Advances in Stochastic Processes with Machine Learning Applications

Lorenzo Lucchese

Supervised by
Prof. Almut E. D. Veraart
Prof. Mikko S. Pakkanen

This thesis is submitted for the degree of Doctor of Philosophy
of Imperial College London and the Diploma of Imperial College

IMPERIAL

Department of Mathematics
Imperial College London
180 Queen's Gate, London SW7 2BZ

December 2024

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

© 2024 Lorenzo Lucchese

Statement of Originality

I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Lorenzo Lucchese
December 2024

Abstract

Stochastic processes play a fundamental role in our attempt to model the world around us. This thesis explores several topics within this field, with a particular focus on prediction, estimation, and inference. We begin by addressing the empirical question of whether order book-based financial markets exhibit predictability, using deep learning architectures to extract information from their microstructural dynamics. Moving from discrete-time to continuous-time modeling, we then consider continuous-time autoregressive processes, developing consistent and asymptotically normal estimators for their drift parameters under both continuous and discrete observations. Finally, we present new limiting results for the empirical expected signature, a powerful non-parametric statistic of path-valued random variables, and review its application in various machine learning algorithms, demonstrating the practical benefits of our findings. A recurring theme across all three projects is the use of high-frequency data, an essential bridge between theory and practice.

Acknowledgments

First and foremost I would like to thank my supervisors, Almut and Mikko, without whom this thesis would have never come to life. Their kindness and dedication will always remain in my heart.

I would also like to thank all the fellow graduate students who shared this journey with me. At WCM (which we all know is much more than a PhD office), we faced struggles and celebrated achievements together, forming long-lasting friendships. My gratitude extends to everyone in the Random Systems CDT: in one way or another, each of you contributed to shaping this thesis.

Next, I would like to thank my colleagues at GSA for their understanding and respect for my double commitment. Their proactive support was essential to effectively balance my time between industry and academia during the final year of my studies.

I am also deeply grateful to my friends. Whether we've known each other for a lifetime or we only recently met, I feel incredibly lucky for all the lighthearted moments we've shared. A special mention goes to my *legendary* flatmates for putting up with me over the past three years.

Finally, I would like to thank my family: my parents, Sonia and Enrico, for their unconditional love and support, and Gaia, for filling my life with love and joy—you truly bring meaning to my days.

A Gaia

Preface

Every PhD journey is marked by randomness, and mine was no exception. Usually, as a research project matures and new results are established, one unavoidably deviates from the original proposal. Sometimes the change of course can be even more abrupt and one ends up branching out into new research topics. In my case though, randomness was not only a defining characteristic of my path of academic self-discovery but also comprised the primary subject of my research.

Throughout the three years I spent pursuing this degree I was part of the Mathematics of Random Systems Center for Doctoral Training, an EPSRC-funded cohort of doctoral students investigating the most disparate projects in probabilistic modeling and stochastic analysis. Covering topics from random combinatorial structures to stochastic partial differential equations, with applications spanning physics, finance, and healthcare, this served as an endless source of research inspiration.

Motivated by insightful discussions with fellow students and my supervisors, I embraced this unique opportunity and soon found myself engaged in three distinct research projects. This thesis is a collection of the results I achieved in these areas. Given this premise, the reader might thus find the following to be a less informative but more genuine title:

Random Topics in Random Systems.

Contents

1 Introduction **5**

1.1 Modeling stochastic processes . . . 5

1.2 Outline . . . 7

2 The Short-Term Predictability of Order Book Markets **9**

2.1 Introduction . . . 9

2.1.1 Financial markets and exchanges: setting the scene . . . 9

2.1.2 Order book predictability: asking the right questions . . . 10

2.1.3 Related work and contributions . . . 11

2.2 The space of models under consideration . . . 12

2.2.1 The order book . . . 12

2.2.2 Predictability of returns . . . 14

2.2.3 Deep learning models . . . 18

2.2.4 DeepVOL . . . 22

2.2.5 Multi-horizon models . . . 26

2.3 Dataset . . . 28

2.3.1 LOBSTER data . . . 29

2.3.2 Processed data . . . 30

2.4 Empirical results . . . 33

2.4.1 Do high-frequency returns display predictability? . . . 36

2.4.2 Which order book representations perform the best? . . . 37

2.4.3 Can we use a single model across multiple horizons? . . . 38

2.4.4 Can we use a single model across multiple stocks? . . . 39

2.5 Conclusions and outlook . . . 42

Appendix **44**

2.A Deep learning . . . 44

2.A.1 Deep neural network architectures . . . 44

2.A.2 Training deep neural networks . . . 46

2.B Model confidence sets . . . 49

2.C Data processing . . . 51

2.C.1 Volume data processing . . . 52

2.D Dataset descriptive statistics: ATVI deep dive . . . 52

2.D.1 Features . . . 53

2.D.2 Responses . . . 53

2.E Simpler predictive models: an empirical auto-regressive specification . . . 54

2.F Defining returns . . . 57

2.G Infrastructure latency . . . 60

2.H Stock selection . . . 63

2.I Model details . . . 63

3 Continuous-time Autoregressive Processes 66
3.1 Introduction . . . 66
3.1.1 Motivation: from OU processes to CAR processes . . . 66
3.1.2 Contributions and literature review . . . 68
3.2 Multivariate continuous-time autoregressive processes . . . 70
3.3 Continuous observations: explicit maximum likelihood . . . 73
3.3.1 Deriving the likelihood . . . 74
3.3.2 Maximum likelihood estimation . . . 79
3.4 Discrete observations: high-frequency estimators . . . 83
3.4.1 Approximating the integrals . . . 86
3.4.2 Approximating the derivatives . . . 88
3.4.3 Thresholding the jumps . . . 92
3.4.4 Simulation study . . . 101
3.4.5 Practical considerations . . . 102
3.5 Application: GrCAR . . . 106
3.6 Empirical application: light curve modeling . . . 110
3.7 Conclusions and outlook . . . 115

Appendix 117
3.A Preliminaries . . . 117
3.A.1 Lévy processes . . . 117
3.A.2 Stochastic integration . . . 118
3.A.3 Canonical space . . . 120
3.B Properties of the state-space representation of the MCAR(p) process . . . 121
3.C Continuous observations: proofs . . . 122
3.C.1 Proof of Proposition 3.3.3 . . . 122
3.C.2 Proof of Lemma 3.3.8 . . . 129
3.C.3 Proof of Theorem 3.3.9 . . . 130
3.D Discrete observations: proofs . . . 132
3.D.1 Proof of Theorem 3.4.5 . . . 132
3.D.2 Proof of Lemma 3.4.8 . . . 134
3.D.3 Proof of Theorem 3.4.14 . . . 137
3.D.4 Proof of Theorem 3.4.19 . . . 140
3.D.5 Proof of Theorem 3.4.23 . . . 145
3.E GrCAR: proofs . . . 156
3.E.1 Proof of Theorem 3.5.3 . . . 156
3.F Finite difference approximations of derivatives . . . 157
3.G Simulating MCAR(p) processes . . . 160
3.G.1 An approximate simulation method: Euler-Maruyama scheme . . . 160
3.G.2 An exact simulation method: the finite jump activity case . . . 161
3.H Practical considerations: details and simulation study . . . 163
3.H.1 Disentangling Brownian and jump component of a Lévy process . . . 163

3.H.2 Joint estimation . . . 164
3.H.3 Simulation study with unknown $\mathbf{b}$, Σ and F . . . 165

4 Expected Signature Estimation **172**
4.1 Introduction . . . 172
4.2 A simple motivating example: Brownian motion . . . 175
4.2.1 Variance reduction via control variates . . . 181
4.3 Canonical geometric stochastic processes . . . 185
4.3.1 Semimartingales: convergence in probability of signatures . . . 188
4.4 Estimating the expected signatures of a canonical geometric stochastic process . . . 189
4.4.1 In-fill asymptotics . . . 190
4.4.2 Long-span asymptotics . . . 205
4.5 Simulations . . . 226
4.5.1 Consistency . . . 228
4.5.2 Asymptotic normality . . . 230
4.6 Estimating the expected signature of a martingale . . . 235
4.6.1 Martingale correction: variance reduction via control variates . . . 235
4.6.2 Machine learning applications . . . 243
4.7 Conclusions and outlook . . . 260

Appendix **262**
4.A Brownian motion: proofs . . . 262
4.A.1 Proof of Theorem 4.2.1 . . . 262
4.A.2 Proof of Theorem 4.2.3 . . . 263
4.A.3 Proof of Corollary 4.2.5 . . . 264
4.A.4 Proof of Theorem 4.2.6 . . . 264
4.A.5 Proof of Theorem 4.2.8 . . . 267
4.A.6 Proof of Theorem 4.2.10 . . . 269
4.A.7 Proof of Proposition 4.2.12 . . . 271
4.A.8 Proof of Proposition 4.2.13 . . . 273
4.B Semimartingales as canonical geometric stochastic processes: proofs . . . 275
4.B.1 Proof of Theorem 4.3.8 . . . 275
4.C In-fill asymptotics: proofs . . . 276
4.C.1 Proof of Theorem 4.4.4 . . . 278
4.C.2 Proof of Lemma 4.4.6 . . . 282
4.C.3 Proof of Theorem 4.4.7 . . . 283
4.D Long-span asymptotics: proofs . . . 288
4.D.1 Proof of Proposition 4.4.23 . . . 288
4.D.2 Proof of Proposition 4.4.24 . . . 289
4.D.3 Proof of Proposition 4.4.28 . . . 290
4.D.4 Proof of Theorem 4.4.34 . . . 291
4.E Martingale correction: proofs . . . 295
4.E.1 Proof of Lemma 4.6.2 . . . 295
4.F Controlled linear regression . . . 301

4.F.1 Controlled Ordinary Least Squares (OLS) estimation 302
4.F.2 Simulation study 307

Bibliography **311**

Chapter 1

Introduction

1.1 Modeling stochastic processes

The way humans perceive randomness has radically evolved over the centuries. Once associated with unquestionable fate, randomness has since become a subject of rigorous inquiry, with researchers developing tools to understand stochasticity and its influence on the world. A prime example is its central role in quantum mechanics, which revolutionized humanity's view of the universe in the early twentieth century. A key breakthrough that enabled mathematicians to describe random systems was Kolmogorov's axiomatic study of probability (Kolmogorov, 1950). Building on this solid mathematical foundation allows us to better understand systems displaying unpredictability, whether randomness is interpreted as pure chance or lack of information.

The modeling process usually starts with a specific task one would like to accomplish based on an observable quantity. Examples in this thesis include developing a high-frequency trading strategy, analyzing a star's light curve, and describing the dynamics of an ideal gas. A model is then constructed to represent reality under stochasticity: the observable is endowed with a probability measure, and the observed data is treated as one of its possible realizations. Based on the available knowledge, the model can incorporate different levels of detail by defining functional relationships governed by free parameters. Using real-world measurements, the model is tuned – a process known as estimation or fitting – to closely approximate reality. Once tuned, the model serves as a proxy for the true data generating process[1], enabling inferences about the observable's dynamics or predictions of its future behavior. This thesis explores a range of modeling tasks under different levels of domain knowledge: from the constrained dynamics of continuous-time autoregressive (CAR) processes in Chapter 3, to the over-parametrized deep learning models in Chapter 2, and the fully non-parametric approaches based on expected signatures in Chapter 4.

In this thesis, we focus on observables indexed by time, known as stochastic processes[2]. Steering clear of extensive philosophical debates[3], one might argue that all observable quantities are in constant evolution, making this a natural – if not the only – choice. From a modeling perspective, the time-indexing may be taken to be a discrete set or a continuum of values. In discrete-time,

[1] Estimation error can often be quantitatively accounted for in this step.

[2] Although the indexing set need not always represent time, this assumption suffices for our purposes and is arguably more intuitive.

[3] To paraphrase Heraclitus: *panta rei* (everything flows).

one can then build, estimate, and analyze models using tools from time series analysis. In the parametric setting these range from the specific functional relations of the ARMA(p, q) (Yule, 1927; Slutzky, 1937) and ARCH models (Engle, 1982) to over-parametrized deep learning architectures for sequential data like LSTMs (Hochreiter and Schmidhuber, 1997), discussed in Chapter 2. Non-parametric approaches often analyze the spectral properties of the time series (Priestley, 1982), as touched upon in Section 3.6. Continuous-time models, in contrast, require arguably more sophisticated tools from stochastic calculus. These find their origins in the pioneering work of Itô (Itô, 1944) followed by significant contributions by Doob, Meyer and Dellacherie (Doob, 1962; Dellacherie and Meyer, 1978) and the more recent expositions by Protter, Jacod, Shiraev (Jacod and Shiryaev, 1987; Protter, 1990) and many others. In this setting, estimation techniques such as Sørensen (1991), which inspired the results of Chapter 3, can then be employed to fit parametric continuous-time models to data. More recently, rough path theory (Lyons, 1998) has emerged as a model-agnostic alternative to classical Itô calculus[4], giving rise to the non-parametric methods discussed in Chapter 4.

Natural connections between discrete-time and continuous-time models often exist. On one hand, a continuous-time model may arise as the limit of an appropriately rescaled discrete-time model and, on the other, a discrete-time model is always obtained when sampling a latent continuous-time model at a discrete set of times. This connection, for example, underlies the relationship between the continuous-time autoregressive (CAR) models of Chapter 3 and classic discrete-time autoregressive (AR) models. From an estimation point of view, when dealing with continuous-time models, a practical limitation is that observations can only be collected in discrete-time[5]. While often theoretically useful, statistics based on continuous-time processes thus cannot be computed from observed data and hence need to be approximated by their discrete-time versions. Estimation results are then derived under in-fill asymptotics, an approach common to Chapter 3 and Chapter 4.

As with any statistical procedure, a necessary prerequisite is the availability of data. Over the past few decades, rapid technological advancements have made it increasingly simple to store and handle large amounts of information in digital format[6]. Fueled by these technological advances, data on an increasingly finer scale – commonly known as high-frequency data[7] – can now be collected and analyzed. Chapter 2 leverages high-frequency tick data to analyze the efficiency of order

[4]We refer to Friz and Victoir (2010) for a detailed exposition of rough path theory in the context of stochastic procsses.

[5]At this point the reader might wonder why one would want to consider continuous-time models in the first place. Specifying a continuous-time latent model yields a single consistent description of the process across time scales, a particularly helpful property when observations are collected over irregularly spaced grids.

[6]Just a few years ago, many of the numerical experiments presented in this thesis would have been impossible or prohibitively time-consuming. Fortunately, the aforementioned technological advances, combined with the resources provided by the Imperial High-Performance Computing (HPC) cluster (Imperial College Research Computing Service, 2022), made it possible to manage terabytes of (real and simulated) data and complete all experiments within a reasonable timeframe, ensuring the timely submission of this thesis.

[7]The archetypal example of high-frequency data is financial tick data which, at the time of writing, is timestamped at nanosecond precision by most major financial exchanges. The light-curve data considered in Section 3.6, which is sampled at 5 minute intervals, can also be considered to be high-frequency relative to the time scale of the astronomical phenomena being modeled.

book-based markets. Similarly, high-frequency data underpins the theoretical results in Chapter 3 and Chapter 4, where estimation relies on in-fill asymptotics.

1.2 Outline

In the first chapter, we conduct a systematic large-scale analysis of order book-driven predictability of high-frequency returns by leveraging deep learning techniques. First, we introduce a new and robust representation of the order book, the volume representation. Next, we carry out an extensive empirical experiment to address various questions regarding predictability. We investigate if and how far ahead there is predictability, the importance of a robust data representation, the advantages of multi-horizon modeling, and the presence of universal trading patterns. We use model confidence sets, which provide a formalized statistical inference framework particularly well suited to answer these questions. Our findings show that the class of deep learning models considered in our work can consistently identify predictability in mid-price returns. The performance of the models is strongly dependent on the choice of order book representation, and in this respect, the volume representation appears to have multiple practical advantages.

The second chapter aims to develop estimation and inference methods for the drift parameters of multivariate Lévy-driven continuous-time autoregressive processes. Starting from a continuous-time observation of the process, we develop consistent and asymptotically normal maximum likelihood estimators. We then relax the unrealistic assumption of continuous-time observation by considering natural discretizations based on a combination of Riemann-sum, finite difference, and thresholding approximations. The resulting estimators are also proven to be consistent and asymptotically normal under a general set of conditions, allowing for both finite and infinite jump activity in the driving Lévy process. When discretizing the estimators, allowing for irregularly spaced observations is of great practical importance. In this regard, CAR(p) models are not just relevant for "true" continuous-time processes: a CAR(p) specification provides a natural continuous-time interpolation for modeling irregularly spaced data – even if the observed process is inherently discrete. When the multivariate observation is known to possess a graphical structure, we can naturally restrict the degrees of freedom of the drift parameters. We refer to such a process as GrCAR and discuss the corresponding drift estimators and their properties. The finite sample behavior of all theoretical asymptotic results is empirically assessed by extensive simulation experiments. Finally, as a practical application, we model a high-frequency light curve with a CAR model, characterizing its variability via spectral analysis.

In the third and final chapter, we develop theoretical results for the estimation of the expected signature of a stochastic process. The signature transform of a path is an infinite but countable sequence of its "iterated integrals" summarizing the data in a top-down fashion, meaning that the informational content of the signature terms decays exponentially. When the path is a random variable, i.e. a stochastic process, the expected signature can fully characterize its law (Chevyrev and

Lyons, 2016; Chevyrev and Oberhauser, 2018), in the same way that moments can characterize the law of a finite-dimensional random variable. This "model-free" property of the expected signature has been successfully leveraged to build machine-learning algorithms for time series and sequential data. Usually, to estimate the expected signature, one collects samples of the path over a discrete observation partition, which leads to two sources of error: (1) the error due to the sample signatures being computed from piecewise linear approximations of the path, and (2) the error originating from the expectation being computed using the empirical measure. In this respect, the way the samples are collected plays a crucial role, with the most straightforward assumptions being an equally-spaced partition and i.i.d. samples. In many practically relevant applications though, for example in the financial domain, one usually has access to a single – possibly irregular – data stream that needs to be "chopped up" into shorter paths. In the third chapter of this thesis, we thus aim to explore general conditions on the law of the stochastic process ensuring consistency and asymptotic normality of the expected signature estimator in the in-fill and large-sample limit under both sampling schemes. All theoretical results are complemented by extensive simulations evaluating the estimator's finite sample properties. Moreover, when the process is a martingale, we suggest a modification of the estimator with significantly lower mean squared error. The modified estimator empirically improves multiple machine learning algorithms employing expected signature computations, suggesting both theoretical and practical advantages.

Chapter 2

The Short-Term Predictability of Order Book Markets

We begin this journey by addressing an empirical question: do order-book markets exhibit predictability? While the question is simple to state, it lends itself to multiple interpretations and approaches. A significant part of this chapter will thus be devoted to formalizing the experimental setup and introducing the relevant statistical methodology.

We will be restricting ourselves to order book-driven predictability, which can be described as the market's reaction to past sequences of orders and trades – a dynamic often exemplified by market impact. An order book-based market will be considered predictable if there exists a prediction map, i.e. a function that uses the order book's past dynamics to predict the mid-price's future direction, statistically outperforming an appropriately chosen unpredictive benchmark.

Motivated by universal approximation theorems, we aim to remain relatively agnostic about the underlying functional relationship and instead parametrize the class of prediction maps using a deep neural architecture. To account for the sequential nature of the data, the model will be strongly characterized by a recurrent design.

Without further ado, let us begin with the basics and introduce order book-based financial markets.

2.1 Introduction

2.1.1 Financial markets and exchanges: setting the scene

A financial market is an ensemble of market agents willing to buy or sell a certain financial security, such as a stock, bond, or derivative. The natural infrastructure for a financial market is an exchange, a physical or virtual place that brings together buyers and sellers, facilitating the occurrence of transactions. Multiple types of exchanges exist, ranging from open outcry exchanges, where agents physically meet to discuss prices, to electronic exchanges, where transactions occur almost instantaneously via the internet.

Today most transactions occur on electronic exchanges. The way trades are executed on an electronic exchange may vary: in an auction market, trades occur directly between buyers and sellers, while in a dealer market, trades go through a specifically designated intermediary, known as a market maker. Throughout this chapter, we will be focusing on listed U.S. equities. These are

This chapter is based on the published article: Lucchese, Pakkanen, and Veraart (2024).

often traded on two main exchanges: the NYSE, a hybrid (floor and electronic) auction market accounting for 20% of the U.S. equities market transactions, and the Nasdaq, a fully electronic dealer market executing about 16% of such trades (as of time of writing, source: Cboe (2023)). Both of these markets allow traders to access live order book information, i.e. the collection of all standing orders for a given security. In theory, this allows for symmetric information across traders, which should all have access to the same data regarding market depth, liquidity, and price discovery dynamics.

In practice, traders have access to different technology, receiving market data and submitting orders at different latencies. In the quest to exploit the advantages gained by faster access to markets, almost two decades ago, a new market participant appeared. These market players are today known as High-Frequency Traders (HFTs) and, over the years, have rapidly grown to represent a significant share of the market (Abergel, Lehalle, and Rosenbaum, 2014). Their role has since been the object of a fierce debate: their critics claim HFTs engage in predatory – and sometimes illegal – behavior, while their supporters believe HFTs to be overall beneficial to the market by providing liquidity, reducing spreads, and helping price discovery dynamics. In this chapter, we will not delve into questions regarding the legitimacy of HFT practices, but we will instead aim to independently analyze the value of faster access to order book information. Over the past few decades, HFT companies have engaged in a fierce race to zero latency, making vast economic efforts to reduce their latency by just a few microseconds. What we aim to explore in this work is one of the possible reasons why such a race happened in the first place. Specifically, we will be analyzing the predictive value of order book data, i.e. to what extent can a trader with immediate access to the order book predict the future direction of the market?

It is important to note that electronic auction and dealer markets are the predominant form of exchanges, but alternative market structures exist, for example, decentralized securities can be traded on decentralized exchanges (DEXs) where orders are executed through an automated market maker. It is also important to note that not all trades occur on exchanges, recently there has been a surge in alternative trading systems (ATS) such as dark pools and ECNs, which now count for 43% of the volume of traded exchange-listed US equities (Cboe, 2023). In general, over the past couple of decades, the trading process has become increasingly complex due to market fragmentation, availability of new technologies such as smart order routers (SORs), and – sometimes controversial – regulation, e.g. RegNMS (US Securities and Exchange Commission, 2005) and MiFID (European Parliament and Council, 2004). While practices that exploit statistical arbitrage between competing trading venues exist, in this work, we will assume the trader has access to a single electronic order book-based exchange, namely the Nasdaq.

2.1.2 Order book predictability: asking the right questions

As discussed in the opening paragraph, we would like to explore whether, contrary to low-frequency returns, ultra-high-frequency returns tend to display predictability. Empirical studies (Sirignano and

Cont, 2019) have shown that price formation dynamics, i.e. next mid-price moves, are predictable. In this chapter, we will try to understand whether such predictability persists at longer horizons. Intuitively, predictability in high-frequency returns may be understood to arise simply from the way an order book market is structured, i.e. the side of the order book with less liquidity is more likely to erode faster, resulting in a price increase/decrease, or as the product of recurring trading patterns in response to liquidity information. The approaches considered in the literature for forecasting high-frequency returns from order book data can be roughly divided into two categories: relatively simple models built on carefully handcrafted features (Aït-Sahalia et al., 2022) and more sophisticated architectures applied directly to raw order book data (Zhang, Zohren, and Roberts, 2019; Zhang and Zohren, 2021; Kolm, Turiel, and Westray, 2021). In this work, we will focus on the latter class of models, leveraging the ability of deep learning techniques to learn complex dependence structures. We will consider a specific class of deep learning models, introduced by Zhang, Zohren, and Roberts (2019), designed to extract features from order book data. There is empirical evidence (Bengio, Courville, and Vincent, 2013) which suggests that, although deep learning models can extract complex features, the way data is represented may have a significant impact on model performance. We will hence explore how model performance varies when changing the way the order book data is fed as input. Equipped with this class of models, we will aim to investigate the following questions, which naturally arise from our preceding discussion:

1. Do high-frequency returns display order book-driven predictability? And if so, how far ahead can we predict?

2. Which order book representations perform best?

3. Can we use a single model across multiple horizons?

4. Can we use a single model across multiple stocks?

We aim to answer all these questions in a formalized statistical inference framework based on model confidence sets (Hansen, Lunde, and Nason, 2011).

2.1.3 Related work and contributions

It is important to note that mathematical modeling of order book dynamics is a very broad and active area of research: ranging from Hawkes process models for order and trade events in continuous time (Bacry and Muzy, 2014; Large, 2007) to discrete-time synthetic data-driven order book generation (Byrd, Hybinette, and Balch, 2020; Coletta et al., 2022). The work presented in this chapter can be seen as contributing to two parallel research streams in the literature for order book-driven mid-price predictions. On one hand, we expand on the deep learning ideas discussed in Zhang, Zohren, and Roberts (2019), Zhang and Zohren (2021), and Kolm, Turiel, and Westray (2021) by introducing new data representations and carrying out a disciplined comparison

between the specifications. On the other, we explore a set of questions related to short-term price predictability in a similar spirit to Aït-Sahalia et al. (2022) but under a different class of models.

In the broader context of the questions addressed in this chapter, related works are those of Sirignano and Cont (2019), exploring the universality of order book dynamics, and Wu et al. (2021), advocating for robust representations of order books. We base our experimental procedure on model confidence sets, introduced by Hansen, Lunde, and Nason (2011). We believe this formalized statistical inference framework perfectly suits our aim of addressing questions that require comparisons between multiple models and benchmarks.

The two main contributions of this chapter are summarized as follows. First, we introduce a deep learning model for mid-price forecasting based on a more robust representation of the order book, which we will refer to as deepVOL. This representation allows us to easily adapt the model to the setting where more granular L3 data[1] is available. Second, we provide new empirical results addressing essential questions regarding short-term price predictability in a disciplined experimental framework.

2.2 The space of models under consideration

2.2.1 The order book

At a given point in time, an order book contains all the (visible) buy and sell orders placed for a given security on a specific exchange. The lowest ask price, resp. the highest bid price, is known as the first ask order book level, resp. the first bid order book level. Subsequent levels are defined accordingly: the l-th ask/bid level is the l-th lowest/highest price at which someone is willing to sell/buy the security. An example of a 10-level order book snapshot, i.e. the collection of orders at the best 10 bid/ask prices at a given point in time, is displayed in Figure 2.1. We note that in electronic exchanges, orders can be submitted on an evenly spaced discrete set of prices, known as ticks. The smallest price increment is known as the tick size, for most U.S. traded stocks, it corresponds to \$0.01. It is important to note that not all tick prices may have standing orders; therefore, the first 10-level ask/bid prices may not coincide with the first 10 ask/bid tick prices.

There are three main types of actions that traders can request on an exchange:

- a limit order: an order to buy a given quantity of the security at a given price;
- a market order: an order to buy a given quantity of the security at the best available price;
- a cancellation or partial deletion: an order to fully or partially delete a standing limit order.

Note that while market orders are always immediately executed once posted, passive limit orders, i.e. orders which do not cross the spread, will sit in the order book until they are matched. A limit order entering the market at a tick price where other limit orders are already present will be added

[1] See Section 2.2.1 for the definitions of L1, L2 and L3 order book data.

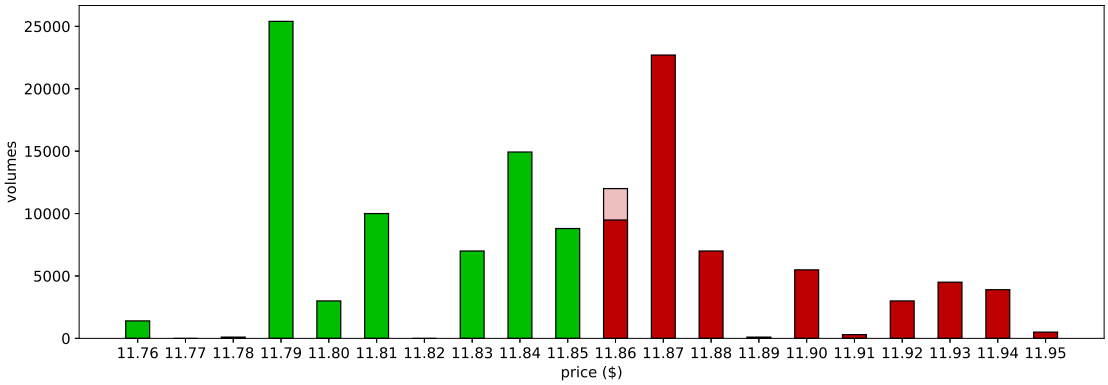


Figure 2.1: A sample snapshot of an order book. Ask (resp. bid) volumes are denoted by red (resp. green) bars. The lighter red shaded area represents the change in the order book shape when some of the liquidity at the best ask price is removed.

at the end of the standing queue. Trades occur each time a market order or aggressive limit order[2] is posted, the requested volume is matched to the standing limit orders according to price-time priority. We will not delve into the detailed characteristics of all the different order types which exist, but it suffices to point out that the three actions described above represent the fundamental drivers of the evolution of order books. For example, in Figure 2.1, the change in order book shape may be due to either a (partial) deletion of a first-level ask limit order or the execution of a buy order (either a market order or an aggressive limit order).

During trading hours, electronic exchanges operate continuously, and for this reason, order books are sometimes referred to as continuous books. Exchanges such as the Nasdaq timestamp each event at nanosecond precision. Between different stocks, the level of trading activity may vary significantly, and the time elapsed between consecutive events may differ by orders of magnitude. For this reason, we define an alternative time clock: a discrete order book clock, which increments by one each time an action is executed on the order book. Throughout our discussion, we will explore questions of predictability with respect to this order book clock, which is the same as the one considered in Ntakaris et al. (2018). Some authors, for example, Aït-Sahalia et al. (2022) consider alternative order book-driven time clocks, such as transaction clocks and volume clocks.

We understand that models based on order book-specific clocks might be challenging to use in practical trading applications. On one hand this approach tends to discard some important microstructural information, such as event inter-arrival time. On the other, when constructing target returns using an event-based clock, one might identify predictable instances which are not practically tradable due to physical latency limitations, especially in periods of intense trading activity[3]. While acknowledging these limitations, we still believe order book-based clocks provide

[2] A limit order that crosses the spread.

[3] In practice one might thus want to analyze the latency sensitivity of the predictions produced by these models, i.e. understand how predictability of the returns vanishes as one increases physical latency in the construction of the target return as detailed in Appendix 2.G.

a more natural measure of time than physical wall clocks for exploring predictability as they are more suitable for comparing results across stocks: a 100ms time horizon has a significantly different meaning for stocks with different levels of trading activity.

Another important observation is that order book data might not be accessed by all market participants at the same level of granularity. The Nasdaq Quotation Dissemination Service makes the following distinctions:

- L1 data: the best bid and ask prices and corresponding volumes;
- L2 data: all available bid and ask prices and corresponding volumes;
- L3 data: all available bid and ask prices and corresponding volumes split among the orders in the queue.

We will be comparing the performance of models when different levels of data granularity are available[4]. Figure 2.2 provides a visual comparison of L1, L2, and L3 data.

We introduce the following notation: at time $t \in \mathbb{N}$ we denote by

- $p^{(l)}_{x,t}, v^{(l)}_{x,t}$ for $l = 1, 2, \ldots$ the l-th level ask/bid price and volume, for $x \in \{a, b\}$;
- $m_t = \dfrac{p^{(1)}_{b,t} + p^{(1)}_{a,t}}{2}$ the mid-price;
- $\pi^{(j)}_{x,t}, s^{(j)}_{x,t}$ for $j = 1, 2, \ldots$ the j-th ask/bid tick price from the mid and corresponding volume for $x \in \{a, b\}$, i.e. this is defined recursively as

$$\pi^{(1)}_{x,t} = \begin{cases} m_t & \text{if } m_t \in \mathcal{T}, \\ m_t \pm \frac{\vartheta}{2} & \text{if } m_t \notin \mathcal{T}, \end{cases}$$
$$\pi^{(j)}_{x,t} = \pi^{(j-1)}_{x,t} \pm \vartheta \quad \text{for } j = 2, 3, \ldots,$$

  where ϑ is the tick size, $\mathcal{T} = \{k\vartheta : k \in \mathbb{N}\}$ is the set of possible tick prices and $\pm$ depends on $x \in \{a, b\}$.

- $q^{(i,k)}_{x,t}$ for $k = 1, 2, \ldots$ the queue corresponding to volume $s^{(i)}_{x,t}$ and ordered by time priority.

2.2.2 Predictability of returns

We will explore models that aim to identify short-term predictability in returns. We first introduce the familiar regression framework for return predictions. We then rephrase the task in terms of a classification problem, extending the definition of predictability to this setting.

[4]In order book-based markets the most granular view available is a time-sequenced data feed that provides every single event occurring in the order book (on the Nasdaq this is known as the ITCH feed). This feed contains information that is discarded even from the L3 view of the order book, such as trade information, order history (modifications and partial fills) and executions of hidden quantities. In Table 2.2 we give a small sample of what one (decoded) message in an ITCH feed might look like.

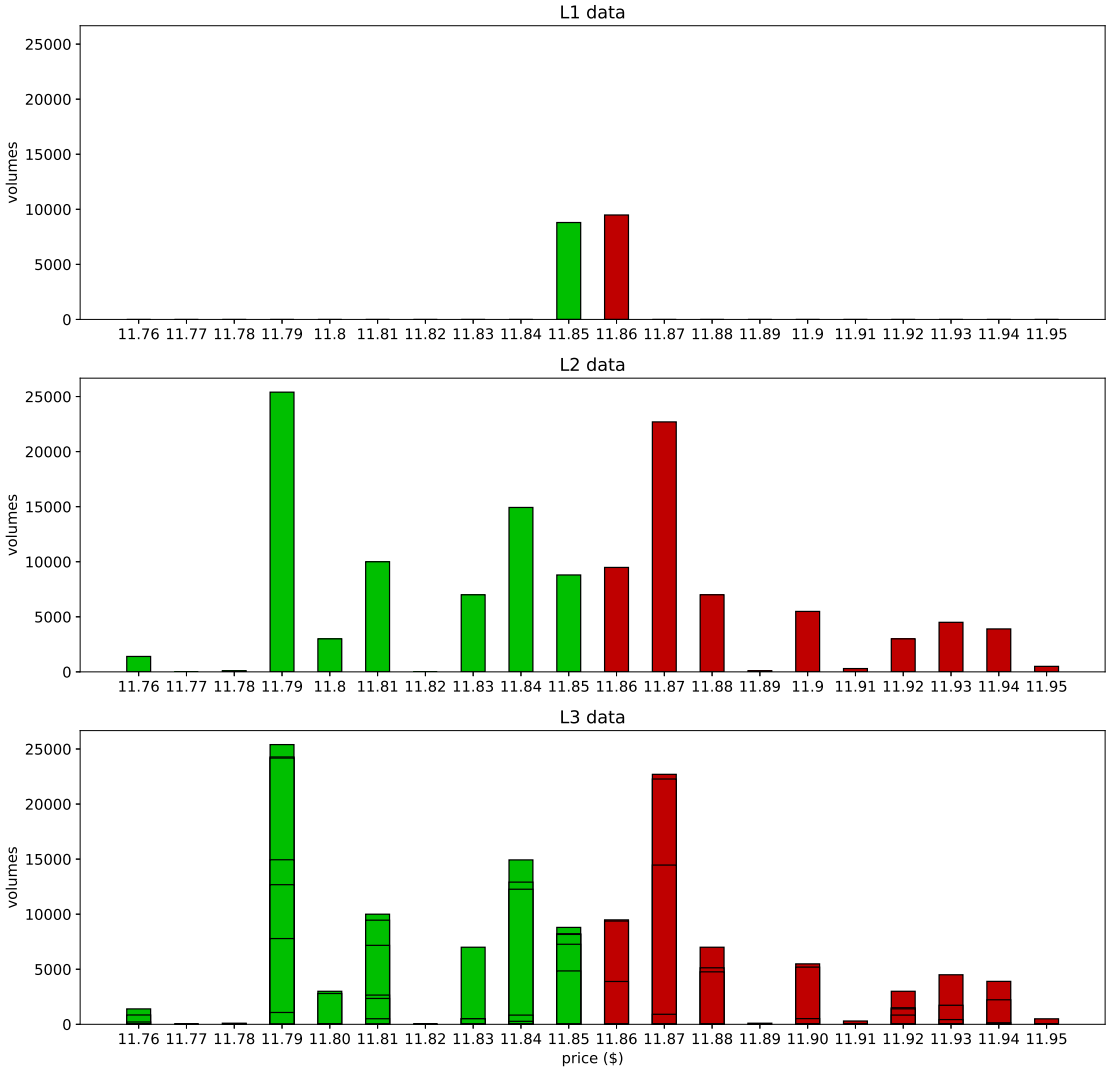


Figure 2.2: L1, L2 and L3 representations of the order book in Figure 2.1.

We will be exploring predictability arising from past order book data and thus we define the information σ-algebra $\mathcal{F}_t$ to be

$$\mathcal{F}_t = \sigma(\mathbf{x}_t, \ldots, \mathbf{x}_{t-T+1}),$$

where $\mathbf{x}_t, \ldots, \mathbf{x}_{t-T+1}$ are order book derived features at times $t, \ldots, t-T+1$ for some look-back window of length T.

Predictions in the regression framework

Let us first consider the regression setting. At time t we denote by $r_{t,t+h} \in \mathbb{R}$ the h-step ahead mid-price return, as defined in Section 2.3.2. Given order book information at time t, $\mathcal{F}_t$, there exists a measurable function g such that

$$\mathbb{E}[r_{t,t+h}|\mathcal{F}_t] = g(\mathbf{x}_t, \ldots, \mathbf{x}_{t-T+1}),$$

or, equivalently,

$$r_{t,t+h} = g(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}) + \epsilon_t,$$

for some mean-zero noise variable ϵ_t orthogonal to the space of $\mathcal{F}_t$-measurable random variables. A prediction is defined to be any $\mathcal{F}_t$-measurable random variable and the best prediction is the $\mathcal{F}_t$-measurable random variable $r^{(t)}_{t,t+h}$ which minimizes the expected cost

$$\mathbb{E}[C(r^{(t)}_{t,t+h}, r_{t,t+h})],$$

for some appropriate cost function $C : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. In the case of quadratic cost $C(r_1, r_2) = (r_1 - r_2)^2$, we have

$$r^{(t)}_{t,t+h} = \mathbb{E}[r_{t,t+h}|\mathcal{F}_t] = g(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}).$$

Different choices of cost function are possible, for example, when C is absolute cost, i.e. $C(r_1, r_2) = |r_1 - r_2|$, the best prediction is the conditional median of $r_{t,t+h}$ given $\mathcal{F}_t$. Given a parametric family of models $\{g_\theta(\cdot)\}_{\theta\in\Theta}$ and an observed training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}, r_{t,t+h})\}_{t\in\mathcal{I}_{\text{train}}}$ one can first learn a function $g_{\hat{\theta}}$ approximating g and then produce the return predictions

$$\hat{r}^{(t)}_{t,t+h} = g_{\hat{\theta}}(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}),$$

for test data points $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}, r_{t,t+h})\}_{t\in\mathcal{I}_{\text{test}}}$. Assuming returns to be stationary, we say that there is order book-driven predictability if the learned predictions outperform an *unpredictive benchmark* prediction on the testing set with respect to the chosen cost function $C(\cdot, \cdot)$.

Predictions in the classification framework

In this work, we will discretize the space of returns by grouping mid-price movements as downward, no-change, and upward[5]. We hence introduce the discretized return random variable

$$c_{t,t+h} = \begin{cases} \downarrow & \text{if } r_{t,t+h} \in (-\infty, -\gamma), \\ = & \text{if } r_{t,t+h} \in [-\gamma, +\gamma], \\ \uparrow & \text{if } r_{t,t+h} \in (+\gamma, +\infty), \end{cases}$$

for an appropriate choice of $\gamma > 0$, cf. Section 2.3.2. Instead of modeling only the expected conditional return, in the classification setting, one aims to approximate the whole conditional

[5]We believe that, for high-frequency returns, a classification framework is more natural: mid-prices move in (half)-tick increments, and hence the corresponding returns live on a discrete lattice (see Section 2.3.2). Moreover, grouping returns in three classes (downward, no-change, and upward) significantly reduces the effects of idiosyncratic noise, i.e. large price changes unrelated to microstructural dynamics, while preserving the ability to investigate questions regarding predictability.

distribution, i.e. find measurable functions $p_\downarrow, p_=, p_\uparrow$ such that

$$\mathbb{P}(c_{t,t+h} = *|\mathcal{F}_t) = p_*(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}),$$

for $* \in \{\downarrow, =, \uparrow\}$. The discretized return prediction for an unobserved sample is then given by the minimizer of the expected misclassification cost:

$$\hat{c}_{t,t+h} = \underset{*\in\{\downarrow,=,\uparrow\}}{\operatorname{argmin}} \sum_{\star\in\{\downarrow,=,\uparrow\}} c_{*|\star} p_{\star,\hat{\theta}}(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}),$$

where the parametric conditional probabilities $p_{\downarrow,\hat{\theta}}, p_{=,\hat{\theta}}, p_{\uparrow,\hat{\theta}}$ are learnt from a training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}, c_{t,t+h})\}_{t\in\mathcal{I}_{\text{train}}}$ and $c_{*|\star}$ is the misclassification cost of a $*$ observation classified as a $\star$. Assuming the return process is stationary, we say that there is order book-driven predictability if the learned predictions outperform an *unpredictive benchmark* prediction on a test set in terms of total misclassification cost.

Specifying an appropriate cost function is task-dependent. For example, a trader using the predictions as trading signals might be more interested in the correctness of up and down predictions than no-change ones. On the other hand, a market maker might prioritize the correctness of no-change predictions when using these to decide whether to tighten quoted spreads. In both cases, the consequences of different types of errors are asymmetric and heavily impact the choice of the cost matrix $C = \{c_{*|\star}\}_{*,\star\in\{\downarrow,=,\uparrow\}}$. In the classification framework an alternative approach is to compare the predicted conditional distributions, $\{p_*(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1})\}_{*\in\{\downarrow,=,\uparrow\}}$ to the realized outcomes directly. This approach does not require specifying a cost matrix C nor a return prediction. The predicted conditional distribution (the output of our model) is directly compared with the observed data via a suitable "distance" on the space of probability measures $\mathcal{P}(\{\downarrow, =, \uparrow\})$, where realized returns are encoded as Dirac measures. A natural choice of such "distance" is given by the categorical cross-entropy, as this can be interpreted (under appropriate assumptions) as the (log)-likelihood of the test set:

$$-\log \mathbb{P}(\mathcal{D}_{\text{test}}|\theta) \propto -\frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{t\in\mathcal{I}_{\text{test}}} \sum_{*\in\{\downarrow,=,\uparrow\}} \mathbb{1}_{\{c_{t,t+h}=*\}} \log p_{*,\hat{\theta}}(\mathbf{x}_t, \dots, \mathbf{x}_{t-T+1}),$$

where the model parameters $\hat{\theta}$ are learnt from the training set $\mathcal{D}_{\text{train}}$. As this work does not target a specific trading strategy but is more concerned with general predictability questions, we will be evaluating our models based on the categorical cross-entropy loss. In the following, we will thus say that there is order book-driven predictability if the learned conditional distributions outperform an *unpredictive benchmark* distribution on a test set relative to categorical cross-entropy loss. As discussed in Appendix 2.A, the natural choice for the loss used in training will also be categorical cross-entropy.

Note that in both the regression and classification framework, we compare the learned predic-

tions with those obtained from an *unpredictive benchmark* model to determine whether there is predictability by using a score/cost function. Simply comparing point estimates of the scores does not provide a sound statistical argument answering the question of whether there is predictability: the difference in score may simply be due to statistical variability. This is where the model confidence set (Hansen, Lunde, and Nason, 2011) procedure comes in. As we will discuss in Section 2.4 this provides a statistical testing framework to determine whether the learned models statistically outperform the *unpredictive benchmark*, i.e. whether there is predictability.

Remark 2.2.1. It is clear that the definition of predictability is intrinsically tied to that of *unpredictive benchmark*. Different choices for *unpredictive benchmark* models are possible, for example, in the regression framework, a natural choice is given by (a version of) the efficient market hypothesis[6] (EMH) (Fama, 1970). In this case, the unpredictive hypothesis assumes the conditional expected return to be zero. Under the classification framework, the simple EMH does not translate into a clear-cut model for the conditional distribution of returns. In this setting, we consider as a natural unpredictive hypothesis a slightly stronger version of the EMH: under the *unpredictive benchmark*, returns are assumed to be IID and independent of any information up to time t. The *unpredictive benchmark* conditional distribution will therefore be given by the empirical distribution of the training set. As a side remark, we believe it is important to note the EMH was originally proposed in a very different market environment and at significantly lower frequencies than the ones considered in this work. Nevertheless, we believe the EMH provides a natural unpredictive benchmark hypothesis when testing for predictability.

As mentioned in Section 2.1.2, there are two main approaches considered in the literature for exploring the short-term predictability of returns in order book markets: the first considers carefully handcrafted features $\mathbf{x}_t, \dots, \mathbf{x}_{t-T}$ and relatively simple specifications for the prediction functions g or $(p_\downarrow, p_=, p_\uparrow)$, e.g. linear specifications or decision trees, the second – which we will explore in this chapter – uses raw order book features $\mathbf{x}_t, \dots, \mathbf{x}_{t-T}$ as inputs to more complex prediction functions, e.g. deep neural network architectures. We will assume the reader is familiar with deep learning techniques and thus give a brief overview of the relevant concepts in Appendix 2.A, more detailed expositions can be found in Goodfellow, Bengio, and Courville (2016).

