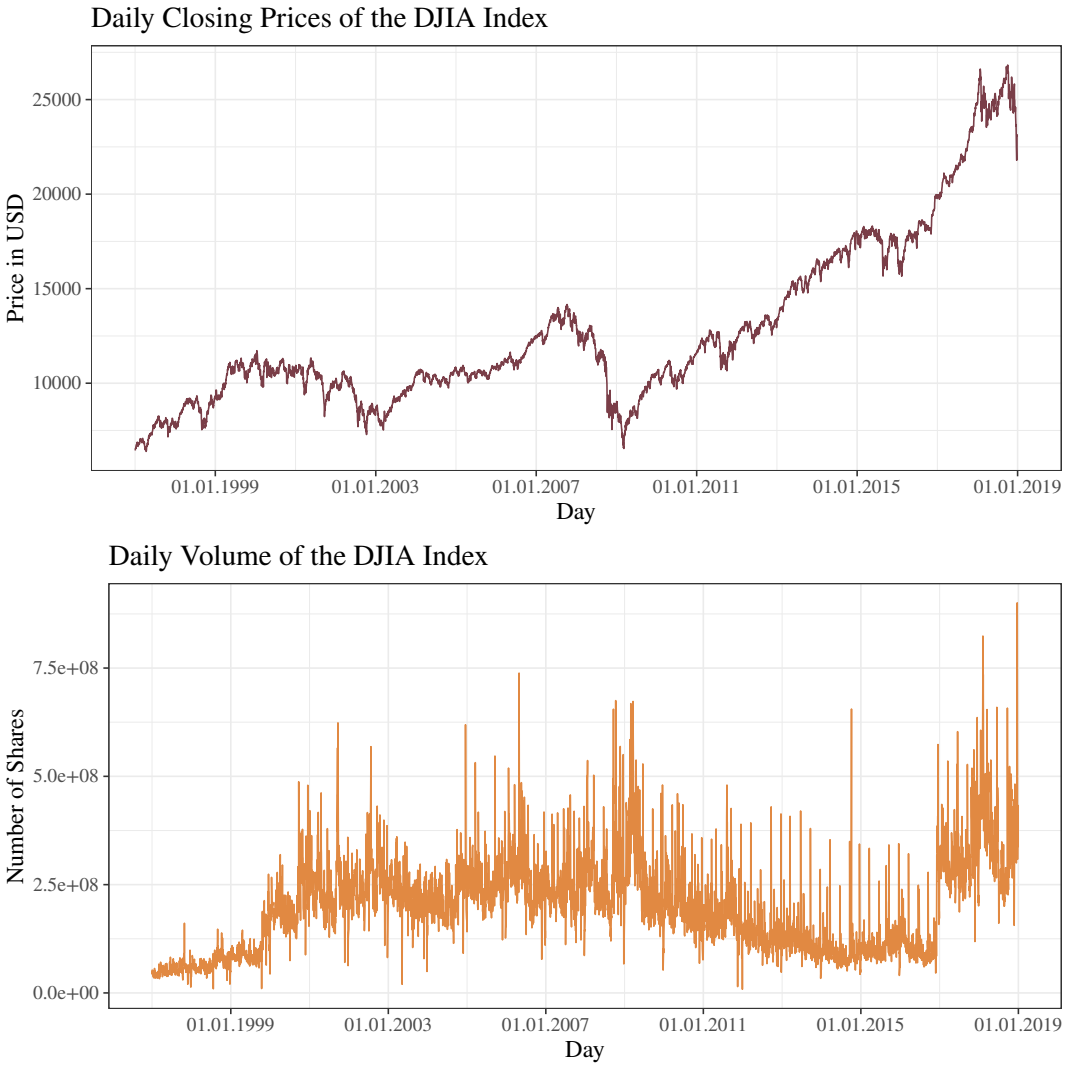

Figure A.3: Daily closing prices adjusted for dividends and splits and daily number of shares traded of the DJIA index from January, 1997 to December, 2018.

NYSE:BP NYSE:CVX NYSE:E NYSE:PSX

NYSE:RDS-A NYSE:TOT NYSE:XOM

Figure A.4: List of 7 stocks representing the Big Oil companies.

- Appendix B -

High-Frequency Data Literature

We take a brief look at some statistics of the high-frequency data literature according to Elsevier (2019). Table B.1 lists the most productive authors in terms of the number of papers with keywords related to high-frequency data. Table B.2 lists the most prominent journals publishing papers with related keywords.

Author	Affiliation	Articles
Aït-Sahalia, Yacine	Princeton University,USA	10
Andersen, Torben G.	Northwestern University, USA	11
Barndorff-Nielsen, Ole E.	Aarhus University, Denmark	17
Bollerslev, Tim	National Bureau of Economic Research, USA	16
Corsi, Fulvio	University of London, United Kingdom	11
Degiannakis, Stavros	Panteion University, Greece	13
Hansen, Peter Reinhard	Aarhus University, Denmark	10
Hwang, Eunju J.	Gachon University, South Korea	10
Jacod, Jean	Pierre and Marie Curie University, France	14
Kong, Xin-Bing	Nanjing Audit University, China	10
Liu, Zhi	University of Macau, Macao	14
Lunde, Asger	Aarhus University, Denmark	11
Ma, Feng	Southwest Jiaotong University, China	11
McAleer, Michael	National Tsing Hua University, Taiwan	13
Meddahi, Nour	Toulouse School of Economics, France	11
Mykland, Per Aslak	University of Illinois at Chicago, USA	18
Podolskij, Mark	Aarhus University, Denmark	23
Russo, Francesco	University of Paris-Saclay, France	16
Shephard, Neil	Harvard University, USA	16
Shin, Dong Wan	Ewha Womans University, South Korea	16
Todorova, Neda	Griffith University, Australia	14
Vetter, Mathias	University of Kiel, Germany	11
Wang, Yazhen	University of Wisconsin Madison, USA	10
Wei, Yu	Yunnan University of Finance and Economics, China	10
Zhang, Lan	University of Illinois at Chicago, USA	10

Table B.1: List of authors with 10 or more scientific articles containing the term integrated volatility, integrated variance, quadratic variation, realized volatility, realized variance or microstructure noise in the title, abstract or keywords as of December 31, 2018.

ISSN	Journal	Articles
0091-1798	Annals of Probability	15
0090-5364	Annals of Statistics	14
0361-0926	Applied Economics	13
0960-3107	Applied Financial Economics	14
1350-486X	Applied Mathematical Finance	16
1350-7265	Bernoulli	22
0167-9473	Computational Statistics and Data Analysis	20
0361-0926	Communications in Statistics - Theory and Methods	13
0747-4938	Econometric Reviews	23
0266-4666	Econometric Theory	17
0012-9682	Econometrica	10
0264-9993	Economic Modelling	21
0165-1765	Economics Letters	15
0140-9883	Energy Economics	20
1351-847X	European Journal of Finance	11
0949-2984	Finance and Stochastics	14
1544-6123	Finance Research Letters	19
0029-5981	International Journal for Numerical Methods in Engineering	11
0169-2070	International Journal of Forecasting	29
0219-0249	International Journal of Theoretical and Applied Finance	17
1057-5219	International Review of Financial Analysis	17
0883-7252	Journal of Applied Econometrics	16
0378-4266	Journal of Banking and Finance	39
0735-0015	Journal of Business and Economic Statistics	27
0304-4076	Journal of Econometrics	80
0927-5398	Journal of Empirical Finance	37
1479-8409	Journal of Financial Econometrics	45
0304-405X	Journal of Financial Economics	19
0277-6693	Journal of Forecasting	29
0270-7314	Journal of Futures Markets	48
1042-4431	Journal of International Financial Markets, Institutions and Money	10
0261-5606	Journal of International Money and Finance	11
0162-1459	Journal of the American Statistical Association	15
1062-9408	North American Journal of Economics and Finance	10
0378-4371	Physica A: Statistical Mechanics and Its Applications	34
0178-8051	Probability Theory and Related Fields	10
1469-7688	Quantitative Finance	46
0275-5319	Research in International Business and Finance	10
0167-7152	Statistics and Probability Letters	15
0736-2994	Stochastic Analysis and Applications	11
0304-4149	Stochastic Processes and Their Applications	64
1000-6788	System Engineering Theory and Practice	16

Table B.2: List of journals with 10 or more scientific articles containing the term integrated volatility, integrated variance, quadratic variation, realized volatility, realized variance or microstructure noise in the title, abstract or keywords as of December 31, 2018.

- Appendix C -

High-Frequency Data Analysis in R

This appendix closely follows Holý (2018c). We review capabilities of statistical software R developed by R Core Team (2019) in financial high-frequency data analysis. For statistical analysis, R offers a tremendous amount of user-created packages. Growth of the number of R packages published on CRAN over the years is shown in Figure C.1. Underlying data for Figure C.1 are retrieved using a script by Daróczi (2017). Many packages are directly related to academic research as indicated by Table C.1. Underlying data for Table C.1 are retrieved from Elsevier (2019). Table C.2 lists R packages useful in analysis of financial high-frequency data while Table C.3 lists some generally useful packages.

We describe the functionality of selected packages throughout this appendix. The package `xts` (Ryan et al., 2018a) handles time series recorded at very fine scale and the `quantmod` (Ryan et al., 2018b) package offers basic tools for financial data. The package `highfrequency` (Boudt et al., 2018) is a general source of various high-frequency methods while the `ACDm` (Belfrage, 2016) package focuses on the ACD model, the `midasr` (Kvedaras and Zemlys, 2016) package focuses on the MIDAS model and the `rugarch` (Ghalanos, 2018) package focuses on the GARCH model. Packages `PortfolioEffectEstim` (Kostin et al., 2016) and `PortfolioEffectHFT` (Kostin et al., 2017) provide an interface to cloud service PortfolioEffect specializing in financial analysis but can be utilized for client-side high-frequency data as well. For a more detailed description of the packages, we refer to the documentation included in the individual packages. For other relevant packages, we refer to R CRAN task views *Time Series Analysis* (Hyndman, 2019) and *Empirical Finance* (Eddelbuettel, 2019)

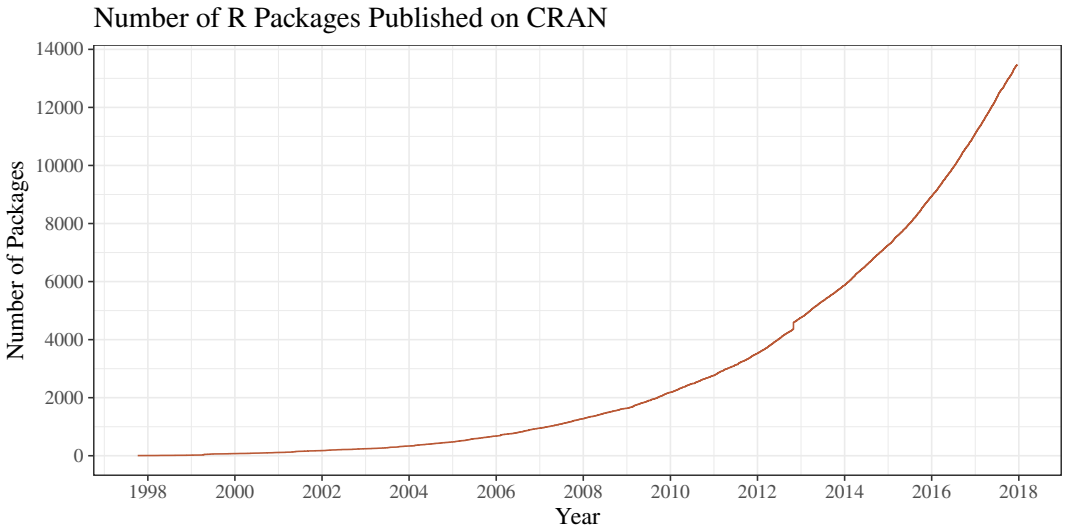

Figure C.1: Growth of the number of R packages published on CRAN.

ISSN	Journal	Articles
1367-4803	Bioinformatics	340
1471-2105	BMC Bioinformatics	239
1471-2164	BMC Genomics	33
0167-9473	Computational Statistics and Data Analysis	50
1061-8600	Journal of Computational and Graphical Statistics	43
1548-7660	Journal of Statistical Software	486
2041-210X	Methods in Ecology and Evolution	100
1755-0998	Molecular Ecology Resources	36
0305-1048	Nucleic Acids Research	33
1932-6203	PLOS One	126
2073-4859	The R Journal	161
0277-6715	Statistics in Medicine	40

Table C.1: List of journals with 30 or more scientific articles containing the term R package in the title, abstract or keywords as of December 31, 2017.

Package	Title
`ACDm`	Tools for Autoregressive Conditional Duration Models
`highfrequency`	Tools for Highfrequency Data Analysis
`midasr`	Mixed Data Sampling Regression
`PortfolioEffectEstim`	High Frequency Price Estimators by PortfolioEffect
`PortfolioEffectHFT`	High Frequency Portfolio Analytics by PortfolioEffect
`quantmod`	Quantitative Financial Modelling Framework
`rugarch`	Univariate GARCH Models
`xts`	eXtensible Time Series

Table C.2: List of R packages useful in financial high-frequency data analysis.

Package	Title
`data.table`	Extension of 'data.frame'
`dplyr`	A Grammar of Data Manipulation
`ggplot2`	Create Elegant Data Visualisations Using the Grammar of Graphics
`gridExtra`	Miscellaneous Functions for "Grid" Graphics
`knitr`	A General-Purpose Package for Dynamic Report Generation in R
`Rcpp`	Seamless R and C++ Integration
`reshape2`	Flexibly Reshape Data: A Reboot of the Reshape Package
`rio`	A Swiss-Army Knife for Data I/O
`shiny`	Web Application Framework for R
`tikzDevice`	R Graphics Output in LaTeX Format
`xtable`	Export Tables to LaTeX or HTML

Table C.3: List of generally useful R packages.

Data Preprocessing

In practice, it is necessary to clean high-frequency data before the actual analysis. Brownlees and Gallo (2006) and Barndorff-Nielsen et al. (2009) discuss individual steps in data cleaning. Preprocessing of high-frequency data can be done using the `highfrequency` package. The cleaning procedure for trade data includes removing transactions with zero prices (the function `noZeroPrices`), restricting transactions to a single exchange (the function `selectExchange`), removing transactions with abnormal sales condition (the function `salesCondition`), merging transactions with the same timestamp (the function `mergeTradesSameTimestamp`) and removing outliers in transactions based on quote data (the function `rmTradeOutliers`). It is also possible to use wrapper functions `tradesCleanup` and `tradesCleanupFinal`. The cleaning procedure for quote data includes removing quotes with zero prices (the function `noZeroQuotes`), restricting quotes to a single exchange (the function `selectExchange`), removing quotes with a large spread (the function `rmLargeSpread`), removing quotes with a negative spread (the function `rmNegativeSpread`), merging quotes with the same timestamp (the function `mergeQuotesSameTimestamp`) and removing outliers in quotes (the function `rmOutliers`). It is also possible to use wrapper function `quotesCleanup`.

In some cases, it may also be suitable to aggregate data to a given frequency. Although this naturally leads to a data loss, it can be very useful specially in multivariate analysis due to trading asynchronicity of different assets. The calendar sampling may be accomplished by the function `aggregatets` of the `highfrequency` package.

Market Microstructure

It is often assumed (specially from the volatility analysis perspective) that there exists a theoretical efficient price with continuous values. This price is, however, unobserved mainly due to discretness of the observed price and bid-ask spread. A common approach is to model the observed price as the sum of the efficient price and the market microstructure noise which captures all the undesirable effects (Hansen and Lunde, 2006). Properties of the market microstructure noise can be investigated using the `PortfolioEffectEstim` package. Variance of the noise is estimated by functions `noise_acnv`, `noise_rnv`, `noise_urnv` and `noise_uznv`. The noise to signal ratio is estimated by the `noise_nts` function.

Duration Analysis

As high-frequency data are irregularly spaced, it is natural to investigate behavior of times between successive transactions denoted as trade durations. Other financial durations such as price durations (times until the price process changes by a given level) and volume durations (times until traded volume reaches a given level) are subject of duration analysis as well. Engle and Russell (1998) proposed the autoregressive conditional duration (ACD) model, which became the standard in financial duration modeling. The package `ACDm` is centered around the ACD model and its extensions. The function `computeDurations` computes trade, price and volume durations. As financial durations exhibit strong intraday patterns, it is suitable to remove diurnal patterns using the `diurnalAdj` function. The function `acdFit` fits the ACD model based on various dynamics and distributions.