2.2.3 Deep learning models

In the previous section, we set up the learning framework for the return prediction task. We now discuss how the neural networks covered in Appendix 2.A can be combined to model price formation mechanisms and predict h-step ahead high-frequency returns. We will consider a specific class of deep learning models based on the deepLOB architecture (Zhang, Zohren, and Roberts, 2019).

[6] In an efficient market prices always fully reflect available information and hence the conditional expectation of a price return given a public information set is zero.

deepLOB, (Zhang, Zohren, and Roberts, 2019)

This network acts on raw order book input. A CNN module and an inception module feed into an LSTM layer which produces the final classification output. The CNN and inception module aim to extract short-term spatio-temporal features in the data, while the LSTM module works on longer-term dependencies.

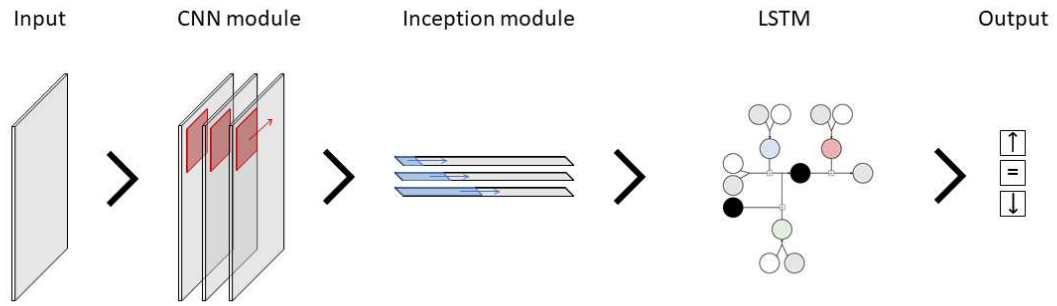


Figure 2.3: Core modules of network architectures.

The deepLOB architecture is summarized in Figure 2.3 and is made up of the following components:

- **Input**. The first L level raw order book information - price and volume - with a look-back window of length T is used as input. The input at time t is thus a $(T \times 4L)$ array given by

$$\{\mathbf{x}_{t-\tau}\}_{\tau=0}^{T-1} = \left\{ \left\{ \left(p_{a,t-\tau}^{(l)}, v_{a,t-\tau}^{(l)}, p_{b,t-\tau}^{(l)}, v_{b,t-\tau}^{(l)} \right) \right\}_{\tau=0}^{T-1} \right\}_{l=1}^{L} \in \mathbb{R}^{T \times 4L}.$$

  A feature-wise rolling window z-score standardization is applied to the input, i.e. $v_{a,t-\tau}^{(1)}$ is standardized using the mean and standard deviation of the first level ask volumes over the previous five days.

- **CNN module**. Convolutions are applied to the data in both the spatial and temporal dimensions. The spatial convolutions aim to aggregate information across order book levels, and the temporal convolutions can be understood as smoothing and/or differencing operations. The CNN module is summarized in Figure 2.4.

- **Inception module**. This module up-samples the convoluted data by applying various temporal convolutions with different filter lengths (time-window). Each temporal convolution can be interpreted as a (weighted) moving average. It is similar in spirit to the computation of technical indicators, but the frequencies at which this is applied are substantially different.

- **LSTM**. The Long Short-Term Memory layer takes the multidimensional time series produced by the inception module and feeds it through a recurrent network structure aimed at extracting

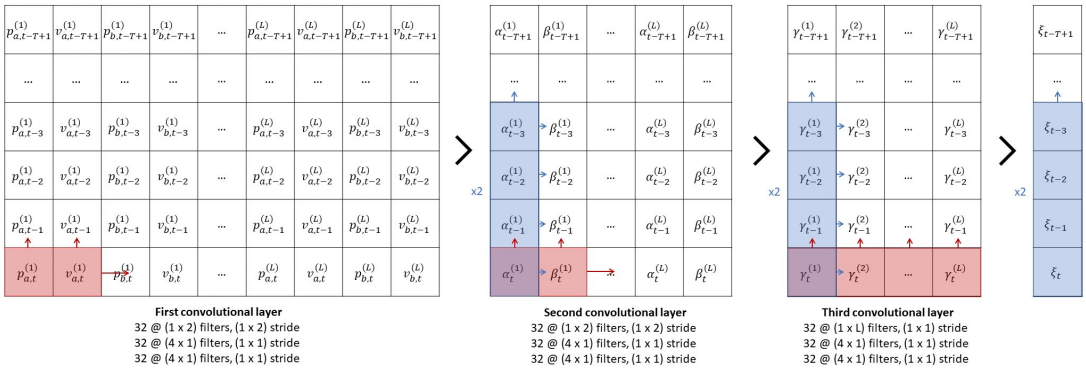


Figure 2.4: CNN module deepLOB.

longer-term dependencies among the data. The last hidden state of the LSTM is passed through a dense layer with a softmax activation function to produce the h-step ahead return prediction $\downarrow, =$ or $\uparrow$.

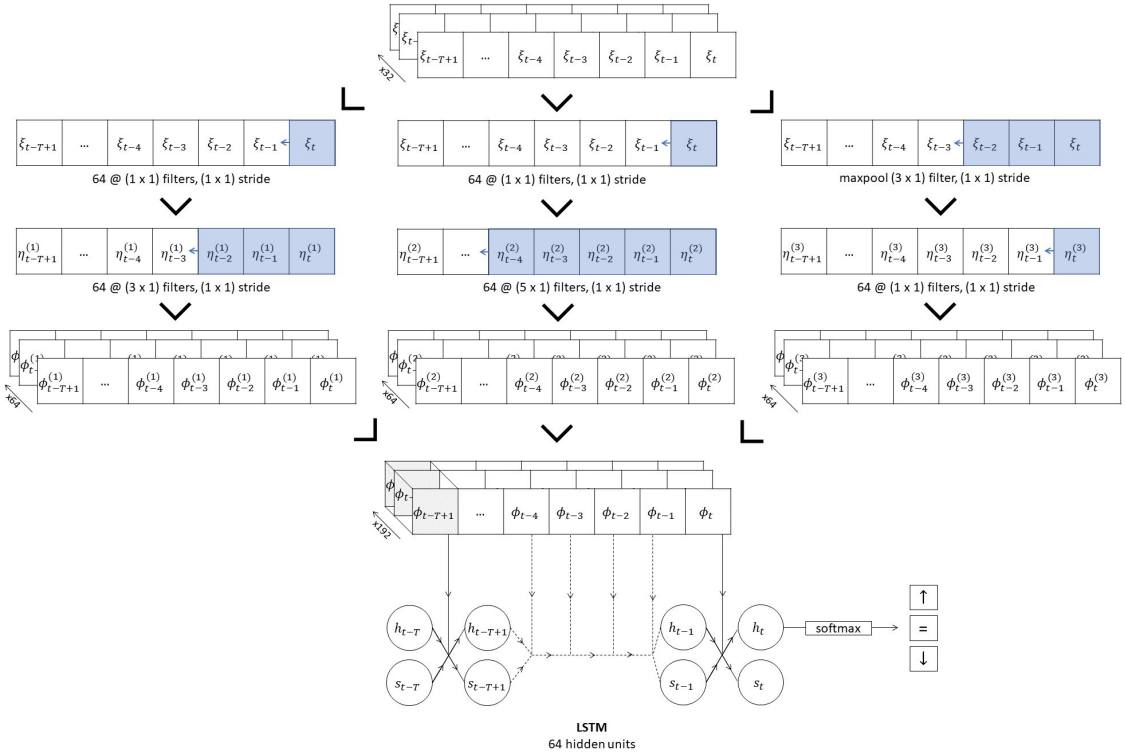


Figure 2.5: Inception module and LSTM layer.

The exact details of the deepLOB architecture can be found in Table 2.I.1.

deepOF, (Kolm, Turiel, and Westray, 2021)

In the deepLOB model order book states, which are non-stationary, are mapped to a stationary quantity, returns. While in theory, this shouldn't be a problem (due to universality property of

deep neural networks), Kolm, Turiel, and Westray (2021) argue that using some form of stationary input might improve model performance by facilitating the training procedure.

A stationary[7] order book quantity is order flow. First-level order flow describes the net flow of orders at the best bid and ask. This was introduced in Cont, Kukanov, and Stoikov (2013) to explore the price impact of order book events. This single quantity parsimoniously models the instantaneous effect of order book events on prices. The first-level bid and ask order flows corresponding to the order book event occurring at time t (in order book time) are defined by:

$$bOF_t^{(1)} = \begin{cases} v_{b,t}^{(1)}, & \text{if } p_{b,t}^{(1)} > p_{b,t-1}^{(1)}, \\ v_{b,t}^{(1)} - v_{b,t-1}^{(1)}, & \text{if } p_{b,t}^{(1)} = p_{b,t-1}^{(1)}, \\ -v_{b,t-1}^{(1)}, & \text{if } p_{b,t}^{(1)} < p_{b,t-1}^{(1)}, \end{cases} \quad \text{and} \quad aOF_t^{(1)} = \begin{cases} -v_{a,t-1}^{(1)}, & \text{if } p_{a,t}^{(1)} > p_{a,t-1}^{(1)}, \\ v_{a,t}^{(1)} - v_{a,t-1}^{(1)}, & \text{if } p_{a,t}^{(1)} = p_{a,t-1}^{(1)}, \\ v_{b,t}^{(1)}, & \text{if } p_{a,t}^{(1)} < p_{a,t-1}^{(1)}. \end{cases}$$

The difference between the two is known as order flow imbalance:

$$OFI_t^{(1)} = bOF_t^{(1)} - aOF_t^{(1)}.$$

Bid order flow corresponds to the net change in volume at the best bid level; the three cases in the definition can be understood as:

- a new bid order being placed at a higher price than the current best bid;
- the order volume at the best bid price increasing or decreasing;
- the entire volume at the best bid price being consumed, thus decreasing the best bid price.

Note that, as discussed in Section 2.2.1, volumes can increase or decrease due to orders being placed, executed, or canceled. A similar interpretation holds for ask order flow.

In Xu, Gould, and Howison (2019) the authors explore flows of volumes at deeper levels in the order book, by introducing multi-level order flow. The definitions are exactly as above with superscript (1) replaced by general (l). We note that in both Cont, Kukanov, and Stoikov (2013) and Xu, Gould, and Howison (2019) the authors investigate the explanatory power of (multi-level) order flow imbalance for price changes, i.e. the relationship between contemporaneous order flow imbalance and price changes.

In our setting, as in Kolm, Turiel, and Westray (2021), we are instead interested in exploring the predictive power of order flow, i.e. the relationship between past order flow and future price changes. We will hence refer to deepOF as the deepLOB architecture with stationary order flow

[7] Generally, prices are deemed to be non-stationary while their changes, i.e. returns, are modeled under the assumption of stationarity. Similarly, order flow can be understood as representing the stationary change in best bid/ask volume. Whether a time series is indeed stationary is a difficult statistical property to test and usually requires imposing some quite specific model on the data. In this context, we thus refrain from making the notion of stationarity more rigorous than the previous simple argument by analogy.

input, i.e. a $(T \times 2L)$ array given by

$$\{\mathbf{x}_{t-\tau}\}_{\tau=0,\dots,T-1} = \left\{ \left\{ \left(aOF_{t-\tau}^{(l)}, bOF_{t-\tau}^{(l)} \right) \right\}_{\tau=0}^{T-1} \right\}_{l=1}^{L} \in \mathbb{R}^{T \times 2L}.$$

The order flow input enters the CNN module in Figure 2.4 at the second convolutional layer. The rest of the architecture is exactly the same as deepLOB, as detailed in Table 2.I.1. We note that the first convolutions applied to the order flow input, i.e. the second convolutional layer in Figure 2.4, aggregate information across bid and ask order flows, essentially computing a weighted order flow imbalance.

Remark 2.2.2. The original deepOF specification in Kolm, Turiel, and Westray (2021) was structured as a (multi-horizon) regression task. In their setting, the last layer of the deep neural network maps each prediction to $\mathbb{R}$ instead of $\mathcal{P}(\{\downarrow, =, \uparrow\})$. Moreover, as discussed in Section 2.2.5, the way multi-horizon predictions are produced in the original work does not rely on the encoder-decoder structure we use for our multi-horizon models. Another slight difference with the experiments in Kolm, Turiel, and Westray (2021) is in the standardization procedure: instead of standardizing each feature by its training mean and standard deviation, we use a rolling window approach.

2.2.4 DeepVOL

The need for a robust representation

As discussed in the previous section the main difference between deepLOB and deepOF is in the way the data is fed into the model. While deepLOB uses raw order book data as input, deepOF uses a derived quantity, order flow. In general, the success of deep learning tasks is highly dependent on the way the data is represented. The task of predicting returns in order book markets is no exception. In order to achieve the best possible results, one should adopt a robust representation of the data. Wu et al. (2021) identify five main desiderata for a robust representation of order book data:

- Region of interest: the entire order book may contain a wide range of prices, the data representation should select a region of interest based on a complexity-performance trade-off.
- Efficiency: the data representation should avoid excessive dimensionality.
- Validity: the data representation should have a simple definition of valid manipulations.
- Smoothness: the data representation should be robust to small perturbations.
- Compatibility: the data representation should be compatible with the deep learning architecture.

In other words, a robust representation of the order book is a parsimonious data representation well suited for the downstream modeling task. We note the order book representations used in

deepLOB and deepOF do not conform to these desiderata. Order book states organized by 'level' do not have a simple validity (price and volume information are intrinsically tangled and would loose their significance if treated separately in a black box algorithm), are not robust to small perturbation (small orders added at empty ticks completely change the order book feature vector), and are incompatible with deepLOB's CNN module (the spatial structure is not homogeneous as there is no fixed interval between levels). It turns out that while the 'level' representation of the order book may be easily understandable by humans it is less so for statistical models. In addition to not satisfying the desiderata, this representation does not respect the following fundamental but implicit assumption of deep learning models: signals at the same entry of the input should come from the same source. In the 'level' representation, as new order book events happen, the same signal (i.e. a posted order) may move between levels.

We, therefore, introduce volume features, which provide a robust representation of order book data. Fixing a window of size $W > 0$ we define:

$$\{\mathbf{x}_{t-\tau}\}_{\tau=0}^{T-1} = \left\{\left(s_{b,t-\tau}^{(W)}, \ldots, s_{b,t-\tau}^{(1)}, s_{a,t-\tau}^{(1)}, \ldots, s_{a,t-\tau}^{(W)}\right)\right\}_{\tau=0}^{T-1},$$

where $s_{x,t-\tau}^{(j)}$ for $x \in \{a, b\}$ are the bid/ask volumes corresponding to the j-th price from the mid $\pi_{x,t-\tau}^{(j)}$, as defined in Section 2.2.1. We note that the volume representation indeed satisfies the five desiderata: a region of interest is identified (via the window $W > 0$), it is efficient (for the same dimension of input it may convey more or less information than the 'level' representation, depending on how sparse the orders are placed in the order book), it has a simple validity (all entries are in the same units), it is robust to small perturbations (new orders at empty levels minimally affect the feature) and is compatible with the CNN architecture (the spatial structure of the volumes is homogeneous). An intuitive visualization of this representation as a one-dimensional gray-scale strip is given in Figure 2.6, when including a time dimension this naturally becomes a two-dimensional gray-scale image. The main drawback of the volume representation is that when orders are placed far apart in the order book it is sparse and a larger window $W > 0$ may be required. Our definition of volume features is similar to the mid-price-centered moving window representation of Wu et al. (2021), with the latter living in $\mathbb{R}^{2W+1}$ instead of $\mathbb{R}^{2W}$ and using $\pm$ signs to distinguish between bid and ask volumes. The need for a new, more robust representation of the order book was reached independently.

1400	0	100	25400	3000	10000	0	7000	14930	8800	12000	22700	7000	100	5490	300	3000	4500	3900	500
1400	0	100	25400	3000	10000	0	7000	14930	8800	9484	22700	7000	100	5490	300	3000	4500	3900	500

Figure 2.6: Gray-scale visualization of the volume representation of the order book in Figure 2.1.

In order to use the deepLOB architecture with this new data representation, we must adapt the

CNN module. In particular, we fold the volume input into a three-dimensional array

$$\left\{\left\{\left\{s^{(j)}_{x,t-\tau}\right\}_{\tau=0}^{T-1}\right\}_{j=1}^{W}\right\}_{x\in\{a,b\}} \in \mathbb{R}^{T\times W\times 2},$$

and feed this into a three-dimensional convolutional layer with a $(2 \times 2 \times 1)$ filter and $(1 \times 1 \times 1)$ stride. This layer aims to extract imbalances in the order book by comparing volumes on the two sides of the mid-price. The CNN module with the appropriate changes is depicted in Figure 2.7. The rest of the deepVOL architecture is exactly the same as deepLOB with one slight difference in the way the data is normalized. Thinking of the volume representation as a gray-scale image a natural choice of normalization is

$$\left\{s^{(j)}_{x,t-\tau}\right\}_{\tau,j,x} \mapsto \frac{1}{\max\limits_{\tau,j,x} s^{(j)}_{x,t-\tau}} \left\{s^{(j)}_{x,t-\tau}\right\}_{\tau,j,x},$$

instead of the rolling window standardization we apply to deepLOB and deepOF features.

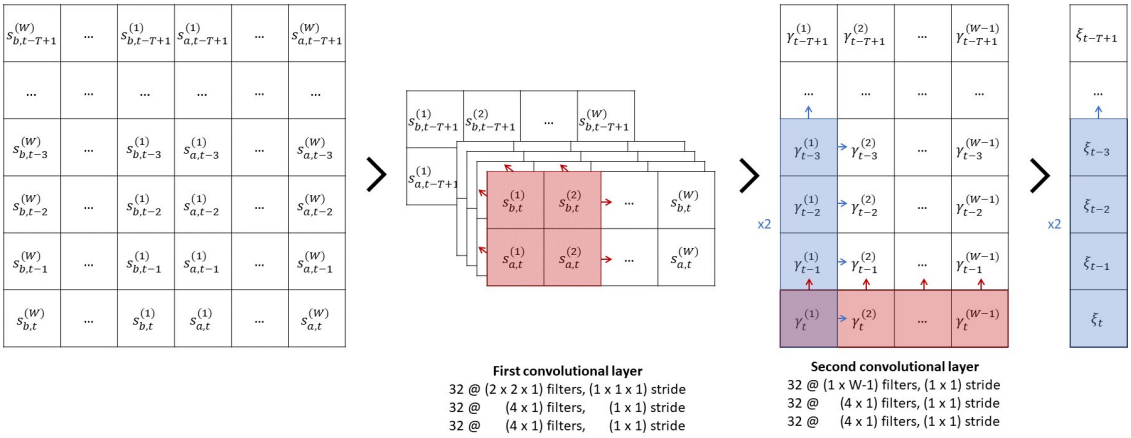


Figure 2.7: CNN module deepVOL.

Remark 2.2.3. One could define a volume flow quantity based on the distance from the mid, in the same way order flow describes the flow of orders based on the level. Following the same motivation for considering order flow in deepOF, one could consider using a volume flow quantity as input to the deep learning architectures with the desired robust representation properties.

L3 volume features

All models considered so far, deepLOB, deepOF, and deepVOL use L2 data as input. If one has access to more granular data, i.e. L3 data breaking down each volume queue into single orders, one might be interested in trying to leverage this information to obtain better predictive performance. We thus define a natural extension of the volume representation considered in the previous section.

Let us denote by $(q^{(j,1)}_{x,t}, q^{(j,2)}_{x,t}, \ldots) \in \mathbb{R}^{\mathbb{N}}$ for $x \in \{a, b\}$ the queue at the j-th bid/ask price from

the mid $\pi^{(j)}_{x,t}$ as introduced in Section 2.2.1. Here $q^{(j,k)}_{x,t}$ denotes the volume of the k-th order in the queue sorted by time priority and is set to zero if there is no such order. The aggregated volume at $\pi^{(j)}_{x,t}$ is given by:

$$s^{(j)}_{x,t} = \sum_{k\geq 1} q^{(j,k)}_{x,t}.$$

A natural extension of the volume representation would therefore be to consider

$$\left\{\left\{\left\{\left(q^{(j,1)}_{x,t-\tau}, q^{(j,2)}_{x,t-\tau}, \dots\right)\right\}_{\tau=0}^{T-1}\right\}_{j=1}^{W}\right\}_{x\in\{a,b\}} \in \mathbb{R}^{T\times W\times 2\times\mathbb{N}}.$$

Unfortunately, this is an infinite-dimensional array that cannot be directly fed into machine learning models. We therefore cut off the queue at a given depth level. In order to avoid discarding precious information, we aggregate all orders sitting past the maximum depth level at the end of the queue. For a given depth level $D > 0$, we thus consider the L3 volume feature

$$\left\{\left\{\left\{\left(q^{(j,1)}_{x,t-\tau}, \dots, q^{(j,D-1)}_{x,t-\tau}, \sum_{k\geq D} q^{(j,k)}_{x,t-\tau}\right)\right\}_{\tau=0}^{T-1}\right\}_{j=1}^{W}\right\}_{x\in\{a,b\}} \in \mathbb{R}^{T\times W\times 2\times D}.$$

Remark 2.2.4. A similar approach to the one used for cutting off the queue might be applied to aggregate volumes sitting deeper in the order book when deriving L2 representations. This would give a better idea of the total liquidity in the order book.

In order to extract relevant information from the queue, we add an initial convolutional layer, which maps each queue to a weighted sum of the order sizes. The weighted aggregated volumes are then fed into the deepVOL architecture as above. The resulting CNN module is summarized in Figure 2.8.

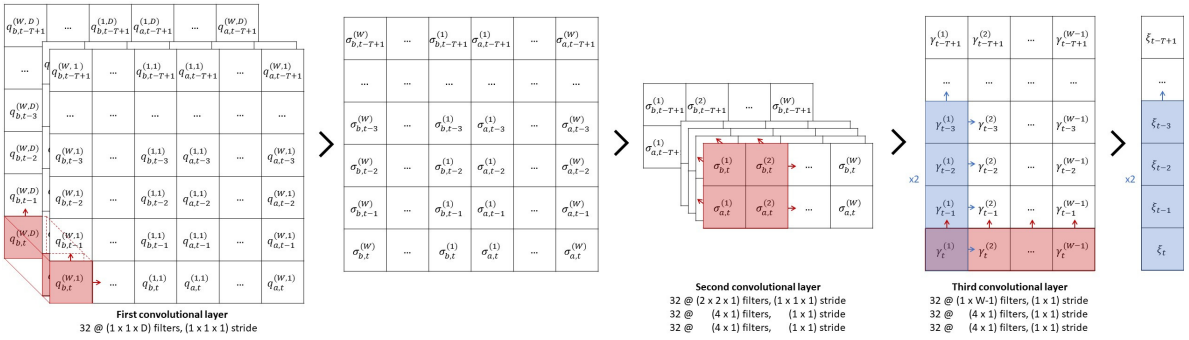


Figure 2.8: CNN module deepVOL(L3). Abusing notation slightly we denote $\sum_{k\geq D} q^{(j,k)}_{x,t-\tau}$ by $q^{(j,D)}_{x,t-\tau}$.

The full architectures for deepVOL and deepVOL(L3) are detailed in Table 2.I.2.

Remark 2.2.5. Note that even with the most granular L3 representation, some economically significant information from the ITCH feed is lost. For example, in the L1/L2/L3 data representations of all models considered in this work, a cancellation and a trade affecting the same orders appear

identical. However, in practice, it might be reasonable to expect the market to react differently to a trade than to a cancellation.

2.2.5 Multi-horizon models

So far we considered single horizon models, i.e. order book input x_t is mapped to a three class distribution $(p_\downarrow, p_=, p_\uparrow)$ corresponding to the discretized return $c_{t,t+h}$ at fixed horizon h. In the following, we consider a generalization of the architectures considered thus far to the setting where the forecasting horizon is a vector $\mathbf{h} = (h_1, \ldots, h_K)$. In this case the modeling task consists in predicting the distributions of the discretized returns $c_{t,t+\mathbf{h}} = (c_{t,t+h_1}, \ldots, c_{t,t+h_K})$.

The simplest way of adjusting the current models to the multi-horizon framework would be to replace the last softmax layer in Figure 2.5 with K parallel dense softmax layers. The last hidden state of the LSTM module would hence be mapped to an array of size $3 \times K$, corresponding to K distributions over three classes. A similar architecture – though in the regression framework – is considered in Kolm, Turiel, and Westray (2021). While this approach is perfectly valid it does not make use of the sequential nature of the task, potentially neglecting an important structural feature of the data.

In this work, we will be leveraging architectures inspired by machine translation which are naturally suited for sequential forecasting. Specifically, we will be considering encoder-decoder models: an encoder maps the input data to a latent summary state (also known as context vector) and then a decoder rolls forward predictions sequentially. In this context let $z_{t-T+1}, \ldots, z_t$ denote the last T hidden states of an encoder at time t, then a decoder rolls forward predictions by:

$$
\begin{aligned}
&z'_0 = z_t, \;\; p_0 = \hat{p}_0, \\
&c_k = h(z_{t-T+1}, \ldots, z_t), \;\; z'_k = f([z'_{k-1}, c_k], p_{k-1}), \;\; p_k = g(z'_{k-1}, c_k),
\end{aligned}
$$

for $k = 1, \ldots, K$. Here $h(\cdot)$ is a function acting on the hidden states of the encoder to extract the context vector c_k (this may possibly depend on other inputs as well, such as previous hidden states of the decoder), $f(r, p)$ is a recurrent layer with recurrent input r and exogenous input p, $g(\cdot)$ is an output layer depending on both the decoder hidden state and the context. The general mechanism of such encoder-decoder architecture is visualized in Figure 2.9.

In our experiments, we will consider a sequence-to-sequence decoder (Cho et al., 2014), the simplest example of such architecture. In this setting we set $c_k \equiv z_t$, implicitly assuming that at every forecasting horizon, the last hidden state of the encoder summarizes all the relevant information required to make the prediction. More complex architectures exist, for example, Luong, Pham, and Manning (2015) introduce an attention-based decoder that uses a weighted combination of all the hidden states of the encoder as context vector. In this setting different weights are used at different forecasting horizons, selectively accessing hidden states of the encoder during decoding. While attention-based networks have been successfully applied for high-frequency mid-

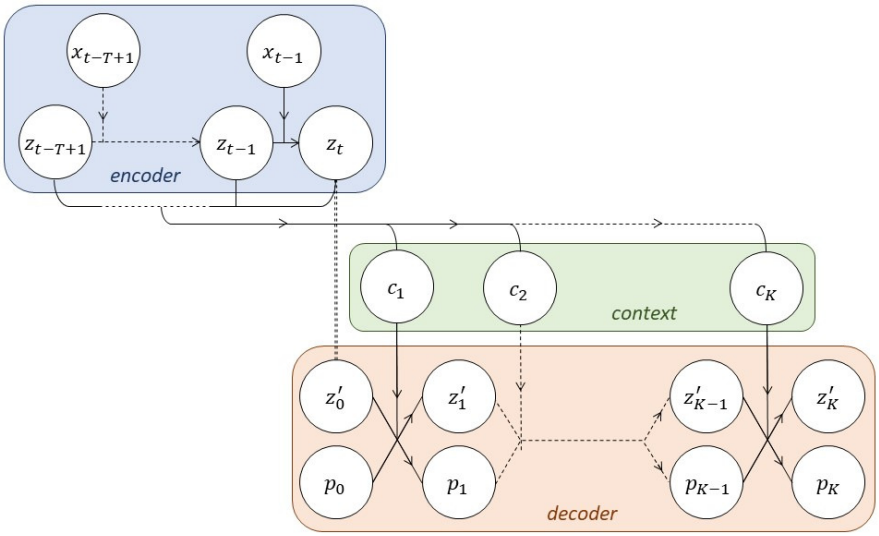


Figure 2.9: A general encoder-decoder architecture.

price predictions (Zhang and Zohren, 2021; Tran et al., 2019), in this work we wish to only exemplify the potential of multi-horizon models and thus restrict ourselves to simple seq2seq decoders.

In the experiments, we set $f(\cdot)$ to be an LSTM and $g(\cdot)$ to be a dense layer with softmax activation. Moreover p_0 is initialized at $\hat{p}_0 = (0, 1, 0)$.

Using multi-horizon forecasting was first proposed for deepLOB in Zhang and Zohren (2021): the LSTM module of the deepLOB architecture acts as an encoder, mapping order book states to the LSTM final latent vector $z_t = [h_t, s_t]$. A seq2seq decoder then rolls forward the prediction producing distributional forecasts at horizons $\mathbf{h} = (h_1, \ldots, h_K)$. The LSTM module with seq2seq decoder is illustrated in Figure 2.10. Clearly, there is nothing stopping us from applying the same multi-horizon structure to the output of deepOF and deepVOL convolutional modules.

Full details of the multi-horizon architectures can be found in Table 2.I.1 and Table 2.I.2.

Remark 2.2.6. We adopt the same multi-horizon framework as in Zhang and Zohren (2021), where the response $c_{t,t+h_k}$ at horizon $h_k \in \{h_1, \ldots, h_K\}$ corresponds to the return from time t to time $t+h_k$. Alternatively, one might wish to consider as multi-horizon responses the subsequent incremental returns $c_{t+h_{k-1},t+h_k}$, i.e. the returns between time $t + h_{k-1}$ and time $t + h_k$. By choosing evenly spaced h_k's one would obtain more consistent responses across prediction time steps $\{1, \ldots, K\}$ and possibly improve the performance of the multi-horizon models.

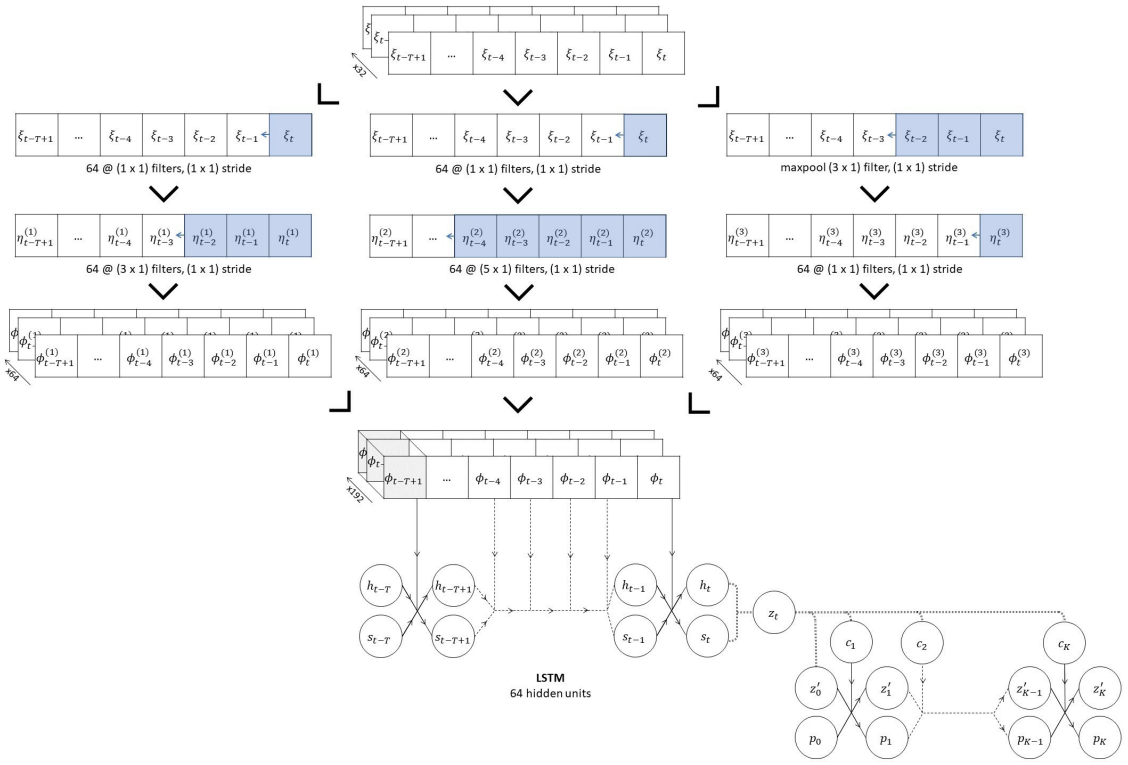


Figure 2.10: Inception module and LSTM layer with seq2seq decoder for multi-horizon forecasting.

2.3 Dataset

In this section, we introduce the dataset used in the experiments presented in Section 2.4. First, we briefly describe how LOBSTER order book data is related to the Nasdaq ITCH feed. Next, we give details of how we process LOBSTER data to obtain features and responses for the experiments. We also provide some descriptive statistics of the data.

We consider the same universe of stocks and trading period as in Kolm, Turiel, and Westray (2021): through LOBSTER (Huang and Polak, 2011) we access one year of open (9:30 EST) to close (16:00 EST) trading data for 115 Nasdaq tickers from January 2, 2019, to January 31, 2020. To produce results in feasible computational time, we select a subset of 10 stocks, trying to preserve a sufficiently varied set of liquidity characteristics. The ten tickers and their liquidity characteristics are summarized in Table 2.1. For full details on how the 10 stocks were selected, see Appendix 2.H. Note that, as previously discussed, on the Nasdaq the tick size is fixed at \$0.01 independently of the value of the stock, meaning the size of the tick relative to the price/spread can have major consequences on market behaviors (Bouchaud et al., 2018). Taking the tick-to-spread ratio as a discriminant we can define small tick stocks as stocks with wide spreads relative to the tick size and large tick stocks as those with narrow spreads relative to the tick (often equal to one tick). In our universe, cf. Table 2.1, we can categorize QRTEA, EXC, AAL, WBA, ATVI as large tick stocks (average spread between 1 and 2 ticks), LILAK, XRAY, PCAR, AAPL as medium tick stocks (average spread between 2 and 3 ticks) and CHTR as a small tick stock (average spread

greater than 3 ticks). Usually large tick stocks display dense "locked" order books with long queues while small tick stocks have sparse order books with many empty price ticks. This observation will be of particular relevance when considering universal models in Section 2.4.4.

Ticker	Updates (000)	Trades (000)	Price Changes (000)	Price (USD)	Spread (bps)	Volume (USD MM)
LILAK	49.89	1.81	3.77	18.09	15.92	2.63
QRTEA	121.32	5.23	2.79	13.54	9.75	9.53
XRAY	83.07	5.41	7.11	52.32	4.35	21.67
CHTR	80.97	6.75	19.21	400.68	6.22	111.71
PCAR	131.38	6.85	13.03	70.63	3.82	34.43
EXC	298.32	7.58	7.21	47.35	2.45	37.55
AAL	398.13	13.15	10.99	30.65	3.88	50.29
WBA	328.5	13.2	15.18	57.81	2.54	75.19
ATVI	423.01	17.97	20.25	49.87	2.81	91.82
AAPL	1137.36	64.19	127.86	215.8	0.99	1156.3

Table 2.1: Selected stocks' characteristics, daily averages.

2.3.1 LOBSTER data

In Section 2.2.1 we described the basic mechanisms governing electronic exchanges such as the Nasdaq. Every day, trading activity on the Nasdaq alone results in hundreds of thousands of order book updates for each stock, cf. Table 2.1. To keep track of all events occurring on the exchange and communicate them efficiently to (subscribed) market participants the Nasdaq uses the TotalView-ITCH protocol. For efficiency, instead of streaming the entire state of the order book after each update, only information on the event changing the order book is sent out to market participants. An example of a (decoded) ITCH message is reported in Table 2.2, note these are timestamped at nanosecond precision. It is up to each market participant to store the current state of the order book and update it each time a new message arrives.

Event ID	Timestamp (sec)	Event Type	Order ID	Size (shares)	Price	Direction
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1312	34714.133632201	deletion	206833312	2516	$11.86	sell
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.2: Example message from a LOBSTER message file. This example corresponds to the order book dynamics shown in Figure 2.1.

The service LOBSTER provides to academic researchers is to reconstruct the historic order book from Nasdaq's Historical TotalView-ITCH data. For each selected ticker and date, LOBSTER returns message and order book files subsampled at a given granularity. In practical terms, a 10-level data granularity yields the set of messages corresponding to order book updates in the

first 10 levels along with the corresponding reconstructed order book 'chopped at 10 levels'. The evolution of the order book determined by the message in Table 2.2 is reported in Table 2.3. More information on the order book reconstruction algorithm used by LOBSTER can be found in Huang and Polak (2011) and on the website `www.lobsterdata.com`.

Event ID	**Timestamp (sec)**	$p_a^{(1)}$	$v_a^{(1)}$	$p_b^{(1)}$	$v_b^{(1)}$	$\cdots$	$p_a^{(10)}$	$v_a^{(10)}$	$p_b^{(10)}$	$v_b^{(10)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1311	34713.685155243	\$11.86	12000	\$11.85	8800	$\cdots$	\$11.95	500	\$11.73	5500
1312	34714.133632201	\$11.86	9484	\$11.85	8800	$\cdots$	\$11.95	500	\$11.73	5500
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 2.3: Example order book update from a LOBSTER order book file. This example corresponds to the order book dynamics shown in Figure 2.1.

2.3.2 Processed data

From the historic LOBSTER order book data we build the features $\mathbf{x}_t, \ldots, \mathbf{x}_{t-T+1}$ and the target responses $c_{t,t+h}$. In the experiments in Section 2.4 we apply the learning framework of Section 2.2.2 with the deep learning models introduced in Section 2.2.3 to these feature-response pairs, using the model confidence set (MCS) procedure for model comparison. As previously discussed in Section 2.2.1 we will be measuring time t (and prediction horizons h) using an order book-driven clock, ticking every time an event occurs on the order book. This clock corresponds to the Event ID in Table 2.2 and Table 2.3. Note we access LOBSTER data up to level 10, thus our order book clock is conditional on the updating event being in the first 10 levels. By construction, the data contains all information on new limit orders, market orders, and cancellations restricted to the first 10 levels. We apply some minor pre-processing steps to the LOBSTER data, summarized in Appendix 2.C.

Features: order book, order flow and volume

We start by discussing how to derive the features $\mathbf{x}_t$ at each time point t from the raw LOBSTER data, i.e. from Table 2.2 and Table 2.3. While it is quite simple to build L1/L2 order book, order flow and volume features, reconstructing L3 volume data requires a bit more work.

The raw order book input $\mathbf{x}_t$ used in deepLOB (Zhang, Zohren, and Roberts, 2019) simply corresponds to the LOBSTER data in Table 2.3, i.e. at $t = 1312$ the 10-level order book feature is given by

$$\mathbf{x}_t = \left(p_{a,t}^{(1)}, v_{a,t}^{(1)}, p_{b,t}^{(1)}, v_{b,t}^{(1)}, \ldots, p_{b,t}^{(10)}, v_{b,t}^{(10)}\right) = (11.86, 9484, 11.85, 8800, \ldots, 11.73, 5500).$$

The features are standardized using a 5-day rolling window.

To compute the order flow input $\mathbf{x}_t$ used in deepOF (Kolm, Turiel, and Westray, 2021) we compute the bid and ask order flows from the LOBSTER data in Table 2.3, i.e. at $t = 1312$ the 10-level order flow feature is given by

$$\mathbf{x}_t = \left(aOF_t^{(1)}, bOF_t^{(1)}, \ldots, bOF_t^{(10)}\right) = (-2516, 0, \ldots, 0).$$

Again, the features are standardized using a 5-day rolling window.

To construct the volume input $\mathbf{x}_t$ at L2 granularity for the deepVOL model described in Section 2.2.4, we select only the volume information from Table 2.3 adding in zeros corresponding to empty price ticks, i.e. at $t = 1312$ the volume feature with window size $W = 10$ is given by

$$\mathbf{x}_t = \left(s_{b,t}^{(10)}, \ldots, s_{b,t}^{(1)}, s_{a,t}^{(1)}, \ldots, s_{a,t}^{(10)}\right) = (1400, \ldots, 8800, 9484, \ldots, 500).$$

Note that, in this example, $s_{b,t}^{(10)} = 1400 \neq 5500 = v_{b,t}^{(10)}$ since some bid price ticks are empty, cf. Figure 2.1. As discussed in Section 2.2.4, volume features are normalized using max-scaling over the whole input array $(\mathbf{x}_t, \ldots, \mathbf{x}_{t-T+1})$.

Finally, to construct the L3 volume features, one needs to work with the message file in Table 2.2 to keep track of the queues. For example, the volume queue at the first ask price, i.e. $\pi_{a,t}^{(1)} = \$11.86$, given at time $t = 1311$ by

$$\left(q_{a,t}^{(1,1)}, q_{a,t}^{(1,2)}, q_{a,t}^{(1,3)}, q_{a,t}^{(1,4)}, q_{a,t}^{(1,5)}\right) = (2516, 2000, 1484, 4500, 1500),$$

is updated at time $t = 1312$ to

$$\left(q_{a,t}^{(1,1)}, q_{a,t}^{(1,2)}, q_{a,t}^{(1,3)}, q_{a,t}^{(1,4)}\right) = (2000, 1484, 4500, 1500),$$

by the deletion in Table 2.2. All other queues are left unchanged. For more details on the complexities of reconstructing volume features from LOBSTER data see Appendix 2.C.1. A deep dive into the distributions of the processed order book, order flow, and volume features is carried out in Appendix 2.D.1.

Responses: categorical mid-price returns

In this chapter, we are interested in answering questions regarding the predictability of market returns. Inevitably, the way returns, i.e. the target responses, are defined has a profound effect on this analysis. Here, in line with the related literature, we treat the mid-price as the "true" price and define returns relative to it, but it is important to note that, by definition, this is not a tradable price. We define the return at horizon h as

$$r_{t,t+h} = \frac{\overline{m}_{t+h}^{(k)} - m_t}{m_t}, \quad \overline{m}_{t+h}^{(k)} = \frac{1}{2k+1} \sum_{i=-k}^{k} m_{t+h+i},$$

where m_t denotes the mid-price at time t and k is a fixed smoothing window. The mid-price m_t is computed from Table 2.3 by $\frac{1}{2}\left(p^{(1)}_{b,t} + p^{(1)}_{a,t}\right)$, i.e. at times $t = 1311$ and $t = 1312$ we have $m_t = \$11.855$. This definition is subject to two possible interpretations. Treating the smoothed mid-price as a de-noised estimate of the true (latent) price h steps ahead, we can understand the return as the percentage change of the true (latent) price relative to the current mid. Alternatively, the return can be understood as the average return one would experience by entering a position at the current mid and exiting it roughly h steps ahead (assuming mid-mid trading). For all the horizons $h \in \{10, 20, 30, 50, 100, 200, 300, 500, 1000\}$ which are considered in Section 2.4 we fix $k = 5$. In Appendix 2.F we discuss other methods for defining mid-price returns and their shortcomings.

Remark 2.3.1. When defining the returns we assume immediate access to the order book. In practice though, hardware and software constraints lead to non-zero time lags when receiving messages and sending orders to the exchange. While, in our setting, such latencies have a negligible impact on the definition of returns, we discuss how their presence could be more precisely accounted for in Appendix 2.G.

All our experiments are carried out in a classification framework where the discretized returns are defined as

$$c_{t,t+h} = \begin{cases} \downarrow, & \text{if } r_{t,t+h} \in (-\infty, -\gamma), \\ =, & \text{if } r_{t,t+h} \in [-\gamma, +\gamma], \\ \uparrow, & \text{if } r_{t,t+h} \in (+\gamma, +\infty), \end{cases}$$

for some $\gamma > 0$. In order to make the three classes roughly symmetric and as balanced[8] as possible, we empirically choose γ from the training set $\mathcal{D}_{\text{train}}$ by

$$\hat{\gamma}_h = \frac{|\hat{Q}_h(0.33)| + \hat{Q}_h(0.66)}{2},$$

where $\hat{Q}_h$ is the empirical quantile function of the training set returns $\{r_{t,t+h}\}_{t \in \mathcal{I}_{\text{train}}}$. As discussed in Section 2.4, we will be splitting our data $\mathcal{D}$ in disjoint windows $\mathcal{D}_w = \mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}} \cup \mathcal{D}_{w,\text{test}}$ for $w = 1, \ldots, W$. The choice of γ will, therefore, be window w and horizon h specific. Descriptive statistics of the target return labels for the first window $w = 1$ are reported in Appendix 2.D.2.

We note that, since stocks trade on a discrete grid of prices determined by the tick size[9] ϑ, also the mid-price m_t evolves on a discrete price (with steps of size $\vartheta/2$). One could thus define the dollar return from t to $t + h$ by the number of (half) ticks the mid-price moves, i.e.

$$R_{t,t+h} = m_{t+h} - m_t \in \{\ldots, -\vartheta, -\vartheta/2, 0, \vartheta/2, \vartheta, \ldots\} \equiv \{\ldots, -2, -1, 0, +1, +2, \ldots\}.$$

This return is by definition discretized, and thus one could directly apply classification models

[8] Note that, especially for short horizons h, many of the returns $r_{t,t+h}$ can be exactly zero, leading to $\hat{Q}_h(0.33) = \hat{Q}_h(0.66) = 0$. When this happens the three classes are unbalanced.

[9] $\vartheta = \$0.01$ for many US listed stocks and all the stocks considered in this work.

(grouping large negative and positive returns to obtain a finite number of classes). A similar approach is used in Sirignano and Cont (2019) when predicting the next change in mid-price. In our work, we consider the estimate for the "true" mid-price to be

$$\overline{m}_{t+h}^{(k)} = \frac{1}{2k+1}\sum_{i=-k}^{k} m_{t+h+i},$$

which lives on a much finer grid than m_{t+h}: over the, possibly quite long, time horizon h multiple changes to the mid-price might occur. In this case the smallest change has little meaning and so we group the dollar returns

$$R_{t,t+h} = \overline{m}_{t+h}^{(k)} - m_t \in \left\{\ldots, -\frac{\vartheta}{2k+1}, -\frac{\vartheta}{4k+2}, 0, \frac{\vartheta}{4k+2}, \frac{\vartheta}{2k+1}, \ldots\right\},$$

into larger classes given by

$$(-\infty, -m_t\gamma), [-m_t\gamma, +m_t\gamma], (+m_t\gamma, +\infty).$$

2.4 Empirical results

In this section, we explore the four questions introduced in Section 2.1.2 and attempt to answer them via the statistical framework provided by model confidence sets (MCS). This inference procedure allows to compare a set of competing models $\mathcal{M}_0$ based on observed loss time series $\{L_{i,w}\}_{w=1}^{W}$ where $L_{i,w}$ denotes the loss of model $i \in \mathcal{M}_0$ at time $w \in W$. To apply the model confidence set procedure to the data described in Section 2.3, we divide the 55 weeks from January 14, 2019 to January 31, 2020 into $W = 11$ five-week periods[10], as represented in Figure 2.11. Each window of data $\mathcal{D}_w$ is divided into a training-validation set $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, the first four weeks, and a test set $\mathcal{D}_{w,\text{test}}$, the fifth week. The joint training-validation set is then further split into training and validation sets by randomly selecting 5 days out of the four weeks for validation. First, the training dataset $\mathcal{D}_{w,\text{train}}$ is used to choose the γ threshold for defining the return labels, as detailed in Section 2.3.2 (note that the choice of γ is specific to the choice of window, horizon, and ticker). Then, the joint training-validation dataset $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$ is used to train the models. For the *unpredictive benchmark* model, this simply means determining the empirical distribution of returns in $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, cf. Appendix 2.D.2. For the deep neural network architectures, this amounts to finding the optimal parameters which minimize the training weighted cross-entropy loss. To do so, we use Adam optimization with validation-based early stopping as described in Appendix 2.A. Once the model has been trained we compute the out-of-sample losses on $\mathcal{D}_{w,\text{test}}$, i.e. for period

[10] Data from the week starting on January 7, 2019, is used to start the rolling window standardization required for deepLOB and deepOF architectures

$w \in \{1, \ldots, 11\}$ and model $i \in \mathcal{M}_0$

$$L_{i,w} = -\text{cce}(\hat{\mathbf{p}}^i_{w,\text{test}}, \mathbf{c}_{w,\text{test}}) = -\frac{1}{|\mathcal{I}_{w,\text{test}}|} \sum_{t \in \mathcal{I}_{w,\text{test}}} \sum_{* \in \{\downarrow, =, \uparrow\}} \mathbb{1}_{\{c_{t,w,\text{test}} = *\}} \log \hat{p}^i_{*,t},$$

is the categorical cross-entropy loss corresponding to the estimated probabilities $\hat{\mathbf{p}}^i_{w,\text{test}}$ when compared with the observations $\mathbf{c}_{w,\text{test}}$. $\hat{\mathbf{p}}^i_{w,\text{test}}$ are the class probabilities produced by model i on the testing set $\mathcal{D}_{w,\text{test}}$ after being trained on the training-validation set $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$. The time series of test losses $\{L_{i,w}\}_{w=1}^{11}$ are then "fed through" the MCS procedure described in Appendix 2.B to obtain the set of MCS p-values $\{p_i^{\text{MCS}}\}_{i \in \mathcal{M}_0}$. The intuitive interpretation of these p-values is: if model $i \in \mathcal{M}_0$ has an MCS p-value lower than a prescribed confidence level then it is deemed statistically inferior to other models in $\mathcal{M}_0$ at that confidence level. This naturally justifies the following definition of order book-driven predictability.

Definition 2.4.1. For $\alpha \in (0, 1)$ we say that there is order book-driven predictability at confidence level $1 - \alpha$ if $p^{\text{MCS}}_{\text{benchmark}} < \alpha$.

In other words, if the *unpredictive benchmark* is deemed to be statistically inferior to some of the other models in $\mathcal{M}_0$ at least one order book-driven model does better than the *unpredictive benchmark*, i.e. such model is predictive.

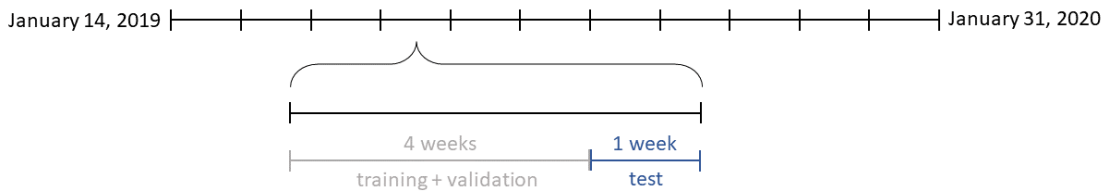


Figure 2.11: Dataset windowing for experiments. We divide the 55 weeks from January 14, 2019 to January 31, 2020 into $W = 11$ five-week periods. In each period, the first four weeks are used for training (a subset of 5 days is randomly selected for holdout validation) and the fifth week is used for testing.

One of the strengths of the way we apply the model confidence set procedure in this context is that it allows to fully account for parameter estimation error, while classic methods for comparing deep learning models do not. To make this statement precise let us make explicit the dependence of the test loss

$$L_{i,w} = -\text{cce}(\mathbf{p}^i(\mathbf{X}_{w,\text{test}}; \hat{\theta}_i(\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}, \epsilon_{i,w})), \mathbf{c}_{w,\text{test}}),$$

on the training-validation data $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, the fitting procedure initialization seed $\epsilon_{i,w}$ and the testing data $\mathcal{D}_{w,\text{test}} = \{\mathbf{X}_{w,\text{test}}, \mathbf{c}_{w,\text{test}}\}$. In the related literature, deep learning models are often compared by fitting the model to the same data $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ starting with different seeds and reporting the out-of-sample average loss and standard error. This accounts only for the uncertainty in ϵ and not in the sampled training and testing datasets (cf. Remark 2.A.1). Our approach instead

naturally considers overall parameter estimation uncertainty, comprised of both statistical and optimization errors, by comparing model losses as functions of the random processes $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, $\epsilon_{w,i}$ and $\mathcal{D}_{w,\text{test}}$. This provides a more robust comparison of model performance, i.e. specification error.

Remark 2.4.2. Note that not all training procedures may converge to the optimal combination of parameters. This is a characteristic of any model learned via numerical optimization methods. We further note that in our experimental setup, no hyperparameter tuning is carried out for any of the models. When using these models in a production setting, one may obtain better results by selecting parameters using cross-validation on the training-validation set. Parameters which one may investigate tuning include:

- Architecture hyperparameters:
  - number of filters in each convolutional layer (we fix 32 channels);
  - number of weighted averages and lengths of averaging windows;
  - number of LSTM hidden nodes (we fix 64 hidden nodes);
  - decoder type for multi-horizon models (we use seq2seq).
- Feature hyper-parameters:
  - number of levels in features (we fix $L = 10$ levels);
  - queue depth in deepVOL(L3) (we fix queue depth $D = 10$);
  - look back window length (we fix $T = 100$);
- Training/Optimization hyper-parameters:
  - Adam learning rate (we fix $\eta = 0.01$);
  - Adam parameters (we fix $\epsilon = 1, \beta_1 = 0.9, \beta_2 = 0.999$);
  - batch size (we fix batch size 256);
  - number of epochs (we fix 50 epochs);
  - early stopping patience (we fix 10 epochs validation patience);
  - training set downscale factor[11] (we subset the training dataset by a factor of 10).

Here we are not interested in obtaining the best fit possible for a specific model but in comparing different models/benchmarks on a level playing field. We thus leave questions related to hyper-parameter tuning for future work.

[11]It is worth noting that in all our experiments, we subset the training dataset by a factor of 10, i.e. subsequent feature-response observations in the dataset are obtained by sliding forward by 10 order book events. This serves multiple purposes: on one hand, it significantly reduces the computational effort required to train the models, on the other, it helps to prevent the risk of overfitting to specific patterns arising in the training dataset. Down-sampling the dataset might also have adverse effects, for example, reducing the available training data of an illiquid stock by a factor of 10 might result in insufficient information to robustly train these data-intensive models.

Once the models are trained, we compare their out-of-sample performance on the test set using the cross-entropy loss, i.e. we have time series $\{L_{i,w}\}_{w=1}^{11}$ for each model $i \in \mathcal{M}_0$ where the model set $\mathcal{M}_0$ will depend on the question/hypothesis investigated. From these we can thus compute the differences in performance $\{d_{ij,w}\}_{w=1}^{11}$ such that $d_{ij,w} = L_{i,w} - L_{j,w}$, which form the basis for the MCS procedure. We will assume the loss difference time series $\{(d_{ij,w})_{i,j\in\mathcal{M}_0}\}_{w=1}^{11}$ to be stationary. We do not conduct multivariate DF or KPSS tests as these are based on asymptotic distributions, and thus, due to the low data regime ($W = 11$), they would not yield statistically robust results.

Remark 2.4.3. The reader with particular attention to detail might argue that losses are computed on different test sample sizes (depending on the number of order book updates occurring in the test set, i.e. $n_{\text{test},w}$) and thus the variance of losses and of their differences may change over time. It is important to note that the sample sizes are random, and there does not seem to be strong evidence against their stationary, so unconditional on sample size, the difference process can still be reasonably assumed to be stationary.

All code is developed in Python with the `tensorflow` library and the `keras` API, with some layers requiring custom `tensorflow` methods. Due to the computationally intensive nature of the experiments, specialized infrastructure was required to store the data (5TB) and train the models (GPUs). All computations were carried out on Imperial College's High-Performance Computing cluster (Imperial College Research Computing Service, 2022), which provides access to several RTX6000 GPUs. The code is available on GitHub at `https://github.com/lorenzolucchese/deepOBs` under the BSD 3-clause license.

The results discussed hereafter are specific to the experimental setup under consideration, i.e. they are specific to the selected stocks, time period, and models. Different experimental setups may lead to different results.

2.4.1 Do high-frequency returns display predictability?

In order to understand whether high-frequency returns display predictability we consider the following set of models:

$$\mathcal{M}_0 = \big\{\text{benchmark},\ \text{deepLOB(L1)},\ \text{deepOF(L1)},\ \text{deepLOB(L2)},\ \text{deepOF(L2)},\ \text{deepVOL(L2)},\ \text{deepVOL(L3)}\big\},$$

where L1 models only use the first level of the order book, while L2 and L3 models use the first 10 levels. For each return horizon $h \in \{10, 20, 30, 50, 100, 200, 300, 500, 1000\}$ and TICKER $\in$ {LILAK, QRTEA, XRAY, CHTR, PCAR, EXC, AAL, WBA, ATVI, AAPL} we determine the MCS p-value of each model in $\mathcal{M}_0$. The MCS p-value for the *unpredictive benchmark* model are reported in Table 2.4. Lighter green shading corresponds to $\alpha = 0.05$ while darker green is $\alpha = 0.01$.

From the results reported in Table 2.4 we see that at high frequencies predictability is systematically present. For most of the stocks under consideration, we were able to identify predictability up to 50 order book events ahead at the 99% confidence level.

Except for LILAK, which is the most illiquid stock, we observe a substantial correlation between

	$h = 10$	$h = 20$	$h = 30$	$h = 50$	$h = 100$	$h = 200$	$h = 300$	$h = 500$	$h = 1000$
LILAK	0.00	0.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00
QRTEA	1.00	0.01	0.01	0.00	0.00	0.00	0.08	0.77	0.54
XRAY	0.00	0.00	0.00	0.01	0.10	0.23	0.48	1.00	0.93
CHTR	0.00	0.00	0.04	0.00	0.42	1.00	1.00	1.00	0.97
PCAR	0.00	0.00	0.00	0.00	1.00	1.00	0.60	1.00	1.00
EXC	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.51	1.00
AAL	0.11	0.00	0.00	0.00	0.00	0.71	1.00	1.00	1.00
WBA	0.00	0.00	0.00	0.00	0.02	0.27	0.13	0.01	1.00
ATVI	0.00	0.00	0.00	0.00	0.00	0.78	1.00	1.00	1.00
AAPL	0.00	0.00	0.00	0.23	1.00	1.00	1.00	1.00	1.00

Table 2.4: MCS p-values of the *unpredictive benchmark* model for the 10 tickers and 9 horizons under consideration. When the p-value is low at least one of the order book-driven models statistically outperforms the *unpredictive benchmark*, i.e. there is order book-driven predictability according to Definition 2.4.1.

the persistence in predictability and the average Updates to Price Changes ratio, cf. Table 2.1. Recalling that in our setting the horizon h is measured in order book updates, this can be interpreted as it being easier for the deep learning models to predict returns that are the result of fewer price changes. One might thus expect to obtain a more consistent maximum predictable horizon across stocks when using a price change-driven clock to measure time.

2.4.2 Which order book representations perform the best?

Having discussed the extent to which the class of models under consideration can identify predictability, it is now natural to ask which of these models performs best. In the MCS framework, this corresponds to determining the specifications which are consistently placed in the set of superior models. We restrict our attention to the horizons and stocks at which predictability is identified and, for each model in consideration, we count the number of times it is identified as a superior model. The results are reported in Table 2.5.

	$\alpha = 0.05$	$\alpha = 0.01$
benchmark	0%	0%
deepLOB(L1)	7%	11%
deepOF(L1)	23%	22%
deepLOB(L2)	5%	11%
deepOF(L2)	88%	89%
deepVOL(L2)	65%	84%
deepVOL(L3)	77%	86%

Table 2.5: % of times the model is in the α-MCS when predictability is identified at the corresponding significance level α.

From the results in Table 2.5 (at the 99% confidence level) we can make the following observations

on the way order book representations influence model performance. When considering deep learning models for short-term return prediction having access to L2 data provides a significant advantage over L1 data. We see that models with L1 data are rarely placed in the set of superior models when predictability is identified. When going from L2 to L3 data instead, the increased granularity does not seem to provide a clear advantage: deepVOL(L2) and deepVOL(L3) display similar performance. Our experiment, therefore, suggests that L3 data might be excessively granular when predicting high-frequency returns from order books when using the deepVOL architecture.

The choice of features used to represent the order book is also crucial when leveraging deep learning methods for return prediction. Using order flow or volume representations provides a significant improvement in performance: the basic deepLOB model ends up being included in the set of best models only 10% of the time and is outperformed even by the model with only first-level (L1) order flow. Volume- and order flow-based models (with L2/L3 data granularity) display comparable performance, being placed in the set of superior models in 85-90% of the predictable horizons.

2.4.3 Can we use a single model across multiple horizons?

In this section, we explore whether using a seq2seq decoder to produce multi-horizon predictions is beneficial. Such models have the clear advantages of having a single set of weights (the network size is only slightly bigger than single horizon networks) and output multiple predictions in very similar run times. This significantly reduces both the memory required to store the models and the time needed to train them. But how do they perform when compared to their single-horizon counterparts?

To answer this question we run the same experiment as in Section 2.4.1 and Section 2.4.2 but enlarge the set of models with the seq2seq specifications. We focus only on prediction horizons $h \in \{10, 20, 30, 50\}$. The results are reported in Table 2.6 and Table 2.7.

From Table 2.7 we note that, for each input type, the seq2seq specifications outperform their single horizon counterparts. We suggest this behavior might be due to the increased availability of information in a multi-horizon setting. For a given input-target pair, when targets are multi-horizon, more information is available on the "order book regime" the inputs should be mapped to. Multi-horizon models can thus learn a more granular map from the input variables to the latent space of "order book regimes", which might be beneficial for producing predictions.

Remark 2.4.4. Note that running the model confidence set procedure again with a larger set of models leads to a counter-intuitive situation where fewer predictable horizons are identified. One would expect that adding models could only increase the number of horizons at which predictability is identified. While this is true in the limit as the number of evaluation periods increases to infinity, it does not hold in the setting where we have access to a finite set of observations[12]. With this observation in mind, we note that results in Table 2.6 are consistent with those in Table 2.4.

[12]This is due to the statistical nature of the MCS procedure: the more models under consideration, the higher the probability that an unpredictive model will outperform the benchmark due to sampling effects.

	$h = 10$	$h = 20$	$h = 30$	$h = 50$
LILAK	0.00	0.00	0.31	0.55
QRTEA	1.00	0.10	0.06	0.00
XRAY	0.00	0.00	0.00	0.00
CHTR	0.00	0.00	0.01	0.01
PCAR	0.00	0.00	0.00	0.00
EXC	0.23	0.00	0.00	0.00
AAL	0.08	0.00	0.00	0.00
WBA	0.00	0.00	0.00	0.00
ATVI	0.01	0.00	0.00	0.00
AAPL	0.00	0.00	0.00	0.07

Table 2.6: MCS p-values of the *unpredictive benchmark* model for the 10 tickers and 9 horizons under consideration when seq2seq models are also considered. When the p-value is low at least one of the order book-driven models statistically outperforms the *unpredictive benchmark*, i.e. there is order book-driven predictability according to Definition 2.4.1.

2.4.4 Can we use a single model across multiple stocks?

In a similar spirit to Sirignano and Cont (2019), we investigate questions regarding the universality of order book dynamics. Intuitively, at a microstructural level, securities which are traded by market participants with similar characteristics may be subject to the same trading patterns, independently of the underlying stock's properties. In Sirignano and Cont (2019), the authors observe that price formation dynamics driven by past order book information, i.e. next mid-price moves, display common patterns across different stocks. In this section, we explore whether similar results hold over longer horizons.

We run the same experiment as in Section 2.4.1 but train the models on multiple stocks simultaneously. Specifically, we split the set of 10 stocks under consideration into two; a first "in-sample" set of stocks given by {QRTEA, CHTR, EXC, WBA, AAPL}, and a second "out-of-sample" set {LILAK, XRAY, PCAR, AAL, ATVI}. For each window w, we use all the training-validation data for the "in-sample" stocks to train (and validate) the models. We then evaluate the trained models on the test data of both the "in-sample" and the "out-of-sample" stocks. The results are reported in Table 2.8.

Focusing on "in-sample" stocks, we note that using universal models leads to results that are partially inconsistent with those in Table 2.4. A possible explanation for this is that equities on the Nasdaq all trade with the same tick size, $\vartheta = \$0.01$, causing some stocks to trade with wider spreads (small tick) and others with smaller spreads (large tick) relative to the tick size. The results of Table 2.8 suggest the universal models are learning the trading dynamics of small tick stocks (as these are more represented in the training data) and hence perform poorly on medium and large tick stocks. These results are intrinsically tied to the observations of Remark 2.4.5.

These observations carry over to the "out-of-sample" stocks. For AAL and ATVI, two large tick stocks, universal models can consistently outperform the *unpredictive benchmark* predictions without ever learning from the stock's past order book data. This provides evidence of the presence

	$\alpha = 0.05$	$\alpha = 0.01$
benchmark	0%	0%
deepLOB(L1)	6%	10%
deepOF(L1)	13%	16%
deepLOB(L2)	6%	13%
deepOF(L2)	63%	71%
deepVOL(L2)	63%	74%
deepVOL(L3)	75%	84%
deepLOB(L1, seq2seq)	13%	26%
deepOF(L1, seq2seq)	16%	16%
deepLOB(L2, seq2seq)	16%	39%
deepOF(L2, seq2seq)	91%	97%
deepVOL(L2, seq2seq)	78%	84%
deepVOL(L3, seq2seq)	81%	87%

Table 2.7: % of times the model is in the α-MCS when predictability is identified at the corresponding significance level α.

of common trading patterns in the order books of different tickers. These result suggest that when developing universal models one should either group stocks based on their microstructural characteristics or appropriately normalize input and target data to account for such heterogeneity, we refer to Briola, Bartolucci, and Aste (2024) for a recent review on the topic.

A natural question to ask is whether the conclusions from Section 2.4.2 carry over to the universal setting. The results in Table 2.9 suggest the superior power of L2 over L1 data is retained for universal models. In this case, though, the increased granularity of L3 data appears to be actually beneficial. Moreover, when considering universal models, the volume representation appears to outperform both order book and order flow inputs. These results thus suggest that some predictive universal patterns can be extracted only from the most granular data. This empirical observation agrees well with the previously suggested explanation that universal models may be picking up trading patterns of large tick stocks where, due to the increased constraint on L1/L2 dynamics, the queue dynamics gain informativeness.

Remark 2.4.5. As discussed in Section 2.3.2, the choice of γ used to define the return classes is stock-specific. This means that up/down labels for different securities may correspond to different numbers of mid-price changes. This is in line with our choice of order book-driven clock, which is by definition also stock-specific. It is important to note that the choices of γ and clock t might not entirely account for structural trading differences of stocks, making it harder to identify universal trading patterns. This contrasts with Sirignano and Cont (2019), where the next mid-price move

	$h = 10$	$h = 20$	$h = 30$	$h = 50$	$h = 100$	$h = 200$	$h = 300$	$h = 500$	$h = 1000$
LILAK	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QRTEA(∗)	0.00	0.00	0.01	0.06	0.00	0.13	0.59	0.34	1.00
XRAY	0.03	1.00	1.00	1.00	1.00	1.00	0.62	0.81	1.00
CHTR(∗)	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00	1.00
PCAR	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EXC(∗)	0.00	0.00	0.00	0.00	0.00	0.39	0.69	1.00	1.00
AAL	0.00	0.00	0.00	0.00	0.00	0.58	0.60	1.00	1.00
WBA(∗)	0.00	0.00	0.00	0.00	0.02	0.58	0.36	0.14	1.00
ATVI	0.00	0.00	0.00	0.00	0.60	0.74	0.88	1.00	1.00
AAPL (∗)	0.07	1.00	1.00	1.00	1.00	0.59	1.00	0.65	1.00

Table 2.8: MCS p-values of the *unpredictive benchmark* model for the 10 tickers and 9 horizons under consideration, universal models only. When the p-value is low at least one of the order book-driven models statistically outperforms the *unpredictive benchmark*, i.e. there is order book-driven predictability according to Definition 2.4.1. (∗) are "in-sample" stocks.

	$\alpha = 0.05$	$\alpha = 0.01$
benchmark	0%	0%
deepLOB(L1, universal)	0%	0%
deepOF(L1, universal)	0%	0%
deepLOB(L2, universal)	8%	24%
deepOF(L2, universal)	42%	62%
deepVOL(L2, universal)	71%	81%
deepVOL(L3, universal)	92%	100%

Table 2.9: % of times each model is in the α-MCS when predictability is identified at the corresponding significance level α.

has a straightforward universal structural interpretation. There are a couple more differences with Sirignano and Cont (2019) we want to point out:

- First, in Sirignano and Cont (2019), the authors consider a much bigger set of stocks, comprised of 500 "in-sample" stocks used for training and 500 "out-of-sample" stocks, making their model more general and thus more universal.

- Second, in that paper, the authors also investigate questions regarding the stationarity of price formation dynamics. This entails using a single long training window: the greater availability of data leads to more stable results. To employ the model confidence set procedure and allow the models to capture patterns specific to different economic regimes, we did not adopt this approach in our work.

- The LSTM model considered in that paper differs from the deepLOB/deepOF/deepVOL architectures. In Sirignano and Cont (2019), the LSTM specification is an online model, i.e. inputs at time t represent the order book state at time t only and are fed through an

LSTM-based architecture updating the stored hidden state and producing output predictions. In our models, instead, there is no storage of hidden units between one prediction and the next, i.e. at each time step t, we input the order book history between $t - T + 1$ and t (in the form of raw order books, order flow or volumes) and, after appropriate convolutional feature extraction, apply an LSTM to the processed sequential data to produce the prediction. When studying dependence on order book history T, Sirignano and Cont (2019) refer to the cut-off horizon used in the backpropagation through time computation of gradients during the training procedure.

2.5 Conclusions and outlook

In this chapter, we explored empirical questions regarding the predictability of mid-price returns driven by order book data in centralized exchanges. The predictability in price formation dynamics, already considered in the sense of next mid-price change in Sirignano and Cont (2019), was found to persist up to 50-300 order book updates ahead, horizons over which multiple mid-price changes may occur. Such predictable horizons might vary from a few milliseconds to nearly half a second, depending on the stock under consideration. These results contrast with low-frequency returns, where predictability is much harder to identify but easier to trade once discovered. In fact, in the high-frequency context, predictability is not always exploitable due to technological limitations and market microstructure issues. A related and more challenging question is, therefore, to understand whether the predictability identified in this chapter is actually tradable.

The experiments were carried out using specific deep learning architectures. Other than identifying predictable horizons, we aimed to answer questions related to the models under consideration. In particular, we found strong empirical evidence for using L2 data over L1 data. But, from the results presented in this chapter, using even more granular order book information (L3) seems to benefit only universal models. The experiments also highlight the importance of carefully choosing a data representation: models based on the basic order book level representation are considerably outperformed by those with order flow or volume inputs. Finally, we found empirical evidence for the presence of universal trading patterns in order book dynamics. Our preliminary results suggest that deep learning architectures may pick up different predictable order book patterns when trained on a pool of stocks instead of a single one. The volume representation has some considerable theoretical advantages, in particular, it is robust to small perturbations, which might explain its superior predictive ability when considering universal models.

Many further theoretical questions regarding predictability which we believe to be of research interest remain unanswered. First, it would be natural to explore whether predictability in returns can be entirely explained by order book structure or if recurring trading patterns play a relevant role. Next, one could compare the predictive performance of the deep learning architectures discussed in

this chapter to that of the models based on carefully engineered features[13] considered in Aït-Sahalia et al. (2022). Finally, only Nasdaq-traded stocks were considered in this work. This is a centralized electronic dealer market on which relatively big companies are listed. It would be interesting to explore whether similar results are obtained for securities traded on exchanges with different market structures. With appropriate experimental setups, we believe the model confidence set procedure used in this chapter provides a solid framework to tackle all these questions.

On the practical side, there are some relevant issues one should consider. First, it is essential to note that in this work we focus on mid-to-mid returns, which are not tradable in practice. When considering specific trading applications, one should thus define the return labels appropriately, for example, working with ask/bid-to-mid returns. Second, in this work, we use an order book-driven clock. In practice, this means the prediction horizon is intrinsically stochastic, which might be a problem in execution. Another critical aspect that would need to be accounted for in practical high-frequency trading is total execution speed. This consists of order book information latency, model prediction run time, and order submission speed. Zhang, Zohren, and Roberts (2019) and Kolm, Turiel, and Westray (2021) argue that their models (which are also considered in this work) are sufficiently fast to be used by traders with good connections: see Appendix 2.G for a discussion on the effects of infrastructure latency.

As with any trading application, one must also account for the impact his own trade will have on the order book. Moreover, if many traders exploit the same inefficiencies, this might erode all predictive power above the lowest tradable latencies. Overall we believe that while the predictability identified in this work might not be directly tradable through a standalone strategy, it might still help some market players, such as market makers, to gauge the direction of the market and adjust their quotes accordingly.

[13] A linear regression on volume imbalance (a signal of market pressure) provides a good basic model for short-term price changes. Understanding whether the deep learning models considered in this work are able to outperform or identify alternative predictable patterns when compared to such economically motivated features is an interesting avenue for future work.

Appendix

2.A Deep learning

2.A.1 Deep neural network architectures

When approximating a function[14] f from a set of models $\{f_\theta\}_{\theta\in\Theta}$, such as in the return prediction task introduced in Section 2.2.2, a quite common approach is to consider a parametric family of deep neural networks. Theoretically motivated by the universality of such models, this approach does not rely on the functional form of f being correctly specified.

The main idea behind deep neural networks is to lift the input into a higher dimensional space to extract relevant latent features before projecting onto output space. Mathematically a neural network consists of a composition of functions, known as layers, $f_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ for $l = 1, \dots, L$ where d_0 is the dimensionality of input space and d_L is the dimensionality of the output space, for example $d_L = 3$ in our three-class classification task. A neural network is considered deep if the number of layers L is large.

In the simplest case the layers f_l's are "activated" affine transformations, i.e. for input $z_{l-1} \in \mathbb{R}^{d_{l-1}}$

$$f_l(z_{l-1}) = \sigma_l(W_l z_{l-1} + b_l),$$

where $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $b_l \in \mathbb{R}^{d_l}$ are learnable parameters, also known as weights, and σ_l is a non-linear activation function applied element-wise. Neural networks of the form $f = f_L \circ \cdots \circ f_1$ such that each f_l is of this form have been shown to be universal for the class of continuous functions (Cybenko, 1989; Hornik, Stinchcombe, and White, 1990), i.e. any continuous function (on a bounded domain) can be approximated arbitrarily well by a large enough network in the supremum norm[15]. Related works have shown the representational benefits of depth: there are functions that deep networks can construct with polynomially many parameters, which instead require exponentially many parameters when considering shallow networks (Telgarsky, 2015). More complex and domain-specific forms of layers f_l exist; we briefly introduce convolutional and recurrent/LSTM layers.

Convolutional layers

Convolutional layers are a natural choice when the input is an image, i.e.

$$z_{l-1} \in \mathbb{R}^{d_{l-1}} = \mathbb{R}^{h_{l-1} \times w_{l-1} \times c_{l-1}},$$

[14] This notation covers both regression $f(\cdot) = g(\cdot)$ and classification $f(\cdot) = (p_\downarrow(\cdot), p_=(\cdot), p_\uparrow(\cdot))$ tasks.
[15] It follows directly that the class of neural networks is dense in $L^p(\mathcal{X})$ space, for bounded $\mathcal{X} \subseteq \mathbb{R}^d$.

where h_{l-1} is the height in pixels, w_{l-1} is the width in pixels and c_{l-1} is the number of channels, for example $c_{l-1} = 3$ if the image is in RGB. A convolutional layer is a specific case of a generic dense layer with parameter restrictions to account for adjacencies in the input's structure. Weights are constrained to be a family kernels, each aiming to extract local information. More specifically, a convolutional layer consists of k_l weight kernels, also known as filters, which are convoluted with the input image to produce the output. Mathematically a 2-dimensional convolution with k_l $(n_l \times m_l)$ filters and stride $(s_l \times t_l)$ is described by

$$[f_l(z_{l-1})]_{i,j,k} = \sum_{n=1}^{n_l} \sum_{m=1}^{m_l} \sum_{c=1}^{c_{l-1}} \left\{ [W_l^{(k)}]_{n,m,c} [z_{l-1}]_{s_l(i-1)+n, t_l(j-1)+m, c} + b_l^{(k)} \right\},$$

where $W_l^{(k)} \in \mathbb{R}^{n_l \times m_l \times c_{l-1}}$, for $k = 1, \ldots, k_l$, are the k_l filters, $b_l^{(k)} \in \mathbb{R}$ are bias terms and $[\cdot]_{i,j,k}$ denotes the i, j, k-th entry of a three dimensional tensor. The output then lives in $\mathbb{R}^{h_l \times w_l \times k_l}$, where $h_l = (h_{l-1} - n_l)/s_l + 1$ and $w_l = (w_{l-1} - m_l)/t_l + 1$, assuming that $s_l | (h_{l-1} - n_l)$ and $t_l | (w_{l-1} - m_l)$. While this equation might look a bit daunting, convolutional layers are often easily understood via a graphical depiction, as in Figure 2.A.1. Many empirical studies (Krizhevsky, Sutskever, and Hinton, 2012; He et al., 2016) have investigated the ability of convolutional layers to extract relevant features from input with grid-like topologies. In Section 2.2.3, we discuss how this ability might be leveraged to extract relevant features from order books.

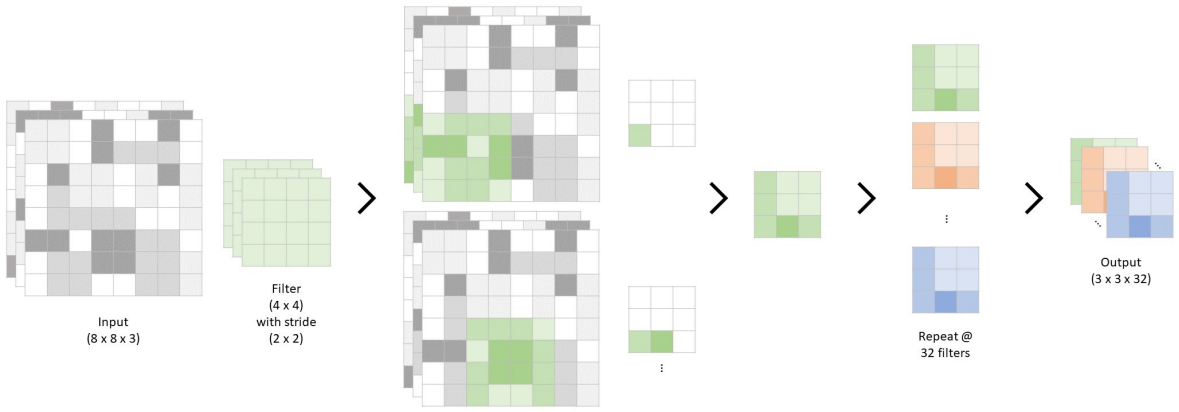


Figure 2.A.1: Convolutional layer, input $z_{l-1} \in \mathbb{R}^{8 \times 8 \times 3}$, filter size $(n_l \times m_l) = (4 \times 4)$, stride $(s_l \times t_l) = (2 \times 2)$ and number of filters $k_l = 32$.

Recurrent/LSTM layers

When the input has a built-in temporal structure, this can be quite naturally accounted for by recurrent layers. This type of layer retains information over time, discovering temporal dependencies in the data. Recurrent layers may be used with streaming data to obtain online predictions or applied to a whole time series yielding a single output. We focus on the latter case, i.e. assuming

the input $z_{l-1} \in \mathbb{R}^{d_{l-1}}$ has a temporal structure of the form

$$z_{l-1} = \left(z_{l-1}^{(1)}, \ldots, z_{l-1}^{(T_{l-1})}\right)^{\mathrm{T}} \in \mathbb{R}^{T_{l-1} \times n_{l-1}},$$

the recurrent layer $z_l = f_l(z_{l-1})$ is given by

$$h_{l-1}^{(0)} \in \mathbb{R}^{m_{l-1}}, \quad h_{l-1}^{(t)} = \phi_l\left(h_{l-1}^{(t-1)}, z_{l-1}^{(t)}\right) \text{ for } t = 1, \ldots, T_{l-1}, \quad z_l = h_{l-1}^{(T_{l-1})},$$

where $\phi_l : \mathbb{R}^{m_{l-1}} \times \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{m_{l-1}}$ is a parameterized recurrent function and the $h_{l-1}^{(t)}$'s are known as hidden states. The most widely used type of recurrent layer is Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). While other types of recurrent layers may suffer from vanishing or exploding gradients when training, the specific structure of LSTMs largely prevents such problems (Hochreiter et al., 2001). In an LSTM layer each hidden state h is augmented with a memory state s, and hence the recurrence becomes

$$\left(h_{l-1}^{(t)}, s_{l-1}^{(t)}\right) = \phi_l\left(h_{l-1}^{(t-1)}, s_{l-1}^{(t-1)}, z_{l-1}^{(t)}\right).$$

The equations governing the LSTM recurrence are given by

$$\begin{aligned} h_{l-1}^{(t)} &= o_{l-1}^{(t)} \odot \tanh(s_{l-1}^{(t)}), \\ s_{l-1}^{(t)} &= g_{l-1}^{(t)} \odot \sigma\left(b_l + U_l z_{l-1}^{(t)} + W_l h_{l-1}^{(t-1)}\right) + f_{l-1}^{(t)} \odot s_{l-1}^{(t-1)}, \end{aligned}$$

where the input, forget and output gates depend on the context, i.e. on $h_{l-1}^{(t-1)}, z_{l-1}^{(t)}$, via:

$$\begin{aligned} g_{l-1}^{(t)} &= \sigma\left(b_l^g + U_l^g z_{l-1}^{(t)} + W_l^g h_{l-1}^{(t-1)}\right), \\ f_{l-1}^{(t)} &= \sigma\left(b_l^f + U_l^f z_{l-1}^{(t)} + W_l^f h_{l-1}^{(t-1)}\right), \\ o_{l-1}^{(t)} &= \sigma\left(b_l^o + U_l^o z_{l-1}^{(t)} + W_l^o h_{l-1}^{(t-1)}\right), \end{aligned}$$

for sigmoid activation function σ, element-wise product $\odot$ and parameters $b_l, b_l^g, b_l^f, b_l^o \in \mathbb{R}^{m_{l-1}}$, $U_l, U_l^g, U_l^f, U_l^o \in \mathbb{R}^{m_{l-1} \times n_{l-1}}$ and $W_l, W_l^g, W_l^f, W_l^o \in \mathbb{R}^{m_{l-1} \times m_{l-1}}$. The network diagram of an LSTM layer is given in Figure 2.A.2. LSTMs summarize relevant context information in the memory cell $s_{l-1}^{(t)}$, performing exceptionally well on tasks with a strong temporal dependency (Melis, Dyer, and Blunsom, 2017).

2.A.2 Training deep neural networks

As discussed in Section 2.2.2, once we have specified a parametric model for approximating the prediction function $f \approx f_\theta$ we wish to learn the best approximation to f from a training set of observed input-output data $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i \in \mathcal{I}_{\text{train}}}$. To do so, we aim to find the parameter

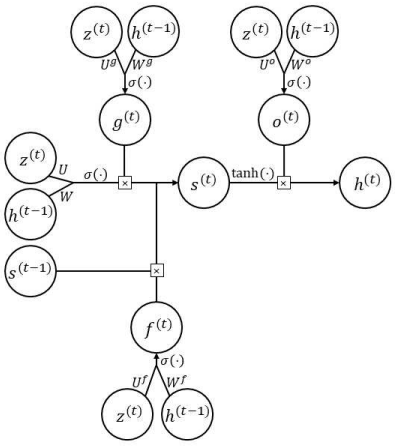


Figure 2.A.2: LSTM layer network diagram.

combination θ which yields predictions closest to the observations:

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, L(\theta|\mathcal{D}_{\text{train}}) = \underset{\theta\in\Theta}{\operatorname{argmin}}\Big\{\frac{1}{|\mathcal{I}_{\text{train}}|}\sum_{i\in\mathcal{I}_{\text{train}}}\mathcal{L}(f_\theta(x_i), y_i)\Big\},$$

where $\mathcal{L}(\cdot,\cdot)$ is a loss function quantifying the distance between the prediction and the observed value, and θ contains all network parameters (weights matrices, bias terms, CNN filters...). The form of $\mathcal{L}$ may be chosen using probabilistic arguments or according to some other task-specific criteria.

When the model f_θ is a deep neural network the loss landscape $L(\theta) = L(\theta|\mathcal{D}_{\text{train}})$ may be very complex and finding a "good" minimizer is often a difficult task. The most widely used approach is to use (some variant of) gradient descent in which the model parameters are iteratively updated by

$$\theta^{(n+1)} = \theta^{(n)} - \eta\nabla_\theta L(\theta^{(n)}),$$

until a convergence criterion is met. Here $\eta > 0$ is a fixed learning rate, and the weights $\theta^{(0)}$ are chosen according to an appropriate initialization rule. When considering deep neural network architectures, the number of parameters is often very large, and therefore second-order methods are often intractable or computationally infeasible. Many variants of the first order gradient descent algorithm exist though. For example, the stochastic gradient descent algorithm updates parameter values based on estimates of the gradient $\nabla_\theta L(\theta^{(n)})$ computed from a random subset (known as a batch) of the training dataset. More advanced first-order optimizers consider momentum and/or use adaptive learning rates. In our experiments, we use the Adam optimizer (Kingma and Ba, 2017) which updates the parameters based on adaptive estimates of first and second-order moments:

$$m^{(n+1)} = \frac{1}{1-\beta_1^n}\Big\{\beta_1 m^{(n)} + (1-\beta_1)\nabla_\theta L(\theta^{(n)})\Big\},$$
$$v^{(n+1)} = \frac{1}{1-\beta_2^n}\Big\{\beta_2 v^{(n)} + (1-\beta_2)[\nabla_\theta L(\theta^{(n)})]^2\Big\},$$

$$\theta^{(n+1)} = \theta^{(n)} - \eta \frac{m^{(n+1)}}{\sqrt{v^{(n+1)}} + \epsilon},$$

for parameters $\eta, \beta_1, \beta_2, \epsilon > 0$.

Remark 2.A.1. Note that, contrary to other model specifications, such as linear models, the fitting procedure of deep neural networks is not closed form. This adds a further layer of uncertainty. Any model is subject to the following two sources of uncertainty: one due to model specification (i.e. $f \approx f_\theta$) and one due to parameter estimation uncertainty (i.e. statistical error arising from estimating θ from a finite sample). Models that do not have closed forms for $\hat{\theta}$ from observed data are further subject to numerical optimization error in the parameter estimation phase. Sometimes, the optimization procedure may not even converge to a (global) minimum. With deep neural networks, we thus trade off smaller model uncertainty (at least theoretically due to the universality property) for larger parameter estimation error.

When the task is a classification problem f_θ is usually chosen so that it maps to $\{p \in [0,1]^C : \|p\|_1 = 1\}$, where C is the number of classes in the set $\mathcal{C}$. This allows to give a probabilistic interpretation of the deep neural network outputs, i.e.

$$f_\theta(x) = \{p_{c,\theta}(x)\}_{c \in \mathcal{C}}, \text{ where } \; p_{c,\theta}(x) = \mathbb{P}(y = c | x, \theta).$$

In this case, a natural choice of loss function to use during training is the cross-entropy loss, which can be understood by maximum likelihood arguments: assuming the responses are independent given the features and the distribution of the features is independent of the parameters θ the negative log-likelihood is

$$L(\theta | \mathcal{D}_{\text{train}}) = -\log \mathbb{P}(\mathcal{D}_{\text{train}} | \theta) \propto -\frac{1}{|\mathcal{I}_{\text{train}}|} \sum_{i \in \mathcal{I}_{\text{train}}} \sum_{c \in \mathcal{C}} \mathbb{1}_{\{y_i = c\}} \log p_{c,\theta}(x_i).$$

When training samples for the classes are unbalanced, say class $\bar{c} \in \mathcal{C}$ has many more samples than all other classes, the optimization algorithm tends to get stuck in the trivial minimum given by the Dirac measure at $\bar{c} \in \mathcal{C}$. One way to mitigate this effect is to re-weight the categorical cross entropy loss such that the network is less incentivized to move towards the trivial minimum. With appropriate weighting, $\bar{c} \in \mathcal{C}$ observations influence the direction of the gradient less, thus reducing the strength of the attraction towards the trivial minimum. Mathematically one uses the following weighted categorical cross-entropy loss

$$L(\theta | \mathcal{D}_{\text{train}}) = -\frac{1}{|\mathcal{I}_{\text{train}}|} \sum_{i \in \mathcal{I}_{\text{train}}} \sum_{c \in \mathcal{C}} w_c \mathbb{1}_{\{y_i = c\}} \log p_{c,\theta}(x_i),$$

where the weights w_c are chosen to be inversely proportional to the class's training set density. For example, in our experiments, we set:

$$w_c = w_c(\mathcal{D}_{\text{train}}) = \frac{|\mathcal{I}_{\text{train}}|}{\#\{c_i \in \mathcal{D}_{\text{train}} : c_i = c\}}.$$

2.B Model confidence sets

The model confidence set procedure introduced by Hansen, Lunde, and Nason (2011) provides a formalized statistical inference framework in which to compare competing models. It does not assume that any of the models is the true data generating process but simply aims to identify a set of models that will contain the best model with a given level of confidence, known as model confidence set, MCS. A model confidence set is, therefore, analogous to a confidence interval for a parameter. To identify the best models, one must compare the model outputs (either as estimated class probabilities or as forecasts) by selecting an appropriate loss/score function.

Let us denote by $\mathcal{M}_0$ the set of models under consideration and by $\{L_{i,w}\}_{w=1}^{W}$ for $i \in \mathcal{M}_0$ the time series of losses. In the MCS framework for models $i, j \in \mathcal{M}_0$ and window $w \in \{1, \dots, W\}$ we define:

$$d_{ij,w} = L_{i,w} - L_{j,w}.$$

Under a stationarity assumption on the $(d_{ij,w})_{w \geq 1}$, we define the set of best models among those in $\mathcal{M}_0$ as

$$\mathcal{M}^* = \{i \in \mathcal{M}_0 : \mathbb{E}[d_{ij}] \leq 0 \ \forall j \in \mathcal{M}_0\}.$$

This is the set we want to identify by using the MCS procedure. The MCS algorithm is then defined as follows.

Definition 2.B.1. Let $\delta_{\mathcal{M}}$ and $e_{\mathcal{M}}$ be an equivalence test and an elimination rule.

Step 0. Initially set $\mathcal{M} = \mathcal{M}_0$.

Step 1. Test $H_{0,\mathcal{M}} : \mathbb{E}[d_{ij}] = 0, \ \forall i, j \in \mathcal{M}$ using $\delta_{\mathcal{M}}$ at level α.

Step 2. If $H_{0,\mathcal{M}}$ is accepted, set $\hat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$; otherwise, use $e_{\mathcal{M}}$ to eliminate an object from $\mathcal{M}$ and repeat the procedure from Step 1.

Under appropriate assumptions on the equivalence test and elimination rule[16], the model confidence set $\hat{\mathcal{M}}^*_{1-\alpha}$ has the following asymptotic properties:

(i) $\liminf_{W \to \infty} \mathbb{P}(\mathcal{M}^* \subset \hat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$,

(ii) $\lim_{W \to \infty} \mathbb{P}(i \in \hat{\mathcal{M}}^*_{1-\alpha}) = 0$ for $i \notin \mathcal{M}^*$,

where $W \to \infty$ denotes the number of sample periods. If furthermore, the equivalence test and elimination rule satisfy a coherency condition[17], we have the following finite sample property:

[16] We require the equivalence test $\delta_{\mathcal{M}}$ to have asymptotic level α and asymptotic power 1, while the elimination rule $e_{\mathcal{M}}$ needs to satisfy $\lim_{W \to 0} \mathbb{P}(e_{\mathcal{M}} \in \mathcal{M}^* | H_{1,\mathcal{M}}) = 0$.