Volatility Analysis

Perhaps the most studied area in the high-frequency literature is the volatility analysis. First, we estimate the integrated variance. The integrated variance is equal to the quadratic variation in the absence of jumps. Simple estimator of the integrated variance is the realized variance (the function `rCov` from the `highfrequency` package and the function `variance_rv` from the `PortfolioEffectEstim` package). However, this estimator is biased in the presence of the jumps and the market microstructure noise. The average realized variance (the function `rAVGCov` from the `highfrequency` package)

can be used to diminish effects of the noise, but it is still a biased estimator. Jump-robust estimators include the realized bi-power variation of Barndorff-Nielsen (2004) (the function `rBPCov` from the `highfrequency` package), the min realized variance of Andersen et al. (2012) (the function `minRV` from the `highfrequency` package) and the median realized variance of Andersen et al. (2012) (the function `medRV` from the `highfrequency` package). Noise-robust estimators include the two scales realized variance of Zhang et al. (2005) (the function `rTSCov` from the `highfrequency` package and the function `variance_tsrv` from the `PortfolioEffectEstim` package), the multi scales realized variance of Zhang (2006) (the function `variance_msrv` from the `PortfolioEffectEstim` package), the realized kernel of Barndorff-Nielsen et al. (2008) (the function `rKenelCov` from the `highfrequency` package and the function `variance_krv` from the `PortfolioEffectEstim` package), the pre-averaging estimator or modulated realized covariance of Hautsch and Podolskij (2013) (the function `MRC` from the `highfrequency` package and the function `variance_mrv` from the `PortfolioEffectEstim` package) and the uncertainty zones realized variance of Robert and Rosenbaum (2012) (the function `variance_uzrv` from the `PortfolioEffectEstim` package). The jump robust modulated realized variance of Podolskij and Vetter (2009) (the function `variance_jrmrv` from the `PortfolioEffectEstim` package) is robust to both jumps and the noise.

Next, we forecast the estimated integrated variance (denoted as the realized measure). A popular model forecasting realized measure is the HAR model of Corsi (2009) (the function `harModel` from the `highfrequency` package). Another model for volatility of returns and realized measure is the HEAVY model of Shephard and Sheppard (2010) (the function `heavyModel` from the `highfrequency` package). Realized measure can also be augmented into the GARCH model resulting in the realized GARCH model of Hansen et al. (2012). Univariate GARCH models (including the realized GARCH model) are provided by the `rugarch` package with the `ugarchspec` function specifying the model, the `ugarchfit` function fitting the model, the `ugarchforecast` function forecasting the model as well as other functions. The MIDAS models of Ghysels et al. (2006) can be estimated using the `midasr` package. Package `PortfolioEffectHFT` also offers a framework for volatility forecasting. Traditional models such as ARIMA in various time series packages may be utilized as well.

Finally, we estimate the spot volatility. The function `spotvol` from the `highfrequency` package provides various methods for the spot volatility estimation. For details, see documentation of the `highfrequency` package.

Higher Moments Analysis

An analogue of the realized variance for the skewness is the realized skewness (the function `rSkew` from the `highfrequency` package) and realized kurtosis (the function `rKurt` from the `highfrequency` package) for the kurtosis. Both estimators are biased in the presence of jumps and the market microstructure noise.

The integrated quarticity can be estimated by the realized quarticity (the function `rQuar` from the `highfrequency` package and the function `quarticity_rq` from the `PortfolioEffectEstim` package). However, this estimator is biased in the presence of the jumps and the market microstructure noise. Jump-robust estimators include the min realized quarticity of Andersen et al. (2012) (the function `minRQ` from the `highfrequency` package), the median realized quarticity of Andersen et al. (2012) (the function `minRQ` from the `highfrequency` package), the realized tripower quarticity of Andersen et al. (2012) (the function `quarticity_rtq` from the `PortfolioEffectEstim` package) and the realized quadpower quarticity of Andersen et al. (2012) (the function `quarticity_rqq` from the `PortfolioEffectEstim` package). Jump-robust and noise-robust estimators include the modulated realized quarticity of Podolskij and Vetter (2009) (the function `quarticity_mrq` from the `PortfolioEffectEstim` package) and the modulated tripower quarticity of Podolskij and Vetter (2009) (the function `quarticity_mtq` from the `PortfolioEffectEstim` package).

A generalization of the integrated variance and the integrated quarticity is the integrated power variation. It can be estimated by the realized multipower variation of Andersen et al. (2012) (the function `rMPV` from the `highfrequency` package). It is biased in the presence of the market miscrostructure noise. However, for some values of powers, it can be robust to jumps. The class of the realized multipower variation includes specifications for the realized variance, the bi-power variation and the realized quarticity.

Jump Analysis

The asset price is often modeled as a process with continuous values and a finite number of jumps. The question is whether the jumps are present in the price process in a given time frame. The package `highfrequency` offers three functions testing for the presence of jumps. The function `AJjumptest` implements the test of Aït-Sahalia and Jacod (2009), the function `BNSjumptest` implements the test of Barndorff-Nielsen and Shephard (2006) and the function `JOjumptest` implements the test of Jiang and Oomen (2008).

Liquidity Analysis

Various liquidity measures can be estimated by the `tqLiquidity` function from the `highfrequency` package. Most of these measures require quote data while some also need transaction data. For details, see documentation of the `highfrequency` package.

Discussion

The software and programming language R offers several packages specializing in financial high-frequency data analysis. One possible drawback is the inconsistency of function structures between packages (and even within the package `highfrequency` itself) as unfortunately expected in the software with so many packages and authors. Overall, the presented packages well cover data preprocessing and market microstructure investigation as well as analysis of durations, jumps, volatility, higher moments and liquidity. For other topics or more deeper analysis, it is needed to write own code in R.

- Appendix D -

Special Functions in Mathematics

In this appendix, we define some lesser-known special functions in mathematics used in the thesis. Namely, we focus on the gamma function, the digamma function and the polygamma function. For a comprehensive list of special functions with definitions, descriptions and figures, we refer to Olver et al. (2010) and National Institute of Standards and Technology (2019).

The *gamma function* is an extension of the factorial function to real and complex numbers. It is defined for all complex numbers except zero and negative integers x as

$$\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} \, \mathrm{d}z. \tag{D.1}$$

For a positive integer n, it is equal to

$$\Gamma(n) = (n-1)!, \tag{D.2}$$

where ! denotes the factorial. The gamma function $\Gamma(x)$ is illustrated in Figure D.1.

The *digamma function* is defined as the derivative of the logarithm of the gamma function

$$\psi(x) = \frac{\partial}{\partial x} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{D.3}$$

The digamma function $\psi(x)$ is illustrated in Figure D.2.

The *polygamma function of order m* is defined as the $(m+1)$th derivative of the logarithm of the gamma function

$$\psi_m(x) = \frac{\partial^{m+1}}{\partial x^{m+1}} \log \Gamma(x) = \frac{\partial^m}{\partial x^m} \psi(x). \tag{D.4}$$

For $m = 0$, it is simply the digamma function, i.e. $\psi_0(x) = \psi(x)$. The trigamma function $\psi_1(x)$, the tetragamma function $\psi_2(x)$, the pentagamma function $\psi_3(x)$ and the hexagamma function $\psi_4(x)$ are illustrated in Figure D.3.

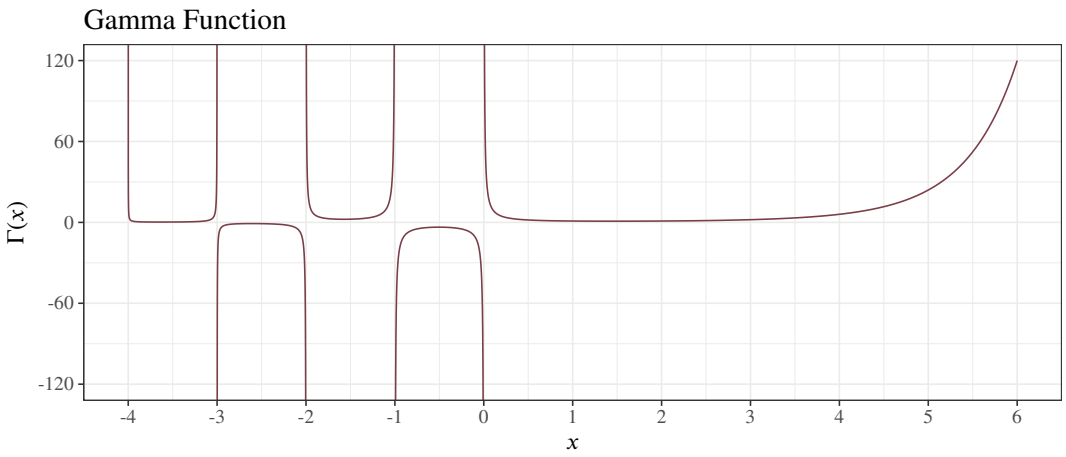

Figure D.1: Gamma function.

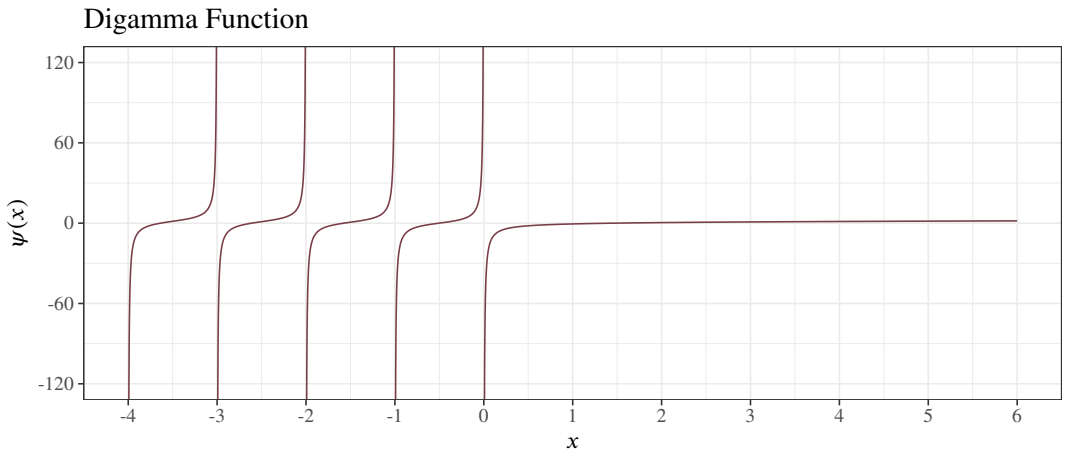

Figure D.2: Digamma function.

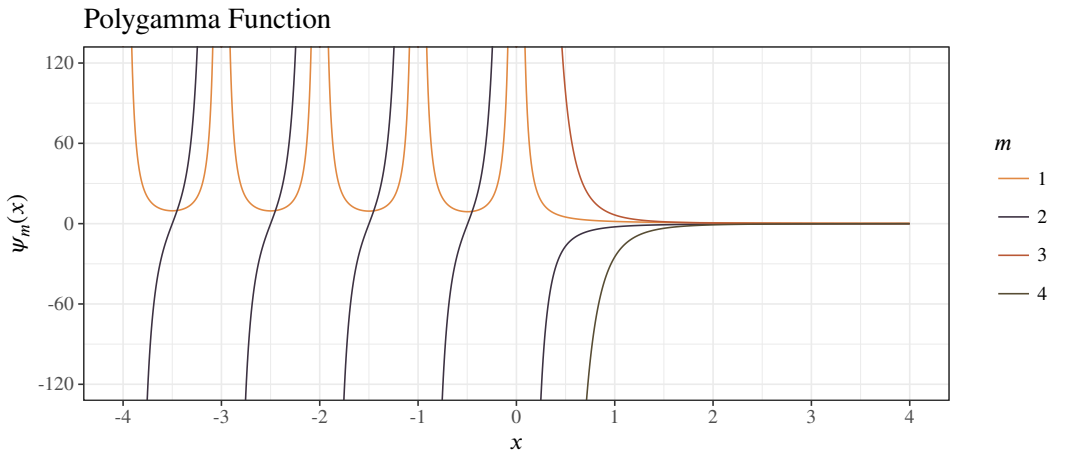

Figure D.3: Polygamma function of order m.

- List of Figures -

- List of Tables -

- Bibliography -

AHN, H.-J., CAI, J., CHEUNG, Y. L. 2005. Price Clustering on the Limit-Order Book: Evidence from the Stock Exchange of Hong Kong. *Journal of Financial Markets*. Volume 8. Issue 4. Pages 421–451. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2005.07.001.

AÏT-SAHALIA, Y., JACOD, J. 2009. Estimating the Degree of Activity of Jumps in High Frequency Data. *The Annals of Statistics*. Volume 37. Issue 5A. Pages 2202–2244. ISSN 0090-5364. https://doi.org/10.1214/08-aos640.

AÏT-SAHALIA, Y., JACOD, J. 2010. Is Brownian Motion Necessary to Model High-Frequency Data? *The Annals of Statistics*. Volume 38. Issue 5. Pages 3093–3128. ISSN 0090-5364. https://doi.org/10.1214/09-aos749.

AÏT-SAHALIA, Y., JACOD, J. 2014. *High-Frequency Financial Econometrics*. First Edition. Princeton. Princeton University Press. ISBN 978-0-691-16143-3. https://doi.org/10.1515/9781400850327.

AÏT-SAHALIA, Y., MANCINI, L. 2008. Out of Sample Forecasts of Quadratic Variation. *Journal of Econometrics*. Volume 147. Issue 1. Pages 17–33. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2008.09.015.

AÏT-SAHALIA, Y., MYKLAND, P. A. 2003. The Effects of Random and Discrete Sampling when Estimating Continuous-Time Diffusions. *Econometrica*. Volume 71. Issue 2. Pages 483–549. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.t01-1-00416.

AÏT-SAHALIA, Y., XIU, D. 2016. *A Hausman Test for the Presence of Market Microstructure Noise in High Frequency Data*. Working Paper. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=2741911.

AÏT-SAHALIA, Y., YU, J. 2009. High Frequency Market Microstructure Noise Estimates and Liquidity Measures. *The Annals of Applied Statistics*. Volume 3. Issue 1. Pages 422–457. ISSN 1932-6157. https://doi.org/10.1214/08-aoas200.

AÏT-SAHALIA, Y., MYKLAND, P. A., ZHANG, L. 2005. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *The Review of Financial Studies*. Volume 18. Issue 2. Pages 351–416. ISSN 0893-9454. https://doi.org/10.1093/rfs/hhi016.

AÏT-SAHALIA, Y., FAN, J., XIU, D. 2010. High-Frequency Covariance Estimates with Noisy and Asynchronous Financial Data. *Journal of the American Statistical Association*. Volume 105. Issue 492. Pages 1504–1517. ISSN 0162-1459. https://doi.org/10.1198/jasa.2010.tm10163.

AÏT-SAHALIA, Y., MYKLAND, P. A., ZHANG, L. 2011. Ultra High Frequency Volatility Estimation with Dependent Microstructure Noise. *Journal of Econometrics*. Volume 160. Issue 1. Pages 160–175. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.028.

AKAIKE, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*. Volume 19. Issue 6. Pages 716–723. ISSN 0018-9286. https://doi.org/10.1109/tac.1974.1100705.

AKNOUCHE, A., GUERBYENNE, H. 2006. Recursive Estimation of GARCH Models. *Communications in Statistics - Simulation and Computation*. Volume 35. Issue 4. Pages 925–938. ISSN 0361-0918. https://doi.org/10.1080/03610910600880328.

ALDRIDGE, I. 2013. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Second Edition. Hoboken. Wiley. ISBN 978-1-118-34350-0. https://doi.org/10.1002/9781119203803.

ALVAREZ, A., PANLOUP, F., PONTIER, M., SAVY, N. 2012. Estimation of the Instantaneous Volatility. *Statistical Inference for Stochastic Processes*. Volume 15. Issue 1. Pages 27–59. ISSN 1387-0874. https://doi.org/10.1007/s11203-011-9062-2.

AMISANO, G., GIACOMINI, R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics*. Volume 25. Issue 2. Pages 177–190. ISSN 0735-0015. https://doi.org/10.1198/073500106000000332.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., LABYS, P. 2000. Great Realisations. *Risk*. Volume 13. Issue 3. Pages 105–108. ISSN 0952-8776. https://www.risk.net/infrastructure/1530292/great-realisations.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., LABYS, P. 2001. The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association*. Volume 96. Issue 453. Pages 42–55. ISSN 0162-1459. https://doi.org/10.1198/016214501750332965.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., LABYS, P. 2003. Modeling and Forecasting Realized Volatility. *Econometrica*. Volume 71. Issue 2. Pages 579–625. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.00418.

ANDERSEN, T. G., BOLLERSLEV, T., MEDDAHI, N. 2011. Realized Volatility Forecasting and Market Microstructure Noise. *Journal of Econometrics*. Volume 160. Issue 1. Pages 220–234. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.032.