[17] The required coherency condition is $\mathbb{P}(\delta_{\mathcal{M}} \text{ rejects } H_{0,\mathcal{M}}, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$.

(iii) $\mathbb{P}(\mathcal{M}^* \subset \hat{\mathcal{M}}^*_{1-\alpha}) \geq 1-\alpha$.

In Hansen, Lunde, and Nason (2011), the authors give multiple practical examples of equivalence tests and elimination rules. We focus on implementing the MCS procedure with tests constructed from t-statistics and bootstrap estimators. This choice of equivalence test and elimination rule has two practical advantages: it does not require estimating a variance-covariance matrix for the time series $\{(d_{ij,w})_{i,j\in\mathcal{M}_0}, w \geq 1\}$ which might be difficult when $|\mathcal{M}_0| \approx W$ and it satisfies the coherency condition required for the finite sample property (iii) of the MCS procedure. Let us assume the following holds:

Assumption 2.B.2. For some $r > 2$ and $\gamma > 0$, it holds that that $\{(d_{ij,w})_{i,j\in\mathcal{M}_0}, w \geq 1\}$ is strictly stationary, α-mixing of order $-r/(r-2)$ and $\mathrm{Var}(d_{ij,w}) > 0$, $\mathbb{E}|d_{ij,w}|^{r+\gamma} < \infty$ for all $i, j \in \mathcal{M}_0$.

For a set of models $\mathcal{M} \subset \mathcal{M}_0$ with time series of relative performances $\{(d_{ij,w})_{i,j\in\mathcal{M}}, w \geq 1\}$ define the following quantities $\forall i, j \in \mathcal{M}$:

$$\bar{d}_{ij} = \frac{1}{W}\sum_{w=1}^{W} d_{ij,w}, \quad \bar{d}_{i\cdot} = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}} \bar{d}_{ij},$$

respectively measuring the sample relative performance of the models i and j and the sample relative performance of model i and all other models in $\mathcal{M}$. We can then introduce the following t-statistics $\forall i \in \mathcal{M}$:

$$t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d}_{i\cdot})}},$$

where $\widehat{\mathrm{Var}}(\bar{d}_{i\cdot})$ are appropriate estimators of the variances of the $\bar{d}_{i\cdot}$'s. The associated test statistic is $T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t_{i\cdot}$. The equivalence test and elimination rule are given by:

($\delta_{\mathcal{M}}$) Reject $H_{0,\mathcal{M}}$ if $T_{\max,\mathcal{M}} > F_{\varrho}^{-1}(1-\alpha)$ where $T_{\max} \sim F_{\varrho}$.

($e_{\mathcal{M}}$) Eliminate $\mathrm{argmax}_{i\in\mathcal{M}} t_{i\cdot}$.

The asymptotic distribution of the test statistic $T_{\max,\mathcal{M}}$ is nonstandard because it depends on the (usually unknown) nuisance parameter ϱ under the null. The value of ϱ is not strictly necessary since we can consistently estimate the distribution of $T_{\max,\mathcal{M}}$ by using a bootstrapping procedure. In this case, the equivalence test compares the observed value $T_{\max,\mathcal{M}}$ with the bootstrap empirical quantile. The resulting procedure preserves the asymptotic[18] properties (i) and (ii) and the finite sample property (iii) under Assumption 2.B.2. For details of the bootstrapping procedure see the Appendix of Hansen, Lunde, and Nason (2011).

Remark 2.B.3. The test based on the t-statistic discussed above relies on the fact that the null hypothesis $H_{0,\mathcal{M}}$ can be equivalently rewritten as

$$H_{0,\mathcal{M}} : \mathbb{E}[d_{i\cdot}] = 0, \ \forall i \in \mathcal{M}, \quad \text{where} \quad d_{i\cdot} = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}} d_{ij}.$$

[18] As both the sample size W and the number of bootstrap samples B increase.

In this chapter, we present results using MCS p-values, which are defined as follows.

Definition 2.B.4. Let $(\delta_{\mathcal{M}}, e_{\mathcal{M}})$ be the equivalence test and elimination rule associated with a MCS procedure as defined in Definition 2.B.1. The elimination rule defines a decreasing sequence of random sets $\mathcal{M}_0 \supset \mathcal{M}_1 \supset \ldots \supset \mathcal{M}_{|\mathcal{M}_0|}$ by successively eliminating models $e_{\mathcal{M}_0}, \ldots, e_{|\mathcal{M}_0|}$. Let $p_{H_{0,\mathcal{M}}}$ denote the p-value associated with the null hypothesis $H_{0,\mathcal{M}}$ under the equivalence test $\delta_{\mathcal{M}}$ with the convention that $p_{H_{0,\mathcal{M}_{|\mathcal{M}_0|}}} \equiv 1$. Then for model $i = e_{\mathcal{M}_k} \in \mathcal{M}_0$ the MCS p-value is defined as

$$p_i^{\text{MCS}} = \max_{k' \leq k} \; p_{H_{0,\mathcal{M}_{k'}}}.$$

Note that the definition of MCS p-value is such that for $i \in \mathcal{M}_0$ $p_i^{\text{MCS}} \geq \alpha \iff i \in \hat{\mathcal{M}}^*_{1-\alpha}$.

Remark 2.B.5. The MCS procedure is quite flexible. It allows to compare models between each other and to compare them to an *unpredictive benchmark* model (by including the *unpredictive benchmark* in $\mathcal{M}_0$). When only two models are considered, i.e. $|\mathcal{M}_0| = 2$, the MCS procedure reduces to testing the null hypothesis $H_0 : \mathbb{E}[d_{i_1 i_2}] = 0$.

The model confidence set procedure relies on the assumption that the loss differences time series $\{(d_{ij,w})_{i,j \in \mathcal{M}_0},\, w \geq 1\}$ is stationary and α-mixing. This means that even when the loss process $\{(L_{i,w})_{i \in \mathcal{M}_0},\, w \geq 1\}$ is non-stationary, e.g. model performance is tied to regime changes, the model confidence set procedure might still be applicable. In Section 2.4, we assume the time series $\{(d_{ij,w})_{i,j \in \mathcal{M}_0},\, w \geq 1\}$ to be stationary based on intuitive arguments, but we discuss some statistical checks which might be carried out to provide further evidence in this context. First of all, it is worth noting that general tests for stationarity are not available: most tests require assuming a parametric form on the time series. The simplest example would be to impose a linear structure on the time series, i.e. assume the difference time series $\{\boldsymbol{d}_w,\, w \geq 1\} = \{(d_{ij,w})_{i,j \in \mathcal{M}_0},\, w \geq 1\}$ follows a Vector Auto-Regressive process of order 1:

$$\boldsymbol{d}_w = \Gamma \boldsymbol{d}_{w-1} + \boldsymbol{\eta}_w,$$

where $\boldsymbol{\eta}_w \sim N(0, \Sigma_\eta)$ are serially uncorrelated. In this setting, one can consider various tests to determine whether the process is stationary. For example, one may apply multivariate homogeneous Dickey-Fuller tests (Harvey and Bates, 2003) to test the null of a unit root against the alternative of stationarity or a multivariate KPSS test to test (short-memory) stationarity as the null and the presence of a unit root as the alternative (Nyblom and Harvey, 2000).

2.C Data processing

From LOBSTER (Huang and Polak, 2011) we obtain the order book and message files introduced in Section 2.3.1. These files are jointly processed to obtain the desired order book features, i.e. raw order books, order flow and volumes, and responses as described in Section 2.3.2. Before extracting the feature-response pairs the data is treated in the following way:

1. order book states with crossed quotes are removed;
2. states occurring at the same time stamp (at nanosecond precision) are collapsed onto the last state, e.g. an aggressive order executing against multiple resting limit orders;
3. the first and last 10 minutes of market activity are dropped;
4. the smoothed returns at the requested horizons – as defined in Section 2.3.2 – are computed and matched to the corresponding features.

The data provided by LOBSTER is easily accessible and does not exhibit significant anomalies. The only data inconsistencies we found were on 2019-11-05 for the WBA ticker after the trading halt occurring between 13:32:50.233001415 and 13:37:50.2335639.

2.C.1 Volume data processing

LOBSTER provides level-based data in a form that can be easily manipulated to derive order book and order flow features (we access the first $L = 10$ levels). Deriving the volume features from LOBSTER data instead is a non-trivial task. If we set a window $W > L$ for the volume representation, some levels past $L = 10$ might have prices that fall in the volume window, i.e. $p_{x,t}^{(L+1)} \in \{\pi_{x,t}^{(1)}, \dots, \pi_{x,t}^{(W)}\}$, and we won't have the required information. To avoid this issue in all our experiments we set $W = L = 10$. Reconstruction of L3 queue data has a further complication: every time a quoting price leaves the 10-level range, all orders submitted/canceled at this price are not recorded. Therefore there is no way of reconstructing what happens at such a price level while it is outside of the 10-level range. To minimize the effect of this shortcoming, we adopt the following systematic approach: each time a quoting price (re-)enters the 10-level range, we initialize the corresponding queue with a single order of size given by the total liquidity available at that price level (even if it may be made up of multiple independent orders).

These issues could be mitigated by accessing data deeper in the order book from LOBSTER, say $L = 50$ or 100, while keeping a relatively small window $W << L$.

2.D Dataset descriptive statistics: ATVI deep dive

In this section, we report some descriptive statistics of the processed data used for the experiments. We focus on one representative stock, Activision Blizzard Inc. (ATVI). First, we explore the distribution of the order book, order flow, and volume features. For each level/price tick, we report the aggregated distributional boxplot. Moreover, in order to understand the temporal behavior of the data, we fix one specific level/price tick and look at how the distribution varies day by day. Finally, we analyze the distribution of the target categorical returns. To do so we fix the first window in the experiment procedure, i.e. $w = 1$ in Section 2.4, and plot the unconditional train, validation, and test distributions at each prediction horizon $h \in \{10, 20, 30, 50, 100, 200, 300, 500, 1000\}$.

2.D.1 Features

In Figures 2.D.1–2.D.5 we plot the aggregate and temporal distributions of the features $\mathbf{x}_t$ described in Section 2.3.2. We note the following:

- As expected, order book prices display significant non-stationarity, cf. Figure 2.D.1.
- The difference between order book volumes $v_{t,x}^{(l)}$, Figure 2.D.2, and volumes $s_{t,x}^{(w)}$, Figure 2.D.4, is that in the former we measure the distance from the mid in levels, i.e. the volume is strictly positive, while in the latter we use all, i.e. also possibly empty, price ticks.
- In Figure 2.D.3 we note that order flow is zero most of the time. By definition, we have non-zero flow only when an event affects exactly that level or the whole order book shifts.
- To provide some insight on the distribution of L3 volume features, we plot queue depth statistics in Figure 2.D.5. We note that most queue depths are less than 9 orders long, this justifies the queue depth cutoff of length 10 chosen in our experiments.
- Finally, looking at Figure 2.D.2, Figure 2.D.4 and Figure 2.D.5, one can clearly note the difference between first level/price tick dynamics and those of the rest of the order book: active price discovery leads to lower volumes and shorter queues.

2.D.2 Responses

In Figure 2.D.6 we plot the unconditional empirical distributions of the target responses defined in Section 2.3.2 for the first window $w = 1$. Note that the threshold $\hat{\gamma}_h$ is based on the training dataset only. As detailed in Section 2.4, the train and validation sets are randomly partitioned from a four week period, i.e. $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, and thus are likely to display similar volatility regimes. On the other hand, the test set $\mathcal{D}_{w,\text{test}}$ corresponds to the following week, cf. Figure 2.11, and thus may exhibit a different volatility profile. By joining the train and validation distribution we obtain the *unpredictive benchmark* described in Remark 2.2.1 and used to define predictability in the experiments in Section 2.4.

In Figure 2.D.7 we explore the dependence structure of the return labels by plotting[19] the previous h-step return $c_{t-h,t}$ against the next h-step return $c_{t,t+h}$. We note that for all horizons $h \in \{10, 20, 30, 50, 100, 200, 300, 500, 1000\}$ the null hypothesis of independence is rejected by a contingency chi-square test. The fact that return labels display some form of persistency has a two-fold interpretation. On one hand, it provides evidence towards the existence of predictability in the data, i.e. against the i.i.d. Efficient Market Hypothesis on which the *unpredictive benchmark* is based, cf. Remark 2.2.1. On the other, it suggests that a simple one-step ahead Markovian model of returns may display predictive power, i.e. a model which predicts the next price move

[19] To be more precise, we plot $c_{t-h-k,t}$ against $c_{t,t+h}$ in order to avoid any kind of overlap between the labels, cf. Section 2.3.2. We drop the k for clarity of exposition and notational convenience.

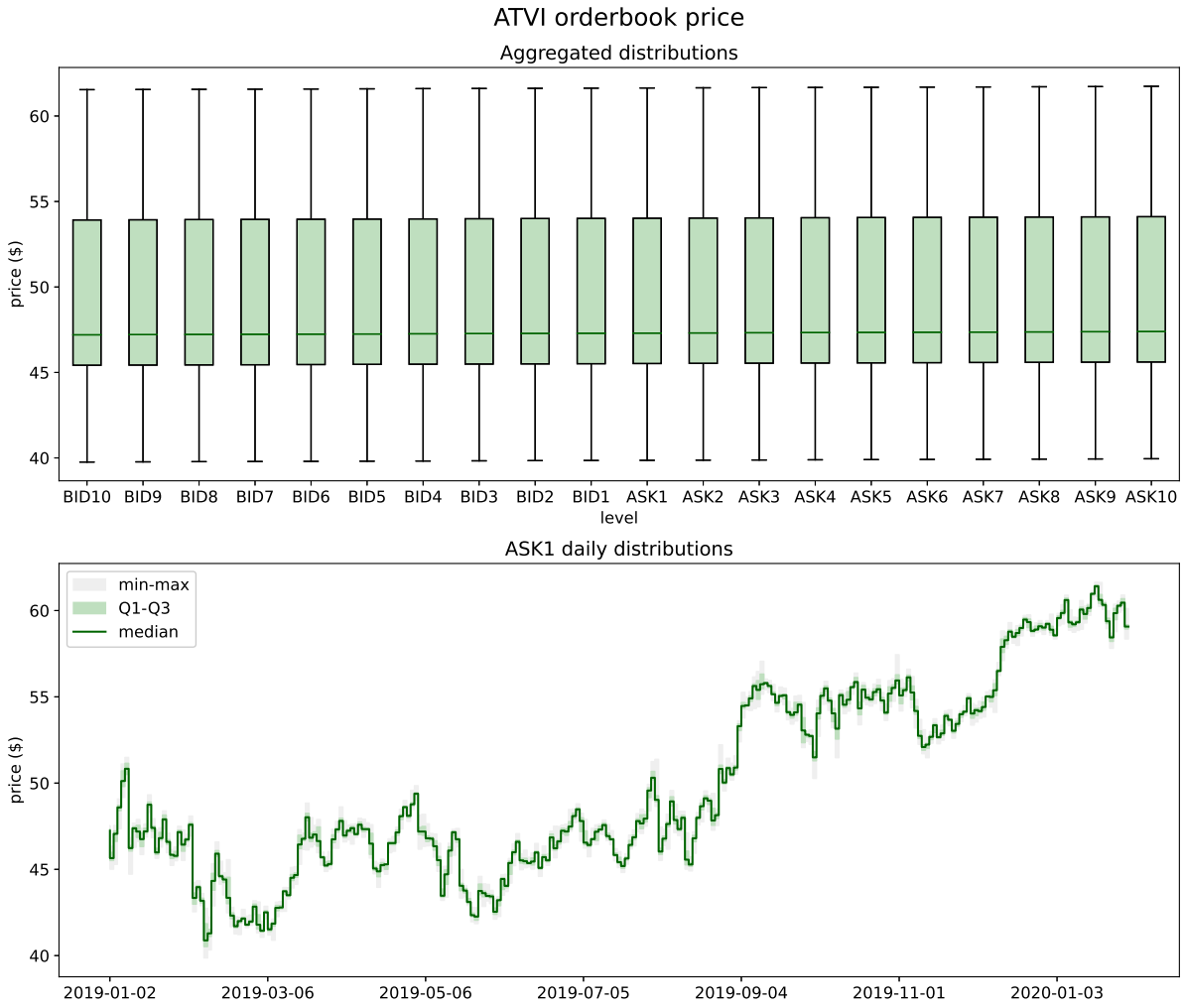


Figure 2.D.1: ATVI order book prices during the period from January 2nd, 2019 to Januray 31st, 2020.

$c_{t,t+h}$ conditional on the last return $c_{t-h,t}$ only. In this respect, we note the return label $c_{t-h,t}$ can be computed from the features $(\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t)$ for T sufficiently large. In view of the universal approximation property of neural networks, any such simple model can thus be considered as 'nested' in the deep learning models treated in this work. In Appendix 2.E we explore how well a specific model based on $c_{t-h,t}$ only performs when compared to the more complex deep learning models considered in this chapter.

2.E Simpler predictive models: an empirical auto-regressive specification

In this section, we explore how a simple model performs when compared to the deep learning architectures considered in this work. We base the model on the simple empirical observation that the target return labels display time-dependence, cf. Appendix 2.D.2 and Figure 2.D.7. We consider a

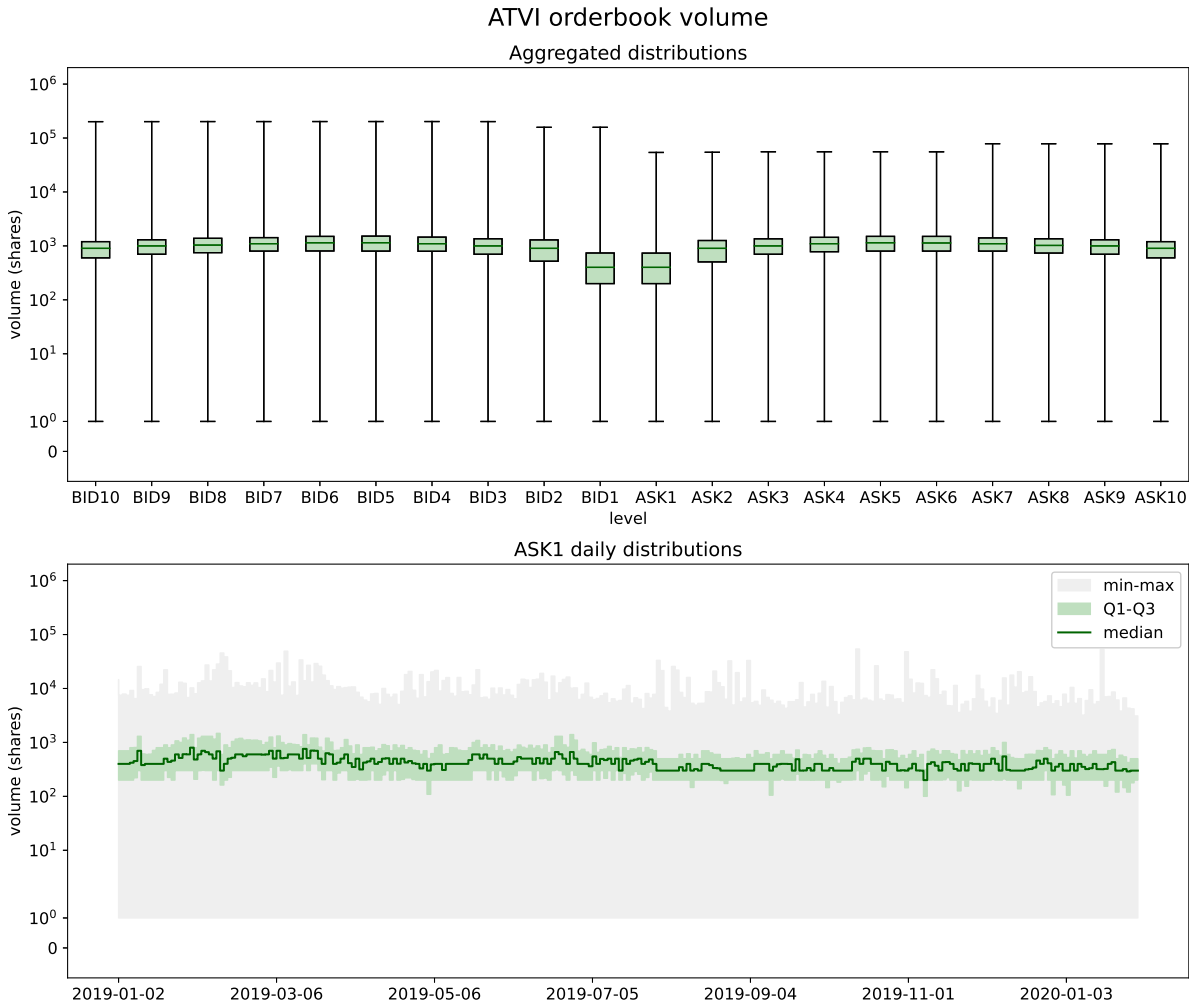


Figure 2.D.2: ATVI order book volumes during the period from January 2nd, 2019 to Januray 31st, 2020.

one-step ahead[20] auto-regressive (AR) model, i.e. we aim to estimate $p_*(\star) = \mathbb{P}(c_{t,t+h} = *|c_{t-h,t} = \star)$ for $*, \star \in \{\downarrow, =, \uparrow\}$. We use a non-parametric approach, setting $\hat{p}_*(\star)$ in period $w \in \{1, \dots, 11\}$ to be the empirical distribution of $c_{t,t+h} = * \in \{\downarrow, =, \uparrow\}$ given the previous return label was $c_{t-h,t} = \star \in \{\downarrow, =, \uparrow\}$ in the training and validation set $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$, i.e. the row normalised matrices displayed in Figure 2.D.7. Formally, this is the maximum likelihood estimator of the conditional probabilities $\mathbb{P}(c_{t,t+h} = *|c_{t-h,t} = \star)$ given the dataset $\mathcal{D}_{w,\text{train}} \cup \mathcal{D}_{w,\text{val}}$. We repeat the experiments of Section 2.4.1 and Section 2.4.2 with this empirical AR model in the set $\mathcal{M}_0$. The updated results are reported in Table 2.E.1 and Table 2.E.2 (we only report results at the 99% confidence level, i.e. $\alpha = 0.01$).

The results reported in Table 2.E.1 are consistent with those in Table 2.4. The main differences are due to some additional predictable horizons identified by including the empirical auto-regressive model. This mostly occurs for stocks with low liquidity (LILAK, QRTEA, and CHTR) where the low data regime makes it more challenging for deep learning models to train effectively.

[20]To avoid any look-ahead bias we actually condition on $c_{t-h-k,t}$ where k is the smoothing factor introduced in Section 2.3.2. We write $c_{t-h,t}$ for clarity of exposition.

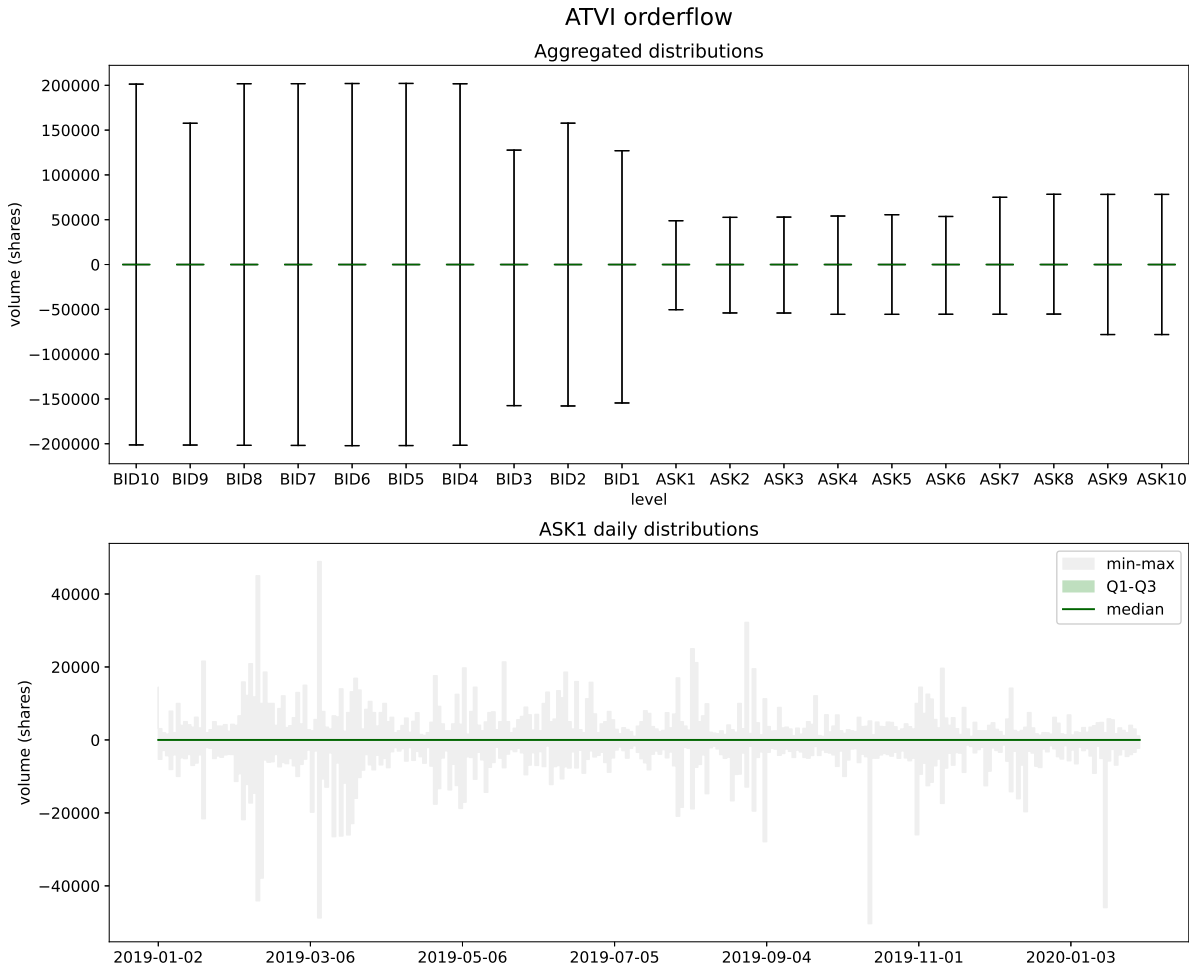


Figure 2.D.3: ATVI order flow during the period from January 2nd, 2019 to Januray 31st, 2020.

In Table 2.E.2 we note the greater expressivity of deep learning models provides a significant advantage compared to the simple empirical auto-regressive model. While the simple AR model does display predictive power, it is often outperformed by the order book-driven deep learning models deepOF(L2) and deepVOL(L2/L3). In Table 2.E.3 we dive deeper into this analysis by reporting for each ticker-horizon combination whether the simple AR model or deepVOL(L3) is in the MCS of best models.

While our results suggest deepOF and deepVOL architectures significantly outperform the simple AR model, a more in-depth study comparing these specifications to the models in Aït-Sahalia et al. (2022) is required to shed some light on the question of whether the expressivity of deep learning techniques is the key to producing good predictive models or if careful feature engineering can be as effective. This analysis is of particular relevance in applications where prediction speed plays a fundamental role, see Appendix 2.G.

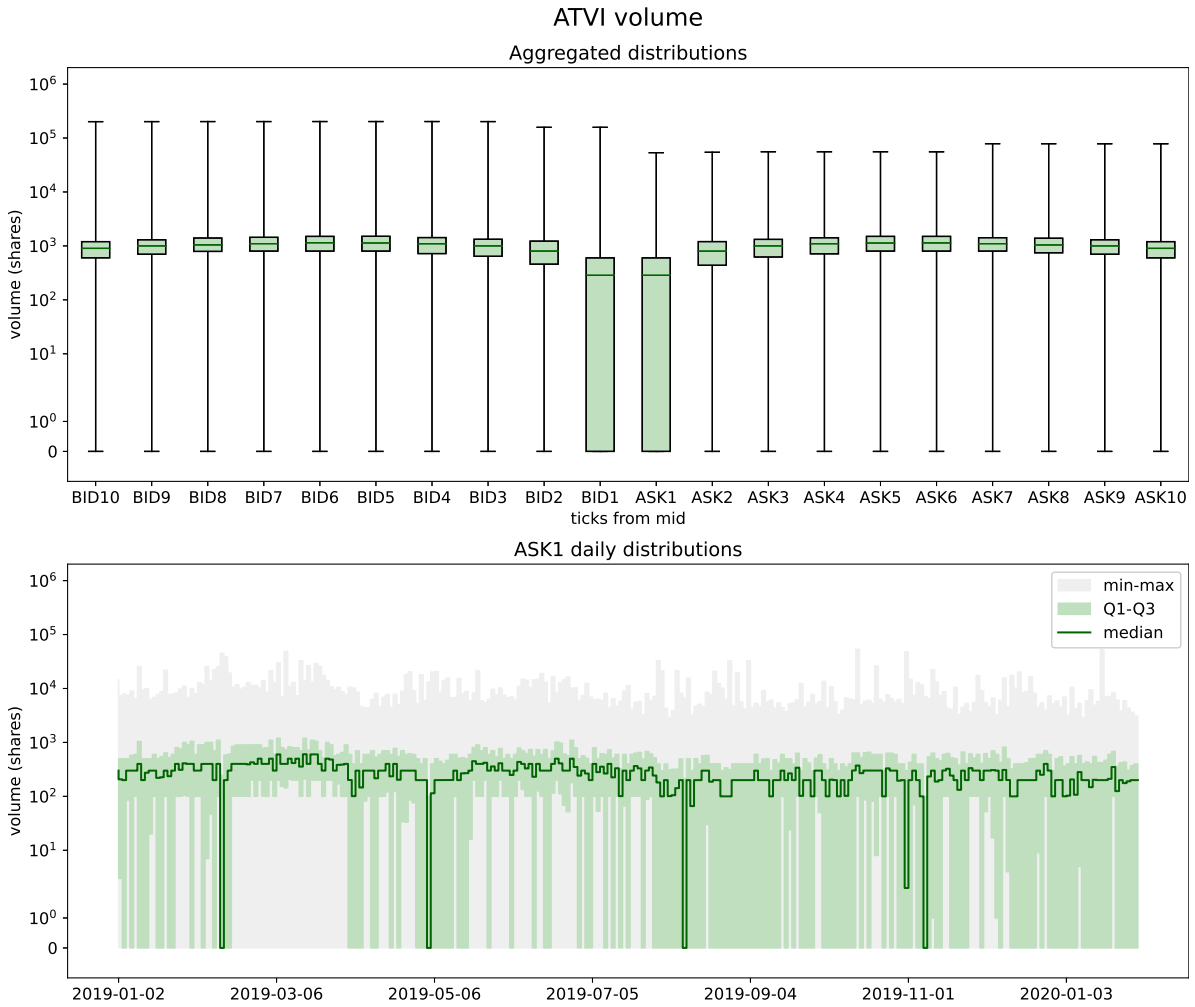


Figure 2.D.4: ATVI volumes during the period from January 2nd, 2019 to Januray 31st, 2020.

2.F Defining returns

The way returns are defined is often an overlooked topic which inevitably has a significant effect on model evaluation. Here we state two alternative definitions to that given in Section 2.3.2 taken from the literature and discuss their properties. In all cases, the mid-price is treated as the "true" price and returns are defined relative to it. Recall this is not a tradable price.

- In Ntakaris et al. (2018) the authors define the return at horizon h as

$$r_{t,t+h} = \frac{\overline{m}_{t,t+h} - m_t}{m_t}, \ \overline{m}_{t,t+h} = \frac{1}{h}\sum_{i=1}^{h} m_{t+i},$$

  where m_t denotes the mid-price at time t. This definition of return compares the smoothed mid-price over the next h time steps to the current mid-price. In this case, the return can be equivalently understood as the average mid-to-mid return over the next h time steps. As the horizon h becomes longer, so does the window over which we smooth the returns. This has a

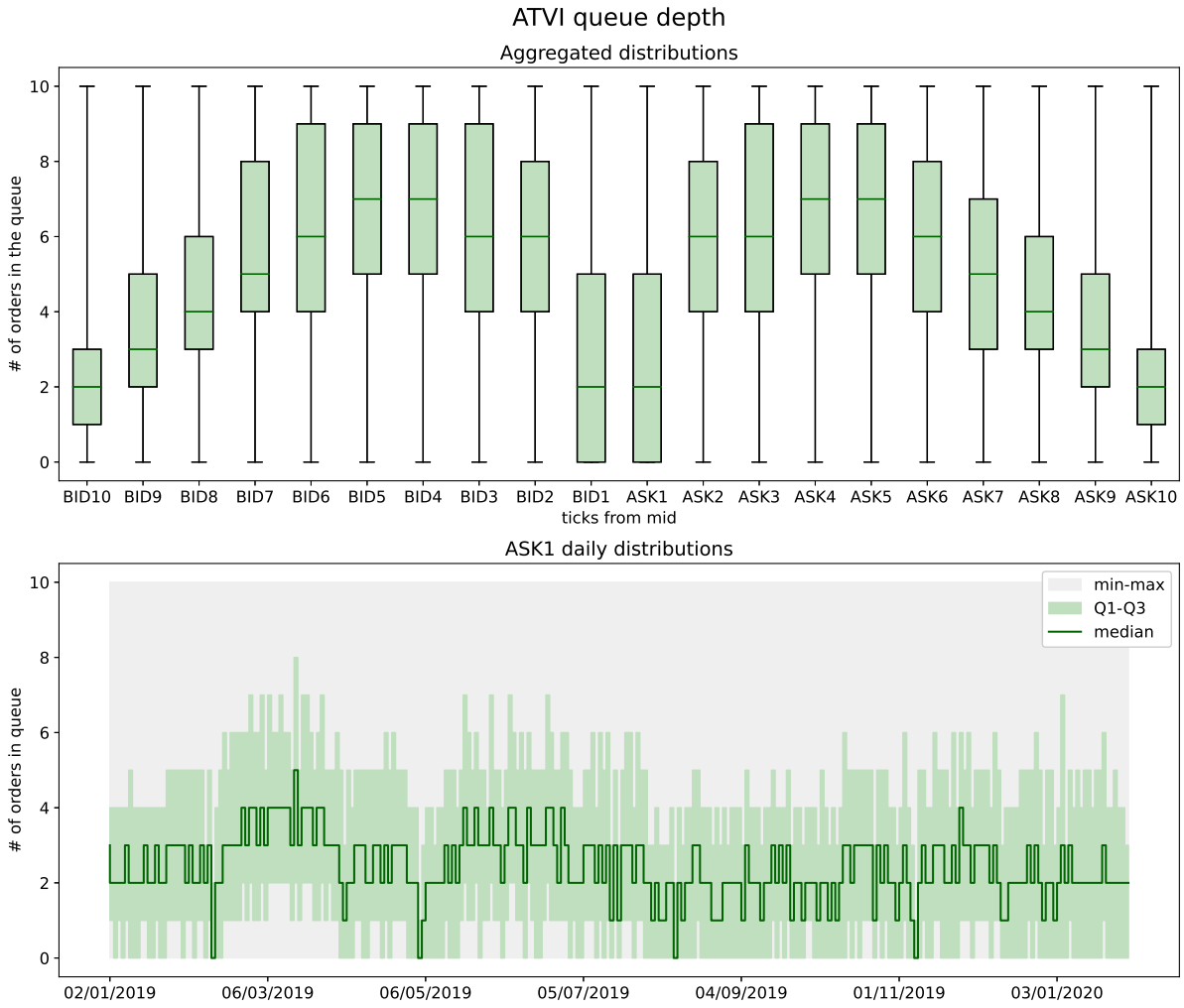


Figure 2.D.5: ATVI queue depths during the period from January 2nd, 2019 to Januray 31st, 2020.

couple of practical disadvantages in terms of the predictability questions we wish to explore. First, predictability over $h = 100$ time steps might be due to changes of the mid-price over the first, say, $h = 50$ time steps, and thus it becomes difficult to discuss questions of persistency in predictability. Second, this definition results in stronger correlation between returns at different horizons, r_{t+h_1} and r_{t+h_2}, inducing a hidden bias in multi-horizon models.

- In Zhang, Zohren, and Roberts (2019) and Zhang and Zohren (2021) the authors define the return at horizon h by

$$r_{t,t+h} = \frac{\overline{m}_{t,t+h} - \overline{m}_{t-h,t}}{\overline{m}_{t-h,t}}, \ \overline{m}_{t,t+h} = \frac{1}{h}\sum_{i=1}^{h} m_{t+i}, \ \overline{m}_{t-h,t} = \frac{1}{h}\sum_{i=0}^{h-1} m_{t-i},$$

where m_t denotes the mid-price at time t. This definition has similar drawbacks to the previous one. But, additionally, we believe it may suffer from look-ahead bias: for example, if the mid-price has been going up in the past h steps (an information which is included in our covariates), then it is more likely the return $r_{t,t+h}$ will be positive.

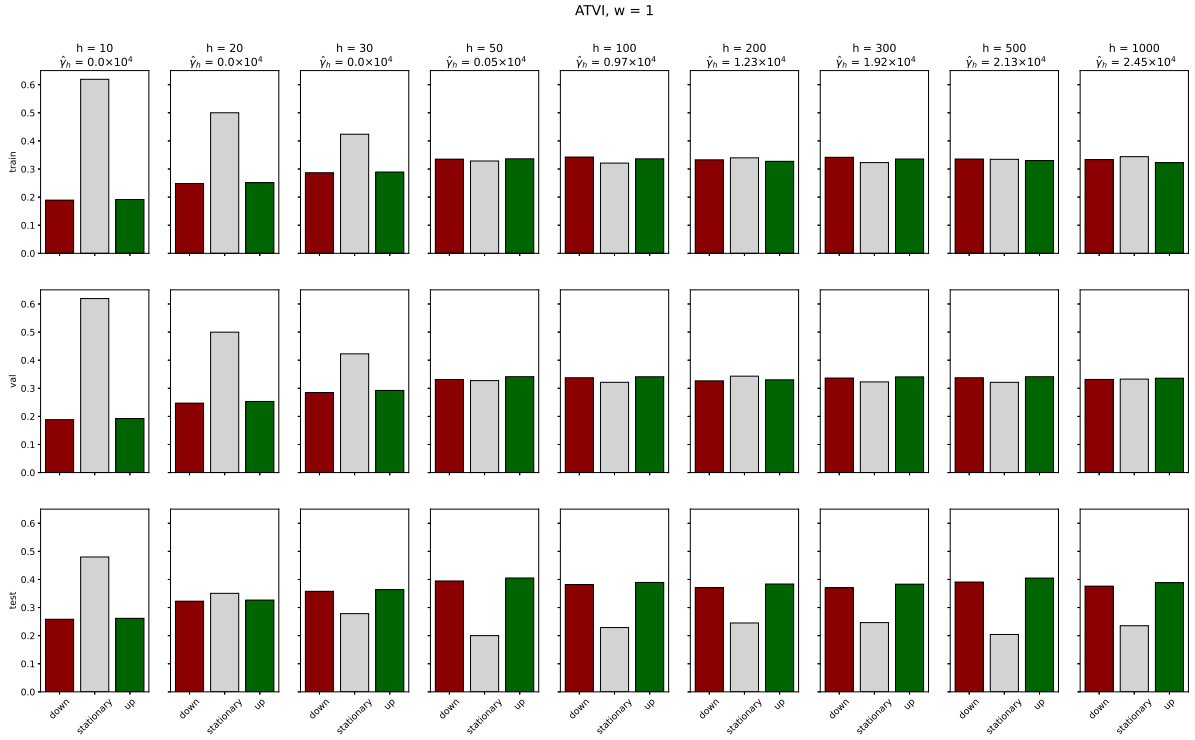


Figure 2.D.6: ATVI return labels during the first window of the experiment, i.e. from January 14th, 2019 to February 15th, 2019.

The definition of returns used in this chapter, cf. Section 2.3.2, does not suffer from the issues identified for these two return specifications.

Remark 2.F.1. The definition of returns intrinsically depends on the clock used to measure time. Note that in Kolm, Turiel, and Westray (2021) the authors use a physical time clock with stock-specific horizon h. In their setting, the authors use simple un-smoothed mid-price returns,

$$r_{t,t+h} = \frac{m_{t+h} - m_t}{m_t}.$$

With our choice of order book-driven clock, even assuming mid-mid trading, there is no way of placing a trade exactly h order book events ahead, i.e. the prediction horizon is random. Smoothing the exit price can thus be understood as averaging out the uncertainty in the execution time. Moreover, in our work, we aim to investigate questions regarding structural market predictability. In our setting, smoothing mid-prices leads to better estimates of the true (latent) prices by weakening idiosyncratic noise effects. As pointed out in the conclusions in Section 2.5 though, in practical trading applications it may be more appropriate to consider a physical time clock and un-smoothed returns.

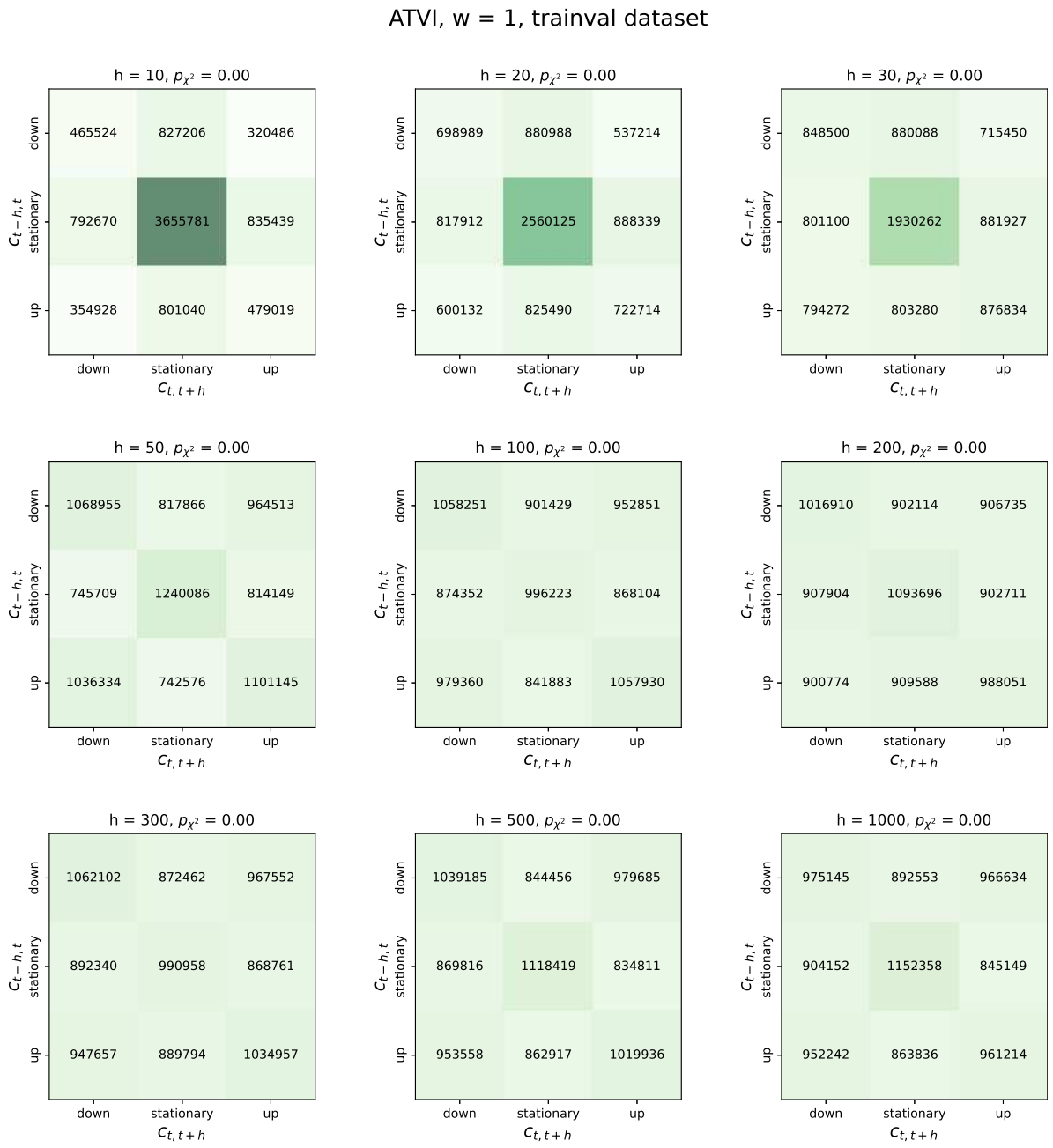


Figure 2.D.7: ATVI return labels dependence in the joint training and validation dataset during the first window of the experiment, i.e. from January 14th, 2019 to February 15th, 2019.

2.G Infrastructure latency

In this work, we explored the predictive value of order book data by assuming instantaneous access to such information. In practice, due to technological constraints, market participants experience varying degrees of delay when engaging with the order book. Let us write $\lambda^{n,\text{view}}$ and $\lambda^{n,\text{act}}$ for the latencies that market participant n experiences when viewing and acting on the order book

	$h = 10$	$h = 20$	$h = 30$	$h = 50$	$h = 100$	$h = 200$	$h = 300$	$h = 500$	$h = 1000$
LILAK	0.00	0.00	0.38	0.40	0.04	0.08	0.00	0.01	1.00
QRTEA	0.00	0.01	0.03	0.00	0.00	0.00	0.06	0.85	0.26
XRAY	0.00	0.00	0.00	0.01	0.09	0.35	0.35	0.73	0.85
CHTR	0.00	0.00	0.03	0.00	0.00	0.00	0.06	0.14	0.56
PCAR	0.00	0.00	0.00	0.00	0.86	0.42	0.31	0.17	1.00
EXC	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.38
AAL	0.04	0.00	0.00	0.00	0.00	0.65	0.70	0.10	0.38
WBA	0.00	0.00	0.00	0.00	0.01	0.23	0.04	0.00	0.18
ATVI	0.00	0.00	0.00	0.00	0.01	0.61	0.34	0.12	0.03
AAPL	0.00	0.00	0.00	0.16	0.08	0.03	0.05	0.06	0.09

Table 2.E.1: MCS p-values of the *unpredictive benchmark* model for the 10 tickers and 9 horizons under consideration. When the p-value is low at least one of the order book-driven models statistically outperforms the *unpredictive benchmark*, i.e. there is order book-driven predictability according to Definition 2.4.1.

	$\alpha = 0.01$
benchmark	0%
empirical AR model	28%
deepOF(L1)	12%
deepOF(L1)	26%
deepLOB(L2)	12%
deepOF(L2)	81%
deepVOL(L2)	77%
deepVOL(L3)	79%

Table 2.E.2: % of times the model is in the α-MCS when predictability is identified at the level $\alpha = 0.01$.

respectively. For $x \in \{\text{view}, \text{act}\}$ we can decompose

$$\lambda^{n,x} = \lambda_0^x + \lambda_+^{n,x},$$

where λ_0^{act} is the time it takes the financial entity running the market, e.g. the Nasdaq, to execute an action on the order book once received and λ_0^{view} is the time it takes the market entity to send out the information regarding an order book update. Market participants make large efforts to reduce their idiosyncratic latencies $\lambda_+^{n,x}$ for $x \in \{\text{view}, \text{act}\}$ by optimizing software and hardware, one prototypical example being co-location. Note that $\lambda_+^{n,\text{act}}$ also includes the time needed to take a trading decision which, in our setting, requires a forward pass through a pre-trained predictive model, e.g. deepLOB/deepOF/deepVOL. Figure 2.G.1 represents how market-wide and idiosyncratic latencies affect market participants.

When considering practical trading applications one should thus take into account the effect these latencies may have on the definition of returns, cf. Section 2.3.2 and Appendix 2.F. For example, denoting by $t = 1, 2, \ldots$ the order book-driven clock considered in this work, one may

	$h = 10$	$h = 20$	$h = 30$	$h = 50$	$h = 100$	$h = 200$	$h = 300$	$h = 500$	$h = 1000$
LILAK									
QRTEA									
XRAY									
CHTR									
PCAR									
EXC									
AAL									
WBA									
ATVI									
AAPL									

Table 2.E.3: Comparison between emprical AR model and deepVOL(L3) MCS results, $\alpha = 0.01$. Shading indicates whether the empirical AR model and/or deepVOL(L3) is in the set of best models. White: neither, light blue: empirical AR model only, blue: deepVOL(L3), dark blue: both. Grey shading indicates the ticker-horizon combination was not deemed predictable at the 99% confidence level.

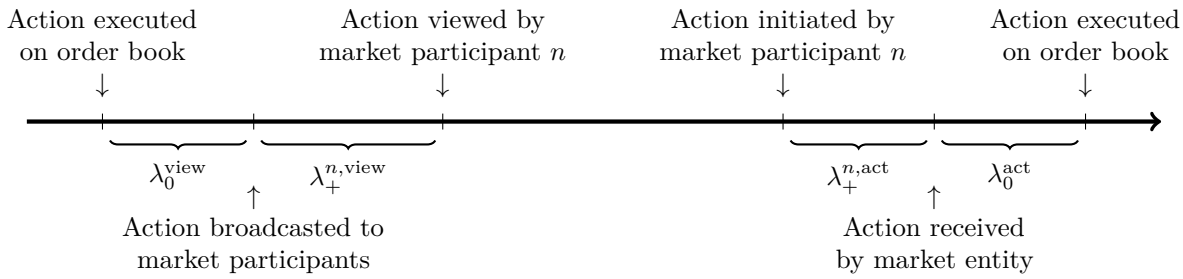


Figure 2.G.1: Market-wide and idiosyncratic latencies.

want to predict the return between physical times $\tau_t + \lambda^{n,\text{tot}}$ and $\tau_{t+h} + \lambda^{n,\text{tot}}$ where

$$\lambda^{n,\text{tot}} = \lambda^{n,\text{view}} + \lambda^{n,\text{act}} = \lambda_0^{\text{tot}} + \lambda_+^{n,\text{tot}},$$

is the total round trip latency time to view and act once the order book clock ticks and $\tau : \mathbb{N} \to [0, \infty)$ is the increasing random function mapping the order book clock to physical time. An interesting analysis in this setting would be to analyze how predictability varies as a function of $\lambda_+^{n,\text{tot}} = \lambda_+^{n,\text{view}} + \lambda_+^{n,\text{act}}$.

In order to quantify the magnitude of these time lags and better understand their impact we briefly summarize some infrastructure tests published by market participants.

- As of 2023, the Nasdaq reports the door-to-door speed of its matching engine, i.e. $\lambda_0^{\text{tot}} = \lambda_0^{\text{act}} + \lambda_0^{\text{view}}$, to be 'sub-40 μs, with the fastest production implementation at 14 μs' (Nasdaq, 2023). Studies from 2015 report the door-to-door latencies of Brazilian and European futures exchanges in the order of a few hundred microseconds (Kirilenko and Lamacie, 2015; Menkveld and Zoican, 2017).

- Leading industry producers of co-located network adapters and feed handlers report the total speed for processing an order book message and sending a trade action, i.e. tick-to-trade

latency $\lambda_+^{n,\text{tot}} = \lambda_+^{n,\text{view}} + \lambda_+^{n,\text{act}}$, to be of the order of a couple of microseconds (μs) (Enyx, 2020; CSPi, 2023). This is achieved via the use of specific hardware known as FPGA cards. These figures assume relatively simple trading decision applications, when working with more sophisticated predictive models, such as the ones considered in this work, one should account for a couple of hundred microseconds in tick-to-trade latency, see Table III in Zhang, Zohren, and Roberts (2019) for unoptimized forward-pass run times.

The total latency $\lambda^{k,\text{tot}}$ is thus less than a fraction of a millisecond (ms) while the average time elapsed between consecutive price changes in our data ranges from 0.18 to 8.38 seconds depending on the liquidity of the stock, cf. Table 2.1. The effect of infrastructural latencies on the definition of returns, cf. Section 2.3.2, should thus not deteriorate results significantly.

2.H Stock selection

Ideally, we would like to work with the same set of stocks as in Kolm, Turiel, and Westray (2021). Due to computational limitations, we were able to conduct experiments only on a subset of 10 tickers from the total 115 NASDAQ stocks considered in the original paper. We selected 10 stocks with diverse liquidity characteristics, hopefully providing a sufficiently representative sub-sample of the whole set.

To choose the sub-sample, we use the stock characteristics provided in Kolm, Turiel, and Westray (2021), Table 6. For each of the liquidity characteristics – Updates, Trades, Price Changes, Spread – we compute a sub-score based on the characteristic's rank. For example, the stock with the most updates is assigned an updates score of 1, while the stock with the least number of updates is assigned an updates score of 0. The characteristic-specific scores are then averaged to obtain a general "liquidity score" for each stock. The 10 chosen stocks correspond to the 10 evenly spaced quantiles of the "liquidity score". The liquidity characteristics of the 10 chosen stocks are reported in Table 2.1.

2.I Model details

In the experiments, we use the model architectures reported in Table 2.I.1 and Table 2.I.2. To speed up the training procedure, we use batch normalization (Ioffe and Szegedy, 2015) (with momentum = 0.6) after every convolutional layer and inception block. To prevent overfitting, we use a dropout layer (Srivastava et al., 2014) (with noise = 0.2) positioned after the inception module.

All the models are implemented with `float32` precision policy: at every layer, `float32` is used as computation and variable data type. Loss and gradient computations are also carried out with `float32` precision.

Model Architectures							
Model		DeepLOB			DeepOF		
		Architecture	Description	Output Size	Architecture	Description	Output Size
Input Layer			Raw order book states, $\{\{(p^{(l)}_{a,t-\tau}, v^{(l)}_{a,t-\tau}, p^{(l)}_{b,t-\tau}, v^{(l)}_{b,t-\tau})\}_{\tau=0}^{T-1}\}_{l=1}^{L}$	$(T, 4L, 1)$		Multi-level order flow, $\{\{(aOF^{(l)}_{t-\tau}, bOF^{(l)}_{t-\tau})\}_{\tau=0}^{T}\}_{l=1}^{L}$	$(T, 2L, 1)$
First Convolutional layer	Spatial convolution	32 (1×2) filters, (1×2) stride	Convolve price and volume information	$(T, 2L, 32)$			
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 2L, 32)$			
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 2L, 32)$			
Second Convolutional layer	Spatial convolution	32 (1×2) filters, (1×2) stride	Convolve bid and ask price-volume information	$(T, L, 32)$	32 (1×2) filters, (1×2) stride	Convolve order flow information	$(T, L, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, L, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, L, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, L, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, L, 32)$
Third Convolutional layer	Spatial convolution	32 (1×L) filters, (1×1) stride	Convolve information through the whole order book	$(T, 1, 32)$	32 (1×L) filters, (1×1) stride	Convolve information through the whole order book	$(T, 1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$
Inception module (parallel blocks)	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
	Temporal convolution	64 (3×1) filters, padding	Convolve through three time-steps, "MA(3)"	$(T, 1, 64)$	64 (3×1) filters, padding	Convolve through three time-steps, "MA(3)"	$(T, 1, 64)$
	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
	Temporal convolution	64 (5×1) filters, padding	Convolve through five time-steps, "MA(5)"	$(T, 1, 64)$	64 (5×1) filters, padding	Convolve through five time-steps, "MA(5)"	$(T, 1, 64)$
	Temporal maxpooling	(3×1) pool size, (1×1) stride & padding	Rolling maximum, three time-steps	$(T, 1, 32)$	(3×1) pool size, (1×1) stride & padding	Rolling maximum, three time-steps	$(T, 1, 32)$
	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (4×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
Concatenate and reshape			Concatenate and reshape the outputs of the inception module to obtain an $\mathbb{R}^{192}$-valued time series of length T	$(T, 192)$		Concatenate and reshape the outputs of the inception module to obtain an $\mathbb{R}^{192}$-valued time series of length T	$(T, 192)$
LSTM module	LSTM	64 hidden units	Capture longer term dependencies	(64)	64 hidden units	Capture longer term dependencies	(64)
(a) Output layer	Dense layer	Softmax activation	Produce classification output	(3)	Softmax activation	Produce classification output	(3)
(b) Decoder module	LSTM decoder (seq2seq/attention)	64 hidden units, softmax activation	Produce sequential classification output	(5, 3)	64 hidden units, softmax activation	Produce sequential classification output	(5, 3)

Table 2.I.1: DeepLOB and DeepOF network description.
(a) Single horizon, (b) Multi-horizon.

Model Architectures							
Model		DeepVOL			DeepVOL L3		
		Architecture	Description	Output Size	Architecture	Description	Output Size
Input Layer			Volume representation, $\{\{\{s^{(j)}_{x,t-\tau}\}_{\tau=0}^{T-1}\}_{j=1}^{W}\}_{x\in\{a,b\}}$	$(T, W, 2, 1)$		L3 volume representation, $\{\{\{\{q^{(j,k)}_{x,t-\tau}\}_{\tau=0}^{T-1}\}_{j=1}^{W}\}_{x\in\{a,b\}}\}_{k=1}^{D}$	$(T, W, 2, D, 1)$
First Convolutional layer	Spatial convolution				32 (1×1×D) filters, (1×1×1) stride	At each price, convolve the queue volumes	$(T, W, 2, 32)$
Second Convolutional layer	Spatial convolution	32 (1×2×2) filters, (1×1×1) stride	Convolve volumes on opposite sides of the mid	$(T, W-1, 32)$	32 (1×2×2) filters, (1×1×1) stride	Convolve volumes on opposite sides of the mid	$(T, W-1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, W-1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, W-1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, W-1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, W-1, 32)$
Third Convolutional layer	Spatial convolution	32 (1×W-1) filters, (1×1) stride	Convolve information through the whole order book	$(T, 1, 32)$	32 (1×W-1) filters, (1×1) stride	Convolve information through the whole order book	$(T, 1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$
	Temporal convolution	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$	32 (4×1) filters, padding	Convolve through four time-steps	$(T, 1, 32)$
Inception module (parallel blocks)	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
	Temporal convolution	64 (3×1) filters, padding	Convolve through three time-steps, "MA(3)"	$(T, 1, 64)$	64 (3×1) filters, padding	Convolve through three time-steps, "MA(3)"	$(T, 1, 64)$
	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
	Temporal convolution	64 (5×1) filters, padding	Convolve through five time-steps, "MA(5)"	$(T, 1, 64)$	64 (5×1) filters, padding	Convolve through five time-steps, "MA(5)"	$(T, 1, 64)$
	Temporal maxpooling	(3×1) pool size, (1×1) stride & padding	Rolling maximum, three time-steps	$(T, 1, 32)$	(3×1) pool size, (1×1) stride & padding	Rolling maximum, three time-steps	$(T, 1, 32)$
	Temporal convolution	64 (1×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$	64 (4×1) filters	Increase dimensionality (Network-in-Network)	$(T, 1, 64)$
Concatenate and reshape			Concatenate and reshape the outputs of the inception module to obtain an $\mathbb{R}^{192}$-valued time series of length T	$(T, 192)$		Concatenate and reshape the outputs of the inception module to obtain an $\mathbb{R}^{192}$-valued time series of length T	$(T, 192)$
LSTM module	LSTM	64 hidden units	Capture longer term dependencies	(64)	64 hidden units	Capture longer term dependencies	(64)
(a) Output layer	Dense layer	Softmax activation	Produce classification output	(3)	Softmax activation	Produce classification output	(3)
(b) Decoder module	LSTM decoder (seq2seq/attention)	64 hidden units, softmax activation	Produce sequential classification output	$(5, 3)$	64 hidden units, softmax activation	Produce sequential classification output	$(5, 3)$

Table 2.I.2: DeepVOL and DeepVOL L3 network description.
(a) Single horizon, (b) Multi-horizon.

Chapter 3

Continuous-time Autoregressive Processes

We now transition from discrete-time to continuous-time modeling. In the previous chapter, we considered a discrete set of events as the time parameter of the underlying process: the LSTM's hidden state was updated whenever an action occurred in the order book. This discrete-time framework constrains the modeling granularity and timescale: (i) there is often no natural way to interpolate between observations, and (ii) the timesteps must have a consistent interpretation. While these constraints posed no issues when modeling order books updated at each tick, many systems involve quantities that could be measured at any point in time but are only observed over a finite collection of (irregularly) spaced points.

In such cases, a continuous-time model offers a unified representation that remains consistent across varying observational timescales. The class of continuous-time autoregressive (CAR) processes explored in this chapter exemplifies this approach. Probably due to their smooth dynamics, CAR processes have received limited research attention. However, as we shall see at the end of this chapter, their behavior closely aligns with certain physical phenomena, including light curves. This chapter focuses on developing the estimation and inference tools needed to analyze such observations under a CAR specification.

3.1 Introduction

3.1.1 Motivation: from OU processes to CAR processes

Let $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ denote a d-dimensional Lévy process with characteristics $(\mathbf{b}, \Sigma, F)$ on the probability space $(\Omega', \mathcal{F}', \mathbb{P}')$, i.e. $\mathbf{b} \in \mathbb{R}^d$, Σ a symmetric positive semi-definite $d \times d$ matrix and F a Lévy measure on $\mathbb{R}^d$, that is $F(\{0\}) = 0$ and $\int_{\mathbb{R}^d}(\|x\|^2 \wedge 1)F(dx) < \infty$. For a detailed treatment of Lévy processes see Sato (1999). The Ornstein-Uhlenbeck (OU) process driven by $\mathbb{L}$ with drift matrix $A \in \mathcal{M}_d(\mathbb{R})$ is the unique strong solution $\mathbb{Y} = \{\mathbf{Y}_t,\ t \geq 0\}$ to the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{Y}_t = -A\mathbf{Y}_t\,\mathrm{d}t + \mathrm{d}\mathbf{L}_t,\ t \geq 0. \tag{3.1.1}$$

This process has been the subject of great research interest from a theoretical and practical point of view. When $\mathbb{L}$ is taken to be a Brownian motion, its properties are now well understood, with

This chapter is based on the submitted article: Lucchese, Pakkanen, and Veraart (2023).

the first studies dating back to Uhlenbeck and Ornstein (1930). OU processes with more general Lévy drivers have also been extensively analyzed, for example, Sato and Yamazato (1984) related stationary OU-type processes with operator self-decomposable distributions while Masuda (2004) studied their transition semigroups, transition densities, and mixing conditions. Such processes have been used for modeling purposes in various fields, including physics, finance, and biology. For a concrete financial application, see Barndorff-Nielsen and Shephard (2001) where a positive stationary OU process is used to model the volatility process in a stochastic volatility model.

Formally, the OU process (3.1.1) satisfies

$$p(D)\mathbf{Y}_t = D\mathbf{L}_t, \ t \geq 0, \tag{3.1.2}$$

where $p(z) = z + A$ is a first order polynomial and D denotes differentiation with respect to time t. Informally, $D\mathbf{L}_t$ represents a continuous-time random noise process and Equation (3.1.2) a stochastically perturbed first-order homogeneous linear differential equation with constant coefficients. If we instead interpret D as a backshift operator we immediately notice why OU processes can be understood as the continuous-time analog of discrete-time autoregressive processes of order 1, AR(1) for short. A deeper connection between the two classes of processes is present: under appropriate conditions on A and $\mathbb{L}$, for any $h > 0$ the h-skeleton of the OU process $\mathbb{Y}$, defined by $\mathbb{Y}_h := \{\mathbf{Y}_{nh}, \ n \in \mathbb{N}\}$, is an AR(1) process, see for example Masuda (2004, Theorem 4.3). In view of this connection, we can naturally generalize OU processes in the same way as AR(p) processes generalize AR(1) processes, i.e. for $p \in \mathbb{N}$ and $A_1, \ldots, A_p \in \mathcal{M}_d(\mathbb{R})$ we consider the $(p-1)$-times "differentiable" process $\mathbb{Y}$ which formally satisfies Equation (3.1.2) where

$$p(z) = z^p + A_1 z^{p-1} + \ldots + A_{p-1} z + A_p,$$

is a so-called autoregressive polynomial and D denotes differentiation with respect to time t. Equation (3.1.2) can now be interpreted as a higher-order homogeneous linear differential equation with constant coefficients subject to exogenous random kicks. Such a process, which will be rigorously defined in Section 3.2, is known as a Continuous Autoregressive (CAR) process and falls in the wider class of Continuous Autoregressive Moving Average (CARMA) processes, the continuous-time counterpart of ARMA time series models. An overview of CARMA processes is given in Brockwell and Davis (2016, Chapter 11.5), focusing mainly on the results for Gaussian CARMA processes obtained by Jones (1981) and Chan and Tong (1987). Properties of Lévy-driven CARMA processes are investigated by Brockwell (2009) and Brockwell and Lindner (2009) in the univariate case and by Marquardt and Stelzer (2007) in the more general multivariate setting. It is important to note that the study of the Lévy-driven OU and CARMA processes is an active area of research, and many researchers have made significant contributions to their understanding. The mentioned works are merely a small fraction of the research conducted in this field and should not be regarded as an exhaustive compilation.

3.1.2 Contributions and literature review

Firstly and most prominently we develop consistent and asymptotically normal estimators for the drift coefficients of multivariate Lévy-driven CAR(p) processes both when continuous-time and (possibly irregularly spaced) discrete-time observations are available. As a necessary step in the discretization procedure, in Section 3.4.2 we provide a rigorous treatment of the finite difference approximation of the derivatives of a differentiable stochastic process which, to the best of our knowledge, is a novelty in the field. Moreover, we believe the proofs of the asymptotic results for the OU drift estimators appearing in Mai (2014) might contain some inconsistencies. By carefully modifying the arguments we show the stated properties hold under slightly different assumptions as a special case of our more general asymptotic results for CAR(p) processes. Finally, when the multivariate observation has a known underlying network structure, we reduce the model parametrization by introducing the GrCAR architecture, a natural extension of the GrOU process introduced in Courgeau and Veraart (2022), and the corresponding estimators in Section 3.5.

Estimation of OU drift coefficients

When $p = 1$ our task reduces to estimating the drift coefficient of the Lévy-driven OU process in Equation (3.1.1). In this setting, a wide range of estimators have been investigated in the literature under different assumptions on the driving Lévy process $\mathbb{L}$, cf. Table 3.1. Other than the well-studied Gaussian case, specific results were developed in the context of non-decreasing Lévy drivers in Jongbloed, Van Der Meulen, and Van Der Vaart (2005) and Brockwell, Davis, and Yang (2007b) for Davis-McCormick M-estimators. As discussed in the previous section, this class of processes is of particular relevance for volatility modeling (Barndorff-Nielsen and Shephard, 2001). Assuming the Lévy driver has a non-degenerate Gaussian component, Mai (2014) develops continuous-time maximum likelihood estimators and analyzes their discrete-time counterparts, naturally allowing for irregularly spaced data. Gushchin, Pavlyukevich, and Ritsch (2020) extend the continuous-time likelihood results when $\mathbb{L}$ has heavier tails and Courgeau and Veraart (2022) consider a specific multivariate extension, imposing a graphical structure on the observations. More general (symmetric) Lévy drivers are considered in Fasen (2013) and Masuda (2010) under a discrete-time uniform sampling scheme, where the authors analyze least squares and least absolute deviation estimators respectively.

In the present work, we aim to revisit the results in Mai (2014), treated in Courgeau and Veraart (2022) in the multivariate case, and extend them to multivariate CAR(p) processes of general order $p \in \mathbb{N}$. Working under similar assumptions, we will consider explicit continuous-time maximum likelihood estimators and their corresponding discretizations.

	Lévy driver $\mathbb{L}$ with characteristics $(\mathbf{b}, \Sigma, F)$			
	Square-integrable[1]		Heavy tails[2]	
	Non-decreasing[3]	$\Sigma > 0$[4]	$\Sigma > 0$[4]	
continuous-time	Brockwell, Davis, and Yang (2007b)	Mai (2014), Courgeau and Veraart (2022)	Gushchin, Pavlyukevich, and Ritsch (2020)	
discrete-time (uniform)	Brockwell, Davis, and Yang (2007b), Jongbloed, Van Der Meulen, and Van Der Vaart (2005)	Mai (2014), Courgeau and Veraart (2022)		Fasen (2013), Masuda (2010)
discrete-time (irregular)		Mai (2014), Courgeau and Veraart (2022)		

Table 3.1: Estimation of OU drift coefficients.
[1] $\int \|x\|^2 F(\mathrm{d}x) < \infty$. [2] F is regularly varying of order $\alpha \in (0, 2)$. [3] $\mathrm{supp}(F) \subseteq (0, \infty)$, $\Sigma = 0$.
[4] Σ strictly positive definite.

Estimation of CAR drift coefficients

Fewer results are available for the estimation of CAR(p) processes of order $p \in \mathbb{N}$, some of which are summarized in Table 3.2. In the more general setting of Lévy-driven CARMA(p, q) models, to the best of our knowledge, all available estimators are based on uniform discrete-time observations and their statistical analysis relies on the properties of the h-skeleton of the process. Following this approach, Brockwell, Davis, and Yang (2011) considers least squares estimators with non-negative Lévy drivers while Schlemm and Stelzer (2012a) and Schlemm and Stelzer (2012b) develop Gaussian quasi maximum likelihood estimation results. As previously mentioned, in this work we will take a diametrically opposed approach by starting from continuous-time maximum likelihood estimators. In this spirit, we aim to extend the results for Gaussian CAR(p) processes presented in Brockwell, Davis, and Yang (2007a) by allowing for general Lévy drivers and rigorously proving the asymptotic properties of the discretized estimators.

Outline

The rest of the chapter is organized as follows. In Section 3.2, we define multivariate CAR models via their state-space representation. In Section 3.3, we develop continuous-time maximum likelihood theory based on fundamental results for more general stochastic processes (Jacod and Mémin, 1976; Jacod and Shiryaev, 1987; Sørensen, 1991; Küchler and Sørensen, 1997). In Section 3.4, we discretize the maximum likelihood estimators and prove the asymptotic properties are preserved. The discretizations are obtained by approximating the integrals by Riemann sums, approximating the

	Lévy driver $\mathbb{L}$ with characteristics $(\mathbf{b}, \Sigma, F)$		
	Square-integrable[1]		
	Gaussian[2]	Non-decreasing[3]	$\Sigma > 0$[4]
continuous-time	Brockwell, Davis, and Yang (2007a)		$(*)$
discrete-time (uniform)	Brockwell, Davis, and Yang (2007a)	Brockwell, Davis, and Yang (2011)	Schlemm and Stelzer (2012b), $(*)$
discrete-time (irregular)			$(*)$

Table 3.2: Estimation of CAR drift coefficients. $(*)$ contribution of this chapter.
[1] $\int \|x\|^2 F(\mathrm{d}x) < \infty$. [2] $F \equiv 0$, $\Sigma > 0$. [3] $\mathrm{supp}(F) \subseteq (0, \infty)$, $\Sigma = 0$.
[4] Σ strictly positive definite.

derivatives by finite differences, and disentangling the continuous and jump parts via thresholding. In Section 3.5, we consider the setting where a high-dimensional observation of a CAR process is known to posses a graphical structure, allowing for a parsimonious parametrization of the drift coefficients. Next, in Section 3.6, we discuss an empirical application to light curve data: we show how the variability in a star's intensity can be characterized using the spectrum of a fitted CAR model. Finally, Section 3.7 summarizes the main findings and discusses possible avenues for future work.

3.2 Multivariate continuous-time autoregressive processes

In the following we denote by $\mathcal{M}_{n,d}(\mathbb{R})$ the set of $n \times d$ matrices over $\mathbb{R}$. When working with square matrices we write $\mathcal{M}_d(\mathbb{R}) = \mathcal{M}_{d,d}(\mathbb{R})$ and denote by $\sigma(A)$ the spectrum of $A \in \mathcal{M}_d(\mathbb{R})$. For a vector space V, we denote by V^k its k-th Cartesian power, i.e. $V^k := \{(v_1, \ldots, v_k) : v_1, \ldots, v_k \in V\}$. We start by making the definition of a multivariate Lévy-driven CAR process rigorous.

Definition 3.2.1 (MCAR process)**.** Let $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ be a d-dimensional Lévy process with characteristics $(\mathbf{b}, \Sigma, F)$. Let $p \in \mathbb{N}$ and $\mathbf{A} = (A_1, \ldots, A_p) \in (\mathcal{M}_d(\mathbb{R}))^p$. Define

$$\mathcal{A}_{\mathbf{A}} = \begin{pmatrix} 0_{d\times d} & I_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} & I_{d\times d} & \cdots & 0_{d\times d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{d\times d} & 0_{d\times d} & 0_{d\times d} & \cdots & I_{d\times d} \\ -A_p & -A_{p-1} & -A_{p-2} & \cdots & -A_1 \end{pmatrix} \in \mathcal{M}_{pd}(\mathbb{R}), \quad \mathcal{E} = \begin{pmatrix} 0_{d\times d} \\ 0_{d\times d} \\ \vdots \\ 0_{d\times d} \\ I_{d\times d} \end{pmatrix} \in \mathcal{M}_{pd,d}(\mathbb{R}). \tag{3.2.1}$$

Then, for any set of matrices $A_1, \ldots, A_p \in \mathcal{M}_d(\mathbb{R})$, we define a d-dimensional multivariate continuous-time autoregressive process of order p, MCAR(p) process for short, driven by $\mathbb{L}$ with coefficient

matrices $A_1, \ldots, A_p \in \mathcal{M}_d(\mathbb{R})$ and initial state-space representation ξ, an $\mathcal{F}_0'$-measurable random variable independent of $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$, to be $\mathbb{Y} = \{\mathbf{Y}_t,\ t \geq 0\}$ such that

$$\mathbf{Y}_t = \begin{pmatrix} I_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \end{pmatrix} \mathbf{X}_t, \quad t \geq 0, \tag{3.2.2}$$

where $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ is the unique pd-dimensional solution-process to the SDE

$$\mathrm{d}\mathbf{X}_t = \mathcal{A}_{\mathbf{A}}\mathbf{X}_t\mathrm{d}t + \mathcal{E}\mathrm{d}\mathbf{L}_t, \quad \mathbf{X}_0 = \xi. \tag{3.2.3}$$

Remark 3.2.2. We stress the following important properties of the MCAR(p) process:

- The MCAR(p) process is well-defined since the stochastic differential Equation (3.2.3) has a unique solution-process $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ with the following representation:

$$\mathbf{X}_t = e^{\mathcal{A}_{\mathbf{A}}t}\xi + \int_0^t e^{\mathcal{A}_{\mathbf{A}}(t-s)}\mathcal{E}\mathrm{d}\mathbf{L}_s, \quad t \geq 0.$$

- If, in particular, we assume
  - $\mathbb{L}$ satisfies the log-moment condition $\int_{\|x\|\geq 1} \log\|x\| F(\mathrm{d}x) < \infty$,
  - the coefficient matrices are such that $\sigma(\mathcal{A}_{\mathbf{A}}) \subseteq (-\infty, 0) + i\mathbb{R}$, and
  - the initial condition ξ has law $\mathcal{L}\left(\int_0^\infty e^{\mathcal{A}_{\mathbf{A}}s}\mathcal{E}\,\mathrm{d}\mathbf{L}_s\right)$,

  then the d-dimensional MCAR(p) process $\mathbb{Y} = \{\mathbf{Y}_t,\ t \geq 0\}$ and its state-space representation $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ are strictly stationary, cf. Masuda (2004, Proposition 2.2) and Sato and Yamazato (1984, Theorem 4.1 and Theorem 4.2). If, moreover, the driving Lévy process is assumed to be r-integrable for some $r > 0$ they are also ergodic, cf. Marquardt and Stelzer (2007, Proposition 3.34).

- If we extend $\mathbb{L}$ to be a two-sided Lévy process, i.e. $\mathbb{L} = \{\mathbf{L}_t,\ t \in \mathbb{R}\}$ defined by

$$\mathbf{L}_t := \mathbf{L}_t^1 \mathbb{1}_{[0,\infty)}(t) - \mathbf{L}_{-t-}^2 \mathbb{1}_{(-\infty,0)}(t), \quad t \in \mathbb{R},$$

  where $\mathbb{L}^1 = \{\mathbf{L}_t^1,\ t \geq 0\}$ and $\mathbb{L}^2 = \{\mathbf{L}_t^2,\ t \geq 0\}$ are independent identically distributed one-sided Lévy processes, we can write the state-space representation of the strictly stationary MCAR(p) process $\mathbb{Y} = \{\mathbf{Y}_t,\ t \in \mathbb{R}\}$ as

$$\mathbf{X}_t = \int_{-\infty}^t e^{\mathcal{A}_{\mathbf{A}}(t-s)}\mathcal{E}\mathrm{d}\mathbf{L}_s, \quad t \in \mathbb{R}.$$

The process $\mathbb{X}$ is known as the pd-dimensional state-space process of $\mathbb{Y}$. Formally, the MCAR(p) process $\mathbb{Y} = \{\mathbf{Y}_t,\ t \geq 0\}$ satisfies

$$p(D)\mathbf{Y}_t = D^p\mathbf{Y}_t + A_1D^{p-1}\mathbf{Y}_t + \ldots + A_{p-1}D\mathbf{Y}_t + A_p\mathbf{Y}_t = D\mathbf{L}_t, \tag{3.2.4}$$

where $p(z) = z^p + A_1 z^{p-1} + \ldots + A_{p-1} z + A_p$ is the autoregressive polynomial and the operator D denotes differentiation with respect to t. Note that $D\mathbf{L}_t$ is in general not defined as a random function and hence we make sense of this formal equation by introducing the $\mathbb{R}^{pd}$-dimensional state-space representation $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$. Note that, due to this construction, from a realization of the CAR(p) process $\mathbb{Y}$ we can recover its state-space representation via right differentiation.

Proposition 3.2.3. *Let D denote right differentiation, i.e for a function $f : [0, \infty) \to \mathbb{R}^d$ we say f is right differentiable at $t \in [0, \infty)$ if the limit*

$$\lim_{\delta \downarrow 0} \frac{f(t+\delta) - f(t)}{\delta},$$

exists, in which case we call this the right derivative of f at $t \in [0, \infty)$ and denote it by $Df(t)$. We say f is right differentiable if it is right differentiable at every $t \in [0, \infty)$. Let $\mathbb{Y}$ be an MCAR(p) process with state-space representation $\mathbb{X} = (\mathbb{X}^{(1),\mathrm{T}}, \ldots, \mathbb{X}^{(p),\mathrm{T}})^{\mathrm{T}}$. Then we have that almost surely $\mathbb{Y}$ is $(p-1)$-times right differentiable and

$$\mathbb{X}^{(j+1)} = D^j \mathbb{Y},\ \text{for } j = 0, \ldots, p-1.$$

Proof of Proposition 3.2.3. Note that if $f(t) = \int_0^t g(s)\,\mathrm{d}s$ for a càdlàg function g (the integral is well-defined in the Lebesgue sense since g has countably many discontinuities) then f is right differentiable and at any $t \in [0, \infty)$ we have $Df(t) = g(t)$. This can be seen as follows: fix $\epsilon > 0$, and choose $\delta > 0$ such that if $s \in [t, t+\delta)$ then $|g(t) - g(s)| < \epsilon$. Then for any $0 < \delta' \leq \delta$,

$$\left| \frac{f(t+\delta') - f(t)}{\delta'} - g(t) \right| = \left| \frac{1}{\delta'} \int_t^{t+\delta'} g(s)\,\mathrm{d}s - g(t) \right| \leq \frac{1}{\delta'} \int_t^{t+\delta'} |g(s) - g(t)|\,\mathrm{d}s \leq \epsilon.$$

Thus, we have that, by definition of the MCAR process $\mathbb{Y}$,

$$\mathbf{X}_t^{(1)} = \mathbf{Y}_t,\ t \geq 0,$$

and, inductively, for $j \geq 1$

$$D^{j-1}\mathbf{Y}_t = \mathbf{X}_t^{(j)} = \int_0^t \mathbf{X}_s^{(j+1)}\,\mathrm{d}s,\ t \geq 0 \implies \mathbf{X}_t^{(j+1)} = D^j \mathbf{Y}_t,\ t \geq 0.$$

In particular, we have that $D^{p-1}\mathbb{Y}$ can have (infinitely many) jumps, since it jumps every time $\mathbb{L}$ does and by the same magnitude, while $D^j\mathbb{Y}$ is almost surely continuous for all $j < p-1$. □

Throughout this chapter, we make the following assumptions on the driving Lévy process.

Assumption 3.2.4. The driving Lévy process $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ has characteristics $(\mathbf{b}, \Sigma, F)$ and:

- is square-integrable, i.e. $\mathbb{E}[\|\mathbf{L}_1\|^2] < \infty$;

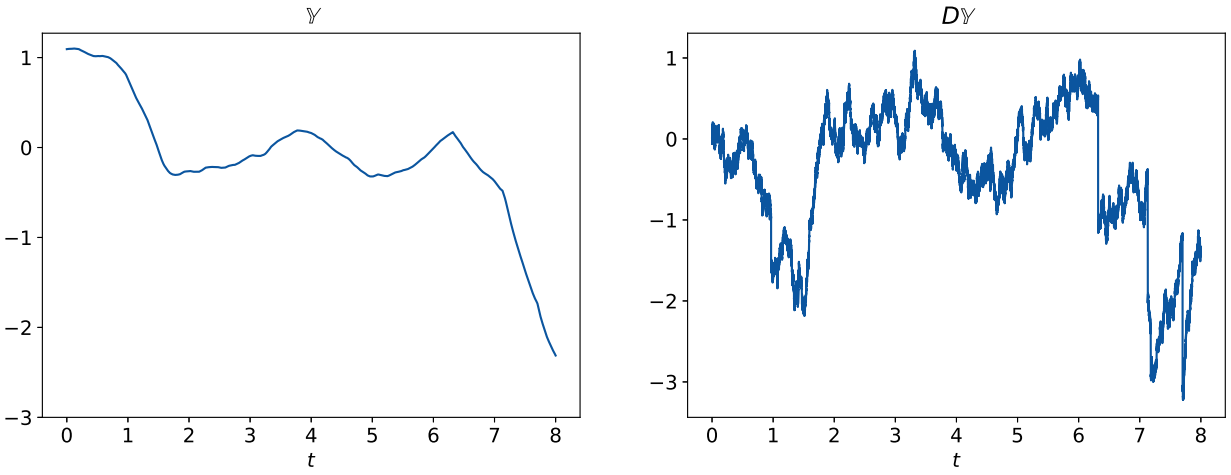


Figure 3.1: State-space path realization $\mathbb{X} = (\mathbb{Y}, D\mathbb{Y})^{\mathrm{T}}$ of a $d = 1$ dimensional MCAR process of order $p = 2$ with drift parameter $\mathbf{A} = (A_1, A_2) = (1, 2)$ and Lévy triplet $(b = 0, \Sigma = 1, F(dx) = \phi(x)dx)$ where $\phi(x)$ is a $N(0, 1)$ probability density. The path is simulated over a discrete time grid by using the exact scheme described in Appendix 3.G.2.

- has a non-degenerate Gaussian part, i.e. Σ is strictly positive definite.

Remark 3.2.5. Assuming Σ is strictly positive definite, we denote by $\Sigma^{1/2}$ its unique symmetric positive definite square root. This assumption is necessary to obtain equivalence between measures on path space and develop a likelihood theory. Note our results will thus not be applicable to non-decreasing driving Lévy processes, i.e. Lévy subordinators, which are considered instead in Brockwell, Davis, and Yang (2011).

3.3 Continuous observations: explicit maximum likelihood

In order to develop a likelihood theory for MCAR processes we first need to understand the probability measures these processes induce on the path space and how such measures are related to each other via a likelihood function. Given an observation of an MCAR process $\mathbb{Y}_{[0,t]} := \{\mathbf{Y}_s,\ s \in [0, t]\}$ one can then estimate the drift coefficients $\mathbf{A}^* \in (\mathcal{M}(\mathbb{R}^d))^p$ by the set of parameters $\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) \in (\mathcal{M}(\mathbb{R}^d))^p$ corresponding to the measure which makes the observation "most likely", i.e. which maximize the likelihood.

Fixing an initial condition for the state-space process $\mathbf{x}_0 \in \mathbb{R}^{pd}$ and a set of coefficient matrices $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$, by applying the definition of an MCAR process and Proposition 3.2.3 we note there is a one-to-one relationship between the MCAR probability measure $\mu_{\mathbf{A},\mathbf{x}_0}$ on the space of $\mathbb{R}^d$-valued paths and the state-space probability measure $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ on the space of càdlàg $\mathbb{R}^{pd}$-valued paths. Let $\mathfrak{C}^{p-1}([0, \infty); \mathbb{R}^d)$ denote the set of $(p-1)$-times right differentiable càdlàg functions so that the support of $\mu_{\mathbf{A},\mathbf{x}_0}$ is contained in $\mathfrak{C}^{p-1}([0, \infty); \mathbb{R}^d) \subseteq D([0, \infty); \mathbb{R}^d)$. Then

$$\mathbb{P}_{\mathbf{A},\mathbf{x}_0} = f_\star \mu_{\mathbf{A},\mathbf{x}_0} \text{ and } \mu_{\mathbf{A},\mathbf{x}_0} = g_\star \mathbb{P}_{\mathbf{A},\mathbf{x}_0},$$

where $f : \mathfrak{C}^{p-1}([0,\infty);\mathbb{R}^d) \to D([0,\infty);\mathbb{R}^{pd})$ maps

$$\{\mathbf{y}_t,\ t \geq 0\} \mapsto \{(\mathbf{y}_t^{\mathrm{T}}, D\mathbf{y}_t^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{y}_t^{\mathrm{T}})^{\mathrm{T}},\ t \geq 0\},$$

and $g : D([0,\infty);\mathbb{R}^{pd}) \to D([0,\infty);\mathbb{R}^d)$ maps

$$\{(\mathbf{x}_t^{(1),\mathrm{T}}, \mathbf{x}_t^{(2),\mathrm{T}}, \ldots, \mathbf{x}_t^{(p),\mathrm{T}})^{\mathrm{T}},\ t \geq 0\} \mapsto \{\mathbf{x}_t^1,\ t \geq 0\}.$$

In view of this correspondence, we can work directly with the probability measures of the state-space representation of the MCAR process, i.e. $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$. These turn out to be easier to analyze since they are solution-measures of the $\mathbb{R}^{pd}$-valued Lévy-driven OU SDE (3.2.3). The main issues in the analysis arise from the degeneracy of the driving Lévy noise $\mathcal{E}\mathbb{L} = \{\mathcal{E}\mathbf{L}_t,\ t \geq 0\}$.

3.3.1 Deriving the likelihood

For a parameter $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$ consider the pd-dimensional state-space representation $\mathbb{X}_{\mathbf{A},\mathbf{x}_0} = \{\mathbf{X}_{\mathbf{A},\mathbf{x}_0,t},\ t \geq 0\}$ of a MCAR(p) process driven by a Lévy process $\mathbb{L}$ with characteristics $(\mathbf{b}, \Sigma, F)$, parameter $\mathbf{A}$ and deterministic initial condition $\xi = \mathbf{x}_0$ on an arbitrary stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}_t',\ t \geq 0\}, \mathbb{P}')$. This satisfies the stochastic differential Equation (3.2.3) which, using the Lévy-Itô representation of $\mathbb{L}$, we can write as

$$\begin{aligned} \mathrm{d}\mathbf{X}_t = (\mathcal{A}_{\mathbf{A}}\mathbf{X}_t + \mathcal{E}\mathbf{b})\mathrm{d}t + \mathcal{E}\Sigma^{1/2}\mathrm{d}\mathbf{W}_t \\ + \int_{\|z\|\leq 1} \mathcal{E}z\{\mu(\mathrm{d}t, \mathrm{d}z) - F(\mathrm{d}z)\mathrm{d}t\} + \int_{\|z\|>1} \mathcal{E}z\mu(\mathrm{d}t, \mathrm{d}z), \end{aligned} \tag{3.3.1}$$

for a d-dimensional standard Wiener process $\mathbb{W} = \{\mathbf{W}_t,\ t \geq 0\}$ and a Poisson random measure $\mu(\mathrm{d}t, \mathrm{d}z)$ on $\mathbb{R}_+ \times \mathbb{R}^d \setminus \{0\}$ with compensator $F(\mathrm{d}z)\mathrm{d}t$. It is easy to check that the coefficients of the SDE satisfy local Lipschitz and linear growth conditions. We can thus combine Jacod and Shiryaev (1987, Theorem II.2.32 and Theorem II.2.33) to deduce that $\forall \mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ there exists a unique solution-measure of the SDE (3.3.1) on the canonical state-space $(\Omega = D([0,\infty), \mathbb{R}^{pd}), \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$ for any initial condition $\mathbf{x}_0 \in \mathbb{R}^{pd}$. We denote this measure by $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$. Under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ the canonical process $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ defined by $\mathbf{X}_t(\omega) = \omega(t),\ \forall t \geq 0$ is a solution of SDE (3.3.1).

Remark 3.3.1. We note that $c_t := \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}$ is not strictly positive definite and thus Sørensen (1991, Condition C) (which is the same as the hypothesis of Jacod and Shiryaev (1987, Theorem II.2.34)) does not hold. This condition is used by Sørensen (1991) to show uniqueness of the solution-measure to SDEs for more general diffusions with jumps.

In the analysis of measures induced by general semimartingales, such as $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$, the characteristics (Jacod and Shiryaev, 1987, Definition II.2.6) provide a valuable tool. It is a standard result (Jacod and Shiryaev, 1987, Theorem III.2.26) that the set of all solution-measures to SDE (3.3.1) with initial condition $\mathbf{X}_0 = \mathbf{x}_0$ is the set of solutions to the martingale problem on the canonical space

$(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$ with deterministic initial condition $\eta = \delta_{\mathbf{x}_0}$ and semimartingale characteristics $(\mathbf{B}_{\mathbf{A}}, C, \nu)$ where for $\omega \in \Omega$, $t \geq 0$

$$\begin{aligned} \mathbf{B}_{\mathbf{A},t}(\omega) &= \int_0^t (\mathcal{A}_{\mathbf{A}}\omega(s) + \mathcal{E}\mathbf{b})\mathrm{d}s = \mathcal{A}_{\mathbf{A}} \int_0^t \omega(s)\mathrm{d}s + \mathcal{E}bt, \\ C_t(\omega) &= \int_0^t \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}\mathrm{d}s = \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}t, \\ \nu(\omega; \mathrm{d}t, \mathrm{d}x) &= \mathrm{d}t \int_{\mathbb{R}^d} \mathbb{1}_{\{\mathrm{d}x\}}(\mathcal{E}z)F(\mathrm{d}z) = \mathrm{d}t\, \delta_0(\mathrm{d}x_{-p})\, F(\mathrm{d}x_p), \end{aligned} \tag{3.3.2}$$

where we write $x = (x_1^{\mathrm{T}}, \ldots, x_p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{pd}$ and $x_{-p} = (x_1^{\mathrm{T}}, \ldots, x_{p-1}^{\mathrm{T}})^{\mathrm{T}}$, for $x_1, \ldots, x_p \in \mathbb{R}^d$. Under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ the canonical process $\mathbb{X}$ is a diffusion with jumps with local characteristics $(\mathbf{B}_{\mathbf{A}}, C, \nu)$.