ANDERSEN, T. G., DOBREV, D., SCHAUMBURG, E. 2012. Jump-Robust Volatility Estimation Using Nearest Neighbor Truncation. *Journal of Econometrics*. Volume 169. Issue 1. Pages 75–93. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2012.01.011.

ANDREOU, E., GHYSELS, E., KOURTELLOS, A. 2010. Regression Models with Mixed Sampling Frequencies. *Journal of Econometrics*. Volume 158. Issue 2. Pages 246–261. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.01.004.

ANDRESEN, A., BENTH, F. E., KOEKEBAKKER, S., ZAKAMULIN, V. 2014. The CARMA Interest Rate Model. *International Journal of Theoretical and Applied Finance*. Volume 17. Issue 2. Pages 1450008/1–1450008/27. ISSN 0219-0249. https://doi.org/10.1142/S0219024914500083.

AUDRINO, F., FENGLER, M. R. 2015. Are Classical Option Pricing Models Consistent with Observed Option Second-Order Moments? Evidence from High-Frequency Data. *Journal of Banking and Finance*. Volume 61. Pages 46–63. ISSN 0378-4266. https://doi.org/10.1016/j.jbankfin.2015.08.018.

AVELLANEDA, M., LEE, J. H. 2010. Statistical Arbitrage in the US Equities Market. *Quantitative Finance*. Volume 10. Issue 7. Pages 761–782. ISSN 1469-7688. https://doi.org/10.1080/14697680903124632.

AWARTANI, B., CORRADI, V., DISTASO, W. 2009. Assessing Market Microstructure Effects via Realized Volatility Measures with an Application to the Dow Jones Industrial Average Stocks. *Journal of Business & Economic Statistics*. Volume 27. Issue 2. Pages 251–265. ISSN 0735-0015. https://doi.org/10.1198/jbes.2009.0018.

AZRAK, R., MÉLARD, G. 2006. Asymptotic Properties of Quasi-Maximum Likelihood Estimators for ARMA Models with Time-Dependent Coefficients. *Statistical Inference for Stochastic Processes*. Volume 9. Issue 3. Pages 279–330. ISSN 1387-0874. https://doi.org/10.1007/s11203-005-1055-6.

AŞÇIOĞLU, A., COMERTON-FORDE, C., MCINISH, T. H. 2007. Price Clustering on the Tokyo Stock Exchange. *Financial Review*. Volume 42. Issue 2. Pages 289–301. ISSN 0732-8516. https://doi.org/10.1111/j.1540-6288.2007.00172.x.

BABBS, S. H., NOWMAN, K. B. 1999. Kalman Filtering of Generalized Vasicek Term Structure Models. *Journal of Financial and Quantitative Analysis*. Volume 34. Issue 1. Pages 115–130. ISSN 0022-1090. https://doi.org/10.2307/2676248.

BALL, C. A., ROMA, A. 1994. Target Zone Modelling and Estimation for European Monetary System Exchange Rates. *Journal of Empirical Finance*. Volume 1. Issue 3-4. Pages 385–420. ISSN 0927-5398. https://doi.org/10.1016/0927-5398(94)90010-8.

BANDI, F. M., RENÒ, R. 2018. Nonparametric Stochastic Volatility. *Econometric Theory*. Volume 34. Issue 6. Pages 1207–1255. ISSN 0266-4666. https://doi.org/10.1017/S0266466617000457.

BANDI, F. M., RUSSELL, J. R. 2005. *Realized Covariation, Realized Beta and Microstructure Noise*. Working Paper. https://www.researchgate.net/publication/253266961.

BANDI, F. M., RUSSELL, J. R. 2006. Separating Microstructure Noise from Volatility. *Journal of Financial Economics*. Volume 79. Issue 3. Pages 655–692. ISSN 0304-405X. https://doi.org/j.jfineco.2005.01.005.

BANDI, F. M., RUSSELL, J. R. 2008. Microstructure Noise, Realized Variance, and Optimal Sampling. *Review of Economic Studies*. Volume 75. Issue 2. Pages 339–369. ISSN 0034-6527. https://doi.org/10.1111/j.1467-937X.2008.00474.x.

BANDI, F. M., RUSSELL, J. R. 2011. Market Microstructure Noise, Integrated Variance Estimators, and the Accuracy of Asymptotic Approximations. *Journal of Econometrics*. Volume 160. Issue 1. Pages 145–159. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.027.

BANDI, F. M., RUSSELL, J. R., YANG, C. 2008a. Realized Volatility Forecasting and Option Pricing. *Journal of Econometrics*. Volume 147. Issue 1. Pages 34–46. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2008.09.002.

BANDI, F. M., RUSSELL, J. R., ZHU, Y. 2008b. Using High-Frequency Data in Dynamic Portfolio Choice. *Econometric Reviews*. Volume 27. Issue 1-3. Pages 163–198. ISSN 0747-4938. https://doi.org/10.1080/07474930701870461.

BANULESCU, D. G., COLLETAZ, G., HURLIN, C., TOKPAVI, S. 2016. Forecasting High-Frequency Risk Measures. *Journal of Forecasting*. Volume 35. Issue 3. Pages 224–249. ISSN 0277-6693. https://doi.org/10.1002/for.2374.

BAO, Y., LEE, T. H., SALTOĞLU, B. 2007. Comparing Density Forecast Models. *Journal of Forecasting*. Volume 26. Issue 3. Pages 203–225. ISSN 0277-6693. https://doi.org/10.1002/for.1023.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2001. Non-Gaussian Ornstein-Uhlenbeck-Based Models and Some of Their Uses in Financial Economics. *Journal of the Royal Statistical Society: Series B (Methodological)*. Volume 63. Issue 2. Pages 167–241. ISSN 1369-7412. https://doi.org/10.2307/2680596.

BARNDORFF-NIELSEN, O. E. 2004. Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*. Volume 2. Issue 1. Pages 1–37. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbh001.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2002a. Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society: Series B (Methodological)*. Volume 64. Issue 2. Pages 253–280. ISSN 1369-7412. https://doi.org/10.1111/1467-9868.00336.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2002b. Estimating Quadratic Variation Using Realized Variance. *Journal of Applied Econometrics*. Volume 17. Issue 5. Pages 457–477. ISSN 0883-7252. https://doi.org/10.1002/jae.691.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2004. Econometric Analysis of Realised Covariation: High Frequency Covariance, Regression and Correlation in Financial Economics. *Econometrica*. Volume 72. Issue 3. Pages 885–925. ISSN 1556-5068. https://doi.org/10.2139/ssrn.305583.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2006. Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation. *Journal of Financial Econometrics*. Volume 4. Issue 1. Pages 1–30. ISSN 0304-405X. https://doi.org/doi.org/10.1093/jjfinec/nbi022.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. 2007. Variation, Jumps and High Frequency Data in Financial Econometrics. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*. Volume 3. London. Pages 328–372. ISBN 978-0521692106. https://doi.org/10.1017/ccol0521871549.010.

BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A., SHEPHARD, N. 2008. Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*. Volume 76. Issue 6. Pages 1481–1536. ISSN 0012-9682. https://doi.org/10.3982/ecta6495.

BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A., SHEPHARD, N. 2009. Realized Kernels in Practice: Trades and Quotes. *Econometrics Journal*. Volume 12. Issue 3. Pages 1–32. ISSN 1368-4221. https://doi.org/10.1111/j.1368-423X.2008.00275.x.

BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A., SHEPHARD, N. 2011. Subsampling Realised Kernels. *Journal of Econometrics*. Volume 160. Issue 1. Pages 204–219. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.031.

BARUNÍK, J., VÁCHA, L. 2015. Realized Wavelet-Based Estimation of Integrated Variance and Jumps in the Presence of Noise. *Quantitative Finance*. Volume 15. Issue 8. Pages 1347–1364. ISSN 1469-7688. https://doi.org/10.1080/14697688.2015.1032550.

BARUNÍK, J., KŘEHLÍK, T., VÁCHA, L. 2016. Modeling and Forecasting Exchange Rate Volatility in Time-Frequency Domain. *European Journal of Operational Research*. Volume 251. Issue 1. Pages 329–340. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2015.12.010.

BAUWENS, L. 2006. Econometric Analysis of Intra-Daily Trading Activity on the Tokyo Stock Exchange. *Monetary and Economic Studies*. Volume 24. Issue 1. Pages 1–24. ISSN 0288-8432. http://www.imes.boj.or.jp/research/abstracts/english/me24-1-1.html.

BAUWENS, L., GIOT, P. 2000. The Logarithmic ACD Model: An Application to the Bid-Ask Quote Process of Three NYSE Stocks. *Annales d'Économie et de Statistique*. Volume 60. Pages 117–149. ISSN 0769-489X. https://doi.org/10.2307/20076257.

BAUWENS, L., GIOT, P. 2003. Asymmetric ACD Models: Introducing Price Information in ACD Models. *Empirical Economics*. Volume 28. Issue 4. Pages 709–731. ISSN 0377-7332. https://doi.org/10.1007/s00181-003-0155-7.

BAUWENS, L., HAUTSCH, N. 2009. Modelling Financial High Frequency Data Using Point Processes. In *Handbook of Financial Time Series*. Berlin, Heidelberg. Springer. Pages 953–979. ISBN 978-3-540-71296-1. https://doi.org/10.1007/978-3-540-71297-8.

BAUWENS, L., GIOT, P., GRAMMIG, J., VEREDAS, D. 2004. A Comparison of Financial Duration Models via Density Forecasts. *International Journal of Forecasting*. Volume 20. Issue 4. Pages 589–609. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2003.09.014.

BEE, M., DUPUIS, D. J., TRAPIN, L. 2016. Realizing the Extremes: Estimation of Tail-Risk Measures from a High-Frequency Perspective. *Journal of Empirical Finance*. Volume 36. Pages 86–99. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2016.01.006.

BEHME, A., LINDNER, A. 2012. Multivariate Generalized Ornstein-Uhlenbeck Processes. *Stochastic Processes and Their Applications*. Volume 122. Issue 4. Pages 1487–1518. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2012.01.002.

BELFRAGE, A. M. 2016. *Package 'ACDm'*. https://cran.r-project.org/package=ACDm.

BELTRATTI, A., MORANA, C. 1999. Computing Value at Risk with High Frequency Data. *Journal of Empirical Finance*. Volume 6. Issue 5. Pages 431–455. ISSN 0927-5398. https://doi.org/10.1016/s0927-5398(99)00008-0.

BENTH, F. E., RÜDIGER, B., SÜSS, A. 2018. Ornstein-Uhlenbeck Processes in Hilbert Space with Non-Gaussian Stochastic Volatility. *Stochastic Processes and Their Applications*. Volume 128. Issue 2. Pages 461–486. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2017.05.005.

BERTRAM, W. K. 2009. Optimal Trading Strategies for Itô Diffusion Processes. *Physica A: Statistical Mechanics and Its Applications*. Volume 388. Issue 14. Pages 2865–2873. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2009.04.004.

BERTRAM, W. K. 2010. Analytic Solutions for Optimal Statistical Arbitrage Trading. *Physica A: Statistical Mechanics and Its Applications*. Volume 389. Issue 11. Pages 2234–2243. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2010.01.045.

BHATTI, C. R. 2010. The Birnbaum-Saunders Autoregressive Conditional Duration Model. *Mathematics and Computers in Simulation*. Volume 80. Issue 10. Pages 2062–2078. ISSN 0378-4754. https://doi.org/10.1016/j.matcom.2010.01.011.

BIAIS, B., GLOSTEN, L., SPATT, C. 2005. Market Microstructure: A Survey of Microfoundations, Empirical Results, and Policy Implications. *Journal of Financial Markets*. Volume 8. Issue 2. Pages 217–264. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2004.11.00.

BIBI, A., FRANCQ, C. 2003. Consistent and Asymptotically Normal Estimators for Cyclically Time-Dependent Linear Models. *Annals of the Institute of Statistical Mathematics*. Volume 55. Issue 1. Pages 41–68. ISSN 0020-3157. https://doi.org/10.1007/bf02530484.

BLASQUES, F. 2017. *Advanced Econometric Methods for Complex Dynamic Models*. Technical Report. https://www.vu.nl/en/study-guide/2018-2019/master/e-g/econometrics-operations-research/index.aspx?view=module&id=50052021.

BLASQUES, F., KOOPMAN, S. J., LUCAS, A. 2014. *Maximum Likelihood Estimation for Score-Driven Models*. Working Paper. https://www.ssrn.com/abstract=2404276.

BLASQUES, F., HOLÝ, V., TOMANOVÁ, P. 2018. *Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros*. Working Paper. https://arxiv.org/abs/1812.07318.

BLAU, B. M., GRIFFITH, T. G. 2016. Price Clustering and the Stability of Stock Prices. *Journal of Business Research*. Volume 69. Issue 10. Pages 3933–3942. ISSN 0148-2963. https://doi.org/10.1016/j.jbusres.2016.06.008.

BLAZSEK, S., ESCRIBANO, A. 2016. Score-Driven Dynamic Patent Count Panel Data Models. *Economics Letters*. Volume 149. Pages 116–119. ISSN 0165-1765. https://doi.org/10.1016/j.econlet.2016.10.026.

BODENHAM, D. A., ADAMS, N. M. 2017. Continuous Monitoring for Changepoints in Data Streams Using Adaptive Estimation. *Statistics and Computing*. Volume 27. Issue 5. Pages 1257–1270. ISSN 0960-3174. https://doi.org/10.1007/s11222-016-9684-8.

BOGOMOLOV, T. 2013. Pairs Trading Based on Statistical Variability of the Spread Process. *Quantitative Finance*. Volume 13. Issue 9. Pages 1411–1430. ISSN 1469-7688. https://doi.org/10.1080/14697688.2012.748934.

BOLLERSLEV, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*. Volume 31. Issue 3. Pages 307–327. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(86)90063-1.

BOLLERSLEV, T., ZHANG, B. Y. B. 2003. Measuring and Modeling Systematic Risk in Factor Pricing Models Using High-Frequency Data. *Journal of Empirical Finance*. Volume 10. Issue 5. Pages 533–558. ISSN 0927-5398. https://doi.org/10.1016/s0927-5398(03)00004-5.

BOROVKOV, K., NOVIKOV, A. 2008. On Exit Times of Levy-Driven Ornstein-Uhlenbeck Processes. *Statistics & Probability Letters*. Volume 78. Issue 12. Pages 1517–1525. ISSN 0167-7152. https://doi.org/10.1016/j.spl.2008.01.017.

BORTOLUZZO, A. B., MORETTIN, P. A., TOLOI, C. M. C. 2010. Time-Varying Autoregressive Conditional Duration Model. *Journal of Applied Statistics*. Volume 37. Issue 5. Pages 847–864. ISSN 0266-4763. https://doi.org/10.1080/02664760902914458.

BOS, C. S., JANUS, P., KOOPMAN, S. J. 2012. Spot Variance Path Estimation and Its Application to High-Frequency Jump Testing. *Journal of Financial Econometrics*. Volume 10. Issue 2. Pages 354–389. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbr013.

BOSWELL, M., PATIL, G. P. 1970. Chance Mechanisms Generating the Negative Binomial Distribution. In *Random Counts in Models and Structures*. Volume 1. Penn State University Press. Pages 3–22. http://www.psupress.org/books/titles/0-271-00114-3.html.

BOUDT, K., CORNELISSEN, J., PAYSEUR, S., NGUYEN, G., SCHERMER, M. 2018. *Package 'highfrequency'*. https://cran.r-project.org/package=highfrequency.

BOUGEROL, P. 1993. Kalman Filtering with Random Coefficients and Contractions. *SIAM Journal on Control and Optimization*. Volume 31. Issue 4. Pages 942–959. ISSN 0363-0129. https://doi.org/10.1137/0331041.

BOWEN, D., HUTCHINSON, M. C., O'SULLIVAN, N. 2010. High-Frequency Equity Pairs Trading: Transaction Costs, Speed of Execution, and Patterns in Returns. *Journal of Trading*. Volume 5. Issue 3. Pages 31–38. ISSN 1559-3967. https://doi.org/10.3905/jot.2010.5.3.031.

BOWEN, D. A., HUTCHINSON, M. C. 2016. Pairs Trading in the UK Equity Market: Risk and Return. *European Journal of Finance*. Volume 22. Issue 14. Pages 1363–1387. ISSN 1351-847X. https://doi.org/10.1080/1351847X.2014.953698.

BOX, T., GRIFFITH, T. 2016. Price Clustering Asymmetries in Limit Order Flows. *Financial Management*. Volume 45. Issue 4. Pages 1041–1066. ISSN 0046-3892. https://doi.org/10.1111/fima.12136.