Remark 3.3.2. See Jacod and Shiryaev (1987, Definition II.2.4 and Definition II.2.24) for the definitions of solutions to martingale problems and solution-measures. Two remarks:

- Note that a solution-measure for an SDE as (3.3.1) defined on a general stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \geq 0\}, \mathbb{P}')$ is a probability measure on the canonical space $(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$. As we vary $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ in (3.3.1) the original stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \geq 0\}, \mathbb{P}')$ does not change but the solution measure $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ on $(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$ does indeed.
- Note that solution-measures and solutions to martingale problems depend on the initial condition, i.e. the initial distribution $\eta(\cdot) = \mathbb{P}'(\mathbf{X}_0 \in \cdot)$. In the case where this is non-deterministic the canonical space needs to be enlarged so that $\mathcal{F}_0$ contains the σ-algebra generated by $\mathbf{X}_0$. In the following though, we will only need to consider deterministic initial conditions $\mathbf{x}_0 \in \mathbb{R}^{pd}$.

Under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$, $\mathbb{X}^c_{\mathbf{A},\mathbf{x}_0} = \{\mathbf{X}^c_{\mathbf{A},\mathbf{x}_0,t},\ t \geq 0\}$ defined for $t \geq 0$ by

$$\mathbf{X}^c_{\mathbf{A},\mathbf{x}_0,t} = \mathbf{X}_t - \mathbf{x}_0 - \sum_{s \leq t} \Delta\mathbf{X}_s \mathbb{1}_{\{\|\Delta\mathbf{X}_s\| \geq 1\}} - \mathbf{B}_{\mathbf{A},t} - \int_0^t \int_{\|x\| \leq 1} x(\mu_{\mathbb{X}} - \nu)(\mathrm{d}t, \mathrm{d}x), \tag{3.3.3}$$

is a continuous local martingale, where μ is the jump measure over $\mathbb{R}_+ \times \mathbb{R}^{pd} \setminus \{0\}$ associated with the canonical process $\mathbb{X}$, i.e.

$$\mu_{\mathbb{X}}(\omega; \mathrm{d}t, \mathrm{d}x) = \sum_{s \geq 0 : \Delta\mathbf{X}_s(\omega) \neq 0} \delta_{(s, \Delta\mathbf{X}_s(\omega))}(\mathrm{d}t, \mathrm{d}x), \tag{3.3.4}$$

and $\Delta\mathbb{X} = \{\Delta\mathbf{X}_t,\ t \geq 0\}$ is the jump process of $\mathbb{X}$, i.e. $\Delta\mathbf{X}_t = \mathbf{X}_t - \mathbf{X}_{t-}$. Note that by construction of the state-space representation of the MCAR(p) process, jumps will only occur in the last d entries of $\mathbb{X}$. Let

$$\mathbf{A}^{(0)} = (0_{d\times d}, \ldots, 0_{d\times d}) \in (\mathcal{M}_d(\mathbb{R}))^p,$$

and note that the jump measure ν does not depend on $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$, i.e. $Y_{\mathbf{A}}(t, x, y) = 1$ in Sørensen (1991) and Jacod and Mémin (1976).

Proposition 3.3.3. *Assume Assumption 3.2.4 holds. Fix $t > 0$ and let $\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}$ denote the restrictions of $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ to $\mathcal{F}_t$ respectively. Then $\forall \mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$*

$$\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0} \sim \mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0},$$

and $\mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}$-almost surely

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}} &= \mathbb{E}^{\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\middle|\mathcal{F}_t\right] \\ &= \exp\Big\{-\int_0^t \mathbf{X}^{\mathrm{T}}_{s-}\tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\,\mathrm{d}\mathbf{X}^{c}_{\mathbf{A}^{(0)},s} - \frac{1}{2}\int_0^t \mathbf{X}^{\mathrm{T}}_s\tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{X}_s\mathrm{d}s\Big\},\end{aligned} \tag{3.3.5}$$

where

$$\tilde{\mathcal{A}}_{\mathbf{A}} = \begin{pmatrix} 0_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ -A_p & -A_{p-1} & \cdots & -A_1 \end{pmatrix}, \tag{3.3.6}$$

and

$$\tilde{\Sigma}^{-1} := \mathcal{E}\Sigma^{-1}\mathcal{E}^{\mathrm{T}} = \begin{pmatrix} 0_{d\times d} & \cdots & 0_{d\times d} & 0_{d\times d} \\ \vdots & \ddots & \vdots & \vdots \\ 0_{d\times d} & \cdots & 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & \cdots & 0_{d\times d} & \Sigma^{-1} \end{pmatrix}. \tag{3.3.7}$$

Proof of Proposition 3.3.3. We follow a similar proof strategy as that in Sørensen (1991, Theorem 2.1), relying on the results of Jacod and Mémin (1976) and Jacod and Shiryaev (1987), but we note that our process does not satisfy Sørensen (1991, Condition C) since $c_t = \tilde{\Sigma} = \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}$ is not strictly positive definite. The outline of the proof is as follows:

- We start by considering the state-space representation of an MCAR process $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}$ satisfying (3.3.1) on an arbitrary stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \geq 0\}, \mathbb{P}')$ and the associated unique solution-measure $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ on the canonical state-space $(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$, i.e. the unique solution to the martingale problem $(\mathbb{X}, (\mathbf{B}_{\mathbf{A}}, C, \nu), \delta_{\mathbf{x}_0})$.

- Next, fixing $t > 0$, we construct the unique solution $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ to the martingale problem with stopped characteristics $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$, i.e. the solution measure of SDE (3.3.1) stopped at $t > 0$. By construction, $\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0} = \mathbb{Q}^t_{\mathbf{A},\mathbf{x}_0}$.

- Fixing a reference measure corresponding to the parameter $\mathbf{A}^{(0)} = (0_{d\times d}, \ldots, 0_{d\times d})$, we then apply Jacod and Mémin (1976, Theorem 4.2.(c)) to $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0} \in (\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0} \in (\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}^{(0)}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ to deduce equivalence of the measures. This step crucially

relies on the specific structure of the problem as we will be able to write

$$\mathbf{B}^t_{\mathbf{A},s} = \mathbf{B}^t_{\mathbf{A}^{(0)},s} + \int_0^s \mathrm{d}C^t_u \mathbf{Z}_{\mathbf{A},u}, \tag{3.3.8}$$

where $\mathbb{Z}_{\mathbf{A}} = \{\mathbf{Z}_{\mathbf{A},s},\ s \geq 0\}$ is the predictable process defined by $\mathbf{Z}_{\mathbf{A},s} = \tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{X}_{s-}$. To apply the theorem we check the assumptions:

1. "τ_n-uniqueness of the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$". We prove this by showing the stronger local uniqueness property. In Sørensen (1991), under the assumption that c_t is strictly positive definite, local uniqueness is deduced by a simple application of Jacod and Shiryaev (1987, Corollary III.2.41). In our setting though this property does not hold and we need to work a bit harder to prove local uniqueness. We check directly the conditions of the more general result Jacod and Shiryaev (1987, Theorem III.2.40) for the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$:
   (a) Jacod and Shiryaev (1987, Assumption III.2.13) is satisfied because we are working on the canonical space.
   (b) Jacod and Shiryaev (1987, Assumption III.2.39) is related to the Markovian structure of our martingale problem.
   (c) The solution to the martingale problem defines a transition kernel as the characteristics are shifted in time and the initial condition is shifted in space: this is obtained via an application of the $\pi - \lambda$ theorem.
   (d) Uniqueness of the original martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ has already been checked.
2. Equivalence of initial conditions: we work with fixed initial condition $\delta_{\mathbf{x}_0}$.
3. The random variable $A^t_{\mathbf{A},\infty}$, given by the limit as $s \uparrow \infty$ of

$$A^t_{\mathbf{A},s} = \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\mathrm{d}C^t_u\mathbf{Z}_{\mathbf{A},u} = \int_0^{s\wedge t} \mathbf{X}^{\mathrm{T}}_u \tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{X}_u\mathrm{d}u, \tag{3.3.9}$$

   is $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ a.s. finite. This follows from bounds obtained from the assumption of square integrability of the driving Lévy process and explicit formulas for the second moments of $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}$.

- Finally, we apply Jacod and Mémin (1976, Theorem 4.5.(b)) to deduce the form of the conditional Radon–Nikodym derivative of $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ w.r.t. $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$. We check the assumptions:

1. $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ satisfies the martingale representation condition as it is the unique solution of the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$, this follows from Jacod and Mémin (1976, Theorem 4.4).

2. $\mathbb{A}^t_{\mathbf{A}} = \{A^t_{\mathbf{A},s},\ s \geq 0\}$ is $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ a.s. continuous and its limit is finite: this simplifies the form of the conditional Radon–Nikodym process as we can drop any dependency on the localizing sequence $(\tau_n)_n$.

For full details, see Appendix 3.C.1. □

Remark 3.3.4. We make some important remarks on the proof:

- The proof crucially relies on being able to write Equation (3.3.8), for which we had to leverage the specific forms of $\mathcal{A}_{\mathbf{A}}$ and $\tilde{\Sigma}$. Our proof, therefore, does not immediately generalize to OU processes with arbitrary matrix coefficient $\mathcal{A}$ and non-negative definite Gaussian covariance $\tilde{\Sigma}$ of the driving Lévy process. In the particular case where $\tilde{\Sigma}$ is strictly positive definite, though, it is a standard result that the same result holds for any matrix coefficient $\mathcal{A}$.

- In the proof it is essential to stop the SDE (3.3.1) at $t > 0$ since, if instead we attempt to apply the theorems by Jacod and Mémin (1976) directly to $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$, then we cannot guarantee that $A_{\mathbf{A},\infty} < \infty$, $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$- and $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$-almost surely, where for $t \geq 0$

$$A_{\mathbf{A},s} := \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u} \mathrm{d}C_u \mathbf{Z}_{\mathbf{A},u} = \int_0^s \mathbf{X}^{\mathrm{T}}_u \tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}} \tilde{\Sigma}^{-1} \tilde{\mathcal{A}}_{\mathbf{A}} \mathbf{X}_u \mathrm{d}u.$$

Remark 3.3.5. The proof of the previous statement can be alternatively obtained by applying Girsanov's theorem to the background Lévy process directly. Starting with a background probability measure $\mathbb{P}'_{\mathbf{0}}$ on a filtered probability space $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \in [0,T]\})$ with respect to which the process $\{\mathbf{L}_t,\ t \in [0,T]\}$ is a Lévy process with Lévy triplet $(\mathbf{b}, \Sigma, F)$, we define the process $\{\mathbf{X}_t,\ t \in [0,T]\}$ as the unique strong solution to

$$\mathrm{d}\mathbf{X}_t = \mathcal{A}_{\mathbf{0}} \mathbf{X}_t \mathrm{d}t + \mathcal{E} \mathrm{d}\mathbf{L}_t, \quad \mathbf{X}_0 = \mathbf{x}_0,$$

i.e. the state space representation of the MCAR process of order p with coefficients $\mathbf{0} \in (\mathcal{M}(\mathbb{R}^d))^p$. For any $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$, we can then apply Girsanov's theorem (Jacod and Shiryaev, 1987, Section III.5) to define the measure $\mathbb{P}'_{\mathbf{A}}$ such that $\mathbb{P}'_{\mathbf{A}} \sim \mathbb{P}'_{\mathbf{0}}$ and

$$\mathbb{E}'_{\mathbf{0}}\left[\frac{\mathrm{d}\mathbb{P}'_{\mathbf{A}}}{\mathrm{d}\mathbb{P}'_{\mathbf{0}}}\middle|\mathcal{F}_T\right] = \exp\Big\{-\int_0^T \langle \mathbf{D}^{\mathbf{A}}_t, \Sigma^{1/2} \mathrm{d}\mathbf{W}_t\rangle - \frac{1}{2}\int_0^T \langle \mathbf{D}^{\mathbf{A}}_t, \Sigma \mathbf{D}^{\mathbf{A}}_t\rangle \mathrm{d}t\Big\},$$

where $\{\mathbf{W}_t,\ t \in [0,T]\}$ is the Brownian motion associated to $\{\mathbf{L}_t,\ t \in [0,T]\}$ under $\mathbb{P}_{\mathbf{0}}$ and the drift is

$$\mathbf{D}^{\mathbf{A}}_t = \Sigma^{-1} \mathcal{E}^{\mathrm{T}} \tilde{\mathcal{A}}_{\mathbf{A}} \mathbf{X}_{t-}, \quad t \in [0,T].$$

Under $\mathbb{P}'_{\mathbf{A}}$, the process

$$\mathbf{L}^{\mathbf{A}}_t = \mathbf{L}_t - \int_0^t \Sigma\, \mathbf{D}^{\mathbf{A}}_s \,\mathrm{d}s, \quad t \in [0,T],$$

is a Lévy process with characteristic triplet $(\mathbf{b}, \Sigma, F)$ and hence $\{\mathbf{X}_t,\ t \in [0,T]\}$ satisfies

$$\mathrm{d}\mathbf{X}_t = (\mathcal{A}_{\mathbf{0}}\mathbf{X}_t + \mathcal{E}^{\mathrm{T}}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{X}_{t-})\mathrm{d}t + \mathcal{E}\mathrm{d}\mathbf{L}_t^{\mathbf{A}} = \mathcal{A}_{\mathbf{A}}\mathbf{X}_t\mathrm{d}t + \mathcal{E}\mathrm{d}\mathbf{L}_t^{\mathbf{A}}, \quad \mathbf{X}_0 = \mathbf{x}_0,$$

i.e. it is the state space representation of the MCAR process of order p with coefficients $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$. The statement of Proposition 3.3.3 is then obtained by mapping $\mathbb{P}'_{\mathbf{0}}$ and $\mathbb{P}'_{\mathbf{A}}$ to the corresponding path space probability measures. In this work we keep the path space construction as we believe it provides a natural setup when developing estimation results (we assume the observations are samples of the path space Ω). Moreover, the tools used in the proof of Proposition 3.3.3 are slightly more general, providing some additional insight into likelihood theory over path space.

3.3.2 Maximum likelihood estimation

Note that Proposition 3.3.3 yields the likelihood of parameter $\mathbf{A}$ with respect to reference parameter $\mathbf{A}^{(0)}$ as the Radon-Nikodym derivative of two measures on the canonical space of $\mathbb{R}^{pd}$-valued paths $(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$ restricted to $\mathcal{F}_t$, that is an $\mathcal{F}_t$-measurable random variable. This is thus a mapping from paths in $D([0,t];\mathbb{R}^{pd})$ to $\mathbb{R}$, i.e. a function of $\mathbb{X}_{[0,t]}$. Setting $\mathbb{Y}_{[0,t]} = \{\mathbf{Y}_s,\ s \in [0,t]\} = \mathbb{X}^{(1)}_{[0,t]}$, understood as an observation of the MCAR process, in view of Proposition 3.2.3 we can write the likelihood (3.3.5) as a function of $\mathbb{Y}_{[0,t]}$ (and its right derivatives). This is formalized by the following lemma in which we define the vectorization operator for elements $\mathbf{V} \in (\mathcal{M}(\mathbb{R}^d))^p$ by $\mathrm{vec} : (\mathcal{M}(\mathbb{R}^d))^p \mapsto \mathbb{R}^{pd^2}$ as

$$\mathbf{V} = (V_1, \ldots, V_p) \mapsto \mathrm{vec}(\mathbf{V}) := (V_1^{(1,1)}, \ldots, V_1^{(d,1)}, \ldots, V_1^{(1,d)}, \ldots, V_1^{(d,d)}, V_2^{(1,1)}, \ldots, V_p^{(d,d)})^{\mathrm{T}}.$$

Lemma 3.3.6. *Assume Assumption 3.2.4 holds. Let $\mathbb{Y}_{[0,t]} = \{\mathbf{Y}_s,\ s \in [0,t]\}$ be the observation of a MCAR(p) process. Suppose the process starts in $\mathbf{y}_0 \in \mathbb{R}^d$ with initial derivatives $D^1\mathbf{y}_0, \ldots, D^{p-1}\mathbf{y}_0 \in \mathbb{R}^d$, i.e. $\mathbf{x}_0 = (\mathbf{y}_0^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{y}_0^{\mathrm{T}})^{\mathrm{T}}$. Let*

$$D^{p-1}\mathbb{Y}^c_{\mathbf{A}^{(0)},[0,t]} = \{D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s}\ s \in [0,t]\},$$

be the last d entries of $\mathbb{X}^c_{\mathbf{A}^{(0)},\mathbf{x}_0}$, the continuous martingale part of the state-space representation with parameter $\mathbf{A}^{(0)} = (0_{d\times d}, \ldots, 0_{d\times d})$ as defined in (3.3.3) – note we drop dependence on $\mathbf{x}_0$ as $D^{p-1}\mathbb{Y}^c_{\mathbf{A}^{(0)},[0,t]}$ will only appear as an integrator and thus the initial condition becomes superfluous. Then the likelihood function of the MCAR(p) parameters $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$ given the observation $\mathbb{Y}_{[0,t]}$ is

$$\mathcal{L}(\mathbf{A}; \mathbb{Y}_{[0,t]}) := \mathbb{E}^{\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\middle|\mathcal{F}_t\right] = \exp\left\{\mathrm{vec}(\mathbf{A})^{\mathrm{T}}\mathbf{H}_t - \frac{1}{2}\mathrm{vec}(\mathbf{A})^{\mathrm{T}}[\mathbf{H}]_t\mathrm{vec}(\mathbf{A})\right\}, \quad (3.3.10)$$

where

$$\mathbf{H}_t = -\begin{pmatrix} \int_0^t D^{p-1}Y_{s-}^{(1)}\Sigma^{-1}\mathrm{d}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} \\ \vdots \\ \int_0^t D^{p-1}Y_{s-}^{(d)}\Sigma^{-1}\mathrm{d}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} \\ \vdots \\ \int_0^t Y_{s-}^{(1)}\Sigma^{-1}\mathrm{d}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} \\ \vdots \\ \int_0^t Y_{s-}^{(d)}\Sigma^{-1}\mathrm{d}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} \end{pmatrix}, \tag{3.3.11}$$

with quadratic covariation matrix

$$[\mathbf{H}]_t = \begin{pmatrix} \int_0^t D^{p-1}Y_{s-}^{(1)}\Sigma^{-1}D^{p-1}Y_{s-}^{(1)}\,\mathrm{d}s & \cdots & \int_0^t D^{p-1}Y_{s-}^{(1)}\Sigma^{-1}Y_{s-}^{(d)}\,\mathrm{d}s \\ \vdots & \ddots & \vdots \\ \int_0^t Y_{s-}^{(d)}\Sigma^{-1}D^{p-1}Y_{s-}^{(1)}\,\mathrm{d}s & \cdots & \int_0^t Y_{s-}^{(d)}\Sigma^{-1}Y_{s-}^{(d)}\,\mathrm{d}s \end{pmatrix}. \tag{3.3.12}$$

Proof of Lemma 3.3.6. The form of the likelihood follows by expanding (3.3.5) and grouping in terms of $A_k^{(i,j)}$ for $k = 1,\dots,p$ and $i,j = 1,\dots,d$. To determine that $[\mathbf{H}]_t$ is indeed the quadratic covariation of $\mathbb{H} = \{\mathbf{H}_t,\ t \geq 0\}$ we use the fact that $[D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}]_t = \Sigma t$ and the multidimensional version of the Kunita-Watanabe identity (Protter, 1990, Theorem II.29), i.e. if $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$, $\mathbb{Z} = \{\mathbf{Z}_t,\ t \geq 0\}$ are semimartingales and $\{\mathbf{G}_t,\ t \geq 0\}$, $\{\mathbf{K}_t,\ t \geq 0\}$ are càglàd adapted processes then

$$\left[\int_0^\cdot \mathbf{G}_s\mathrm{d}\mathbf{X}_s, \int_0^\cdot \mathbf{K}_s\mathrm{d}\mathbf{Z}_s\right]_t = \int_0^t \mathbf{G}_s\mathrm{d}[\mathbf{X},\mathbf{Z}]_s\mathbf{K}_s^{\mathrm{T}}.$$

□

Remark 3.3.7. Note that, given the observation $\mathbb{Y}_{[0,t]}$, the process $D^{p-1}\mathbb{Y}^c_{\mathbf{A}^{(0)},[0,t]}$ can be expressed as

$$\begin{aligned} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} = D^{p-1}\mathbf{Y}_s - D^{p-1}\mathbf{y}_0 - \mathbf{b}s - \sum_{0\leq u\leq s} \Delta D^{p-1}\mathbf{Y}_u \mathbb{1}_{\{\|\Delta D^{p-1}\mathbf{Y}_u\|\geq 1\}} \\ - \int_0^s \int_{\|y\|\leq 1} x\{\tilde{\mu}(\mathrm{d}u,\mathrm{d}y) - \mathrm{d}u\,F(\mathrm{d}y)\}, \quad s \in [0,t], \end{aligned} \tag{3.3.13}$$

where $\tilde{\mu}$ is the jump measure of $D^{p-1}\mathbb{Y}_{[0,t]} = \{D^{p-1}\mathbf{Y}_s,\ s \in [0,t]\}$, i.e.

$$\tilde{\mu}(\mathrm{d}s,\mathrm{d}y) = \sum_{0\leq u\leq t:\Delta D^{p-1}\mathbf{Y}_u\neq 0} \delta_{(u,\Delta D^{p-1}\mathbf{Y}_u)}(\mathrm{d}s,\mathrm{d}y).$$

Under $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$, this process is a Brownian motion with covariance Σ. For any other $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ we can write

$$D^{p-1}\mathbf{Y}^c_{\mathbf{A},s} = D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} + \sum_{j=1}^p \int_0^s A_j D^{p-j}\mathbf{Y}_u\,\mathrm{d}u, \quad s \in [0,t],$$

which is a Brownian motion with covariance Σ under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$.

We can now use the explicit form of the likelihood for the MCAR model to obtain a maximum likelihood estimator when continuous-time observations are available. In order to prove its asymptotic properties, we follow the approach outlined in Küchler and Sørensen (1997, Section 11.7), which first requires establishing the following result for the score process.

Lemma 3.3.8. *Under Assumption 3.2.4, for any* $\mathbf{A} \in \mathcal{M}_d(\mathbb{R})$ *the score*

$$\nabla_{\mathrm{vec}(\mathbf{A}^*)} \log \mathcal{L}(\mathbf{A}; \mathbb{Y}_{[0,t]}) = \mathbf{H}_t - [\mathbf{H}]_t \mathrm{vec}(\mathbf{A}^*), \ t \geq 0, \tag{3.3.14}$$

defines a mean-zero continuous square-integrable $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$*-martingale with quadratic variation* $\{[\mathbf{H}]_t, \ t \geq 0\}$.

Proof of Lemma 3.3.8. See Appendix 3.C.2. □

The main results of this section are given in the following theorem which establishes the existence and uniqueness of the continuous-time maximum likelihood estimator as well as its consistency and asymptotic normality.

Theorem 3.3.9. *Assume Assumption 3.2.4 holds and let*

$$\mathfrak{A} = \{\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p : \sigma(\mathcal{A}_\mathbf{A}) \subseteq (-\infty, 0) + i\mathbb{R}\}. \tag{3.3.15}$$

Define the maximum likelihood estimator

$$\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) = \mathrm{argmax}_{\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p} \mathcal{L}(\mathbf{A}; \mathbb{Y}_{[0,t]}),$$

then the MLE exists and is uniquely defined $\forall t > 0$ *by*

$$\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) = \mathrm{vec}^{-1}([\mathbf{H}]_t^{-1} \mathbf{H}_t), \tag{3.3.16}$$

where $\mathbf{H}_t$ *and* $[\mathbf{H}]_t$ *are given in Equation (3.3.11) and (3.3.12) respectively. Moreover, if* $\mathbf{A}^* \in \mathfrak{A}$ *then the MLE is*

- *(weakly) consistent, i.e.*
$$\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) \xrightarrow{\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}} \mathbf{A}^*, \quad t \to \infty;$$
- *asymptotically normal, i.e.*
$$\sqrt{t}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$
where $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ *is a symmetric positive definite matrix such that*
$$t^{-1}[\mathbf{H}]_t \xrightarrow{\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}} \mathcal{H}_\infty, \quad t \to \infty.$$

Remark 3.3.10. The asymptotic normality result immediately yields a feasible version of the limit

$$[\mathbf{H}]_t^{1/2}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, I_{pd^2}), \quad t \to \infty.$$

Remark 3.3.11. We can quite easily extend these results to the setting where the initial conditions for $\mathbb{Y}_{[0,t]}$ and its first $p-1$ derivatives are assumed to be random with distribution given by the stationary distribution of the MCAR(p) state-space representation with parameters $\mathbf{A}^* \in \mathfrak{A}$, i.e.

$$\mathcal{L}([\mathbf{Y}_0^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{Y}_0^{\mathrm{T}}]^{\mathrm{T}}) = \mathcal{L}\left(\int_0^\infty e^{s\mathcal{A}_{\mathbf{A}^*}}\mathcal{E}\,\mathrm{d}\mathbf{L}_s\right).$$

Under such measure $\mathbb{P}_{\mathbf{A}^*}$ the observed MCAR(p) process $\mathbb{Y}$ and its state-space representation are stationary and ergodic, cf. Remark 3.2.2. All other assumptions withstanding, we can show consistency and asymptotic normality hold under $\mathbb{P}_{\mathbf{A}^*}$ by first showing Lemma 3.3.8 also holds under $\mathbb{P}_{\mathbf{A}^*}$ and then proceeding as in Appendix 3.C.3.

Remark 3.3.12. Strong consistency of the MLE may be obtainable by applying a martingale strong law of large numbers, e.g., Sørensen (1991, Theorem A.7.6). In the following section, as we progress toward our ultimate goal of developing a feasible estimator based on discrete observations of the process, we will only be able to show that subsequent approximations preserve weak consistency (cf. Lemma 3.4.3). Therefore, it suffices for our purposes to establish weak consistency of the continuous-time MLE.

Proof of Theorem 3.3.9. The proof follows standard arguments for exponential families of stochastic processes, see Küchler and Sørensen (1997). The log-likelihood

$$\ell(\mathrm{vec}(\mathbf{A}); \mathbb{Y}_{[0,t]}) = \mathrm{vec}(\mathbf{A})^{\mathrm{T}}\mathbf{H}_t - \frac{1}{2}\mathrm{vec}(\mathbf{A})^{\mathrm{T}}[\mathbf{H}]_t\mathrm{vec}(\mathbf{A}),$$

is a quadratic form and $[\mathbf{H}]_t$ is $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$-a.s. symmetric positive definite. Note that we can make pointwise sense of the (log)likelihood by using the pathwise construction of the stochastic integral given in Karandikar (1995) for the integral[1] appearing in $\mathbf{H}_t$, ensuring we can define the MLE pointwise (i.e. for any $\omega \in \Omega$). We can thus write

$$\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})) = [\mathbf{H}]_t^{-1}\mathbf{H}_t = \mathrm{vec}(\mathbf{A}^*) + [\mathbf{H}]_t^{-1}(\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*)),$$

where the score $\{\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*),\ t \geq 0\}$ is a mean-zero continuous square-integrable $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$-martingale with quadratic variation $[\mathbf{H}]_t$ by Lemma 3.3.8. The desired results then follow by martingale CLTs. For the full details, see Appendix 3.C.3. □

[1] The integrand is the left-limit process of a càdlàg process and hence Karandikar (1995, Theorem 2) applies.

3.4 Discrete observations: high-frequency estimators

In the previous section, we worked under the unrealistic assumption of a continuous-time observation of the MCAR process $\mathbb{Y}_{[0,t]} = \{\mathbf{Y}_s,\ s \in [0,t]\}$. Instead, we now assume that we observe the process on a discrete (possibly irregularly spaced) time grid. For each $t > 0$, let $N_t \in \mathbb{N}$ and

$$\mathcal{P}_t := \{0 = s_0 < s_1 < \ldots < s_{N_t} = t\},$$

be a partition of $[0,t]$ with mesh $\Delta_{\mathcal{P}_t} := \sup_{0 \le i < N_t}(s_{i+1} - s_i)$. The discretely observed process is then given by

$$\mathbb{Y}_{\mathcal{P}_t} := \{\mathbf{Y}_s,\ s \in \mathcal{P}_t\}.$$

As before, we assume the driving Lévy process parameters $(\mathbf{b}, \Sigma, F)$ are given. In this section, we work under the probability measure $\mathbb{P}_{\mathbf{A}^*}$ corresponding to a stationary and ergodic MCAR(p) process with parameter $\mathbf{A}^* \in \mathfrak{A}$. We start by noting that the maximum likelihood estimator

$$\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) = \mathrm{vec}^{-1}([\mathbf{H}]_t^{-1}\mathbf{H}_t),$$

is a function of integrals of the form

$$\int_0^t D^l Y_{s-}^{(i)} \mathrm{d}D^{p-1}Y_{\mathbf{A}^{(0)},s}^{c,(j)} \text{ and } \int_0^t D^l Y_s^{(i)} D^k Y_s^{(j)}\,\mathrm{d}s, \tag{3.4.1}$$

for $i, j \in \{1, \ldots, d\}$ and $l, k \in \{0, \ldots, p-1\}$ where

$$\mathrm{d}D^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)},s}^{c} = \mathrm{d}D^{p-1}\mathbf{Y}_s - \mathbf{b}\,\mathrm{d}s - \mathrm{d}\mathbf{J}_s - \mathrm{d}\mathbf{M}_s, \quad s \in [0,t], \tag{3.4.2}$$

and the jump process $\mathbb{J}_{[0,t]} = \{\mathbf{J}_s,\ s \in [0,t]\}$ and $\mathbb{M}_{[0,t]} = \{\mathbf{M}_s,\ s \in [0,t]\}$ are defined by

$$\mathbf{J}_s = \sum_{0 \le u \le s} \Delta D^{p-1}\mathbf{Y}_u \mathbb{1}_{\{\|\Delta D^{p-1}\mathbf{Y}_u\| \ge 1\}}, \tag{3.4.3}$$

$$\mathbf{M}_s = \int_0^s \int_{\|y\| \le 1} y\{\tilde{\mu}(\mathrm{d}u, \mathrm{d}y) - \mathrm{d}u\, F(\mathrm{d}y)\}. \tag{3.4.4}$$

To write an estimator based only on $\mathbb{Y}_{\mathcal{P}_t}$ we thus need to make the following approximations:

1. Approximate the integrals appearing in Equation (3.4.1) by Riemann sums computed using discretized observations of the processes, i.e. replace

$$\int_0^t *_s \mathrm{d}\,\cdot_s \text{ with } \sum_{n=0}^{N_t-1} *_{s_n}(\cdot_{s_{n+1}} - \cdot_{s_n});$$

2. Approximate the time derivatives in Equation (3.4.1) and Equation (3.4.2) by finite differences,

i.e. replace

$$D^j \mathbf{Y}_{s_n} \text{ with } \hat{D}^j \mathbf{Y}_{s_n} \text{ for } j = 1, \dots, p-1;$$

3. Filter out the discretized jump increments $\Delta^n_{\mathcal{P}_t}\mathbf{J}$ and $\Delta^n_{\mathcal{P}_t}\mathbf{M}$ from $\Delta^n_{\mathcal{P}_t}\hat{D}^{p-1}\mathbf{Y}$ in Equation (3.4.2) via thresholding, i.e. replace

$$\Delta^n_{\mathcal{P}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \text{ with } \left[\Delta^n_{\mathcal{P}_t}\hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^n_{\mathcal{P}_t}\right] \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^n_{\mathcal{P}_t}\right| \le \boldsymbol{\nu}^n_t\right\}}.$$

As we will see in Section 3.4.2, when introducing finite difference approximations, we will need to work on two granularity levels: a finer partition $\mathcal{P}_t$ to approximate the derivatives and a coarser partition $\mathcal{Q}_t$ to approximate the integrals and filter the jumps. Thus, throughout this section, $\mathcal{P}_t$ will denote the finest grid on which observations are available and, when required, $\mathcal{Q}_t$ an appropriately chosen coarsening.

Remark 3.4.1. Let $\mathbb{L}_{[0,t]}$ denote the background driving Lévy process. If we assume $\mathbb{L}_{[0,t]}$ to have finite jump activity, i.e. $F(\mathbb{R}^d) < \infty$, then we can write the continuous martingale part of $D^{p-1}\mathbb{Y}_{[0,t]}$ as

$$D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} = D^{p-1}\mathbf{Y}_s - D^{p-1}\mathbf{Y}_0 - \tilde{\mathbf{b}}s - \sum_{0\le u\le s} \Delta D^{p-1}\mathbf{Y}_u, \quad s \in [0,t],$$

where $(\tilde{\mathbf{b}}, \Sigma, F)$ is the characteristic triplet when the truncation function $\tau(z) \equiv 0$ is used, i.e.

$$\mathbf{L}_t = \tilde{\mathbf{b}}t + \Sigma^{1/2}\mathbf{W}_t + \sum_{0\le s\le t} \Delta\mathbf{L}_s, \quad t \ge 0,$$

where the sum is almost surely finite. In this case we will work with only one jump process, $\tilde{\mathbb{J}}_{[0,t]} = \{\tilde{\mathbf{J}}_s,\ s \in [0,t]\}$ given by

$$\tilde{\mathbf{J}}_s = \sum_{0\le u\le s} \Delta D^{p-1}\mathbf{Y}_u, \tag{3.4.5}$$

in which case the integrator becomes

$$D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} = D^{p-1}\mathbf{Y}_s - D^{p-1}\mathbf{Y}_0 - \tilde{\mathbf{b}}s - \tilde{\mathbf{J}}_s, \quad s \in [0,t]. \tag{3.4.6}$$

Note that by rearranging the integral form of the formal Equation (3.2.4), i.e. the last d entries of SDE (3.2.3), we can retrieve the background driving Lévy process $\mathbb{L}_{[0,t]} = \{\mathbf{L}_s,\ s \in [0,t]\}$ from $\mathbb{Y}_{[0,t]}$ by

$$\mathbf{L}_s = D^{p-1}\mathbf{Y}_s - D^{p-1}\mathbf{Y}_0 + \sum_{j=1}^{p} \int_0^s A^*_j D^{p-j}\mathbf{Y}_u \, \mathrm{d}u, \quad s \in [0,t]. \tag{3.4.7}$$

Under $\mathbb{P}_{\mathbf{A}^*}$ this is a Lévy process with Lévy triplet $(\mathbf{b}, \Sigma, F)$.

Remark 3.4.2. Recall that $D^{p-1}\mathbb{Y}$ jumps whenever $\mathbb{L}$ does and by the same magnitude, i.e. $\Delta D^{p-1}\mathbb{Y} = \Delta\mathbb{L}$. Assuming we can observe the jumps of the process $D^{p-1}\mathbb{Y}$ is equivalent to assuming we observe the jumps of the driving noise $\mathbb{L}$.

The following lemma establishes the framework to prove the subsequent approximations preserve the asymptotic properties of the estimator.

Lemma 3.4.3. *Let* $\hat{\mathbf{A}}_{1,t} = \text{vec}^{-1}([\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t})$ *with* $\mathbf{H}_{1,t} \in \mathbb{R}^{pd^2}, [\mathbf{H}]_{1,t} \in \mathbb{R}^{pd^2 \times pd^2}$ *be a consistent and asymptotically normal estimator for* $\mathbf{A}^* \in \mathfrak{A}$*, i.e.*

- $\hat{\mathbf{A}}_{1,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \ t \to \infty$;
- $\sqrt{t}\left(\text{vec}(\hat{\mathbf{A}}_{1,t}) - \text{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \ t \to \infty$, *where* $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ *is a symmetric positive definite matrix such that* $t^{-1}[\mathbf{H}]_{1,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \ t \to \infty$.

Suppose that

$$t^{-1/2}\left(\mathbf{H}_{2,t} - \mathbf{H}_{1,t}\right) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0, \quad t \to \infty, \tag{3.4.8}$$

and

$$t^{-1/2}\left([\mathbf{H}]_{2,t} - [\mathbf{H}]_{1,t}\right) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0, \quad t \to \infty, \tag{3.4.9}$$

for $\mathbf{H}_{2,t} \in \mathbb{R}^{pd^2}, [\mathbf{H}]_{2,t} \in \mathbb{R}^{pd^2 \times pd^2}$*. Then the estimator* $\hat{\mathbf{A}}_{2,t} = \text{vec}^{-1}([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t})$ *is also consistent and asymptotically normal for* $\mathbf{A}^* \in \mathfrak{A}$*, i.e.*

- $\hat{\mathbf{A}}_{2,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \quad t \to \infty$;
- $\sqrt{t}\left(\text{vec}(\hat{\mathbf{A}}_{2,t}) - \text{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \ t \to \infty$, *and* $t^{-1}[\mathbf{H}]_{2,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \ t \to \infty$.

Proof of Lemma 3.4.3. We start by noting that Equation (3.4.9) immediately implies

$$t^{-1}\left([\mathbf{H}]_{2,t} - [\mathbf{H}]_{1,t}\right) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0, \quad t \to \infty.$$

Combining this with $t^{-1}[\mathbf{H}]_{1,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \ t \to \infty$, by uniqueness of limits we have

$$t^{-1}[\mathbf{H}]_{2,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \quad t \to \infty. \tag{3.4.10}$$

Next, we show that

$$\sqrt{t}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - [\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}\right) \xrightarrow{\mathcal{L}} 0, \quad t \to \infty. \tag{3.4.11}$$

To do so we split the limiting variable as

$$\begin{aligned}\sqrt{t}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - [\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}\right) = \sqrt{t}\Big([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - [\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{1,t}\Big) \\ + \sqrt{t}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{1,t} - [\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}\right),\end{aligned} \tag{3.4.12}$$

and show that each of these two terms converges to zero in probability. We can write the first term as

$$\sqrt{t}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - [\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{1,t}\right) = \left(t[\mathbf{H}]_{2,t}^{-1}\right) t^{-1/2}\left(\mathbf{H}_{2,t} - \mathbf{H}_{1,t}\right).$$

By Equation (3.4.10), continuity of the inversion operator and invertibility of $\mathcal{H}_\infty$ we can apply the continuous mapping theorem to deduce

$$t[\mathbf{H}]_{2,t}^{-1} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty^{-1}, \quad t \to \infty.$$

Thus by multiplicativity of $\mathbb{P}_{\mathbf{A}^*}$-limits and Equation (3.4.8) the first term in (3.4.12) converges to zero in probability under $\mathbb{P}_{\mathbf{A}^*}$. The second term in (3.4.12) can be written as

$$\sqrt{t}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{1,t} - [\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}\right) = \left(t[\mathbf{H}]_{2,t}^{-1}\right) t^{-1/2}\left([\mathbf{H}]_{1,t} - [\mathbf{H}]_{2,t}\right)\left([\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}\right).$$

We note that:

- $t[\mathbf{H}]_{2,t}^{-1}$ converges in probability under $\mathbb{P}_{\mathbf{A}^*}$ to $\mathcal{H}_\infty^{-1}$;
- $t^{-1/2}\left([\mathbf{H}]_{1,t} - [\mathbf{H}]_{2,t}\right)$ converges in probability under $\mathbb{P}_{\mathbf{A}^*}$ to zero by Equation (3.4.9);
- $[\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t}$ converges in probability under $\mathbb{P}_{\mathbf{A}^*}$ to $\mathrm{vec}(\mathbf{A}^*)$ by consistency of $\hat{\mathbf{A}}_{1,t}$ and continuity of vec.

Thus, by multiplicativity of $\mathbb{P}_{\mathbf{A}^*}$-limits, we get that also the second term converges to zero in probability under $\mathbb{P}_{\mathbf{A}^*}$.

Equation (3.4.11) immediately implies

$$[\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - [\mathbf{H}]_{1,t}^{-1}\mathbf{H}_{1,t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0, \quad t \to \infty,$$

and hence consistency of the estimator $\hat{\mathbf{A}}_{2,t}$ follows from consistency of $\hat{\mathbf{A}}_{1,t}$ and continuity of the vec operator. Combining Equation (3.4.11) with the asymptotic normality of $\hat{\mathbf{A}}_{1,t}$ yields

$$t^{1/2}\left([\mathbf{H}]_{2,t}^{-1}\mathbf{H}_{2,t} - \mathrm{vec}^{-1}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$

which, together with Equation (3.4.10), gives the desired asymptotic normality result for $\hat{\mathbf{A}}_{2,t}$. □

3.4.1 Approximating the integrals

First, we assume we have access to the sampled time series of $\mathbb{Y}_{[0,t]}$ and its first $p-1$ derivatives over the partition $\mathcal{P}_t$, i.e. $D^i\mathbb{Y}_{\mathcal{P}_t} := \{D^i\mathbf{Y}_s,\ s \in \mathcal{P}_t\}$ for $i = 0, \dots, p-1$, as well as the jump processes $\mathbb{J}_{\mathcal{P}_t} := \{\mathbf{J}_s,\ s \in \mathcal{P}_t\}$, $\mathbb{M}_{\mathcal{P}_t} := \{\mathbf{M}_s,\ s \in \mathcal{P}_t\}$. In this setting we can perfectly recover $D^{p-1}\mathbb{Y}^c_{\mathbf{A}^{(0)},\mathcal{P}_t}$ by

$$D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} = D^{p-1}\mathbf{Y}_s - D^{p-1}\mathbf{Y}_0 - \mathbf{b}s - \mathbf{J}_s - \mathbf{M}_s, \quad s \in \mathcal{P}_t.$$

Approximating the integrals in (3.4.1) by

$$\sum_{n=0}^{N_t-1} D^l Y^{(i)}_{s_n}\left(D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)},s_{n+1}} - D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)},s_n}\right) \text{ and } \sum_{n=0}^{N_t-1} D^l Y^{(i)}_{s_n} D^k Y^{(j)}_{s_n}(s_{n+1} - s_n),$$

for $i, j \in \{1, \dots, d\}$ and $l, k \in \{0, \dots, p-1\}$, we obtain natural approximations for $\mathbf{H}_t$ and $[\mathbf{H}]_t$ respectively.

In order to derive the asymptotic properties of this discretized estimator, classic in-fill approximations are not sufficient. Instead, we need to combine in-fill and long-span asymptotics by taking the limits as both the mesh size vanishes and the observation horizon increases to infinity. We consider a countable sequence of partitions $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ with mesh $\Delta_{\mathcal{P}_t} \to 0$ as $t \to \infty$ which satisfy the following "high-frequency sampling condition".

Assumption 3.4.4. The sequence of partitions $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ satisfies $\Delta_{\mathcal{P}_t} t \to 0$ as $t \to \infty$.

Under this assumption, we can prove the discretized estimator preserves the asymptotic properties of $\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})$.

Theorem 3.4.5. *Let $\mathbb{Y}_{\mathcal{P}_t}$ denote the sampled observation of a stationary and ergodic MCAR(p) process with parameters $\mathbf{A}^* \in \mathfrak{A}$ under $\mathbb{P}_{\mathbf{A}^*}$ such that the driving Lévy process satisfies Assumption 3.2.4. Let $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ be a countable sequence of partitions satisfying Assumption 3.4.4. Define the discretized estimator*

$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{P}_t}, \mathbb{M}_{\mathcal{P}_t}) = \operatorname{vec}^{-1}\left([\mathbf{H}]_{\mathcal{P}_t}^{-1}\mathbf{H}_{\mathcal{P}_t}\right), \tag{3.4.13}$$

where

$$\mathbf{H}_{\mathcal{P}_t} = -\sum_{n=0}^{N_t-1} \begin{pmatrix} D^{p-1}Y_{s_n}^{(1)}\Sigma^{-1} \\ \vdots \\ D^{p-1}Y_{s_n}^{(d)}\Sigma^{-1} \\ \vdots \\ Y_{s_n}^{(1)}\Sigma^{-1} \\ \vdots \\ Y_{s_n}^{(d)}\Sigma^{-1} \end{pmatrix} \left(D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}, s_{n+1}} - D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}, s_n}\right), \tag{3.4.14}$$

$$[\mathbf{H}]_{\mathcal{P}_t} = \sum_{n=0}^{N_t-1} \begin{pmatrix} D^{p-1}Y_{s_n}^{(1)}\Sigma^{-1}D^{p-1}Y_{s_n}^{(1)} & \cdots & D^{p-1}Y_{s_n}^{(1)}\Sigma^{-1}Y_{s_n}^{(d)} \\ \vdots & \ddots & \vdots \\ Y_{s_n}^{(d)}\Sigma^{-1}D^{p-1}Y_{s_n}^{(1)} & \cdots & Y_{s_n}^{(d)}\Sigma^{-1}Y_{s_n}^{(d)} \end{pmatrix} (s_{n+1} - s_n). \tag{3.4.15}$$

The estimator is

- *(weakly) consistent, i.e.*

$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{P}_t}, \mathbb{M}_{\mathcal{P}_t}) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \quad t \to \infty;$$

- *asymptotically normal, i.e.*

$$\sqrt{t}\left(\operatorname{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{P}_t}, \mathbb{M}_{\mathcal{P}_t})) - \operatorname{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$

where $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ is a symmetric positive definite matrix such that

$$t^{-1}[\mathbf{H}]_{\mathcal{P}_t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \quad t \to \infty.$$

Proof of Theorem 3.4.5. We apply Lemma 3.4.3 with

$$\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) \quad \text{and} \quad \hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{P}_t}, \mathbb{M}_{\mathcal{P}_t}),$$

by decomposing the integrator as

$$\begin{aligned} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} &= D^{p-1}\mathbf{Y}^c_{\mathbf{A}^*,s} - \sum_{j=1}^{p} \int_0^s A_j^* D^{p-j}\mathbf{Y}_u \, \mathrm{d}u \\ &= \Sigma^{1/2}\mathbf{W}_s - \sum_{j=1}^{p} \int_0^s A_j^* D^{p-j}\mathbf{Y}_u \, \mathrm{d}u, \end{aligned} \tag{3.4.16}$$

where $\mathbb{W} = \{\mathbf{W}_s, \ s \geq 0\}$ is a standard Brownian motion under $\mathbb{P}_{\mathbf{A}^*}$. Under Assumption 3.2.4 by Theorem 3.3.9 (and Remark 3.3.11) we have that $\hat{\mathbf{A}}_{1,t}$ is consistent and asymptotic normal. It thus remains to prove Equation (3.4.8) and Equation (3.4.9). See Appendix 3.D.1 for the full details. □

3.4.2 Approximating the derivatives

Next, we assume we have access to the sampled MCAR(p) process only, i.e. $\mathbb{Y}_{\mathcal{P}_t} := \{\mathbf{Y}_s, \ s \in \mathcal{P}_t\}$, and, as in the previous section, the jump processes $\mathbb{J}_{\mathcal{P}_t}$ and $\mathbb{M}_{\mathcal{P}_t}$. Motivated by Proposition 3.2.3 we approximate the first $p-1$ derivatives of $\mathbb{Y}_{[0,t]}$ at each point in $\mathcal{P}_t$ by forward finite differences. We can define these iteratively by

$$\hat{D}^k\mathbb{Y}_{\mathcal{P}_t} := \{\hat{D}^k\mathbf{Y}_{s_n}, \ n = 0, \dots, N_t - k\} \text{ s.t. } \hat{D}^k\mathbf{Y}_{s_n} = \frac{\hat{D}^{k-1}\mathbf{Y}_{s_{n+1}} - \hat{D}^{k-1}\mathbf{Y}_{s_n}}{s_{n+1} - s_n}, \tag{3.4.17}$$

for $k \in \{1, \dots, p-1\}$ and $\hat{D}^0\mathbb{Y}_{\mathcal{P}_t} = \mathbb{Y}_{\mathcal{P}_t}$. In order to obtain appropriate convergence of the (higher-order) finite difference approximations to the corresponding derivatives we assume the observation partitions $\mathcal{P}_t$ become increasingly uniform. This is what we will refer to as the "convergence to uniform spacing" condition.

Assumption 3.4.6. The sequence of partitions $\{\mathcal{P}_t, \ t \in \mathcal{T}\}$ satisfies

$$|c_{\mathcal{P}_t}^{p-1} - 1| = \mathcal{O}\left(\Delta_{\mathcal{P}_t}^{p-3/2}\right), \quad t \to \infty, \tag{3.4.18}$$

where

$$c_{\mathcal{P}_t} := \frac{\inf_{0 \leq i < N_t}(s_{i+1} - s_i)}{\sup_{0 \leq i < N_t}(s_{i+1} - s_i)}.$$

Remark 3.4.7. To understand this condition we note that

- for $p \geq 2$, it implies $c_{\mathcal{P}_t} \to 1,\ t \to \infty$ and Equation (3.4.18) can be understood as the "speed of convergence";
- when $p = 1$, it is trivially satisfied for any sequence of partitions $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$. This is what one would desire: the resulting estimator does not depend on $\mathcal{P}_t$ as there is no approximation of derivatives.

We will also make use of the following immediate consequences of the definition of $c_{\mathcal{P}_t}$:

- $c_{\mathcal{P}_t} \in [0, 1]$, thus Assumption 3.4.6 implies $|c_{\mathcal{P}_t}^k - 1| = \mathcal{O}\left(\Delta_{\mathcal{P}_t}^{k-1/2}\right),\ t \to \infty$ for all $k = 1, \dots, p-1$, and
- for any interval $[s_n, s_{n+1}] \in \mathcal{P}_t$ one has

$$c_{\mathcal{P}_t}\Delta_{\mathcal{P}_t} \leq |s_{n+1} - s_n| \leq \Delta_{\mathcal{P}_t}.$$

A weaker assumption might be sufficient, where we control the ratio between the smallest and the largest interval in each group of $p-1$ successive intervals, i.e. the maximum length over which a forward difference is defined.

Lemma 3.4.8. *Let $\mathbb{Y}_{\mathcal{P}_t}$ denote the sampled observation of a stationary and ergodic MCAR(p) process with parameters $\mathbf{A}^* \in \mathfrak{A}$ under $\mathbb{P}_{\mathbf{A}^*}$ such that the driving Lévy process satisfies Assumption 3.2.4. Let $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ be a countable sequence of partitions satisfying Assumption 3.4.6. For $k \in \{1, \dots, p-1\}$ let $\hat{D}^k\mathbb{Y}_{\mathcal{P}_t}$ denote the k-th forward finite difference of $\mathbb{Y}_{\mathcal{P}_t}$ defined in Equation (3.4.17). Then for any $n \in \{0, \dots, N_t - k\}$ we have*

$$\mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^k\mathbf{Y}_{s_n} - D^k\mathbf{Y}_{s_n}\right\|^2\right] = \mathcal{O}(\Delta_{\mathcal{P}_t}), \quad t \to \infty. \tag{3.4.19}$$

Proof of Lemma 3.4.8. See Appendix 3.D.2. □

Having approximated the derivatives on the partition $\mathcal{P}_t$, we then proceed as in Section 3.4.1 to approximate the integrals. To preserve the estimator's asymptotic properties, we must work on a coarser partition, as demonstrated by the following simple yet illustrative example.

Example 3.4.9. In this example we deviate slightly from the task of estimating MCAR coefficients. Assume that, for some fixed $T > 0$, we wish to approximate the integral

$$\int_0^T Z_t \,\mathrm{d}DZ_t, \quad Z_t = 1 + \int_0^t W_s \mathrm{d}s,$$

for some standard Brownian motion $\{W_t,\ t \geq 0\}$. Then, for any sequence of partitions of the interval $[0, T]$ with vanishing mesh size, we have

$$\sum_{s_n \in \mathcal{P}} Z_{s_n}(DZ_{s_{n+1}} - DZ_{s_n}) = \sum_{s_n \in \mathcal{P}} Z_{s_n}(W_{s_{n+1}} - W_{s_n}) \xrightarrow{L^2} \int_0^T Z_t \mathrm{d}W_t = \int_0^T Z_t \mathrm{d}DZ_t, \quad \Delta_{\mathcal{P}} \downarrow 0.$$

If we observe the process $\mathbb{Z} = \{Z_t,\ t \in [0,T]\}$ on the evenly-spaced partition $\mathcal{P}$ we can form the finite difference approximations

$$\hat{D}Z_{s_n} = \frac{Z_{s_n+1} - Z_{s_n}}{\Delta_{\mathcal{P}}} = \frac{1}{\Delta_{\mathcal{P}}}\int_{s_n}^{s_{n+1}} W_s \mathrm{d}s, \quad s_n \in \mathcal{P}.$$

But if then we attempt to approximate the integral on the same partition used to compute the finite difference approximations

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{s_n\in\mathcal{P}} Z_{s_n}(\hat{D}Z_{s_{n+1}} - \hat{D}Z_{s_n}) - \sum_{s_n\in\mathcal{P}} Z_{s_n}(DZ_{s_{n+1}} - DZ_{s_n})\Big)^2\Big] \\
&\quad \geq \sum_{s_n\in\mathcal{P}} \mathbb{E}[Z_{s_n}^2]\mathbb{E}\Big[\Big((\hat{D}Z_{s_{n+1}} - \hat{D}Z_{s_n}) - (DZ_{s_{n+1}} - DZ_{s_n})\Big)^2\Big] \geq \sum_{s_n\in\mathcal{P}} \frac{2}{3}\Delta_{\mathcal{P}} = \frac{2}{3}T,
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E}\Big[\Big((\hat{D}Z_{s_{n+1}} - \hat{D}Z_{s_n}) - (DZ_{s_{n+1}} - DZ_{s_n})\Big)^2\Big] \\
&\quad = \frac{1}{\Delta_{\mathcal{P}}^2}\mathbb{E}\Big[\Big(\int_{s_{n+1}}^{s_{n+2}} (W_s - W_{s_{n+1}})\mathrm{d}s - \int_{s_n}^{s_n} (W_s - W_{s_n})\mathrm{d}s\Big)^2\Big] \\
&\quad = \frac{2}{\Delta_{\mathcal{P}}^2}\mathbb{E}\Big[\Big(\int_0^{\Delta_{\mathcal{P}}} W_s \mathrm{d}s\Big)^2\Big] = \frac{2}{3}\Delta_{\mathcal{P}}, \quad s_n \in \mathcal{P}.
\end{aligned}$$

If instead we approximate the integral over a coarsening $\mathcal{Q}$ of $\mathcal{P}$ then

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{u_m\in\mathcal{Q}} Z_{u_m}(\hat{D}Z_{u_{m+1}} - \hat{D}Z_{u_m}) - \sum_{u_m\in\mathcal{P}} Z_{u_m}(DZ_{u_{m+1}} - DZ_{u_m})\Big)^2\Big] \\
&\quad = \sum_{u_m\in\mathcal{Q}} \mathbb{E}[Z_{u_m}^2]\mathbb{E}\Big[\Big((\hat{D}Z_{u_{m+1}} - \hat{D}Z_{u_m}) - (DZ_{u_{m+1}} - DZ_{u_m})\Big)^2\Big] \\
&\quad \leq \Big(1 + \frac{T^3}{3}\Big)\sum_{u_m\in\mathcal{Q}} \frac{2}{3}\Delta_{\mathcal{P}} = \frac{2}{3}\Big(1 + \frac{T^3}{3}\Big)T\Delta_{\mathcal{Q}}^{-1}\Delta_{\mathcal{P}} \to 0, \quad \text{as long as } \Delta_{\mathcal{Q}}^{-1}\Delta_{\mathcal{P}} \to 0,
\end{aligned}$$

by assuming for simplicity that no consecutive points in $\mathcal{P}$ appear in $\mathcal{Q}$.

For all $t \in \mathcal{T}$ we hence let

$$\mathcal{Q}_t = \{0 = u_0 < u_1 < \ldots < u_{N_t'} = t\},$$

be a coarsening of $\mathcal{P}_t$ on which we will approximate the integrals by Riemann sums. Note that $N_t' \leq N_t$ and $\Delta_{\mathcal{P}_t} \leq \Delta_{\mathcal{Q}_t}$. As before, we will assume the sequence $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ satisfies the high-frequency sampling condition Assumption 3.4.4, and, hence, so does $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$, but we will also control the rate at which the number of observations increases.

Assumption 3.4.10. The sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ satisfies $c_{\mathcal{Q}_t} \to c > 0$ as $t \to \infty$.

Remark 3.4.11. This is a slightly stronger condition than $N_t'\Delta_{\mathcal{Q}_t} = \mathcal{O}(t)$ as $t \to \infty$, which is used in Mai (2014) and Courgeau and Veraart (2022). The stronger condition stated in Assumption 3.4.10

will only be needed when jointly approximating the derivatives and thresholding the jumps in Section 3.4.3. In any case, both versions of this condition can always be achieved by subsampling $\mathcal{Q}_t$, i.e. discarding observations.

Finally, we control the relative mesh sizes of $\mathcal{P}_t$ and $\mathcal{Q}_t$ via the following joint condition. This ensures the derivative approximations converge faster than the Riemann sums and the resulting estimator does not degenerate. The relationship between $\mathcal{P}_t$ and $\mathcal{Q}_t$ is illustrated in Figure 3.2.

Assumption 3.4.12. The sequences of partitions $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ and $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ are such that $\Delta_{\mathcal{P}_t} = o(t^{-1}\Delta_{\mathcal{Q}_t}^2)$ as $t \to \infty$.

Remark 3.4.13. For example, since $t\Delta_{\mathcal{Q}_t} \to 0$ as $t \to 0$ by Assumption 3.4.4, a sufficient condition is that $\Delta_{\mathcal{P}_t} \propto \Delta_{\mathcal{Q}_t}^3$.

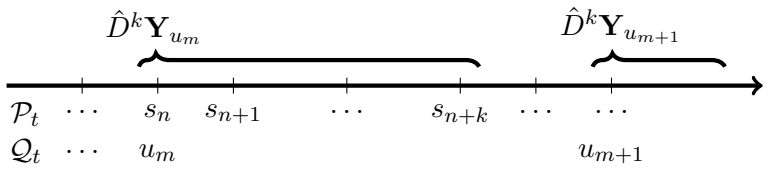


Figure 3.2: Approximating the derivatives on $\mathcal{P}_t$ and the integrals on $\mathcal{Q}_t$.

Theorem 3.4.14. *Let $\mathbb{Y}_{\mathcal{P}_t}$ denote the sampled observation of a stationary and ergodic MCAR(p) process with parameters $\mathbf{A}^* \in \mathfrak{A}$ under $\mathbb{P}_{\mathbf{A}^*}$ such that the driving Lévy process satisfies Assumption 3.2.4. Let $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ be a countable sequence of partitions satisfying Assumption 3.4.6 with refinements $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ satisfying Assumption 3.4.4, Assumption 3.4.10 and Assumption 3.4.12. Define the discretized estimator*

$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t}) := \operatorname{vec}^{-1}\left([\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}^{-1}\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right), \tag{3.4.20}$$

where

$$\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t} = -\sum_{m=0}^{M_t-1}\begin{pmatrix} \hat{D}^{p-1}Y_{u_m}^{(1)}\Sigma^{-1} \\ \vdots \\ \hat{D}^{p-1}Y_{u_m}^{(d)}\Sigma^{-1} \\ \vdots \\ Y_{u_m}^{(1)}\Sigma^{-1} \\ \vdots \\ Y_{u_m}^{(d)}\Sigma^{-1} \end{pmatrix}\left(\hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},u_{m+1}} - \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},u_m}\right), \tag{3.4.21}$$

$$[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t} = \sum_{m=0}^{M_t-1}\begin{pmatrix} \hat{D}^{p-1}Y_{u_m}^{(1)}\Sigma^{-1}\hat{D}^{p-1}Y_{u_m}^{(1)} & \cdots & \hat{D}^{p-1}Y_{u_m}^{(1)}\Sigma^{-1}Y_{u_m}^{(d)} \\ \vdots & \ddots & \vdots \\ Y_{u_m}^{(d)}\Sigma^{-1}\hat{D}^{p-1}Y_{u_m}^{(1)} & \cdots & Y_{u_m}^{(d)}\Sigma^{-1}Y_{u_m}^{(d)} \end{pmatrix}(u_{m+1} - u_m), \tag{3.4.22}$$

are the approximated integrals over the partition $\mathcal{Q}_t$ with the derivatives replaced by the finite difference approximations, $M_t := \max\{m \in \{0, \dots, N'_t\} : u_m \leq s_{N_t - p}\}$ (i.e. need $\hat{D}^{p-1}\mathbf{Y}_{u_{M_t}}$ to be defined) and for $m = 0, \dots, M_t$,

$$\hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}, u_m} = \hat{D}^{p-1}\mathbf{Y}_{u_m} - \hat{D}^{p-1}\mathbf{Y}_0 - \mathbf{b}u_m - \mathbf{J}_{u_m} - \mathbf{M}_{u_m}. \tag{3.4.23}$$

The estimator is

- *(weakly) consistent, i.e.*
$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t}) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \quad t \to \infty;$$
- *asymptotically normal, i.e.*
$$\sqrt{t}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$
where $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ is a symmetric positive definite matrix such that
$$t^{-1}[\mathbf{H}]_{\mathcal{P}_t, \mathcal{Q}_t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \quad t \to \infty.$$

Proof of Theorem 3.4.14. We again apply Lemma 3.4.3, this time with

$$\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{Q}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{Q}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t}) \quad \text{and} \quad \hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t}).$$

Consistency and asymptotic normality of $\hat{\mathbf{A}}_{1,t}$ are given by Theorem 3.4.5 under Assumption 3.2.4 on the driving Lévy process and Assumption 3.4.4 on $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$. As before, it thus remains to prove Equation (3.4.8) and Equation (3.4.9), for which we rely on the result of Lemma 3.4.8. See Appendix 3.D.3 for full details. □

3.4.3 Thresholding the jumps

Finally, we assume that we have access to the sampled MCAR(p) process $\mathbb{Y}_{\mathcal{P}_t}$ only, the most realistic assumption. To remove the dependency on the jump processes $\mathbb{J}_{\mathcal{P}_t}$ and $\mathbb{M}_{\mathcal{P}_t}$ we will use thresholding techniques to disentangle the continuous and jump components of $D^{p-1}\mathbb{Y}_{[0,t]}$. For ease of notation let us write the increments of the partition $\mathcal{Q}_t$ with mesh $\Delta_{\mathcal{Q}_t}$ as

$$\Delta^m_{\mathcal{Q}_t} := u_{m+1} - u_m, \quad m = 0, \dots, M_t - 1,$$

and the increment of a process $\mathbb{Z}$ over the interval $[u_m, u_{m+1}] \in \mathcal{Q}_t$ as

$$\Delta^m_{\mathcal{Q}_t}\mathbf{Z} := (\mathbf{Z}_{u_{m+1}} - \mathbf{Z}_{u_m}), \quad m = 0, \dots, M_t - 1.$$

In the following, we aim to approximate the increments of the process given in Equation (3.4.23), i.e. for $m = 0, \dots, M_t - 1$,

$$\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} = \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} - \Delta^m_{\mathcal{Q}_t} \mathbf{J} - \Delta^m_{\mathcal{Q}_t} \mathbf{M},$$

by the thresholded increments

$$\Delta^m_{\mathcal{Q}_t, \boldsymbol{\nu}^m_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} := \left[\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right] \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \leq \boldsymbol{\nu}^m_t \right\}}, \tag{3.4.24}$$

where $\boldsymbol{\nu}^m_t \in \mathbb{R}^d$ is a thresholding vector, $\mathbb{1}_{\{|\mathbf{x}| \leq \mathbf{y}\}}$ for $\mathbf{y} \in \mathbb{R}^d$ denotes the vector with i-th entry given by $\mathbb{1}_{\{|x_i| \leq y_i\}}$ and $\odot$ denotes element-wise product. The motivation behind the thresholding approach is to interpret the (de-trended) increments

$$\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(i)} - b^{(i)} \Delta^m_{\mathcal{Q}_t},$$

greater than the threshold $\nu^{(i),m}_t$ to be mainly driven by the jump part of $D^{p-1} Y^{(i)}$ and those smaller than $\nu^{(i),m}_t$ to be mainly due to the continuous part of $D^{p-1} Y^{(i)}$. By an appropriate choice of scaling for $\{(\boldsymbol{\nu}^m_t)_{m=0}^{M_t - 1},\ t \in \mathcal{T}\}$ this procedure aims to disentangle the continuous and the jump components of the process as $t \to \infty$. We proceed in a similar way to Mai (2014) and Courgeau and Veraart (2022) where the thresholding technique is used to develop an estimator for the drift of a (multivariate) OU process. We treat the cases where the driving Lévy process has finite and infinite jump activity separately.

Remark 3.4.15. Note that, while in the OU setting Mai (2014) and Courgeau and Veraart (2022) could apply thresholding to $\mathbb{Y}$ directly, here we can only threshold the *approximated* process $\hat{D}^{p-1}\mathbb{Y}$. Moreover, in the following, we allow for the slightly more general case of a (known) non-zero drift $\mathbf{b}$ in the driving Lévy process.

In this section, we thus define the discretized estimator resulting from the combination of Riemann sum, finite difference, and thresholding approximations by

$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t) := \operatorname{vec}^{-1} \left([\mathbf{H}]^{-1}_{\mathcal{P}_t, \mathcal{Q}_t} \mathbf{H}_{\mathcal{P}_t, \mathcal{Q}_t, \boldsymbol{\nu}_t} \right), \tag{3.4.25}$$

where $[\mathbf{H}]^{-1}_{\mathcal{P}_t,\mathcal{Q}_t}$ is given by Equation (3.4.22) and

$$\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} = -\sum_{m=0}^{M_t-1} \begin{pmatrix} \hat{D}^{p-1} Y^{(1)}_{u_m} \Sigma^{-1} \\ \vdots \\ \hat{D}^{p-1} Y^{(d)}_{u_m} \Sigma^{-1} \\ \vdots \\ Y^{(1)}_{u_m} \Sigma^{-1} \\ \vdots \\ Y^{(d)}_{u_m} \Sigma^{-1} \end{pmatrix} \Delta^m_{\mathcal{Q}_t,\boldsymbol{\nu}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}}. \tag{3.4.26}$$

Remark 3.4.16. To show the asymptotic properties are preserved under appropriate conditions, we will apply Lemma 3.4.3 one last time with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$. In this case, it will be sufficient to show that Equation (3.4.8) holds, as the approximator of the quadratic variation $[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}$ does not change and hence (3.4.9) trivially holds.

Finite jump activity

If we assume that the driving Lévy process has finite jump activity then by Remark 3.4.1 we can write

$$\Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} = \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y} - \tilde{\mathbf{b}} \Delta^m_{\mathcal{Q}_t} - \Delta^m_{\mathcal{Q}_t} \tilde{\mathbf{J}}, \tag{3.4.27}$$

where $\tilde{\mathbb{J}}$ is the compound Poisson process given in Equation (3.4.5) with elementwise Poisson counting processes $\mathbf{N}_t = (N^{(1)}_t, \dots, N^{(d)}_t)^{\mathrm{T}}$. We denote by $\lambda^{(i)}$ the jump rate of the process $\{N^{(i)}_t,\ t \geq 0\}$ and by $\tilde{F}^{(i)}(\cdot)$ the probability distribution of jump sizes in the i-th component, i.e. $F^{(i)}(\cdot) = \lambda^{(i)} \tilde{F}^{(i)}(\cdot)$, where $F^{(i)}(\cdot)$ is the i-th marginal of the Lévy measure F. We impose the following joint condition on the sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the thresholding sequences $\{(\boldsymbol{\nu}^m_t)^{M_t-1}_{m=0},\ t \in \mathcal{T}\}$:

Assumption 3.4.17. The sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the sequence of jump thresholds $\{(\boldsymbol{\nu}^m_t)^{M_t-1}_{m=0},\ t \in \mathcal{T}\}$ satisfy for $i \in \{1, \dots, d\}$

(i) $\nu^{(i),m}_t = \left(\Delta^m_{\mathcal{Q}_t}\right)^{\beta^{(i)}}$ with $\beta^{(i)} \in (0,\ 1/2)$;

(ii) $t\Delta^{1-2\beta^{(i)}}_{\mathcal{Q}_t} \to 0, \quad t \to \infty$;

(iii) $t\tilde{F}^{(i)}\left(\left(-2\Delta^{\beta^{(i)}}_{\mathcal{Q}_t},\ 2\Delta^{\beta^{(i)}}_{\mathcal{Q}_t}\right)\right) \to 0, \quad t \to \infty.$

Remark 3.4.18. Note that $\beta^{(i)} \in (0,\ 1/2)$ relates the strength of the thresholding, via Assumption 3.4.17.(i), and the speed at which the mesh $\Delta_{\mathcal{Q}_t}$ vanishes, via Assumption 3.4.17.(ii) and Assumption 3.4.17.(iii).

- Since $1 - 2\beta^{(i)} \in (0,\ 1)$ Assumption 3.4.17.(ii) is strictly stronger than the high-frequency sampling condition for $\mathcal{Q}_t$, i.e. Assumption 3.4.4.

- Intuitively, Assumption 3.4.17.(*ii*) controls the rate at which the probability of observing a continuous increment bigger than the threshold vanishes, while Assumption 3.4.17.(*iii*) controls the rates at which the probability of observing a jump increment smaller than (twice) the threshold vanishes. The "optimal" $\beta^{(i)}$ is thus chosen by appropriately balancing the two conditions. For example, if $\tilde{F}^{(i)}(\cdot)$ has a bounded Lebesgue density, Assumption 3.4.17.(*iii*) becomes
$$t\Delta_{\mathcal{Q}_t}^{\beta^{(i)}} \to 0, \quad t \to \infty,$$
and we see the optimal trade-off is given by $\beta^{(i),*} = 1/3$.

- From a theoretical point of view, given a sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$, one can choose $\beta^{(i)}$ depending on the driving Lévy process such that Assumption 3.4.17.(*ii*) and Assumption 3.4.17.(*iii*) are "optimally" fulfilled. In practice, given the observation partition $\mathcal{Q}_t$, one chooses $\beta^{(i)}$ based on the data, we will discuss this in more detail in Section 3.4.5.

In order to understand the choice of thresholding sequence, we start sketching the main idea of the proof. Let us focus on dimension $i \in \{1, \dots, d\}$ and assume for the moment that we can perfectly observe an increment of $D^{p-1}Y^{(i)}$ over $[u_m, u_{m+1}]$, i.e. the finite difference approximation is close enough to the target value. In this case, one can bound the probability that the thresholding technique fails by considering whether

- an increment is discarded due to a large move in the continuous component, or
- an increment is kept in the presence of a jump.

To do so one requires bounds on

- the probability of continuous increments larger than the threshold $\nu_t^{(i),m} = (\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}$: this is controlled for $\beta^{(i)} \in (0, 1/2)$ by Assumption 3.4.17.(*i*) and Assumption 3.4.17.(*iii*),
- the probability of jumps smaller than $2(\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}$: this is controlled by Assumption 3.4.17.(*ii*).

One can then show that the probability of the thresholding technique being successful approaches one and, on this set, the $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ distance between the original and the thresholded approximator, i.e. $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}$ and $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t}$, approaches zero. The following theorem, along with Theorem 3.4.23 in the infinite activity setting, states the asymptotic properties of the only practically computable, i.e. depending on discrete data $\mathbb{Y}_{\mathcal{P}_t}$ only, estimator for $\mathbf{A}^* \in \mathfrak{A}$, the main result of this section.

Theorem 3.4.19. *Let $\mathbb{Y}_{\mathcal{P}_t}$ denote the sampled observation of a stationary and ergodic MCAR(p) process with parameters $\mathbf{A}^* \in \mathfrak{A}$ under $\mathbb{P}_{\mathbf{A}^*}$ such that the driving Lévy process satisfies Assumption 3.2.4 and has finite jump activity. Let $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ be a countable sequence of partitions satisfying Assumption 3.4.6 with refinements $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ satisfying Assumption 3.4.4, Assumption 3.4.10 and Assumption 3.4.12. Furthermore, assume the sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the thresholding sequences $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$ satisfy Assumption 3.4.17. Then the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ defined in Equation (3.4.25) is*

- *(weakly) consistent, i.e.*

$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \quad t \to \infty;$$

- *asymptotically normal, i.e.*

$$\sqrt{t}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$

where $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ is a symmetric positive definite matrix such that

$$t^{-1}[\mathbf{H}]_{\mathcal{P}_t, \mathcal{Q}_t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \quad t \to \infty.$$

Proof of Theorem 3.4.19. We apply Lemma 3.4.3 with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$. Consistency and asymptotic normality of $\hat{\mathbf{A}}_{1,t}$ are given by Theorem 3.4.14. By Remark 3.4.16 it suffices to show Equation (3.4.8). The arguments of the proof are sketched above, see Appendix 3.D.4 for full details. □

Infinite jump activity

We now consider the more general case, where the driving Lévy process $\mathbb{L}$ might have infinite jump activity, i.e. $F(\mathbb{R}^d) = \infty$. In this case, by Equation (3.4.2), we can write the continuous martingale increments as

$$\Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} = \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} - \Delta^m_{\mathcal{Q}_t} \mathbf{J} - \Delta^m_{\mathcal{Q}_t} \mathbf{M}, \tag{3.4.28}$$

where $\mathbb{J}$ is the finite activity pure jump Lévy process with jumps greater than one given in Equation (3.4.3) and $\mathbb{M}$ is the pure jump martingale with jumps smaller or equal to one given in Equation (3.4.4). Note that $\mathbb{J}$ is a compound Poisson process with elementwise Poisson counting processes $\mathbf{N}_t = (N_t^{(1)}, \ldots, N_t^{(d)})^\mathrm{T}$. We denote by $\lambda^{(i)}$ the jump rate of the process $\{N_t^{(i)},\ t \geq 0\}$ and by $\tilde{F}^{(i)}(\cdot)$ the probability distribution of jump sizes in the i-th component of $\mathbb{J}$, i.e. the i-th marginal of $\tilde{F} := F|_{\{\mathbf{x}:\|\mathbf{x}\|>1\}} / F(\{\mathbf{x} : \|\mathbf{x}\| > 1\})$. We impose the following joint condition on the sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the thresholding sequence $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$:

Assumption 3.4.20. The sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the sequence of jump thresholds $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$ satisfy for $i \in \{1, \ldots, d\}$

(i) $\nu_t^{(i),m} = (\Delta^m_{\mathcal{Q}_t})^{\beta^{(i)}}$ with $\beta^{(i)} \in (0,\ 1/4)$;

(ii) $t\Delta^{1-4\beta^{(i)}}_{\mathcal{Q}_t} \to 0, \quad t \to \infty$;

(iii) $t\tilde{F}^{(i)}\left(\left(-4\Delta^{\beta^{(i)}}_{\mathcal{Q}_t}, 4\Delta^{\beta^{(i)}}_{\mathcal{Q}_t}\right)\right) \to 0,\ t \to \infty$ and $\exists \epsilon^{(i)} > 0$ such that $t\Delta^{2\epsilon^{(i)}}_{\mathcal{Q}_t} \to 0, \quad t \to \infty$ and

$$\mathbb{E}_{\mathbf{A}^*}\left[|M_s^{(i)}| \mathbb{1}_{\left\{|M_s^{(i)}| \leq s^{\beta^{(i)}}\right\}}\right] = \mathcal{O}\left(s^{1+\epsilon^{(i)}}\right) \text{ as } s \downarrow 0.$$

Moreover, the driving Lévy process $\mathbb{L}$ has finite fourth-moment, i.e. $\mathbb{E}[\|\mathbf{L}_1\|^4] < \infty$.

Remark 3.4.21. Some important remarks are in place. If we compare these assumptions to those considered in the finite jump activity case, i.e. Assumption 3.4.17, we note that:

- In Assumption 3.4.20.(i) the upper bound of allowed $\beta^{(i)}$'s is lower. This implies the thresholding cannot be too strong, i.e. enough observations need to be kept to identify the parameters. A choice of $\beta^{(i)} \in (1/4, 1/2)$ would void Assumption 3.4.20.(ii) which is required to control the rate at which the probability of observing increments of $\mathbb{M}$ bigger than the threshold vanishes.

- On the other hand, Assumption 3.4.20.(iii) implicitly requires additional control on $\beta^{(i)}$ to ensure the jump component of the kept increments is negligible. The control on the rate at which the probability of observing a jump increment due to the compound Poisson process $\mathbb{J}$ smaller than (four times) the threshold vanishes – also present in Assumption 3.4.17.(iii) – is complemented by a condition controlling the magnitude of the increments due to the infinite activity jump part $\mathbb{M}$. We note that, in the second part of Assumption 3.4.20.(iii), the existence and choice of $\epsilon^{(i)}$ depends on both $\mathbb{M}$ and $\beta^{(i)}$.

- As in Assumption 3.4.17 the "optimal" $\beta^{(i)}$ is chosen by appropriately balancing conditions (ii) and (iii). For example, if $\tilde{F}^{(i)}(\cdot)$ has a bounded Lebesgue density and we write $\epsilon^{(i)} = f\left(\beta^{(i)}\right)$, the optimal trade-off is given by

$$\beta^{*,(i)} = \underset{\beta^{(i)} \in (0,1/4)}{\operatorname{argmax}} \min\left\{1 - 4\beta^{(i)}, \beta^{(i)}, 2f\left(\beta^{(i)}\right)\right\},$$

  i.e. by balancing the rates $t\Delta_{\mathcal{Q}_t}^{1-4\beta^{(i)}}, t\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}, t\Delta_{\mathcal{Q}_t}^{2f\left(\beta^{(i)}\right)}$. In practice, as for the finite activity case, we usually choose the thresholding sequence from data as discussed in Section 3.4.5.

- When the jump components of $\mathbb{L}$ are independent, the first part of Assumption 3.4.20.(iii) is trivially satisfied since $\tilde{F}^{(i)}([-1, 1]) = 0$. This holds trivially when $d = 1$.

We note that our assumptions differ from those imposed in Mai (2014) to deduce the asymptotic properties of the drift estimator for one-dimensional Lévy-driven OU processes. We had to modify the assumptions after noticing some inconsistencies in the arguments applied at the bottom of Mai (2014, page 944) when splitting the proof of Mai (2014, Equation (35)) into two steps denoted by (i) and (ii) therein. Similar arguments were repeatedly applied in the paper (and ceteris paribus extended in Courgeau and Veraart (2022) to the multivariate setting). Moreover, we wish to underline the following additional differences.

- We will require finiteness of 4-th moments of the driving Lévy process. This was explicitly imposed in Courgeau and Veraart (2022), but was a tacit assumption in Mai (2014).

- The assumptions in Mai (2014) include an additional symmetric condition on $\mathbb{M}$. As discussed in Appendix 3.D.5, including this additional condition allows to "weaken" Assumption 3.4.20.(iii).

Following multiple numerical experiments we believe the assumptions considered here might be stronger than required but, as illustrated in Example 3.4.22, they are not too restrictive.

It is important to explore the attainability of Assumption 3.4.20.(iii), to ensure it is not a void condition. Example 3.4.22 considers the case where the jump component of the driving Lévy process is a Gamma process.

Example 3.4.22. Let $\mathbb{L}$ be a one-dimensional finite variation Lévy process such that

$$L_t = \tilde{b}t + \sigma W_t + G_t,$$

where $\mathbb{G} = \{G_t,\ t \geq 0\}$ is a standard symmetric Gamma processes, i.e. shape $k = 1$ and scale $\theta = 1$. This is an infinite activity process with intensity measure

$$F(\mathrm{d}x) = |x|^{-1} e^{-|x|} \mathbb{1}_{\{x \neq 0\}} \mathrm{d}x.$$

It has infinite activity since

$$F(\mathbb{R}) = 2 \int_0^\infty x^{-1} e^{-x} \mathrm{d}x \geq 2e^{-1} \int_0^1 x^{-1} \mathrm{d}x = 2e^{-1}[\ln(1) - \ln(0)] = \infty,$$

and finite moments for any $n \in \mathbb{N}$ since by Sato (1999, Theorem 25.3) it is sufficient to check that

$$\int_{|x|>1} |x|^n F(\mathrm{d}x) = 2 \int_1^\infty x^{n-1} e^{-x}\, \mathrm{d}x \leq 2(n-1)! < \infty.$$

We can rewrite $\mathbb{L}$ in terms of the Lévy-Itô decomposition with truncation function $\tau(x) = \mathbb{1}_{\{|x| \leq 1\}}$ as

$$L_t = bt + \sigma W_t + J_t + M_t,$$

where, since F is symmetric, $b = \tilde{b}$ and $G_t = M_t + J_t$. In order to check Assumption 3.4.20.(iii) is satisfied we need to control the increments of $\mathbb{M} = \{M_t,\ t \geq 0\}$. Letting $\mathbb{N} = \{N_s,\ s \geq 0\}$ denote the Poisson counting process associated with $\mathbb{J}$ we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[|M_s| \mathbb{1}_{\left\{|M_s| \leq s^\beta\right\}}\right] \\
&\quad \overset{(i)}{=} \mathbb{E}_{\mathbf{A}^*}\left[|M_s| \mathbb{1}_{\left\{|M_s| \leq s^\beta\right\}} | N_s = 0\right] \mathbb{P}_{\mathbf{A}^*}(N_s = 0) + \mathbb{E}_{\mathbf{A}^*}\left[|M_s| \mathbb{1}_{\left\{|M_s| \leq s^\beta\right\}} | N_s \geq 1\right] \mathbb{P}_{\mathbf{A}^*}(N_s \geq 1) \\
&\quad \overset{(ii)}{\leq} \mathbb{E}_{\mathbf{A}^*}\left[|G_s| \mathbb{1}_{\left\{|G_s| \leq s^\beta\right\}} | N_s = 0\right] \mathbb{P}_{\mathbf{A}^*}(N_s = 0) + \mathbb{E}_{\mathbf{A}^*}\left[|M_s| | N_s \geq 1\right] \mathbb{P}_{\mathbf{A}^*}(N_s \geq 1) \\
&\quad \overset{(iii)}{\leq} \mathbb{E}_{\mathbf{A}^*}\left[|G_s| \mathbb{1}_{\left\{|G_s| \leq s^\beta\right\}}\right] + \mathbb{E}_{\mathbf{A}^*}\left[|M_s|\right] \mathbb{P}_{\mathbf{A}^*}(N_s \geq 1) \\
&\quad \overset{(iv)}{=} \mathcal{O}\left(s^{1+\beta}\right) + \mathcal{O}\left(s^{1/2}\right) \mathcal{O}(s) = \mathcal{O}\left(s^{1+\beta}\right),
\end{aligned}$$

using in (i) the law of total expectation, in (ii) the fact that $G_s = M_s$ on $\{N_s = 0\}$ and $|x|\mathbb{1}_A(x) \leq |x|$ for $x \in \mathbb{R}$, in (iii) the law of total expectation and the independence of $\mathbb{M}$ and $\mathbb{N}$, in (iv) the fact that the Lévy process is square integrable, $\mathbb{N}$ is a Poisson process and

$$\begin{aligned}\mathbb{E}_{\mathbf{A}^*}\left[|G_s|\mathbb{1}_{\{|G_s|\leq s^\beta\}}\right] = \int_{-s^\beta}^{s^\beta} |x|p(s,0,x)\,\mathrm{d}x &= 2\int_0^{s^\beta} \frac{1}{2\Gamma(s)}x^s e^{-x}\,\mathrm{d}x \\ &\leq \frac{1}{\Gamma(s)}\int_0^{s^\beta} x^s\,\mathrm{d}x = \frac{1}{\Gamma(s)}\frac{s^{\beta(s+1)}}{s+1} \leq s^{1+\beta},\end{aligned}$$

since $p(s,0,x) = \dfrac{1}{2\Gamma(s)}|x|^{s-1}e^{-|x|}$ is the transition density of the process $\mathbb{G}$ and $\Gamma(s) \geq s^{-1}$ for $s > 0$.
Thus if $\beta \in (0,\, 1/4)$, we can set $\epsilon = \beta$ to obtain

$$\mathbb{E}_{\mathbf{A}^*}\left[|M_s|\mathbb{1}_{\{|M_s|\leq s^\beta\}}\right] = \mathcal{O}\left(s^{1+\epsilon}\right),$$

i.e. Assumption 3.4.20.(iii) is satisfied. In this case, the optimal thresholding power is given by $\beta^* = 1/5$. Similar arguments can be carried out for non-standard (a-)symmetric multidimensional Lévy processes with independent Gamma jump components.

We can now state the consistency and asymptotic normality result for the high-frequency estimator (3.4.25) in the infinite-activity setting.

Theorem 3.4.23. *Let $\mathbb{Y}_{\mathcal{P}_t}$ denote the sampled observation of a stationary and ergodic MCAR(p) process with parameters $\mathbf{A}^* \in \mathfrak{A}$ under $\mathbb{P}_{\mathbf{A}^*}$ such that the driving Lévy process satisfies Assumption 3.2.4. Let $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ be a countable sequence of partitions satisfying Assumption 3.4.6 with refinements $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ satisfying Assumption 3.4.4, Assumption 3.4.10 and Assumption 3.4.12. Furthermore, assume the sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the thresholding sequences $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$ satisfy Assumption 3.4.20. Then the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ defined in Equation (3.4.25) is*

- *(weakly) consistent, i.e.*
$$\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathbf{A}^*, \quad t \to \infty;$$
- *asymptotically normal, i.e.*
$$\sqrt{t}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}), \quad t \to \infty,$$
where $\mathcal{H}_\infty \in \mathcal{M}_{pd^2}(\mathbb{R})$ is a symmetric positive definite matrix such that
$$t^{-1}[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} \mathcal{H}_\infty, \quad t \to \infty.$$

Proof of Theorem 3.4.23. We proceed exactly as in the proof of Theorem 3.4.19 by applying Lemma 3.4.3 with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$. Consistency and asymptotic

normality of $\hat{\mathbf{A}}_{1,t}$ are given by Theorem 3.4.14 and thus, by Remark 3.4.16, it suffices to show Equation (3.4.8). The main idea of the proof is to introduce an auxiliary finite activity process with only the big jumps $\mathbb{J}$. The difference between $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}$ and $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t}$ is then decomposed into three terms which are shown to vanish in $\mathbb{P}_{\mathbf{A}^*}$:

- A first term comparing $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}$ to its approximation obtained by thresholding the auxiliary finite activity process. Similar arguments as in the finite activity setting can thus be applied to this term (since Assumption 3.4.20 is stronger than 3.4.17).

- A second term quantifying the difference between thresholding based on the auxiliary finite activity process and the actual infinite activity process. To show the effect of the small jumps is negligible when thresholding, we require a bound on the probability of observing small jumps with the same magnitude as the threshold, namely Assumption 3.4.20.(*ii*).

- A third term quantifying the difference between the auxiliary finite activity jump process and the actual infinite activity jump process, i.e. the small jumps $\mathbb{M}$, when the increments are kept. This term is controlled by applying the second part of Assumption 3.4.20.(*iii*).

See Appendix 3.D.5 for full details. □

Table 3.3 summarizes how the assumptions of Theorem 3.4.19 and Theorem 3.4.23 relate the driving Lévy process $\mathbb{L}$, the sequences of observation partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$, and the threshold powers $\beta^{(i)},\ i = 1, \ldots, d$.

		$\mathbb{L}$	$\{\mathcal{Q}_t,\ t\in\mathcal{T}\}$	$\{\mathcal{P}_t,\ t\in\mathcal{T}\}$	$\beta^{(i)}$'s
Assumption 3.2.4	$\mathbb{E}[\|\mathbf{L}_1\|^2]<\infty$	■			
	Σ positive definite	■			
Assumption 3.4.4	$\Delta_{\mathcal{Q}_t} t \to 0$		■		
Assumption 3.4.6	$c_{\mathcal{P}_t} \to 1$ "fast enough"			■	
Assumption 3.4.10	$c_{\mathcal{Q}_t} \to c > 0$		■		
Assumption 3.4.12	$\Delta_{\mathcal{P}_t} = o(t^{-1}\Delta^2_{\mathcal{Q}_t})$		■	■	
Assumption 3.4.17	$t\Delta_{\mathcal{Q}_t}^{1-2\beta^{(i)}} \to 0$		■		■
	$t\tilde{F}^{(i)}\Big(\Big(-2\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}, 2\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}\Big)\Big) \to 0$	■	■		■
Assumption 3.4.20	$t\Delta_{\mathcal{Q}_t}^{1-4\beta^{(i)}} \to 0$		■		■
	$t\tilde{F}^{(i)}\Big(\Big(-4\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}, 4\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}\Big)\Big) \to 0$	■	■		■
	control on $\mathbb{M}$ in terms of $\beta^{(i)}, \Delta_{\mathcal{Q}_t}$	■	■		■
	$\mathbb{E}[\|\mathbf{L}_1\|^4]<\infty$	■			

Table 3.3: The assumptions of Theorem 3.4.19 and Theorem 3.4.23 where the thresholds are assumed to be of the form $\nu_t^{(i),m} = (\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}$ for $i = 1,\dots,d$ and $m = 0,\dots,M_t - 1$ for any $t \in \mathcal{T}$.

3.4.4 Simulation study

In this section, we empirically check how well the asymptotic results proved in Theorem 3.4.19 and Theorem 3.4.23 perform in finite samples by carrying out a simulation study. We consider a $d = 1$ dimensional MCAR process of order $p = 2$ with drift parameter $\mathbf{A}^* = (A_1, A_2) = (1, 2)$. The driving Lévy process has characteristic triplet (b, Σ, F) with drift $b = 0$ and variance $\Sigma = 1$. We consider three different jump regimes:

(BM) no jumps, i.e. $F \equiv 0$;

(CP) compound Poisson jump process with rate $\lambda = 1$ and $\mathcal{N}(0,1)$ jump sizes, i.e. $F(\mathrm{d}x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$;

(Γ) symmetric standard Gamma jump process, i.e. $F(\mathrm{d}x) = |x|^{-1} e^{-|x|} \mathrm{d}x$.

In all three cases, the Lévy driver satisfies Assumption 3.2.4. In the first two, the process has finite

jump activity and hence we can apply Theorem 3.4.19, in the last case the process has infinite activity with finite 4th moment and thus Theorem 3.4.23 applies.

We consider a sequence of uniform partitions $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ with step size $\Delta_{\mathcal{P}_t} = t^{-6}$ and uniform refinements $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ with step size $\Delta_{\mathcal{Q}_t} = t^{-2}$ so that Assumption 3.4.4, Assumption 3.4.6, Assumption 3.4.10 and Assumption 3.4.12 hold. Noting that all partitions are refinements of each other we simulate the processes at each point in the finest grid we will use for estimation, i.e. $\mathcal{P}_t$ with $t = 8$. We use the exact procedure described in Appendix 3.G.2 to simulate the finite activity CAR processes and the approximate Euler-Maruyama procedure described in Appendix 3.G.1 to simulate the Gamma-driven CAR process. The code for simulating and estimating MCAR processes is available at `https://github.com/lorenzolucchese/mcar`.

We assume the Lévy triplet $(0, \Sigma, F)$ is known and hence select the theoretically optimal threshold powers $\beta^* = -\infty, 1/3, 1/5$, i.e. no thresholding in the Brownian case and the values given in Remark 3.4.18 and Example 3.4.22 for the finite activity and Gamma drivers. Under these choices of observation partitions and thresholding vectors Assumption 3.4.17 is satisfied by regimes (BM) and (CP) while Assumption 3.4.20 is satisfied by (Γ).

In Figure 3.3 we report the empirical distribution of the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$. We see that as t increases the distributions concentrate around the true values $\mathbf{A}^* = (1, 2)$, validating the consistency result. Moreover, in Figure 3.4, we check how well the pre-asymptotic normal approximation holds for the bivariate statistic $\mathbf{Z}_t$ of the feasible CLT (3.4.29). We see that as t increases the empirical distribution of the $\mathbf{Z}_t$'s approaches a $N(\mathbf{0}, I_{2\times 2})$ distribution.