BROCKWELL, P. J., DAVIS, R. A., YANG, Y. 2007. Estimation for Nonnegative Lévy-Driven Ornstein-Uhlenbeck Processes. *Journal of Applied Probability*. Volume 44. Issue 4. Pages 977–989. ISSN 0021-9002. https://doi.org/10.2307/27595901.

BROUSSARD, J. P., VAIHEKOSKI, M. 2012. Profitability of Pairs Trading Strategy in an Illiquid Market with Multiple Share Classes. *Journal of International Financial Markets, Institutions and Money*. Volume 22. Issue 5. Pages 1188–1201. ISSN 1042-4431. https://doi.org/10.1016/j.intfin.2012.06.002.

BROWN, P., MITCHELL, J. 2008. Culture and Stock Price Clustering: Evidence from The Peoples' Republic of China. *Pacific-Basin Finance Journal*. Volume 16. Issue 1-2. Pages 95–120. ISSN 0927-538X. https://doi.org/10.1016/j.pacfin.2007.04.005.

BROWNLEES, C. T., GALLO, G. M. 2006. Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns. *Computational Statistics & Data Analysis*. Volume 51. Issue 4. Pages 2232–2245. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2006.09.030.

BROWNLEES, C. T., GALLO, G. M. 2010. Comparison of Volatility Measures: A Risk Management Perspective. *Journal of Financial Econometrics*. Volume 8. Issue 1. Pages 29–56. ISSN 1479-8417. https://doi.org/10.1093/jjfinec/nbp009.

BROYDEN, C. G. 1970a. The Convergence of a Class of Double-Rank Minimization Algorithms - 1. General Considerations. *IMA Journal of Applied Mathematics*. Volume 6. Issue 1. Pages 76–90. ISSN 0272-4960. https://doi.org/10.1093/imamat/6.1.76.

BROYDEN, C. G. 1970b. The Convergence of a Class of Double-Rank Minimization Algorithms - 2. The New Algorithm. *IMA Journal of Applied Mathematics*. Volume 6. Issue 3. Pages 222–231. ISSN 0272-4960. https://doi.org/10.1093/imamat/6.3.222.

BUSCH, T., CHRISTENSEN, B. J., NIELSEN, M. O. 2011. The Role of Implied Volatility in Forecasting Future Realized Volatility and Jumps in Foreign Exchange, Stock, and Bond Markets. *Journal of Econometrics*. Volume 160. Issue 1. Pages 48–57. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.014.

CAI, Z. 2007. Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors. *Journal of Econometrics*. Volume 136. Issue 1. Pages 163–188. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2005.08.004.

CALDEIRA, J. F., MOURA, G. V. 2013. Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy. *Brazilian Review of Finance*. Volume 11. Issue 1. Pages 49–80. ISSN 1679-0731. https://bibliotecadigital.fgv.br/ojs/index.php/rbfin/article/view/4785.

CAMERON, A. C., TRIVEDI, P. K. 1986. Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*. Volume 1. Issue 1. Pages 29–53. ISSN 0883-7252. https://doi.org/10.2307/2096536.

CAMERON, A. C., TRIVEDI, P. K. 2013. *Regression Analysis of Count Data*. Second Edition. New York. Cambridge University Press. ISBN 978-1-107-01416-9. https://doi.org/10.1017/ccol0521632013.

CANOVA, F., HANSEN, B. E. 1995. Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability. *Journal of Business & Economic Statistics*. Volume 13. Issue 3. Pages 237–252. ISSN 0735-0015. https://doi.org/10.2307/1392184.

ČECH, F., BARUNÍK, J. 2017. On the Modelling and Forecasting of Multivariate Realized Volatility: Generalized Heterogeneous Autoregressive (GHAR) Model. *Journal of Forecasting*. Volume 36. Issue 2. Pages 181–206. ISSN 0277-6693. https://doi.org/10.1002/for.2423.

ÇELIK, S., ERGIN, H. 2014. Volatility Forecasting Using High Frequency Data: Evidence from Stock Markets. *Economic Modelling*. Volume 36. Pages 176–190. ISSN 0264-9993. https://doi.org/10.1016/j.econmod.2013.09.038.

ČERNÝ, M. 2018. *Narrow Big Data in a Stream: Computational Limitations and Regression*. Technical Report. https://nb.vse.cz/~cernym/tr12018.pdf.

ČERNÝ, M., HLADÍK, M. 2014. The Complexity of Computation and Approximation of the t-Ratio Over One-Dimensional Interval Data. *Computational Statistics & Data Analysis*. Volume 80. Pages 26–43. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2014.06.007.

ČERNÝ, M., SOKOL, O. 2015. Interval Data and Sample Variance: A Study of an Efficiently Computable Case. In *Proceedings of the 33th International Conference Mathematical Methods in Economics*. Cheb. University of West Bohemia, Plzeň. Pages 99–104. ISBN 978-80-261-0539-8. https://mme2015.zcu.cz/conference-proceedings/.

CHENG, X., YU, P. L., LI, W. K. 2011. Basket Trading Under Co-Integration with the Logistic Mixture Autoregressive Model. *Quantitative Finance*. Volume 11. Issue 9. Pages 1407–1419. ISSN 1469-7688. https://doi.org/10.1080/14697688.2010.506445.

CHIRIAC, R., VOEV, V. 2008. Volatility Modelling and Forecasting Multivariate Realized Volatility. *Journal of Applied Econometrics*. Volume 26. Issue 6. Pages 922–947. ISSN 0883-7252. https://doi.org/10.1002/jae.1152.

CHRISTENSEN, K., KINNEBROCK, S., PODOLSKIJ, M. 2010. Pre-Averaging Estimators of the Ex-Post Covariance Matrix in Noisy Diffusion Models with Non-Synchronous Data. *Journal of Econometrics*. Volume 159. Issue 1. Pages 116–133. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.05.001.

CHRISTENSEN, K., OOMEN, R. C. A., PODOLSKIJ, M. 2014. Fact or Friction: Jumps at Ultra High Frequency. *Journal of Financial Economics*. Volume 114. Issue 3. Pages 576–599. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2014.07.007.

CHRISTOFFERSEN, P., FEUNOU, B., JACOBS, K., MEDDAHI, N. 2014. The Economic Value of Realized Volatility: Using High-Frequency Returns for Option Valuation. *Journal of Financial and Quantitative Analysis*. Volume 49. Issue 3. Pages 663–697. ISSN 0022-1090. https://doi.org/10.1017/S0022109014000428.

CHRISTOU, V., FOKIANOS, K. 2014. Quasi-Likelihood Inference for Negative Binomial Time Series Models. *Journal of Time Series Analysis*. Volume 35. Issue 1. Pages 55–78. ISSN 0143-9782. https://doi.org/10.1111/jtsa.12050.

CHUNG, K. L., WILLIAMS, R. 1990. *Introduction to Stochastic Integration*. Second Edition. Boston. Birkhäuser. ISBN 978-1-4612-8837-4. https://doi.org/10.1007/978-1-4612-4480-6.

CHUNG, K. H., VAN NESS, B. F., VAN NESS, R. A. 2004. Trading Costs and Quote Clustering on the NYSE and NASDAQ after Decimalization. *Journal of Financial Research*. Volume 27. Issue 3. Pages 309–328. ISSN 0270-2592. https://doi.org/10.1111/j.1475-6803.2004.00096.x.

CHUNG, K. H., KIM, K. A., KITSABUNNARAT, P. 2005. Liquidity and Quote Clustering in a Market with Multiple Tick Sizes. *Journal of Financial Research*. Volume 28. Issue 2. Pages 177–195. ISSN 0270-2592. https://doi.org/10.1111/j.1475-6803.2005.00120.x.

CLEGG, M., KRAUSS, C. 2018. Pairs Trading with Partial Cointegration. *Quantitative Finance*. Volume 18. Issue 1. Pages 121–138. ISSN 1469-7688. https://doi.org/10.1080/14697688.2017.1370122.

CLEMENTS, M. P., GALVÃO, A. B., KIM, J. H. 2008. Quantile Forecasts of Daily Exchange Rate Returns from Forecasts of Realized Volatility. *Journal of Empirical Finance*. Volume 15. Issue 4. Pages 729–750. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2007.12.001.

CONT, R. 2001. Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance*. Volume 1. Issue 2. Pages 223–236. ISSN 1469-7696. https://doi.org/10.1080/713665670.

CORSI, F. 2009. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*. Volume 7. Issue 2. Pages 174–196. ISSN 1479-8417. https://doi.org/10.1093/jjfinec/nbp001.

CORSI, F., FUSARI, N., LA VECCHIA, D. 2013. Realizing Smiles: Options Pricing with Realized Volatility. *Journal of Financial Economics*. Volume 107. Issue 2. Pages 284–304. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2012.08.015.

COX, D. R. 1981. Statistical Analysis of Time Series: Some Recent Developments. *Scandinavian Journal of Statistics*. Volume 8. Issue 2. Pages 93–108. ISSN 0303-6898. https://doi.org/10.2307/4615819.

COX, D. R. 1962. *Renewal Theory*. First Edition. London. Methuen. ISBN 978-0-412-20570-5. https://books.google.com/books?id=0VxRAAAAMAAJ.

COX, D. R., MILLER, H. D. 1965. *The Theory of Stochastic Processes*. First Edition. London. Methuen. ISBN 978-0-416-23760-3. https://books.google.com/books?id=1f0hYAAACAAJ.

CREAL, D., KOOPMAN, S. J., LUCAS, A. 2008. *A General Framework for Observation Driven Time-Varying Parameter Models*. Working Paper. http://www.tinbergen.nl/discussionpaper/?paper=1416.

CREAL, D., KOOPMAN, S. J., LUCAS, A. 2013. Generalized Autoregressive Score Models with Applications. *Journal of Applied Econometrics*. Volume 28. Issue 5. Pages 777–795. ISSN 0883-7252. https://doi.org/10.1002/jae.1279.

CUMMINS, M., BUCCA, A. 2012. Quantitative Spread Trading on Crude Oil and Refined Products Markets. *Quantitative Finance*. Volume 12. Issue 12. Pages 1857–1875. ISSN 1469-7688. https://doi.org/10.1080/14697688.2012.715749.

CURCI, G., CORSI, F. 2012. Discrete Sine Transform for Multi-Scale Realized Volatility Measures. *Quantitative Finance*. Volume 12. Issue 2. Pages 263–279. ISSN 1469-7688. https://doi.org/10.1080/14697688.2010.490561.

DAHLHAUS, R., NEDDERMEYER, J. C. 2014. Online Spot Volatility-Estimation and Decomposition with Nonlinear Market Microstructure Noise Models. *Journal of Financial Econometrics*. Volume 12. Issue 1. Pages 174–212. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbt008.

DARÓCZI, G. 2017. *Number of R packages submitted to CRAN*. Online. https://gist.github.com/daroczig/3cf06d6db4be2bbe3368.

DAVIS, R. A., DUNSMUIR, W. T. M., STREET, S. B. 2003. Observation-Driven Models for Poisson Counts. *Biometrika*. Volume 90. Issue 4. Pages 777–790. ISSN 0006-3444. https://doi.org/10.1093/biomet/90.4.777.

DAVIS, R. L., VAN NESS, B. F., VAN NESS, R. A. 2014. Clustering of Trade Prices by High-Frequency and Non-High-Frequency Trading Firms. *Financial Review*. Volume 49. Issue 2. Pages 421–433. ISSN 0732-8516. https://doi.org/10.1111/fire.12042.

DE MOURA, C. E., PIZZINGA, A., ZUBELLI, J. 2016. A Pairs Trading Strategy Based on Linear State Space Models and the Kalman Filter. *Quantitative Finance*. Volume 16. Issue 10. Pages 1559–1573. ISSN 1469-7696. https://doi.org/10.1080/14697688.2016.1164886.

DE POOTER, M., MARTENS, M., VAN DIJK, D. 2008. Predicting the Daily Covariance Matrix for S&P 100 Stocks Using Intraday Data: But Which Frequency to Use? *Econometric Reviews*. Volume 27. Issue 1-3. Pages 199–229. ISSN 0747-4938. https://doi.org/10.1080/07474930701873333.

DELATTRE, S., JACOD, J. 1997. A Central Limit Theorem for Normalized Functions of the Increments of a Diffusion Process, in the Presence of Round-Off Errors. *Bernoulli*. Volume 3. Issue 1. Pages 1–28. ISSN 1350-7265. https://doi.org/10.2307/3318650.

DELBAEN, F., SCHACHERMAYER, W. 1994. A General Version of the Fundamental Theorem of Asset Pricing. *Mathematische Annalen*. Volume 300. Issue 1. Pages 463–520. ISSN 0025-5831. https://doi.org/10.1007/bf01450498.

DIEBOLD, F. X., MARIANO, R. S. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*. Volume 13. Issue 3. Pages 253–263. ISSN 0735-0015. https://doi.org/10.1080/07350015.1995.10524599.

DIEBOLD, F. X., STRASSER, G. 2013. On the Correlation Structure of Microstructure Noise: A Financial Economic Approach. *The Review of Economic Studies*. Volume 80. Issue 4. Pages 1304–1337. ISSN 0034-6527. https://doi.org/10.1093/restud/rdt008.

DIKS, C., PANCHENKO, V., DIJK, D. 2011. Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails. *Journal of Econometrics*. Volume 163. Issue 2. Pages 215–230. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2011.04.001.

DIONNE, G., DUCHESNE, P., PACURAR, M. 2009. Intraday Value at Risk (IVaR) Using Tick-by-Tick Data with Application to the Toronto Stock Exchange. *Journal of Empirical Finance*. Volume 16. Issue 5. Pages 777–792. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2009.05.005.

DO, B., FAFF, R. 2010. Does Simple Pairs Trading Still Work? *Financial Analysts Journal*. Volume 66. Issue 4. Pages 83–95. ISSN 0015-198X. https://doi.org/10.2307/25741293.

DO, B., FAFF, R. 2012. Are Pairs Trading Profits Robust to Transaction Costs? *Journal of Financial Research*. Volume 35. Issue 2. Pages 261–287. ISSN 0270-2592. https://doi.org/10.1111/j.1475-6803.2012.01317.x.

DONG, Y., TSE, Y.-K. 2017a. Business Time Sampling Scheme with Applications to Testing Semi-Martingale Hypothesis and Estimating Integrated Volatility. *Econometrics*. Volume 5. Issue 4. Pages 51/1–51/19. ISSN 2225-1146. https://doi.org/10.3390/econometrics5040051.

DONG, Y., TSE, Y.-K. 2017b. On Estimating Market Microstructure Noise Variance. *Economics Letters*. Volume 150. Pages 59–62. ISSN 0165-1765. https://doi.org/10.1016/j.econlet.2016.11.009.

DUFOUR, A., ENGLE, R. 2000. Time and the Price Impact of a Trade. *The Journal of Finance*. Volume 55. Issue 6. Pages 2467–2498. ISSN 0022-1082. https://doi.org/10.1111/0022-1082.00297.

DUNIS, C., LEQUEUX, P. 2000. Intraday Data and Hedging Efficiency in Interest Spread Trading. *The European Journal of Finance*. Volume 6. Issue 4. Pages 332–352. ISSN 1351-847X. https://doi.org/10.1080/13518470050195100.

EDDELBUETTEL, D. 2019. *CRAN Task View: Empirical Finance*. Online. https://cran.r-project.org/view=Finance.

ELLIOTT, R. J., VAN DER HOEK, J., MALCOLM, W. P. 2005. Pairs Trading. *Quantitative Finance*. Volume 5. Issue 3. Pages 271–276. ISSN 1469-7688. https://doi.org/10.1080/14697680500149370.

ELSEVIER. 2019. *Scopus*. Online. https://www.scopus.com.

EMERY, G. W., LIU, Q. 2002. An Analysis of the Relationship between Electricity and Natural-Gas Futures Prices. *Journal of Futures Markets*. Volume 22. Issue 2. Pages 95–122. ISSN 0270-7314. https://doi.org/10.1002/fut.2209.

ENGLE, R. F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*. Volume 50. Issue 4. Pages 987–1007. ISSN 0012-9682. https://doi.org/10.2307/1912773.

ENGLE, R. F. 2000. The Econometrics of Ultra-High-Frequency Data. *Econometrica*. Volume 68. Issue 1. Pages 1–22. ISSN 0012-9682. https://doi.org/10.1111/1468-0262.00091.

ENGLE, R. F. 2002. Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business and Economic Statistics*. Volume 20. Issue 3. Pages 339–350. ISSN 0735-0015. https://doi.org/10.1198/073500102288618487.

ENGLE, R. F., LANGE, J. 2001. Predicting VNET: A Model of the Dynamics of Market Depth. *Journal of Financial Markets*. Volume 4. Issue 2. Pages 113–142. ISSN 1386-4181. https://doi.org/10.1016/S1386-4181(00)00019-7.

ENGLE, R. F., MANGANELLI, S. 2004. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics*. Volume 22. Issue 4. Pages 367–381. ISSN 0735-0015. https://doi.org/10.1198/073500104000000370.

ENGLE, R. F., RUSSELL, J. R. 1998. Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*. Volume 66. Issue 5. Pages 1127–1162. ISSN 0012-9682. https://doi.org/10.2307/2999632.

FALKENBERRY, T. N. 2002. *High Frequency Data Filtering*. Technical Report. https://s3-us-west-2.amazonaws.com/tick-data-s3/pdf/Tick_Data_Filtering_White_Paper.pdf.

FAN, J., WANG, Y. 2008. Spot Volatility Estimation for High-Frequency Data. *Statistics and Its Interface*. Volume 1. Issue 2. Pages 279–288. ISSN 1938-7989. https://doi.org/10.4310/SII.2008.v1.n2.a5.

FASEN, V. 2013. Statistical Estimation of Multivariate Ornstein-Uhlenbeck Processes and Applications to Co-Integration. *Journal of Econometrics*. Volume 172. Issue 2. Pages 325–337. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2012.08.019.

FERNANDES, M., GRAMMIG, J. 2005. Nonparametric Specification Tests for Conditional Duration Models. *Journal of Econometrics*. Volume 127. Issue 1. Pages 35–68. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2004.06.003.

FERNANDES, M., GRAMMIG, J. 2006. A Family of Autoregressive Conditional Duration Models. *Journal of Econometrics*. Volume 130. Issue 1. Pages 1–23. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2004.08.016.

FERSON, S., GINZBURG, L., KREINOVICH, V., LONGPRÉ, L., AVILES, M. 2005. Exact Bounds on Finite Populations of Interval Data. *Reliable Computing*. Volume 11. Issue 3. Pages 207–233. ISSN 1385-3139. https://doi.org/10.1007/s11155-005-3616-1.

FERSON, S., KREINOVICH, V., HAJAGOS, J., OBERKAMPF, W. L., GINZBURG, L. 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Technical Report. https://doi.org/10.2172/910198.

FLEMING, E., KIRBY, C., OSTDIEK, B. 2003. The Economic Value of Volatility Timing Using "Realized" Volatility. *Journal of Financial Economics*. Volume 67. Issue 3. Pages 473–509. ISSN 0304-405X. https://doi.org/10.1016/s0304-405x(02)00259-3.

FLETCHER, R. 1970. A New Approach to Variable Metric Algorithms. *The Computer Journal*. Volume 13. Issue 3. Pages 317–322. ISSN 0010-4620. https://doi.org/0.1093/comjnl/13.3.317.

FORONI, C., MARCELLINO, M., SCHUMACHER, C. 2015. Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials. *Journal of the Royal Statistical Society: Series A (General)*. Volume 178. Issue 1. Pages 57–82. ISSN 0035-9238. https://doi.org/10.1111/rssa.12043.

FRANCQ, C., GAUTIER, A. 2004. Estimation of Time-Varying ARMA Models with Markovian Changes in Regime. *Statistics & Probability Letters*. Volume 70. Issue 4. Pages 243–251. ISSN 0167-7152. https://doi.org/10.1016/j.spl.2004.10.009.

FUKASAWA, M. 2010a. Central Limit Theorem for the Realized Volatility Based on Tick Time Sampling. *Finance and Stochastics*. Volume 14. Issue 2. Pages 209–233. ISSN 0949-2984. https://doi.org/10.1007/s00780-008-0087-3.

FUKASAWA, M. 2010b. Realized Volatility with Stochastic Sampling. *Stochastic Processes and Their Applications*. Volume 120. Issue 6. Pages 829–852. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2010.02.006.

GALENKO, A., POPOVA, E., POPOVA, I. 2012. Trading in the Presence of Cointegration. *The Journal of Alternative Investments*. Volume 15. Issue 1. Pages 85–97. ISSN 1520-3255. https://doi.org/10.3905/jai.2012.15.1.085.

GALLANT, A. R., WHITE, H. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. First Edition. Oxford. Basil Blackwell. ISBN 978-0-631-15765-6. https://books.google.com/books?id=VVOqQgAACAAJ.

GARMAN, M. B. 1976. Market Microstructure. *Journal of Financial Economics*. Volume 3. Issue 3. Pages 257–275. ISSN 0304-405X. https://doi.org/10.1016/0304-405x(76)90006-4.

GATEV, E., GOETZMANN, W. N., ROUWENHORST, K. G. 2006. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *Review of Financial Studies*. Volume 19. Issue 3. Pages 797–827. ISSN 0893-9454. https://doi.org/10.1093/rfs/hhj020.

GATHERAL, J., OOMEN, R. C. A. 2010. Zero-Intelligence Realized Variance Estimation. *Finance and Stochastics*. Volume 14. Issue 2. Pages 249–283. ISSN 0949-2984. https://doi.org/10.1007/s00780-009-0120-1.

GHALANOS, A. 2018. *Package 'rugarch'*. https://cran.r-project.org/package=rugarch.

GHYSELS, E., SINKO, A. 2011. Volatility Forecasting and Microstructure Noise. *Journal of Econometrics*. Volume 160. Issue 1. Pages 257–271. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.035.

GHYSELS, E., SANTA-CLARA, P., VALKANOV, R. 2004. *The MIDAS Touch: Mixed Data Sampling Regression Models*. Working Paper. https://escholarship.org/uc/item/9mf223rs.

GHYSELS, E., SANTA-CLARA, P., VALKANOV, R. 2006. Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies. *Journal of Econometrics*. Volume 131. Issue 1-2. Pages 59–95. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2005.01.004.

GHYSELS, E., SINKO, A., VALKANOV, R. 2007. MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*. Volume 26. Issue 1. Pages 53–90. ISSN 0747-4938. https://doi.org/10.1080/07474930600972467.

GIL-ALANA, L. A. 2000. Mean Reversion in the Real Exchange Rates. *Economics Letters*. Volume 69. Issue 3. Pages 258–288. ISSN 0165-1765. https://doi.org/10.1016/s0165-1765(00)00318-9.

GIOT, P. 2005. Market Risk Models for Intraday Data. *European Journal of Finance*. Volume 11. Issue 4. Pages 309–324. ISSN 1351-847X. https://doi.org/10.1080/1351847032000143396.

GIOT, P., GRAMMIG, J. 2006. How Large is Liquidity Risk in an Automated Auction Market? *Empirical Economics*. Volume 30. Issue 4. Pages 867–887. ISSN 03777332. https://doi.org/10.1007/s00181-005-0003-z.

GIOT, P., LAURENT, S. 2004. Modelling Daily Value-at-Risk Using Realized Volatility and ARCH Type Models. *Journal of Empirical Finance*. Volume 11. Issue 3. Pages 379–398. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2003.04.003.

GIRMA, P. B., PAULSON, A. S. 1999. Risk Arbitrage Opportunities in Petroleum Futures Spreads. *Journal of Futures Markets*. Volume 19. Issue 8. Pages 931–955. ISSN 0270-7314. https://doi.org/10.1002/(sici)1096-9934(199912)19:8<931::aid-fut5>3.0.co;2-l.

GOLDFARB, D. 1970. A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation*. Volume 24. Issue 109. Pages 23–26. ISSN 0025-5718. https://doi.org/10.2307/2004873.

GONÇALVES, S., MEDDAHI, N. 2009. Bootstrapping Realized Volatility. *Econometrica*. Volume 77. Issue 1. Pages 283–306. ISSN 0012-9682. https://doi.org/10.3982/ECTA5971.

GONÇALVES, S., MEDDAHI, N. 2011. Box-Cox Transforms for Realized Volatility. *Journal of Econometrics*. Volume 160. Issue 1. Pages 129–144. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.026.

GÖNCÜ, A., AKYILDIRIM, E. 2016. A Stochastic Model for Commodity Pairs Trading. *Quantitative Finance*. Volume 16. Issue 12. Pages 1843–1857. ISSN 1469-7696. https://doi.org/10.1080/14697688.2016.1211793.

GORGI, P. 2018. Integer-Valued Autoregressive Models with Survival Probability Driven By a Stochastic Recurrence Equation. *Journal of Time Series Analysis*. Volume 39. Issue 2. Pages 150–171. ISSN 0143-9782. https://doi.org/10.1111/jtsa.12272.

GOYENKO, R. Y., HOLDEN, C. W., TRZCINKA, C. A. 2009. Do Liquidity Measures Measure Liquidity? *Journal of Financial Economics*. Volume 92. Issue 2. Pages 153–181. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2008.06.002.

GRAMMIG, J., MAURER, K.-O. 2000. Non-Monotonic Hazard Functions and the Autoregressive Conditional Duration Model. *The Econometrics Journal*. Volume 3. Issue 1. Pages 16–38. ISSN 1368-4221. https://doi.org/10.1111/1368-423x.00037.

GRAMMIG, J., WELLNER, M. 2002. Modeling the Interdependence of Volatility and Inter-Transaction Duration Processes. *Journal of Econometrics*. Volume 106. Issue 2. Pages 369–400. https://doi.org/10.1016/S0304-4076(01)00105-1.

GREENE, W. H. 1994. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*. Working Paper. http://ssrn.com/abstract=1293115.

GRIFFIN, J. E., STEEL, M. F. J. 2006. Inference with Non-Gaussian Ornstein-Uhlenbeck Processes for Stochastic Volatility. *Journal of Econometrics*. Volume 134. Issue 2. Pages 605–644. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2005.07.007.

GRIMSHAW, S. D., MCDONALD, J., MCQUEEN, G. R., THORLEY, S. 2005. Estimating Hazard Functions for Discrete Lifetimes. *Communications in Statistics - Simulation and Computation*. Volume 34. Issue 2. Pages 451–463. ISSN 0361-0918. https://doi.org/10.1081/SAC-200055732.

GUO, M.-Y., ZHANG, S.-Y. 2008. Study on CVaR Forecasts Based on Weighted Realized Volatility. In *Proceedings of the 15th International Conference on Management Science and Engineering*. Long Beach. IEEE. Pages 91–96. ISBN 978-1-4244-2387-3. https://doi.org/10.1109/icmse.2008.4668899.

GUTIERREZ, J. A., TSE, Y. 2011. Illuminating the Profitability of Pairs Trading: A Test of the Relative Pricing Efficiency of Markets for Water Utility Stocks. *The Journal of Trading*. Volume 6. Issue 2. Pages 50–64. ISSN 1559-3967. https://doi.org/10.3905/jot.2011.6.2.050.

HANSEN, P. R., LUNDE, A. 2006. Realized Variance and Market Microstructure Noise. *Journal of Business & Economic Statistics*. Volume 24. Issue 2. Pages 127–161. ISSN 0735-0015. https://doi.org/10.1198/073500106000000071.

HANSEN, P. R., HUANG, Z., SHEK, H. H. 2012. Realized GARCH: A Joint Model for Returns and Realized Measures of Volatility. *Journal of Applied Econometrics*. Volume 27. Issue 6. Pages 877–906. ISSN 0883-7252. https://doi.org/10.1002/jae.1234.

HANSEN, P. R., LUNDE, A., VOEV, V. 2014. Realized Beta GARCH: A Multivariate GARCH Model with Realized Measures of Volatility. *Journal of Applied Econometrics*. Volume 29. Issue 5. Pages 774–799. ISSN 0883-7252. https://doi.org/10.1002/jae.2389.

HÄRDLE, W. K., HAUTSCH, N., MIHOCI, A. 2012. Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics. *Journal of Empirical Finance*. Volume 19. Issue 4. Pages 610–625. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2012.04.002.

HARVEY, A. C. 2013. *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. First Edition. New York. Cambridge University Press. ISBN 978-1-107-63002-4. https://doi.org/0.1017/cbo9781139540933.

HAUTSCH, N. 2001. *Modelling Intraday Trading Activity Using Box-Cox ACD Models*. Working Paper. https://ssrn.com/abstract=289643.

HAUTSCH, N. 2003. Assessing the Risk of Liquidity Suppliers on the Basis of Excess Demand Intensities. *Journal of Financial Econometrics*. Volume 1. Issue 2. Pages 189–215. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbg010.

HAUTSCH, N. 2011. *Econometrics of Financial High-Frequency Data*. First Edition. Berlin, Heidelberg. Springer. ISBN 978-3-642-21924-5. https://doi.org/10.1007/978-3-642-21925-2.

HAUTSCH, N., HUANG, R. 2012. The Market Impact of a Limit Order. *Journal of Economic Dynamics and Control*. Volume 36. Issue 4. Pages 501–522. ISSN 0165-1889. https://doi.org/10.1016/j.jedc.2011.09.012.

HAUTSCH, N., PODOLSKIJ, M. 2013. Preaveraging-Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence. *Journal of Business & Economic Statistics*. Volume 31. Issue 2. Pages 165–183. ISSN 0735-0015. https://doi.org/10.1080/07350015.2012.754313.

HAUTSCH, N., MALEC, P., SCHIENLE, M. 2014. Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes. *Journal of Financial Econometrics*. Volume 12. Issue 1. Pages 89–121. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbt002.

HAUTSCH, N., KYJ, L. M., MALEC, P. 2015. Do High-Frequency Data Improve High-Dimensional Portfolio Allocations? *Journal of Applied Econometrics*. Volume 30. Issue 2. Pages 263–290. ISSN 1099-1255. https://doi.org/10.1002/jae.2361.

HAYASHI, T., YOSHIDA, N. 2005. On Covariance Estimation of Non-Synchronously Observed Diffusion Processes. *Bernoulli*. Volume 11. Issue 2. Pages 359–379. ISSN 1350-7265. https://doi.org/10.3150/bj/1116340299.

HENDERSHOTT, T., MENKVELD, A. J. 2014. Price Pressures. *Journal of Financial Economics*. Volume 114. Issue 3. Pages 405–423. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2014.08.001.

HENDRYCH, R., CIPRA, T. 2018. Self-Weighted Recursive Estimation of GARCH Models. *Communications in Statistics - Simulation and Computation*. Volume 47. Issue 2. Pages 315–328. ISSN 0361-0918. https://doi.org/10.1080/03610910600880328.

HERRERA, R., SCHIPP, B. 2013. Value at Risk Forecasts by Extreme Value Models in a Conditional Duration Framework. *Journal of Empirical Finance*. Volume 23. Pages 33–47. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2013.05.002.

HILBE, J. M. 2011. *Negative Binomial Regression*. Second Edition. New York. Cambridge University Press. ISBN 978-0-521-19815-8. https://doi.org/10.1017/cbo9780511811852.

HOFMANN, K. F., SCHULZ, T. 2016. A General Ornstein-Uhlenbeck Stochastic Volatility Model with Lévy Jumps. *International Journal of Theoretical and Applied Finance*. Volume 19. Issue 8. Pages 1650044/1–1650044/23. ISSN 0219-0249. https://doi.org/10.1142/S0219024916500448.

HØG, E., LUNDE, A. 2003. *Wavelet Estimation of Integrated Volatility*. Working Paper. http://ideas.repec.org/p/sce/scecf3/274.html.

HOLÝ, V. 2016. Impact of Microstructure Noise on Integrated Variance Estimators: A Simulation Study. In *Proceedings of the 34th International Conference Mathematical Methods in Economics*. Liberec. Technical University of Liberec. Pages 289–294. ISBN 978-80-7494-296-9. http://mme2016.tul.cz/index.php?page=conferenceproceedings.

HOLÝ, V. 2017a. Market Microstructure Noise. In *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze*. Praha. Oeconomica. Pages 96–104. ISBN 978-80-245-2199-2. https://fis.vse.cz/studium/doktorske-studium/den-doktorandu/den-doktorandu-2017.

HOLÝ, V. 2017b. Estimating Integrated Variance in the Presence of Microstructure Noise Using Linear Regression. In *Proceedings of the 14th International Conference of Numerical Analysis and Applied Mathematics*. AIP Conference Proceedings. Volume 1863. Rhodes. American Institute of Physics. Pages 560053/1–560053/4. ISBN 978-0-7354-1538-6. https://doi.org/10.1063/1.4992736.

HOLÝ, V. 2017c. Comparison of Integrated Variance Forecasts. In *Proceedings of the 1st International Conference on Applied Statistics and Econometrics*. Tirana. Epoka University. Pages 52–58. ISBN 978-9928-135-20-9. http://icase.epoka.edu.al/2017/category-proceedings-1729.html.

HOLÝ, V. 2017d. Combining Estimates of Industry Production with the Structure of Input-Output Table. In *Proceedings of the 35th International Conference Mathematical Methods in Economics*. Hradec Králové. University of Hradec Králové. Pages 231–235. ISBN 978-80-7435-678-0. https://fim2.uhk.cz/mme/index.php?page=conferenceproceedings.

HOLÝ, V. 2017e. Quadratic Estimators of Quadratic Variation. In *Sborník mezinárodního vědeckého semináře Nové trendy v ekonometrii a operačním výzkumu*. Bratislava. Ekonóm. Pages 32–39. ISBN 978-80-225-4455-9. https://www.fhi.sk/files/netrinecop/Praha2017.pdf.

HOLÝ, V. 2018a. Why Use High-Frequency Data in Quantitative Finance? In *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze*. Praha. Oeconomica. Pages 44–51. ISBN 978-80-245-2259-3. https://fis.vse.cz/studium/doktorske-studium/den-doktorandu/den-doktorandu-2018.

HOLÝ, V. 2018b. How Big Is the Rounding Error in Financial High-Frequency Data? In *Proceedings of the 15th International Conference of Numerical Analysis and Applied Mathematics*. AIP Conference Proceedings. Volume 1978. Thessaloniki. American Institute of Physics. Pages 470076/1–470076/4. ISBN 978-0-7354-1690-1. https://doi.org/10.1063/1.5044146.

HOLÝ, V. 2018c. Analyzing Financial High-Frequency Data in R. In *Proceedings of the 7th International Conference on Management*. Prešov. Bookman s.r.o. Pages 734–739. ISBN 978-80-8165-301-8. http://www.managerconf.com/.

HOLÝ, V., ČERNÝ, M. 2017. Finite-Sample Comparison of Integrated Variance Estimators. In *Proceedings of the 61st World Statistics Congress*. In Press. https://isi-web.org/index.php/publications/proceedings.

HOLÝ, V., SOKOL, O. 2018. Interval Approach in Quadratic Variation Estimation. In *Proceedings of the 36th International Conference Mathematical Methods in Economics*. Jindřichův Hradec. MatfyzPress. Pages 145–150. ISBN 978-80-7378-371-6. https://mme2018.fm.vse.cz/wp-content/uploads/2018/09/MME2018-Electronic_proceedings.pdf.

HOLÝ, V., TOMANOVÁ, P. 2018. *Estimation of Ornstein-Uhlenbeck Process Using Ultra-High-Frequency Data with Application to Intraday Pairs Trading Strategy*. Working Paper. https://arxiv.org/abs/1811.09312.

HU, B., JIANG, C., MCINISH, T., ZHOU, H. 2017. Price Clustering on the Shanghai Stock Exchange. *Applied Economics*. Volume 49. Issue 28. Pages 2766–2778. ISSN 0003-6846. https://doi.org/10.1080/00036846.2016.1248284.

HUANG, H., LEE, T.-H. 2013. Forecasting Value-at-Risk Using High-Frequency Information. *Econometrics*. Volume 1. Issue 1. Pages 127–140. ISSN 2225-1146. https://doi.org/10.3390/econometrics1010127.

HUANG, X., TAUCHEN, G. 2005. The Relative Contribution of Jumps to Total Price Variance. *Journal of Financial Econometrics*. Volume 3. Issue 4. Pages 456–499. ISSN 0304-405X. https://doi.org/10.1093/jjfinec/nbi025.

HUANG, Z., LIU, H., WANG, T. 2016. Modeling Long Memory Volatility Using Realized Measures of Volatility: A Realized HAR GARCH Model. *Economic Modelling*. Volume 52. Pages 812–821. ISSN 0264-9993. https://doi.org/10.1016/j.econmod.2015.10.018.

HUCK, N. 2013. The High Sensitivity of Pairs Trading Returns. *Applied Economics Letters*. Volume 20. Issue 14. Pages 1301–1304. ISSN 1350-4851. https://doi.org/10.1080/13504851.2013.802121.

HUCK, N. 2015. Pairs Trading: Does Volatility Timing Matter? *Applied Economics*. Volume 47. Issue 57. Pages 6239–6256. ISSN 0003-6846. https://doi.org/10.1080/00036846.2015.1068923.

HULL, J., WHITE, A. 1990. Pricing Interest-Rate-Derivative Securities. *Review of Financial Studies*. Volume 3. Issue 4. Pages 573–592. ISSN 0893-9454. https://doi.org/10.2307/2962116.

HYNDMAN, R. J. 2019. *CRAN Task View: Time Series Analysis*. Online. https://cran.r-project.org/view=TimeSeries.

HYNDMAN, R. J., KHANDAKAR, Y. 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*. Volume 27. Issue 3. Pages 1–22. ISSN 1939-0068. https://doi.org/10.18637/jss.v069.i12.

IKEDA, S. S. 2015. Two-Scale Realized Kernels: A Univariate Case. *Journal of Financial Econometrics*. Volume 13. Issue 1. Pages 126–165. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbt024.

IKENBERRY, D. L., WESTON, J. P. 2008. Clustering in US Stock Prices After Decimalisation. *European Financial Management*. Volume 14. Issue 1. Pages 30–54. ISSN 1354-7798. https://doi.org/10.1111/j.1468-036X.2007.00410.x.

JACOBS, H. 2015. What Explains the Dynamics of 100 Anomalies? *Journal of Banking and Finance*. Volume 57. Pages 65–85. ISSN 0378-4266. https://doi.org/10.1016/j.jbankfin.2015.03.006.

JACOBS, H., WEBER, M. 2015. On the Determinants of Pairs Trading Profitability. *Journal of Financial Markets*. Volume 23. Pages 75–97. ISSN 1386-4181. https://doi.org/10.1016/j.finmar.2014.12.001.

JACOD, J. 1996. La variation quadratique moyenne du brownien en presence d'erreurs d'arrondi. *Astérisque*. Volume 236. Pages 155–161. ISSN 0303-1179. http://smf4.emath.fr/Publications/Asterisque/1996/236/html/.

JACOD, J., MYKLAND, P. A. 2015. Microstructure Noise in the Continuous Case: Approximate Efficiency of the Adaptive Pre-Averaging Method. *Stochastic Processes and Their Applications*. Volume 125. Issue 8. Pages 2910–2936. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2015.02.005.

JACOD, J., ROSENBAUM, M. 2013. Quarticity and Other Functionals of Volatility: Efficient Estimation. *The Annals of Statistics*. Volume 41. Issue 3. Pages 1462–1484. ISSN 0090-5364. https://doi.org/10.1214/13-aos1115.

JACOD, J., LI, Y., MYKLAND, P. A., PODOLSKIJ, M., VETTER, M. 2009. Microstructure Noise in the Continuous Case: The Pre-Averaging Approach. *Stochastic Processes and Their Applications*. Volume 119. Issue 7. Pages 2249–2276. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2008.11.004.

JACOD, J., LI, Y., ZHENG, X. 2017. Statistical Properties of Microstructure Noise. *Econometrica*. Volume 85. Issue 4. Pages 1133–1174. ISSN 0012-9682. https://doi.org/10.3982/ecta13085.

JASIAK, J. 1998. Persistence in Intertrade Durations. *Finance*. Volume 19. Pages 166–195. ISSN 1556-5068. https://doi.org/10.2139/ssrn.162008.

JEON, S., CHANG, W., PARK, Y. 2016. An Option Pricing Model Using High Frequency Data. *Procedia Computer Science*. Volume 91. Pages 175–179. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2016.07.035.

JEYASREEDHARAN, N., ALLEN, D. E., YANG, J. W. 2014. Yet Another ACD Model: The Autoregressive Conditional Directional Duration (ACDD) Model. *Annals of Financial Economics*. Volume 9. Issue 1. Pages 1450004/1–1450004/20. ISSN 2010-4952. https://doi.org/10.1142/S2010495214500043.

JIANG, G. J., OOMEN, R. C. 2008. Testing for Jumps When Asset Prices Are Observed with Noise - A "Swap Variance" Approach. *Journal of Econometrics*. Volume 144. Issue 2. Pages 352–370. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2008.04.009.

JIANG, W., RUAN, Q., LI, J., LI, Y. 2018. Modeling Returns Volatility: Realized GARCH Incorporating Realized Risk Measure. *Physica A: Statistical Mechanics and Its Applications*. Volume 500. Pages 249–258. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2018.02.018.

JING, B. Y., KONG, X. B., LIU, Z. 2012a. Modeling High-Frequency Financial Data by Pure Jump Processes. *The Annals of Probability*. Volume 40. Issue 2. Pages 759–784. ISSN 0090-5364. https://doi.org/10.2307/41713654.

JING, B.-Y., KONG, X.-B., LIU, Z., MYKLAND, P. 2012b. On the Jump Activity Index for Semimartingales. *Journal of Econometrics*. Volume 166. Issue 2. Pages 213–223. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2011.09.036.

JING, B.-Y., LIU, Z., KONG, X.-B. 2017. Estimating Volatility Functionals with Multiple Transactions. *Econometric Theory*. Volume 33. Issue 2. Pages 331–365. ISSN 0266-4666. https://doi.org/10.1017/S0266466615000420.

JOHNSON, S. G. 2019. *The NLopt nonlinear-optimization package*. Online. http://ab-initio.mit.edu/nlopt.

KABASINSKAS, A., SAKALAUSKAS, L., SUN, E. W., BELOVAS, I. 2012. Mixed-Stable Models for Analyzing High-Frequency Financial Data. *Journal of Computational Analysis and Applications*. Volume 14. Issue 7. Pages 1210–1226. ISSN 1521-1398. https://www.researchgate.net/publication/259479011.

KANAMURA, T., RACHEV, S. T., FABOZZI, F. J. 2010. A Profit Model for Spread Trading with an Application to Energy Futures. *The Journal of Trading*. Volume 5. Issue 1. Pages 48–62. ISSN 1559-3967. https://doi.org/10.3905/jot.2010.5.1.048.

KARATZAS, I., SHREVE, S. E. 1991. *Brownian Motion and Stochastic Calculus*. Second Edition. New York. Springer. ISBN 978-0-387-97655-6. https://doi.org/10.1007/BF00046894.

KELLY, M. A., CLARK, S. P. 2011. Returns in Trading Versus Non-Trading Hours: The Difference is Day and Night. *Journal of Asset Management*. Volume 12. Issue 2. Pages 132–145. ISSN 1470-8272. https://doi.org/10.1057/jam.2011.2.

KENMOE, R. N., SANFELICI, S. 2014. An Application of Nonparametric Volatility Estimators to Option Pricing. *Decisions in Economics and Finance*. Volume 37. Issue 2. Pages 393–412. ISSN 1593-8883. https://doi.org/10.1007/s10203-013-0150-1.

KEVEI, P. 2018. Ergodic Properties of Generalized Ornstein-Uhlenbeck Processes. *Stochastic Processes and Their Applications*. Volume 128. Issue 1. Pages 156–181. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2017.04.010.

KLEBANER, F. C. 2005. *Introduction to Stochastic Calculus with Applications*. Second Edition. London. Imperial College Press. ISBN 978-1-86094-555-7. https://doi.org/10.1142/p386.

KONG, X.-B. 2012. Confidence Interval of the Jump Activity Index Based on Empirical Likelihood Using High Frequency Data. *Journal of Statistical Planning and Inference*. Volume 142. Issue 6. Pages 1378–1387. ISSN 0378-3758. https://doi.org/10.1016/j.jspi.2011.12.016.

KOOPMAN, S. J., LIT, R. 2017. *Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models*. Working Paper. http://www.tinbergen.nl/discussionpaper/?paper=2797.

KOOPMAN, S. J., JUNGBACKER, B., HOL, E. 2005. Forecasting Daily Variability of the S&P 100 Stock Index Using Historical, Realised and Implied Volatility Measurements. *Journal of Empirical Finance*. Volume 12. Issue 3. Pages 445–475. ISSN 09275398. https://doi.org/10.1016/j.jempfin.2004.04.009.

KOOPMAN, S. J., RUTGER, L., LUCAS, A. 2015. *Intraday Stock Price Dependence Using Dynamic Discrete Copula Distributions*. Working Paper. http://www.tinbergen.nl/discussionpaper/?paper=2461.

KOOPMAN, S. J., LUCAS, A., SCHARTH, M. 2016. Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models. *Review of Economics and Statistics*. Volume 98. Issue 1. Pages 97–110. ISSN 0034-6535. https://doi.org/10.1162/rest_a_00533.

KOOPMAN, S. J., LIT, R., LUCAS, A., OPSCHOOR, A. 2018. Dynamic Discrete Copula Models for High-Frequency Stock Price Changes. *Journal of Applied Econometrics*. Volume 33. Issue 7. Pages 966–985. ISSN 0883-7252. https://doi.org/10.1002/jae.2645.

KOSTIN, A., ZEMNITSKIY, A., NECHAEV, O. 2016. *Package 'PortfolioEffectEstim'*. https://cran.r-project.org/package=PortfolioEffectEstim.

KOSTIN, A., ZEMNITSKIY, A., NECHAEV, O. 2017. *Package 'PortfolioEffectHFT'*. https://cran.r-project.org/package=PortfolioEffectHFT.

KOSULAJEFF, P. 1937. Sur la répartition de la partie fractionnaire d'une variable. *Matematičeskij sbornik, Novaja serija*. Volume 2. Issue 5. Pages 1017–1019. ISSN 0368-8666. http://mi.mathnet.ru/eng/msb5641.

KRAUSS, C. 2017. Statistical Arbitrage Pairs Trading Strategies: Review and Outlook. *Journal of Economic Surveys*. Volume 31. Issue 2. Pages 513–545. ISSN 0950-0804. https://doi.org/10.1111/joes.12153.

KREINOVICH, V. 1996. Maximum Entropy and Interval Computations. *Reliable Computing*. Volume 2. Issue 1. Pages 63–79. ISSN 1385-3139. https://doi.org/10.1007/bf02388188.

KREINOVICH, V., LONGPRE, L. 2004. *Computing Higher Central Moments for Interval Data*. Technical Report. http://digitalcommons.utep.edu/cs_techrep/374.

KRISTENSEN, D. 2010. Nonparametric Filtering of the Realized Spot Volatility: A Kernel-Based Approach. *Econometric Theory*. Volume 26. Issue 1. Pages 60–93. ISSN 0266-4666. https://doi.org/10.1017/s0266466609090616.

KRUSE, R. 2006. *Can Realized Volatility Improve the Accuracy of Value-at-Risk Forecasts?* Working Paper. https://pdfs.semanticscholar.org/1b9f/eb31a8e902dc3fef5b69ddf5f824cef166f9.pdf.

KULLBACK, S., LEIBLER, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*. Volume 22. Issue 1. Pages 79–86. ISSN 0003-4851. https://doi.org/10.2307/2236703.

KVEDARAS, V., ZEMLYS, V. 2016. *Package 'midasr'*. https://cran.r-project.org/package=midasr.

KWIATKOWSKI, D., PHILLIPS, P. C. B., SCHMIDT, P., SHIN, Y. 1992. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *Journal of Econometrics*. Volume 54. Issue 1-3. Pages 159–178. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(92)90104-y.

LAHALLE, E., BAILI, H., OKSMAN, J. 2008. Online Estimation of Time-Varying Volatility Using a Continuous-Discrete LMS Algorithm. *EURASIP Journal on Advances in Signal Processing*. Volume 2008. Issue 1. Pages 532760/1–532760/8. ISSN 1687-6172. https://doi.org/10.1155/2008/532760.

LAMBERT, D. 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. Volume 34. Issue 1. Pages 1–14. ISSN 0040-1706. https://doi.org/10.2307/1269547.

LARGE, J. 2011. Estimating Quadratic Variation When Quoted Prices Change by a Constant Increment. *Journal of Econometrics*. Volume 160. Issue 1. Pages 2–11. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.03.007.

LEHMANN, E. L., CASELLA, G. 1998. *Theory of Point Estimation*. Second Edition. New York. Springer. ISBN 978-0-387-98502-2. https://doi.org/10.1007/b98854.

LEIVA, V., SAULO, H., LEÃO, J., MARCHANT, C. 2014. A Family of Autoregressive Conditional Duration Models Applied to Financial Data. *Computational Statistics & Data Analysis*. Volume 79. Pages 175–191. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2014.05.016.

LEUNG, T., LI, X. 2015. Optimal Mean Reversion Trading With Transaction Costs and Stop-Loss Exit. *International Journal of Theoretical and Applied Finance*. Volume 18. Issue 3. Pages 1550020/1–1550020/31. ISSN 0219-0249. https://doi.org/10.1142/S021902491550020X.

LI, J., LI, L., ZHANG, G. 2017. Pure Jump Models for Pricing and Hedging VIX Derivatives. *Journal of Economic Dynamics and Control*. Volume 74. Pages 28–55. ISSN 0165-1889. https://doi.org/10.1016/j.jedc.2016.11.001.

LI, M. L., CHUI, C. M., LI, C. Q. 2014. Is Pairs Trading Profitable on China AH-Share Markets? *Applied Economics Letters*. Volume 21. Issue 16. Pages 1116–1121. ISSN 1350-4851. https://doi.org/10.1080/13504851.2014.912030.

LI, W., BAI, Z. D. 2011. Analysis of Accumulated Rounding Errors in Autoregressive Processes. *Journal of Time Series Analysis*. Volume 32. Issue 5. Pages 518–530. ISSN 0143-9782. https://doi.org/10.1111/j.1467-9892.2010.00710.x.

LI, Y., MYKLAND, P. A. 2015. Rounding Errors and Volatility Estimation. *Journal of Financial Econometrics*. Volume 13. Issue 2. Pages 478–504. ISSN 1479-8417. https://doi.org/10.1093/jjfinec/nbu005.

LIN, Y.-X., MCCRAE, M., GULATI, C. 2006. Loss Protection in Pairs Trading Through Minimum Profit Bounds: A Cointegration Approach. *Journal of Applied Mathematics and Decision Sciences*. Volume 2006. Pages 1–14. ISSN 1173-9126. https://doi.org/10.1155/jamds/2006/73803.

LINDNER, A., MALLER, R. 2005. Lévy Integrals and the Stationarity of Generalised Ornstein-Uhlenbeck Processes. *Stochastic Processes and Their Applications*. Volume 115. Issue 10. Pages 1701–1722. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2005.05.004.

LIU, B., CHANG, L.-B., GEMAN, H. 2017. Intraday Pairs Trading Strategies on High Frequency Data: The Case of Oil Companies. *Quantitative Finance*. Volume 17. Issue 1. Pages 87–100. ISSN 1469-7688. https://doi.org/10.1080/14697688.2016.1184304.

LIU, L. Y., PATTON, A. J., SHEPPARD, K. 2015. Does Anything Beat 5-Minute RV? A Comparison of Realized Measures Across Multiple Asset Classes. *Journal of Econometrics*. Volume 187. Issue 1. Pages 293–311. ISSN 1872-6895. https://doi.org/10.1016/j.jeconom.2015.02.008.

LIU, Q., LIU, Y., LIU, Z. 2018a. Estimating Spot Volatility in the Presence of Infinite Variation Jumps. *Stochastic Processes and Their Applications*. Volume 128. Issue 6. Pages 1958–1987. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2017.08.015.

LIU, Q., LIU, Y., LIU, Z., WANG, L. 2018b. Estimation of Spot Volatility with Superposed Noisy Data. *North American Journal of Economics and Finance*. Volume 44. Pages 62–79. ISSN 1062-9408. https://doi.org/10.1016/j.najef.2017.11.004.

LIU, Q. 2009. On Portfolio Optimization: How and When Do We Benefit from High-Frequency Data? *Journal of Applied Econometrics*. Volume 24. Issue 4. Pages 560–582. ISSN 0883-7252. https://doi.org/10.2307/40206292.

LIU, S. M., CHOU, C. H. 2003. Parities and Spread Trading in Gold and Silver Markets: A Fractional Cointegration Analysis. *Applied Financial Economics*. Volume 13. Issue 12. Pages 899–911. ISSN 0960-3107. https://doi.org/10.1080/0960310032000129626.

LIU, Z., KONG, X. B., JING, B. Y. 2018c. Estimating the Integrated Volatility Using High-Frequency Data with Zero Durations. *Journal of Econometrics*. Volume 204. Issue 1. Pages 18–32. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2017.12.008.

LOUZIS, D. P., XANTHOPOULOS-SISINIS, S., REFENES, A. P. 2011. *Are Realized Volatility Models Good Candidates for Alternative Value at Risk Prediction Strategies?* Working Paper. https://ssrn.com/abstract=1814171.

LUCAS, A. 2019. *Generalized Autoregressive Score Models*. Online. http://www.gasmodel.com.

LUNDE, A. 1999. *A Generalized Gamma Autoregressive Conditional Duration Model*. Working Paper. https://www.researchgate.net/publication/228464216.

MADHAVAN, A. 2000. Market Microstructure: A Survey. *Journal of Financial Markets*. Volume 3. Issue 3. Pages 205–258. ISSN 1386-4181. https://doi.org/10.1016/s1386-4181(00)00007-0.

MALLIAVIN, P., MANCINO, M. E. 2002. Fourier Series Method for Measurement of Multivariate Volatilities. *Finance and Stochastics*. Volume 6. Issue 1. Pages 49–61. ISSN 0949-2984. https://doi.org/10.1007/s780-002-8400-6.

MANCINI, C. 2013. Measuring the Relevance of the Microstructure Noise in Financial Data. *Stochastic Processes and Their Applications*. Volume 123. Issue 7. Pages 2728–2751. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2013.04.003.

MANCINI, C., MATTIUSSI, V., RENÒ, R. 2015. Spot Volatility Estimation Using Delta Sequences. *Finance and Stochastics*. Volume 19. Issue 2. Pages 261–293. ISSN 0949-2984. https://doi.org/10.1007/s00780-015-0255-1.

MANCINO, M. E., SANFELICI, S. 2012. Estimation of Quarticity with High-Frequency Data. *Quantitative Finance*. Volume 12. Issue 4. Pages 607–622. ISSN 1469-7688. https://doi.org/10.1080/14697688.2012.664936.

MANSKI, C. F. 2003. *Partial Identification of Probability Distributions*. First Edition. New York. Springer. ISBN 978-0-387-00454-9. https://doi.org/10.1007/b97478.

MARCELLINO, M., SCHUMACHER, C. 2010. Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP. *Oxford Bulletin of Economics and Statistics*. Volume 72. Issue 4. Pages 518–550. ISSN 0305-9049. https://doi.org/10.1111/j.1468-0084.2010.00591.x.

MARKOWITZ, H. 1952. Portfolio Selection. *The Journal of Finance*. Volume 7. Issue 1. Pages 77–91. ISSN 0022-1082. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x.

MASUDA, H. 2004. On Multidimensional Ornstein-Uhlenbeck Processes Driven by a General Lévy Process. *Bernoulli*. Volume 10. Issue 1. Pages 97–120. ISSN 1350-7265. https://doi.org/10.3150/bj/1077544605.

MCALEER, M., MEDEIROS, M. C. 2008a. Realized Volatility: A Review. *Econometric Reviews*. Volume 27. Issue 1-3. Pages 10–45. ISSN 0747-4938. https://doi.org/doi.org/10.1080/07474930701853509.

MCALEER, M., MEDEIROS, M. C. 2008b. A Multiple Regime Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries. *Journal of Econometrics*. Volume 147. Issue 1. Pages 104–119. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2008.09.032.

MCMILLAN, D. G., SPEIGHT, A. E., EVANS, K. P. 2008. How Useful Is Intraday Data for Evaluating Daily Value-at-Risk? Evidence from Three Euro Rates. *Journal of Multinational Financial Management*. Volume 18. Issue 5. Pages 488–503. ISSN 1042-444X. https://doi.org/10.1016/j.mulfin.2007.12.003.

MIAO, G. J. 2014. High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach. *International Journal of Economics and Finance*. Volume 6. Issue 3. Pages 96–110. ISSN 1916-971X. https://doi.org/10.5539/ijef.v6n3p96.

MISHRA, A. K., TRIPATHY, T. 2018. Price and Trade Size Clustering: Evidence from the National Stock Exchange of India. *The Quarterly Review of Economics and Finance*. Volume 68. Pages 63–72. ISSN 1062-9769. https://doi.org/10.1016/j.qref.2017.11.006.

MISHRA, A., RAMANATHAN, T. V. 2017. Nonstationary Autoregressive Conditional Duration Models. *Studies in Nonlinear Dynamics and Econometrics*. Volume 21. Issue 4. Pages 1–22. ISSN 1081-1826. https://doi.org/10.1515/snde-2015-0057.

MORI, M., ZIOBROWSKI, A. J. 2011. Performance of Pairs Trading Strategy in the U.S. REIT Market. *Real Estate Economics*. Volume 39. Issue 3. Pages 409–428. ISSN 1080-8620. https://doi.org/10.1111/j.1540-6229.2010.00302.x.

NAGAKURA, D., WATANABE, T. 2015. A State Space Approach to Estimating the Integrated Variance under the Existence of Market Microstructure Noise. *Journal of Financial Econometrics*. Volume 13. Issue 1. Pages 45–82. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbt015.

NASDAQ. 2019. *Stock Reports*. Online. https://www.nasdaq.com/quotes/stock-reports.aspx.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. 2019. *NIST Digital Library of Mathematical Functions*. Online. https://dlmf.nist.gov.

NELDER, J. A., MEAD, R. 1965. A Simplex Method for Function Minimization. *The Computer Journal*. Volume 7. Issue 4. Pages 308–313. ISSN 0010-4620. https://doi.org/10.1093/comjnl/7.4.308.

NEW YORK STOCK EXCHANGE. 2019. *Daily TAQ*. Online. https://www.nyse.com/market-data/historical/daily-taq.

NGO, H.-L., OGAWA, S. 2009. A Central Limit Theorem for the Functional Estimation of the Spot Volatility. *Monte Carlo Methods and Applications*. Volume 15. Issue 4. Pages 353–380. ISSN 0929-9629. https://doi.org/10.1515/mcma.2009.019.

NGUYEN, H. T., KREINOVICH, V., WU, B., XIANG, G. 2012. *Computing Statistics under Interval and Fuzzy Uncertainty*. First Edition. Berlin, Heidelberg. Springer. ISBN 978-3-642-24904-4. https://doi.org/10.1007/978-3-642-24905-1.

NOLTE, I., VOEV, V. 2012. Least Squares Inference on Integrated Volatility and the Relationship Between Efficient Prices and Noise. *Journal of Business & Economic Statistics*. Volume 30. Issue 1. Pages 94–108. ISSN 0735-0015. https://doi.org/10.1080/10473289.2011.637876.

NOURELDIN, D., SHEPHARD, N., SHEPPARD, K. 2012. Multivariate High-Frequency-Based Volatility (HEAVY) Models. *Journal of Applied Econometrics*. Volume 27. Issue 6. Pages 907–933. ISSN 0883-7252. https://doi.org/10.1002/jae.1260.

OCHIAI, T., NACHER, J. C. 2019. VC correlation analysis on the overnight and daytime return in Japanese stock market. *Physica A: Statistical Mechanics and Its Applications*. Volume 515. Pages 537–545. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2018.09.181.

OGAWA, S., SANFELICI, S. 2011. An Improved Two-Step Regularization Scheme for Spot Volatility Estimation. *Economic Notes*. Volume 40. Issue 3. Pages 107–134. ISSN 0391-5026. https://doi.org/10.1111/j.1468-0300.2011.00233.x.

OHTA, W. 2006. An Analysis of Intraday Patterns in Price Clustering on the Tokyo Stock Exchange. *Journal of Banking & Finance*. Volume 30. Issue 3. Pages 1023–1039. ISSN 0378-4266. https://doi.org/10.1016/j.jbankfin.2005.07.017.

OLVER, F. W. J., LOZIER, D. W., BOISVERT, R. F., CLARK, C. W. 2010. *NIST Handbook of Mathematical Functions*. First Edition. Cambridge. Cambridge University Press. ISBN 978-0-521-14063-8. www.cambridge.org/9780521140638.

OOMEN, R. C. A. 2005. Properties of Bias-Corrected Realized Variance Under Alternative Sampling Schemes. *Journal of Financial Econometrics*. Volume 3. Issue 4. Pages 555–577. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbi027.

OOMEN, R. C. A. 2006. Properties of Realized Variance Under Alternative Sampling Schemes. *Journal of Business & Economic Statistics*. Volume 24. Issue 2. Pages 219–237. ISSN 0735-0015. https://doi.org/10.2307/27638871.

PACURAR, M. 2008. Autoregressive Conditional Duration Models in Finance: A Survey of the Theoretical and Empirical Literature. *Journal of Economic Surveys*. Volume 22. Issue 4. Pages 711–751. ISSN 0950-0804. https://doi.org/10.1111/j.1467-6419.2007.00547.x.

PAKKANEN, M. S., SOTTINEN, T., YAZIGI, A. 2017. On the Conditional Small Ball Property of Multivariate Lévy-Driven Moving Average Processes. *Stochastic Processes and Their Applications*. Volume 127. Issue 3. Pages 749–782. ISSN 0304-4149. https://doi.org/10.1016/j.spa.2016.06.025.

PATTON, A. J. 2006. Modelling Asymmetric Exchange Rate Dependence. *International Economic Review*. Volume 47. Issue 2. Pages 527–556. ISSN 0020-6598. https://doi.org/10.1111/j.1468-2354.2006.00387.x.

PATTON, A. J., SHEPPARD, K. 2015. Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility. *Review of Economics and Statistics*. Volume 97. Issue 3. Pages 683–697. ISSN 0034-6535. https://doi.org/10.1162/REST_a_00503.

PEIRIS, M. S. 1986. On Prediction with Time Dependent ARMA Models. *Communications in Statistics - Theory and Methods*. Volume 15. Issue 12. Pages 3659–3668. ISSN 0361-0926. https://doi.org/10.1080/03610928608829339.

PENG, Y.-J., FU, M. C., HU, J.-Q. 2016. Gradient-Based Simulated Maximum Likelihood Estimation for Stochastic Volatility Models Using Characteristic Functions. *Quantitative Finance*. Volume 16. Issue 9. Pages 1393–1411. ISSN 1469-7688. https://doi.org/10.1080/14697688.2016.1185142.

PERLIN, M. S. 2009. Evaluation of Pairs-Trading Strategy at the Brazilian Financial Market. *Journal of Derivatives and Hedge Funds*. Volume 15. Issue 2. Pages 122–136. ISSN 1753-9641. https://doi.org/10.1057/jdhf.2009.4.

PETERS, G. W., KANNAN, B., LASSCOCK, B., MELLENY, C., GODSILL, S. 2011. Bayesian Cointegrated Vector Autoregression Models Incorporating Alpha-Stable Noise for Inter-Day Price Movements Via Approximate Bayesian Computation. *Bayesian Analysis*. Volume 6. Issue 4. Pages 755–792. ISSN 1931-6690. https://doi.org/10.1214/11-BA628.

PIKHART, M., HOLÝ, V. 2018. Modeling E-Sports Matches Using Generalized Autoregressive Score Model. In *Proceedings of the 36th International Conference Mathematical Methods in Economics*. Jindřichův Hradec. MatfyzPress. Pages 428–432. ISBN 978-80-7378-371-6. https://mme2018.fm.vse.cz/wp-content/uploads/2018/09/MME2018-Electronic_proceedings.pdf.

PODOLSKIJ, M., VETTER, M. 2009. Estimation of Volatility Functionals in the Simultaneous Presence of Microstructure Noise and Jumps. *Bernoulli*. Volume 15. Issue 3. Pages 634–658. ISSN 1350-7265. https://doi.org/10.3150/08-BEJ167.

PÖTSCHER, B. M., PRUCHA, I. R. 1997. *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. First Edition. Berlin, Heidelberg. Springer. ISBN 978-3-540-62857-6. https://doi.org/10.1007/978-3-662-03486-6.

PROTTER, P. E. 2004. *Stochastic Integration and Differential Equations*. Second Edition. Berlin, Heidelberg. Springer. ISBN 978-3-540-00313-7. https://doi.org/10.1007/978-3-662-10061-5.

PUSPANINGRUM, H., LIN, Y., GULATI, C. M. 2010. Finding the Optimal Pre-Set Boundaries for Pairs Trading Strategy Based on Cointegration Technique. *Journal of Statistical Theory and Practice*. Volume 4. Issue 3. Pages 391–419. ISSN 1559-8608. https://doi.org/10.1080/15598608.2010.10411994.

R CORE TEAM. 2019. *R: A Language and Environment for Statistical Computing*. Online. https://www.r-project.org.

RICCIARDI, L. M., SATO, S. 1988. First-Passage-Time Density and Moments of the Ornstein-Uhlenbeck Process. *Journal of Applied Probability*. Volume 25. Issue 1. Pages 43–57. ISSN 0021-9002. https://doi.org/10.2307/3214232.

RINNE, K., SUOMINEN, M. 2017. How Some Bankers Made a Million by Trading Just Two Securities? *Journal of Empirical Finance*. Volume 44. Pages 304–315. ISSN 0927-5398. https://doi.org/10.1016/j.jempfin.2016.12.001.

ROBERT, C. Y., ROSENBAUM, M. 2011. A New Approach for the Dynamics of Ultra-High-Frequency Data: The Model with Uncertainty Zones. *Journal of Financial Econometrics*. Volume 9. Issue 2. Pages 344–366. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbq023.

ROBERT, C. Y., ROSENBAUM, M. 2012. Volatility and Covariation Estimation When Microstructure Noise and Trading Times Are Endogenous. *Mathematical Finance*. Volume 22. Issue 1. Pages 133–164. ISSN 0960-1627. https://doi.org/10.1111/j.1467-9965.2010.00454.x.

ROSENBAUM, M. 2009. Integrated Volatility and Round-Off Error. *Bernoulli*. Volume 15. Issue 3. Pages 687–720. ISSN 1350-7265. https://doi.org/10.3150/08-bej170.

ROSENBAUM, M. 2011. A New Microstructure Noise Index. *Quantitative Finance*. Volume 11. Issue 6. Pages 883–899. ISSN 1469-7688. https://doi.org/10.1080/14697680903514352.

ROWAN, T. H. 1990. *Functional Stability Analysis of Numerical Algorithms*. Doctoral Thesis. The University of Texas at Austin. https://www.researchgate.net/publication/2487989.

RUSSELL, J. R. 1999. *Econometric Modeling of Multivariate Irregularly-Spaced High-Frequency Data*. Working Paper. http://faculty.chicagobooth.edu/jeffrey.russell/research/multi.pdf.

RUSSELL, J. R., ENGLE, R. F. 2005. A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model. *Journal of Business and Economic Statistics*. Volume 23. Issue 2. Pages 166–180. ISSN 0735-0015. https://doi.org/10.1198/073500104000000541.

RYAN, J. A., ULRICH, J. M., BENNETT, R. 2018a. *Package 'xts'*. https://cran.r-project.org/package=xts.

RYAN, J. A., ULRICH, J. M., THIELEN, W., TEETOR, P., BRONDER, S. 2018b. *Package 'quantmod'*. https://cran.r-project.org/package=quantmod.

SANFELICI, S., UBOLDI, A. 2014. Assessing the Quality of Volatility Estimators via Option Pricing. *Studies in Nonlinear Dynamics and Econometrics*. Volume 18. Issue 2. Pages 103–124. ISSN 1081-1826. https://doi.org/10.1515/snde-2012-0075.

SARANJEET, K. B., RAMANATHAN, T. V. 2019. Conditional Duration Models for High-Frequency Data: A Review on Recent Developments. *Journal of Economic Surveys*. Volume 33. Issue 1. Pages 252–273. ISSN 0950-0804. https://doi.org/10.1111/joes.12261.

SCHNEEWEISS, H., KOMLOS, J., AHMAD, A. S. 2010. Symmetric and Asymmetric Rounding: A Review and Some New Results. *AStA Advances in Statistical Analysis*. Volume 94. Issue 3. Pages 247–271. ISSN 1863-8171. https://doi.org/10.1007/s10182-010-0125-2.

SCHWARTZ, E. S. 1997. The Stochastic Behaviour of Commodity Prices: Implication for Valuation and Hedging. *The Journal of Finance*. Volume 52. Issue 3. Pages 923–973. ISSN 0022-1082. https://doi.org/10.1111/j.1540-6261.1997.tb02721.x.

SHANNO, D. F. 1970. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation*. Volume 24. Issue 111. Pages 647–656. ISSN 0025-5718. https://doi.org/10.2307/2004840.

SHAO, X. D., LIAN, Y. J., YIN, L. Q. 2009. Forecasting Value-at-Risk Using High Frequency Data: The Realized Range Model. *Global Finance Journal*. Volume 20. Issue 2. Pages 128–136. ISSN 1044-0283. https://doi.org/10.1016/j.gfj.2008.11.003.

SHEPHARD, N., SHEPPARD, K. 2010. Realising the Future: Forecasting with High-Frequency-Based Volatility (HEAVY) Models. *Journal of Applied Econometrics*. Volume 47. Issue 4. Pages 36–37. ISSN 0883-7252. https://doi.org/10.1002/jae.1158.

SHREVE, S. E. 2004a. *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*. First Edition. New York. Springer. ISBN 978-0-387-40100-3. https://www.springer.com/gp/book/9780387401003.

SHREVE, S. E. 2004b. *Stochastic CalcuIus for Finance II: Continuous-Time Models*. First Edition. New York. Springer. ISBN 978-0-387-40101-0. https://www.springer.com/gp/book/9780387401010.

SHUMWAY, R. H., STOFFER, D. S. 2011. *Time Series Analysis and Its Applications*. Third Edition. New York. Springer. ISBN 978-1-4419-7864-6. https://doi.org/10.1007/978-1-4419-7865-3.

SIMON, D. P. 1999. The Soybean Crush Spread: Empirical Evidence and Trading Strategies. *Journal of Futures Markets*. Volume 19. Issue 3. Pages 271–289. ISSN 0270-7314. https://doi.org/10.1002/(sici)1096-9934(199905)19:3<271::aid-fut2>3.0.co;2-p.

SINGH, S., VIPUL. 2015. Performance of Black-Scholes Model with TSRV Estimates. *Managerial Finance*. Volume 41. Issue 8. Pages 857–870. ISSN 0307-4358. https://doi.org/10.1108/MF-06-2014-0177.

SO, M. K. P., XU, R. 2013. Forecasting Intraday Volatility and Value-at-Risk with High-Frequency Data. *Asia-Pacific Financial Markets*. Volume 20. Issue 1. Pages 83–111. ISSN 1387-2834. https://doi.org/10.1007/s10690-012-9160-1.

SOKOL, O., RADA, M. 2016. Interval Data and Sample Variance: How to Prove Polynomiality of Computation of Upper Bound Considering Random Intervals? In *Proceedings of the 34th International Conference Mathematical Methods in Economics*. Liberec. Technical University of Liberec. Pages 779–784. ISBN 978-80-7494-296-9. http://mme2016.tul.cz/index.php?page=conferenceproceedings.

SONDERMANN, D. 2006. *Introduction to Stochastic Calculus for Finance: A New Didactic Approach*. First Edition. Berlin, Heidelberg. Springer. ISBN 978-3-540-34836-8. https://doi.org/10.1007/3-540-34837-9.

SONG, Q., ZHANG, Q. 2013. An Optimal Pairs-Trading Rule. *Automatica*. Volume 49. Issue 10. Pages 3007–3014. ISSN 0005-1098. https://doi.org/10.1016/j.automatica.2013.07.012.

STACY, E. W. 1962. A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics*. Volume 33. Issue 3. Pages 1187–1192. ISSN 0003-4851. https://doi.org/10.2307/2237889.

STEELE, J. M. 2001. *Stochastic Calculus and Financial Applications*. First Edition. New York. Springer. ISBN 978-0-387-95016-7. https://doi.org/10.1007/978-1-4684-9305-4.

STENTOFT, L. 2008. *Option Pricing Using Realized Volatility*. Working Paper. https://ssrn.com/abstract=114814.

STRAUMANN, D., MIKOSCH, T. 2006. Quasi-Maximum-Likelihood Estimation in Conditionally Heteroscedastic Time Series: A Stochastic Recurrence Equations Approach. *The Annals of Statistics*. Volume 34. Issue 5. Pages 2449–2495. ISSN 0090-5364. https://doi.org/10.1214/009053606000000803.

SUN, Y. 2006. *Best Quadratic Unbiased Estimators of Integrated Variance in the Presence of Market Microstructure Noise*. Working Paper. https://ssrn.com/abstract=1714751.

SWAMY, P. A. V. B., TINSLEY, P. A. 1980. Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients. *Journal of Econometrics*. Volume 12. Issue 2. Pages 103–142. ISSN 0304-4076. https://doi.org/10.1016/0304-4076(80)90001-9.

TANG, C. Y., CHEN, S. X. 2009. Parameter Estimation and Bias Correction for Diffusion Processes. *Journal of Econometrics*. Volume 149. Issue 1. Pages 65–81. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2008.11.001.

TARALDSEN, G. 2011. Analysis of Rounded Exponential Data. *Journal of Applied Statistics*. Volume 38. Issue 5. Pages 977–986. ISSN 0266-4763. https://doi.org/10.1080/02664761003692431.

TAYLOR, N. 2017. Realised Variance Forecasting Under Box-Cox Transformations. *International Journal of Forecasting*. Volume 33. Issue 4. Pages 770–785. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2017.04.001.

TAYLOR, S. J. 2016. *Microstructure Noise Components of the S&P 500 Index: Variation, Persistence and Distributions*. Working Paper. https://ssrn.com/abstract=2435853.

TRIANTAFYLLOPOULOS, K., MONTANA, G. 2011. Dynamic Modeling of Mean-Reverting Spreads for Statistical Arbitrage. *Computational Management Science*. Volume 8. Issue 1-2. Pages 23–49. ISSN 1619-697X. https://doi.org/10.1007/s10287-009-0105-8.

TRICKER, A. R. 1992. Estimation of Parameters for Rounded Data from Non-Normal Distributions. *Journal of Applied Statistics*. Volume 19. Issue 4. Pages 465–471. ISSN 0266-4763. https://doi.org/10.1080/02664769200000041.

TRICKER, T. 1984. Effects of Rounding Data Sampled from the Exponential Distribution. *Journal of Applied Statistics*. Volume 11. Issue 1. Pages 54–87. ISSN 0266-4763. https://doi.org/10.1080/02664768400000007.

TSAI, K.-T., LIH, J.-S., KO, J.-Y. 2012. The Overnight Effect on the Taiwan Stock Market. *Physica A: Statistical Mechanics and Its Applications*. Volume 391. Issue 24. Pages 6497–6505. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2012.07.010.

TSAI, Y.-C., LYUU, Y.-D. 2017. A New Robust Kalman Filter for Filtering the Microstructure Noise. *Communications in Statistics - Theory and Methods*. Volume 46. Issue 10. Pages 4961–4976. ISSN 0361-0926. https://doi.org/10.1080/03610926.2015.1096390.

TUCCI, M. P. 1995. Time-Varying Parameters: A Critical Introduction. *Structural Change and Economic Dynamics*. Volume 6. Issue 2. Pages 237–260. ISSN 0954-349X. https://doi.org/10.1016/0954-349x(94)00010-7.

TUKEY, J. W. 1938. On the Distribution of the Fractional Part of a Statistical Variable. *Matematičeskij sbornik, Novaja serija*. Volume 4. Issue 3. Pages 561–562. ISSN 0368-8666. http://mi.mathnet.ru/eng/msb5767.

UBUKATA, M., OYA, K. 2009. Estimation and Testing for Dependence in Market Microstructure Noise. *Journal of Financial Econometrics*. Volume 7. Issue 2. Pages 106–151. ISSN 1479-8409. https://doi.org/10.1093/jjfinec/nbn021.

UHLENBECK, G. E., ORNSTEIN, L. S. 1930. On the Theory of the Brownian Motion. *Physical Review I*. Volume 36. Issue 5. Pages 823–841. ISSN 0031-899X. https://doi.org/10.1103/physrev.36.823.

VASICEK, O. 1977. An Equilibrium Characterisation of the Term Structure. *Journal of Financial Economics*. Volume 5. Issue 2. Pages 177–188. ISSN 0304-405X. https://doi.org/10.1016/0304-405x(77)90016-2.

VEREDAS, D., RODRÍGUEZ-POO, J. M., ESPASA, A. 2002. *On the (Intradaily) Seasonality and Dynamics of a Financial Point Process: A Semiparametric Approach.* Working Paper. https://ideas.repec.org/p/cor/louvco/2002023.html.

VIDYAMURTHY, G. 2004. *Pairs Trading: Quantitative Methods and Analysis*. First Edition. Hoboken. Wiley. ISBN 978-0-471-46067-1. https://www.wiley.com/en-us/Pairs+Trading{%}3A+Quantitative+Methods+and+Analysis-p-9780471460671.

WAHAB, M., COHN, I., LASHGARI, M. 1994. The Gold-Silver Spread: Integration, Cointegration, Predictability, and Ex-Ante Arbitrage. *Journal of Futures Markets*. Volume 14. Issue 6. Pages 709–756. ISSN 0270-7314. https://doi.org/10.1002/fut.3990140606.

WANG, F., SHIEH, S.-J., HAVLIN, S., STANLEY, H. E. 2009. Statistical Analysis of the Overnight and Daytime Return. *Physical Review E*. Volume 79. Issue 5. Pages 056109:1–056109:7. ISSN 539-3755. https://doi.org/10.1103/PhysRevE.79.056109.

WATANABE, T. 2012. Quantile Forecasts of Financial Returns Using Realized GARCH Models. *Japanese Economic Review*. Volume 63. Issue 1. Pages 68–80. ISSN 1352-4739. https://doi.org/10.1111/j.1468-5876.2011.00548.x.

WHITE, H. 1994. *Estimation, Inference and Specification Analysis*. First Edition. Cambridge. Cambridge University Press. ISBN 978-0-521-57446-4. https://doi.org/10.1017/CCOL0521252806.

WINTENBERGER, O. 2013. Continuous Invertibility and Stable QML Estimation of the EGARCH(1,1) Model. *Scandinavian Journal of Statistics*. Volume 40. Issue 4. Pages 846–867. ISSN 0303-6898. https://doi.org/10.1111/sjos.12038.

WORLD FEDERATION OF EXCHANGES. 2019. *Monthly Reports*. Online. https://www.world-exchanges.org/home/index.php/monthly-reports-tool.

XIANG, G., CEBERIO, M., KREINOVICH, V. 2007. Computing Population Variance and Entropy under Interval Uncertainty: Linear-Time Algorithms. *Reliable Computing*. Volume 13. Issue 6. Pages 467–488. ISSN 1385-3139. https://doi.org/10.1007/s11155-007-9045-6.

XIU, D. 2010. Quasi-Maximum Likelihood Estimation of Volatility with High Frequency Data. *Journal of Econometrics*. Volume 159. Issue 1. Pages 235–250. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2010.07.002.

XU, Y. 2013. *The Lognormal Autoregressive Conditional Duration (LNACD) Model and a Comparison with an Alternative ACD Models*. Working Paper. https://ssrn.com/abstract=2382159.

XUE, Y., GENÇAY, R., FAGAN, S. 2014. Jump Detection with Wavelets for High-Frequency Financial Time Series. *Quantitative Finance*. Volume 14. Issue 8. Pages 1427–1444. ISSN 1469-7696. https://doi.org/10.1080/14697688.2013.830320.

YAHOO! 2019. *Yahoo! Finance*. Online. https://finance.yahoo.com.

ZENG, Z., LEE, C. G. 2014. Pairs Trading: Optimal Thresholds and Profitability. *Quantitative Finance*. Volume 14. Issue 11. Pages 1881–1893. ISSN 1469-7688. https://doi.org/10.1080/14697688.2014.917806.

ZHANG, B., LIU, T., BAI, Z. D. 2010. Analysis of Rounded Data from Dependent Sequences. *Annals of the Institute of Statistical Mathematics*. Volume 62. Issue 6. Pages 1143–1173. ISSN 0020-3157. https://doi.org/10.1007/s10463-009-0224-6.

ZHANG, L. 2006. Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach. *Bernoulli*. Volume 12. Issue 6. Pages 1019–1043. ISSN 1350-7265. https://doi.org/10.2307/25464852.

ZHANG, L., MYKLAND, P. A., AÏT-SAHALIA, Y. 2005. A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of the American Statistical Association*. Volume 100. Issue 472. Pages 1394–1411. ISSN 0162-1459. https://doi.org/10.2307/27590680.

ZHANG, M. Y., RUSSELL, J. R., TSAY, R. S. 2001. A Nonlinear Autoregressive Conditional Duration Model with Applications to Financial Transaction Data. *Journal of Econometrics*. Volume 104. Issue 1. Pages 179–207. ISSN 0304-4076. https://doi.org/10.1016/s0304-4076(01)00063-x.

ZHENG, Y., LI, Y., LI, G. 2016. On Fréchet Autoregressive Conditional Duration Models. *Journal of Statistical Planning and Inference*. Volume 175. Pages 51–66. ISSN 0378-3758. https://doi.org/10.1016/j.jspi.2016.02.009.

ZHOU, B. 1996. High-Frequency Data and Volatility in Foreign-Exchange Rates. *Journal of Business & Economic Statistics*. Volume 14. Issue 1. Pages 45–52. ISSN 0735-0015. https://doi.org/10.2307/1392098.

ŽIKEŠ, F., BARUNÍK, J. 2015. Semi-Parametric Conditional Quantile Models for Financial Returns and Realized Volatility. *Journal of Financial Econometrics*. Volume 14. Issue 1. Pages 185–226. ISSN 1479-8417. https://doi.org/10.1093/jjfinec/nbu029.

ZU, Y., BOSWIJK, H. P. 2014. Estimating Spot Volatility with High-Frequency Financial Data. *Journal of Econometrics*. Volume 181. Issue 2. Pages 117–135. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2014.04.001.