3.4.5 Practical considerations

In this section we focus on the practically computable estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ defined in Equation (3.4.25) and investigate some natural questions which arise when applying it to observed data.

First, we note that in Theorem 3.4.19 and Theorem 3.4.23 the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ depends, more or less explicitly, on the driving Lévy triplet $(\mathbf{b}, \Sigma, F)$ as follows:

- the drift $\mathbf{b} \in \mathbb{R}^d$ is used to de-trend the thresholded increments in Equation (3.4.24);
- the symmetric positive definite covariance matrix $\Sigma \in \mathcal{M}_d(\mathbb{R})$ appears in both $\mathbf{H}_{\mathcal{P}_t, \mathcal{Q}_t, \boldsymbol{\nu}_t}$, cf. Equation (3.4.26), and $[\mathbf{H}]_{\mathcal{P}_t, \mathcal{Q}_t}$, cf. Equation (3.4.22);
- the Lévy measure F is used to determine the activity level of the Lévy driver and thence choose the optimal thresholding powers $\beta^{(i)}$ for $i \in \{1, \dots, d\}$ via Assumption 3.4.17 and Assumption 3.4.20.

In practice, these quantities are not known, but some natural and data-driven choices are possible.

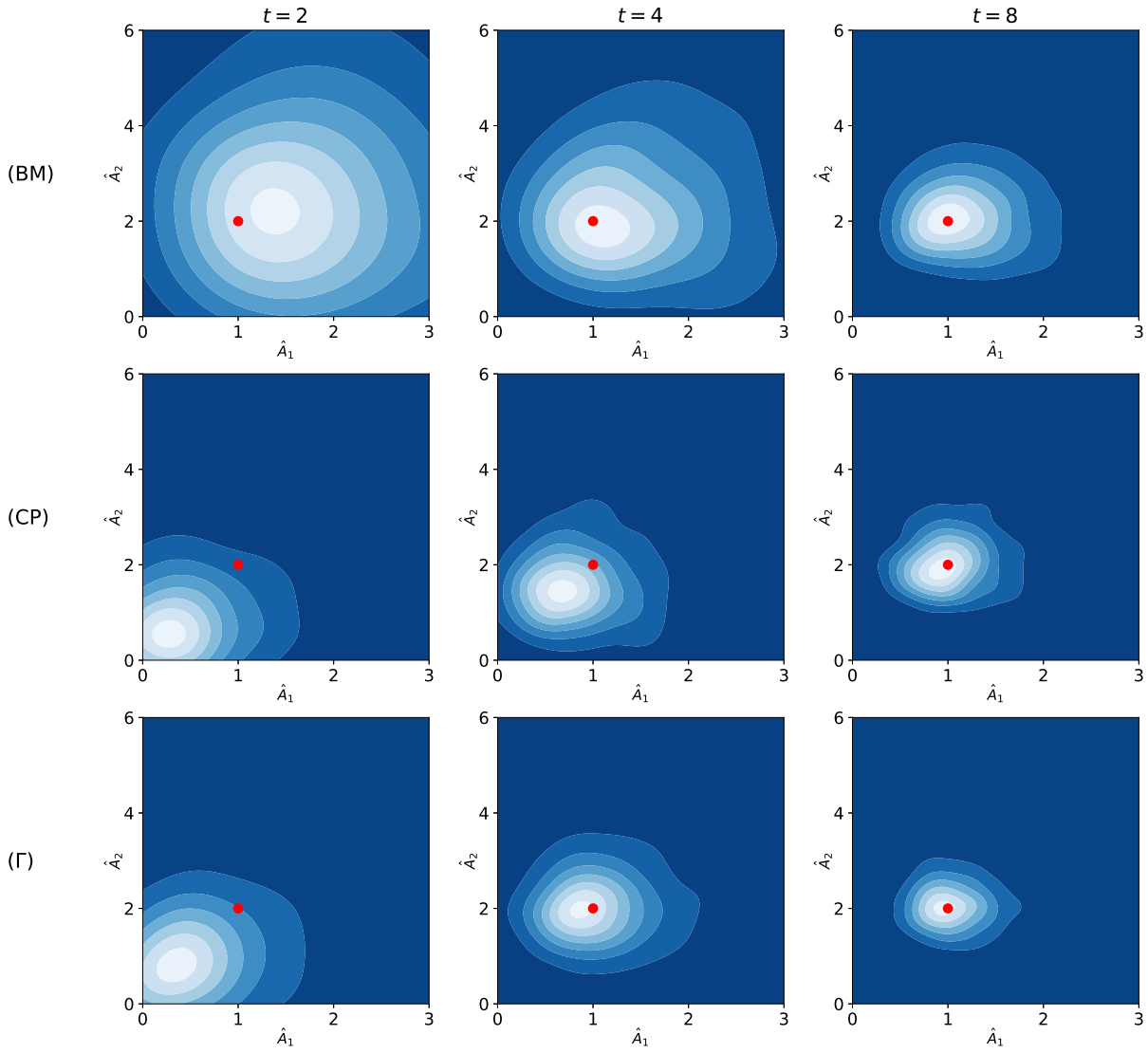


Figure 3.3: Empirical distribution of the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$.

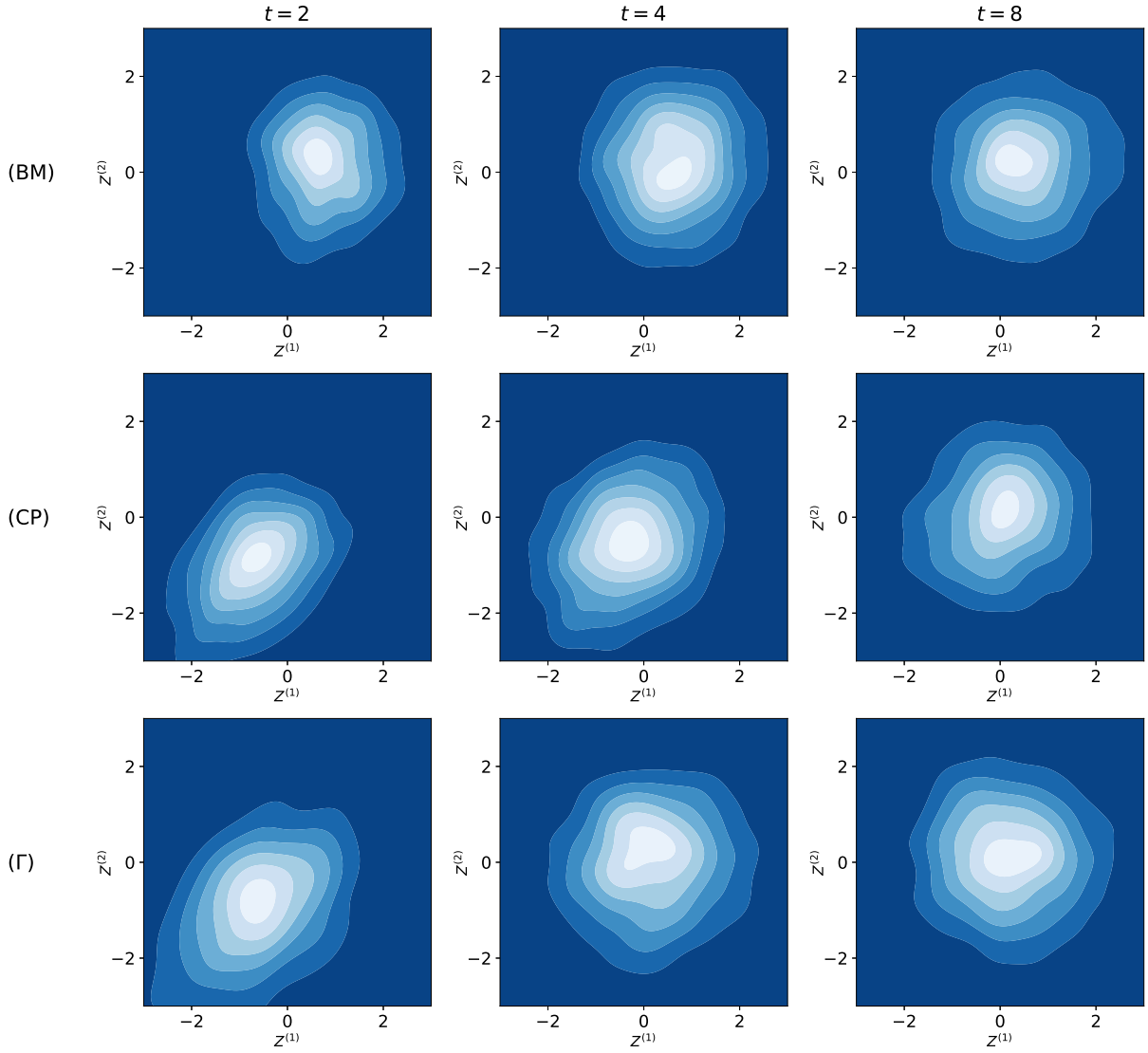


Figure 3.4: Empirical distribution of the statistic $\mathbf{Z}_t$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$.

- A natural choice is to set $\mathbf{b} \equiv 0$. This is exact when the stationary MCAR process is mean-zero and the jumps of the driving Lévy process are symmetric since $\mathbb{E}_{\mathbf{A}}[\mathbf{Y}_1] = A_p^{-1}\mathbb{E}[\mathbf{L}_1] = A_p^{-1}\mathbf{b} = \mathbf{0}$ implies $\mathbf{b} = \mathbf{0}$. More generally, we note that the drift correction has scale $\sim \Delta_{\mathcal{Q}_t}^m$ while the stochastic increment has scale $\sim \sqrt{\Delta_{\mathcal{Q}_t}^m}$, thus, for small $\Delta_{\mathcal{Q}_t}^m$, the choice of $\mathbf{b}$ should have little impact on the estimation of $\mathbf{A}$.

- We note that in the computation of $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ the Σ's appearing in $\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t}$ and $[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}$ cancel out. This removes the dependency on Σ when estimating $\mathbf{A}^*$ but, as discussed later in this section, does not remove the dependency on Σ when making the CLT feasible for inference.

- The thresholding powers $\beta^{(i)}$ for $i \in \{1, \dots, d\}$ can be chosen using a data-driven approach. To be more precise, in practice, we can choose the thresholds $\nu_t^{m,(i)} = (\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}$ directly, dropping their functional form in terms of $\Delta_{\mathcal{Q}_t}^m$ and $\beta^{(i)}$. These thresholds should be selected as small as possible to eliminate the jump part without throwing away too many continuous increments, cf. Aït-Sahalia and Jacod (2014, Section 6.2.2) where increments of a general semimartingale are thresholded to estimate the integrated volatility. Assuming $\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y}$ is close enough to $\Delta_{\mathcal{Q}_t}^m D^{p-1}\mathbf{Y}$, i.e. $\Delta_{\mathcal{P}_t}$ in Lemma 3.4.8 is negligible compared to $\Delta_{\mathcal{Q}_t}^m$, we can write
$$\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} \approx \Delta_{\mathcal{Q}_t}^m \mathbf{L} - \sum_{j=1}^{p} \int_{u_m}^{u_{m+1}} A_j^* D^{p-j}\mathbf{Y}_u \, \mathrm{d}u \approx \Delta_{\mathcal{Q}_t}^m \mathbf{L},$$
whenever $\Delta_{\mathcal{Q}_t}^m$ is small enough. This follows by noting that, for each $j \in \{1, \dots, p\}$, we can apply Jensen's inequality and finiteness of second moments to bound
$$\mathbb{E}_{\mathbf{A}^*}\left[\left\| \int_0^{\Delta_{\mathcal{Q}_t}^m} D^{p-j}\mathbf{Y}_u \, \mathrm{d}u \right\|^2\right] \leq (\Delta_{\mathcal{Q}_t}^m)^2 \mathbb{E}_{\mathbf{A}^*}\left[\int_0^{\Delta_{\mathcal{Q}_t}^m} \frac{\|D^{p-j}\mathbf{Y}_u\|^2}{\Delta_{\mathcal{Q}_t}^m} \, \mathrm{d}u\right] = \mathcal{O}\left((\Delta_{\mathcal{Q}_t}^m)^2\right),$$
which is thus negligible compared to
$$\mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta_{\mathcal{Q}_t}^m \mathbf{L}\right\|^2\right] = \Omega(\Delta_{\mathcal{Q}_t}^m),$$
due to the presence of the Brownian component. We can hence threshold $\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y}$ as if we were disentangling the increments of a Lévy process, cf. Appendix 3.H.1. In this setting, i.e. when choosing thresholds $\nu_t^{m,(i)}$, one may work on each component $\hat{D}^{p-1}Y^{(i)}$ independently.

Next, we discuss how to produce feasible asymptotic inference from Theorem 3.4.19 and Theorem 3.4.23, i.e. how to compute explicit confidence regions for the parameters $\mathbf{A}^* \in \mathfrak{A}$. Under the conditions of the theorems $t^{-1}[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}$ is a consistent estimator for the inverse of the limiting variance-covariance matrix $\mathcal{H}_\infty$. Thus, in the spirit of Remark 3.3.10, we can apply Slutsky's lemma

to write the feasible CLT

$$\mathbf{Z}_t := [\mathbf{H}]^{1/2}_{\mathcal{P}_t,\mathcal{Q}_t}\left(\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t};\mathcal{Q}_t,\boldsymbol{\nu}_t)) - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, I_{pd^2}), \quad t \to \infty. \tag{3.4.29}$$

By working with the distribution of a standard pd^2-dimensional normal random variable one can easily obtain asymptotic confidence regions for the parameters, e.g. elliptic regions centred at $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t};\mathcal{Q}_t,\boldsymbol{\nu}_t)$. The main caveat here is that, as discussed at the beginning of the section, $[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}$ depends on the, usually unknown, covariance matrix Σ. In this case though, having estimated the parameters $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t};\mathcal{Q}_t,\boldsymbol{\nu}_t)$, we can plug these into (3.4.7) along with finite difference and Riemann sum approximations to recover the Lévy increments

$$\Delta^m_{\mathcal{Q}_t}\mathbf{L} \approx \Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y} + \sum_{j=1}^{p}\sum_{s_n\in[u_m,u_{m+1})}\hat{A}_j(\mathbb{Y}_{\mathcal{P}_t};\mathcal{Q}_t,\boldsymbol{\nu}_t)\hat{D}^{p-j}\mathbf{Y}_{s_n}\Delta^n_{\mathcal{P}_t}, \tag{3.4.30}$$

for $m = 0,\dots,M_t-1$ and use standard methods to estimate Σ, for example, one can use the critical region-based estimator discussed in Appendix 3.H.1. The resulting estimator $\hat{\Sigma}$ can then be plugged into $[\mathbf{H}]_{\mathcal{P}_t,\mathcal{Q}_t}$ to compute explicit confidence regions. In Appendix 3.H.3 we repeat the simulation study of Section 3.4.4 assuming Σ and F are unknown. We use the data-driven procedures described above to select the thresholds $(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1}$ and estimate Σ. The simulations yield similar results to those of Section 3.4.4 for the estimator and z-statistic of the drift parameter $\mathbf{A}$ and, moreover, suggest the estimator $\hat{\Sigma}$ is consistent.

Remark 3.4.24. We can go one step further and try to jointly estimate $\mathbf{A}$ and both the parameters of the Lévy process $(\mathbf{b},\Sigma,F)$. To do so we propose an iterative scheme in Appendix 3.H.2 and evaluate the empirical properties of the estimators in Appendix 3.H.3.

Finally, the choice of the hyperparameter p, the order of the MCAR specification, may or may not be motivated by the modeling task at hand. In the latter case, one may tackle such problem from a model selection perspective. For example, one might use a likelihood-based criterion, such as the Aikake Information Criterion (AIC), by plugging in the approximators $\mathbf{H}_{\mathcal{Q}_t,\mathcal{P}_t,\boldsymbol{\nu}_t}$ and $[\mathbf{H}]_{\mathcal{Q}_t,\mathcal{P}_t}$ in the continuous-time likelihood (3.3.5).

3.5 Application: GrCAR

The parametrization of a d-dimensional MCAR process is given by the tuple of coefficient matrices

$$\mathbf{A} = (A_1,\dots,A_p) \in (\mathcal{M}_d(\mathbb{R}))^p.$$

Clearly this means the number of parameters increases quadratically in the dimension d and thus, for a high dimensional observed process, the number of parameters to be estimated can be quite large. When additional structure of the observed process is known, one can significantly reduce such parametrization. For example, as in Courgeau and Veraart (2022), one can interpret the

components of an MCAR process $\mathbb{Y}$ as the nodes of a graph structure linked together through a collection of edges. The edges of the graph are encoded in an adjacency (or graph topology) matrix $A \in \mathcal{M}_d(\{0,1\})$ such that $A^{(i,j)} = 1$ if and only if there exists a link from node i to node j. We set $A^{(i,i)} = 0$ for all $i \in \{1,\dots,d\}$. For any node $i \in \{1,\dots,d\}$ the in-degree is defined as $n_i = 1 \vee \sum_{j=1}^{d} A^{(j,i)}$, i.e. the number of in-going edges to node i. We can thus define the column-normalized adjacency matrix

$$\bar{A} = A\,\mathrm{diag}(n_1^{-1},\dots,n_d^{-1}) \in \mathcal{M}_d(\mathbb{R}).$$

Definition 3.5.1 (Graph CAR process)**.** Let $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ be a d-dimensional Lévy process with characteristics $(\mathbf{b}, \Sigma, F)$. Fix $p \in \mathbb{N}$ and let $\theta \in \mathbb{R}^{p\times 2}$. Define

$$A_k(\theta) = \theta^{(k,1)} I_{d\times d} + \theta^{(k,2)} \bar{A}^{\mathrm{T}} \in \mathcal{M}_d(\mathbb{R}), \quad k \in \{1,\dots,p\}.$$

A graph continuous-time autoregressive process of order p, GrCAR(p) process for short, driven by $\mathbb{L}$ with parameter $\theta \in \mathbb{R}^{p\times 2}$ and adjacency matrix $A \in \mathcal{M}_d(\{0,1\})$ is a is a d-dimensional MCAR(p) process driven by $\mathbb{L}$ with coefficient matrices $\mathbf{A}(\theta) = (A_1(\theta),\dots,A_p(\theta)) \in (\mathcal{M}_d(\mathbb{R}))^p$ and initial state-space representation ξ.

Remark 3.5.2.

- In the definition of the GrCAR process we use the column-normalized matrix $\bar{A}$. This is a modeling choice consistent with the GrOU definition of Courgeau and Veraart (2022) ensuring the strength of the network effect is proportional to the average value of the (derivatives of the) process at the influencing nodes. Alternatively, one could replace $\bar{A}$ by the un-normalized adjacency matrix A, in which case the network effect becomes proportional to the sum of the (derivatives of the) process at the influencing nodes.

- When defining the GrCAR process we transpose the matrix $\bar{A}$ so that if $A^{(i,j)} = 1$ values of node i influence those of node j. In the social network literature instead a link from node i to node j is interpreted as "node i follows node j" and thus the values of node j influence those of node i. By using the row-normalized version of A instead of $\bar{A}^{\mathrm{T}}$ in the construction of the network effect, as in Courgeau and Veraart (2022), the interpretation of the GrCAR parameters is consistent with such literature.

We now leverage the graphical structure of the GrCAR model to write the corresponding explicit estimator for the $\theta \in \mathbb{R}^{p\times 2}$ parameters when continuous-time and discrete-time observations are available. In the following, we consider the vectorization operator for elements $\theta \in \mathbb{R}^{p\times 2}$ given by $\mathrm{vec} : \mathbb{R}^{p\times 2} \mapsto \mathbb{R}^{2p}$ such that

$$\theta \mapsto \mathrm{vec}(\theta) := (\theta^{(1,1)}, \theta^{(1,2)}, \dots, \theta^{(p,1)}, \theta^{(p,2)})^{\mathrm{T}}.$$

Theorem 3.5.3. *Let* $\mathbb{Y}_{[0,t]} = \{\mathbf{Y}_s,\ s \in [0,t]\}$ *be the observation of a stationary and ergodic*

GrCAR(p) process with parameters

$$\theta^* \in \Theta := \{\theta \in \mathbb{R}^{p\times 2} : \sigma(\mathcal{A}_{\mathbf{A}(\theta)}) \subseteq (-\infty, 0) + i\mathbb{R}\}, \tag{3.5.1}$$

and known adjacency matrix $A \in \mathcal{M}(\{0,1\})$ *under* $\mathbb{P}_{\theta^*}$ *such that the driving Lévy process satisfies Assumption 3.2.4. For* $\theta^{(0)} = 0_{p\times 2}$*, let* $D^{p-1}\mathbb{Y}^c_{\theta^{(0)},[0,t]} = \{D^{p-1}\mathbf{Y}^c_{\theta^{(0)},t},\ s \in [0,t]\}$ *be defined as in Equation (3.3.13). Let* $\mathbb{K} = \{\mathbf{K}_t,\ t \geq 0\}$ *be the process defined by*

$$\mathbf{K}_t = -\begin{pmatrix} \int_0^t \langle D^{p-1}\mathbf{Y}_{s-}, \mathrm{d}D^{p-1}\mathbf{Y}^c_{\theta^{(0)},s}\rangle_\Sigma \\ \int_0^t \langle \bar{A}^{\mathrm{T}} D^{p-1}\mathbf{Y}_{s-}, \mathrm{d}D^{p-1}\mathbf{Y}^c_{\theta^{(0)},s}\rangle_\Sigma \\ \vdots \\ \int_0^t \langle \mathbf{Y}_{s-}, \mathrm{d}D^{p-1}\mathbf{Y}^c_{\theta^{(0)},s}\rangle_\Sigma \\ \int_0^t \langle \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}, \mathrm{d}D^{p-1}\mathbf{Y}^c_{\theta^{(0)},s}\rangle_\Sigma \end{pmatrix} \tag{3.5.2}$$

with quadratic covariation matrix

$$[\mathbf{K}]_t = \begin{pmatrix} \int_0^t \langle D^{p-1}\mathbf{Y}_{s-}, D^{p-1}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s & \cdots & \int_0^t \langle D^{p-1}\mathbf{Y}_{s-}, \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s \\ \int_0^t \langle \bar{A}^{\mathrm{T}} D^{p-1}\mathbf{Y}_{s-}, D^{p-1}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s & \cdots & \int_0^t \langle \bar{A}^{\mathrm{T}} D^{p-1}\mathbf{Y}_{s-}, \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s \\ \vdots & \ddots & \vdots \\ \int_0^t \langle \mathbf{Y}_{s-}, D^{p-1}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s & \cdots & \int_0^t \langle \mathbf{Y}_{s-}, \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s \\ \int_0^t \langle \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}, D^{p-1}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s & \cdots & \int_0^t \langle \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}, \bar{A}^{\mathrm{T}}\mathbf{Y}_{s-}\rangle_\Sigma \,\mathrm{d}s \end{pmatrix}. \tag{3.5.3}$$

Then the MLE given by

$$\hat{\theta}(\mathbb{Y}_{[0,t]}) = \mathrm{vec}^{-1}([\mathbf{K}]_t^{-1}\mathbf{K}_t), \tag{3.5.4}$$

is

- *(weakly) consistent, i.e.*
$$\hat{\theta}(\mathbb{Y}_{[0,t]}) \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*, \quad t \to \infty;$$
- *asymptotically normal, i.e.*
$$\sqrt{t}\left(\mathrm{vec}(\hat{\theta}(\mathbb{Y}_{[0,t]})) - \mathrm{vec}(\theta^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{K}_\infty^{-1}), \quad t \to \infty,$$
where $\mathcal{K}_\infty \in \mathcal{M}_{2p}(\mathbb{R})$ *is a symmetric positive definite matrix such that*
$$t^{-1}[\mathbf{K}]_t \xrightarrow{\mathbb{P}_{\theta^*}} \mathcal{K}_\infty, \quad t \to \infty.$$

Next, assume $\mathbb{Y}_{\mathcal{P}_t}$ *is a discretely sampled version of* $\mathbb{Y}_{[0,t]}$ *where* $\{\mathcal{P}_t,\ t \in \mathcal{T}\}$ *is a countable sequence of partitions satisfying Assumption 3.4.6 with refinements* $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ *satisfying Assumption 3.4.4, Assumption 3.4.10 and Assumption 3.4.12. Also, assume the sequence of partitions* $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ *and the thresholding sequences* $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$ *either*

- *satisfy Assumption 3.4.17 and the driving Lévy process has finite jump activity; or*
- *satisfy Assumption 3.4.20.*

Then the estimator

$$\hat{\theta}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t) = \operatorname{vec}^{-1}\left([\mathbf{K}]^{-1}_{\mathcal{P}_t,\mathcal{Q}_t}\mathbf{K}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t}\right), \tag{3.5.5}$$

where

$$\mathbf{K}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} = -\sum_{m=0}^{M_t-1}\begin{pmatrix}\hat{D}^{p-1}\mathbf{Y}_{u_m}^{\mathrm{T}}\\ \hat{D}^{p-1}\mathbf{Y}_{u_m}^{\mathrm{T}}\bar{A}\\ \vdots\\ \mathbf{Y}_{u_m}^{\mathrm{T}}\\ \mathbf{Y}_{u_m}^{\mathrm{T}}\bar{A}\end{pmatrix}\Sigma^{-1}\left[\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t}\right]\odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|\leq\boldsymbol{\nu}_t^m\right\}}, \tag{3.5.6}$$

$$[\mathbf{K}]_{\mathcal{P}_t,\mathcal{Q}_t} = \sum_{m=0}^{M_t-1}\begin{pmatrix}\langle\hat{D}^{p-1}\mathbf{Y}_{u_m},\hat{D}^{p-1}\mathbf{Y}_{u_m}\rangle_\Sigma & \cdots & \langle\hat{D}^{p-1}\mathbf{Y}_{u_m},\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m}\rangle_\Sigma\\ \langle\bar{A}^{\mathrm{T}}\hat{D}^{p-1}\mathbf{Y}_{u_m},\hat{D}^{p-1}\mathbf{Y}_{u_m}\rangle_\Sigma & \cdots & \langle\bar{A}^{\mathrm{T}}\hat{D}^{p-1}\mathbf{Y}_{u_m},\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m}\rangle_\Sigma\\ \vdots & \ddots & \vdots\\ \langle\mathbf{Y}_{u_m},\hat{D}^{p-1}\mathbf{Y}_{u_m}\rangle_\Sigma & \cdots & \langle\mathbf{Y}_{u_m},\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m}\rangle_\Sigma\\ \langle\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m},\hat{D}^{p-1}\mathbf{Y}_{u_m}\rangle_\Sigma & \cdots & \langle\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m},\bar{A}^{\mathrm{T}}\mathbf{Y}_{u_m}\rangle_\Sigma\end{pmatrix}\Delta^m_{\mathcal{Q}_t}, \tag{3.5.7}$$

is

- *(weakly) consistent, i.e.*

$$\hat{\theta}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t) \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*, \quad t\to\infty;$$

- *asymptotically normal, i.e.*

$$\sqrt{t}\left(\operatorname{vec}(\hat{\theta}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)) - \operatorname{vec}(\theta^*)\right) \xrightarrow{\mathcal{L}} \mathbf{Z}\sim N(\mathbf{0}, \mathcal{K}_\infty^{-1}), \quad t\to\infty,$$

  with $t^{-1}[\mathbf{K}]_{\mathcal{P}_t,\mathcal{Q}_t} \xrightarrow{\mathbb{P}_{\theta^*}} \mathcal{K}_\infty, \quad t\to\infty.$

Remark 3.5.4. Note that, unlike the estimator for the MCAR process $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ given in Equation (3.4.25), we cannot easily remove the dependence of $\hat{\theta}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ on Σ. Extending Courgeau and Veraart (2022, Definition 2.6) from the GrOU to the GrCAR setting we can construct an alternative estimator $\hat{\theta}$ for θ from an estimator $\hat{\mathbf{A}}$ for $\mathbf{A}$ by "inverting" the relationship

$$A_k(\theta) = \theta^{(k,1)}I_{d\times d} + \theta^{(k,2)}\bar{A}^{\mathrm{T}} \in \mathcal{M}_d(\mathbb{R}), \quad k\in\{1,\dots,p\}.$$

As long as each node has at least one neighbour the resulting estimator $\hat{\theta}$ depends linearly on $\hat{\mathbf{A}}$, ensuring the asymptotic properties of $\hat{\mathbf{A}}$ are preserved. In particular, by choosing $\hat{\mathbf{A}} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ given in Equation (3.4.25) the resulting estimator $\hat{\theta}$ does not depend on Σ.

Proof of Theorem 3.5.3. The proof follows by similar arguments to Theorem 3.3.9, Theorem 3.4.19 and Theorem 3.4.23. See Appendix 3.E.1. □

In a similar spirit to Section 3.4.4 we carry out a finite sample simulation study. Here, we consider a $d = 5$ dimensional GrCAR process of order $p = 2$ with parameters

$$\theta^* = \begin{pmatrix} \theta^{(1,1)} & \theta^{(1,2)} \\ \theta^{(2,1)} & \theta^{(2,2)} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 0.5 \end{pmatrix},$$

and fully connected underlying adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

The driving Lévy process has characteristics $(\mathbf{0}, \Sigma, F)$ with covariance matrix $\Sigma = I_{5\times 5}$. As in Section 3.4.4 we consider the three jump regimes (BM), (CP) and (Γ), and use the numerical methods described in Appendix 3.G to simulate the processes. Again, we work with uniformly spaced partitions $\mathcal{P}_t$ and $\mathcal{Q}_t$ with mesh sizes $\Delta_{\mathcal{P}_t} = t^{-6}$ and $\Delta_{\mathcal{Q}_t} = t^{-2}$ with given optimal $\beta^* = -\infty, 1/3, 1/5$, so that all the Assumptions of Theorem 3.5.3 are satisfied. The resulting empirical distributions for 1000 Monte Carlo simulations are plotted in $(\theta^{(1,1)}, \theta^{(1,2)})$- and $(\theta^{(2,1)}, \theta^{(2,2)})$-space in Figure 3.5.

3.6 Empirical application: light curve modeling

In this section, we consider a practical application of CAR models and showcase the effectiveness of the estimation procedures developed in this chapter. CAR processes, and more generally CARMA processes, have received significant attention from the astrophysics community as a means to model[2] (one-dimensional) stochastic light curves and quantitatively characterize their variability (Kelly et al., 2014; Moreno et al., 2019). In classic time series analysis one is often interested in breaking down a stationary process into a superposition of random periodic components, quantifying the contribution of each component to the total variability of the process through the power spectral density (PSD). Fundamental results in this context are Wold's theorem and the Wiener-Khinchin theorem which state that the PSD and the autocovariance function form a Fourier transform pair for discrete- and continuous-parameter processes respectively. When observations are from a

[2]This class of processes is chosen for its ability to generate a flexible class of PSDs, specifically sums of Lorentzians. While the CARMA parameters often lack a specific physical interpretation, in some cases – such as modeling light curves from black hole systems – the quasi-periodic oscillation (QPO) features of CARMA PSDs may be driven by astrophysical phenomena in accretion flows (Kelly et al., 2014).

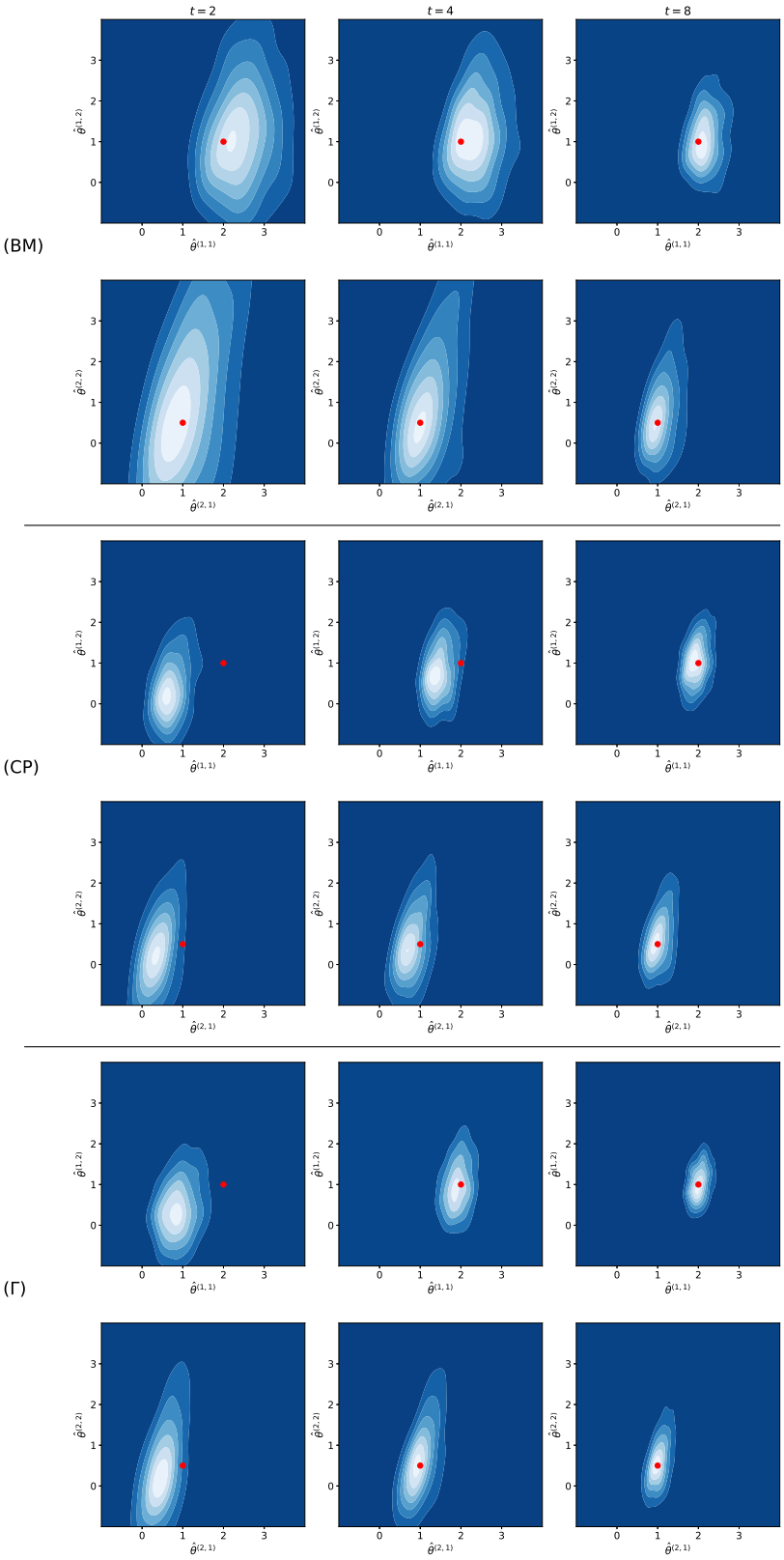


Figure 3.5: Empirical distribution of the estimator $\hat{\theta}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$.

discrete-parameter process, e.g. a regularly sampled stationary process, one can use the periodogram, i.e. the modulus squared of the discrete Fourier transform of the data, to estimate the PSD. Note this is a completely non-parametric approach that provides a reliable PSD estimate (while the simple periodogram is not consistent there exist "windowed" variations with better asymptotic properties). When the observations are irregularly spaced, as is often the case with light curve data, estimation in the frequency domain becomes a bit more challenging. In this case, practitioners often resort to the Lomb-Scargle periodogram to estimate the PSD of the signal, but this method loses robustness for highly non-sinusoidal signals (VanderPlas, 2018). If we are willing to trade off estimation flexibility for robustness resorting to parametric models is a natural approach. Assuming the signal is generated by a CAR process of order p with drift parameters $\mathbf{A} = (A_1, \ldots, A_p) \in \mathbb{R}^p$ and driving Lévy parameters (b, Σ, F) its PSD is given by

$$S(f) = \frac{\sigma_L^2}{|(2\pi i f)^p + \sum_{k=1}^{p} A_k (2\pi i f)^{p-k}|^2}, \quad f \in (-\infty, +\infty), \tag{3.6.1}$$

where $\sigma_L^2 = \mathrm{Var}(L_1) = \Sigma + \int_{\mathbb{R}} x^2 F(\mathrm{d}x) \in (0, \infty)$. We can thus use the estimators developed in this chapter to estimate the CAR parameters to plug into the above expression and obtain an estimate for the power spectral density of the process. A similar approach was introduced in the astrophysics literature in Kelly et al. (2014). In that paper, the authors consider Gaussian CARMA models to model a wide variety of light curves, including an X-ray binary, two active galactic nuclei, a long-period variable star, and an RR Lyrae star. The parameters of the CARMA models are estimated through numerical optimization of the likelihood which, at each gradient descent update, is computed via Kalman filtering. Bayesian inference is carried out through an MCMC posterior sampling algorithm. Both of these methods turn out to be very computationally expensive and subject to optimization error. If one is willing to restrict the set of models under consideration to the class of CAR processes[3] the closed-form estimators (and corresponding inference results) presented in this work provide significant computational speed-ups. For example, the MLE optimization proposed in Kelly et al. (2014) has order $\mathcal{O}(MN)$ where N is the number of observations and M is the number of gradient-descent parameter updates (both Kalman filtering and gradient descent cannot be parallelized) while, assuming parallelization of vectorized operations, the estimator in Equation (3.4.25) can be computed in $\mathcal{O}(\sqrt{N})$. This computational speed-up is especially relevant for the large-scale astronomical datasets that have been and will be collected by massive time-domain optical surveys such as the Vera Rubin Observatory Legacy Survey of Space and Time (LSST).

To showcase the effective applicability of the estimation procedure we consider one of the light curves treated in Kelly et al. (2014): the local active galactic nuclei (AGN) Zw 229-15 in

[3]The estimation and inference methods discussed in Kelly et al. (2014) also allow for the presence of "white noise" measurement errors in the data. The estimators presented in this chapter require instead "direct" observations of a CAR model. We remove the effects of measurement errors via Locally Weighted Scatterplot Smoothing (LOWESS), but more complex joint estimation procedures could be employed.

quarter Q9 from the Kepler observatory, Kepler Input Catalog (KIC) 006932990. The data used in this section can be found in MAST: https://dx.doi.org/10.17909/24fc-ky94. This light curve is almost regularly sampled every 30 minutes over 97 days for a total of 4651 observations, consistently with the notation used so far we will denote this observation partition by $\mathcal{P}$. To remove measurement noise we apply Locally Weighted Scatterplot Smoothing (LOWESS) (Cleveland, 1979) with windowing fraction 0.005. The original data and the smoothed curve are shown in Figure 3.6.

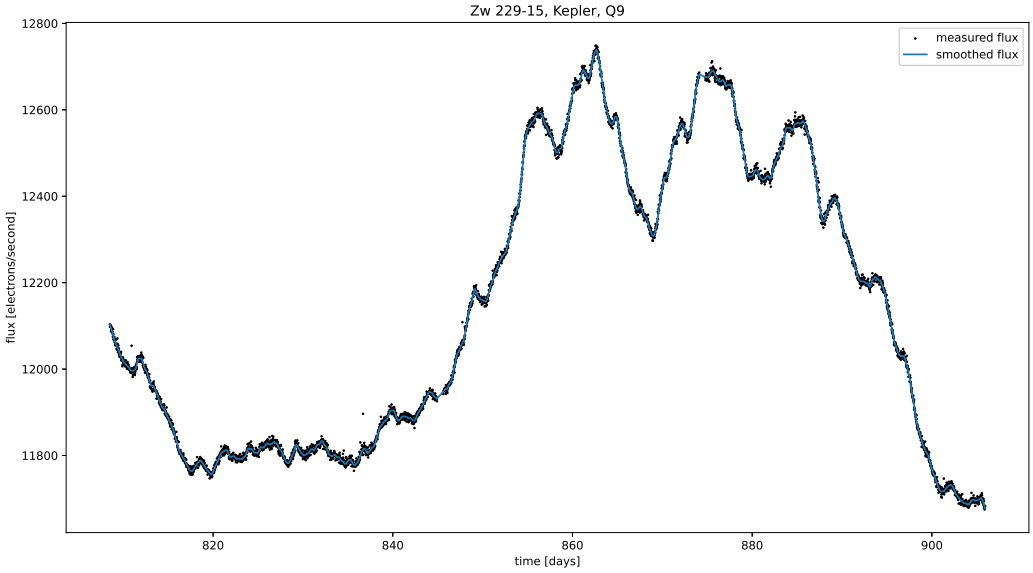


Figure 3.6: Light curve for the AGN Zw 229-15 from the Kepler satellite in quarter Q9 (KIC 006932990), available from MAST: https://dx.doi.org/10.17909/24fc-ky94. The measurement errors are smoothed out by applying LOWESS.

We apply a CAR($p = 2$) model to the smoothed light curve. To estimate the parameters $\mathbf{A} = (A_1, A_2)$ and (b, Σ, F) we use the joint estimation procedure described in Appendix 3.H.2, with the only hyperparameter choice being the coarser partition $\mathcal{Q}$. We select $\mathcal{Q}$ by subsetting $\mathcal{P}$ by a factor of 6. Figure 3.7 gives a graphical visualization of the main steps of the fitting procedure, which include taking finite differences of the process and characterizing the distribution of the background driving Lévy process. The fitted parameters are:

$$\hat{\mathbf{A}} = (\hat{A}_1, \hat{A}_2) = (0.67 \pm 0.23, 0.0004 \pm 0.0012) \quad \text{and} \quad (\hat{b}, \hat{\Sigma}, \hat{F}) = (4.68, 4963.23, 0),$$

where we obtain inference bounds for $\hat{\mathbf{A}}$ as described in Section 3.4.5 and set $\hat{F} \equiv 0$ as no jumps were detected. As in Kelly et al. (2014), albeit using a different CARMA specification, we do not observe evidence against the assumption of the driving noise process being Gaussian.

In Figure 3.8 we report a detailed spectral analysis of the Kepler light curve. In this plot, we compare the (parametric) fitted CAR PSD and the (non-parametric) periodogram. The fitted

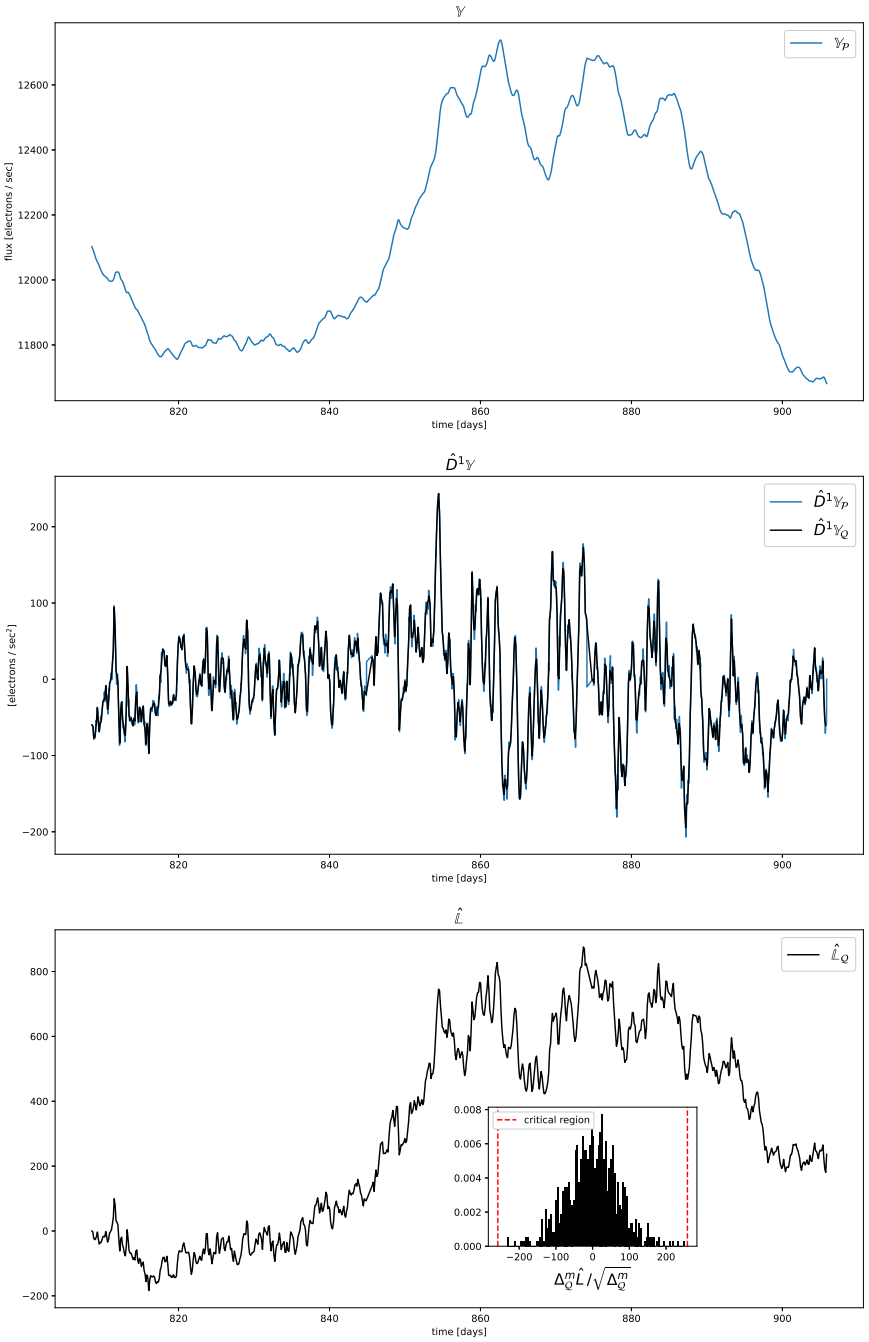


Figure 3.7: Estimation of a CAR($p = 2$) model from the smoothed Kepler light curve Zw 229-15 in Q9. We take finite differences of the process $\mathbb{Y}$ observed over $\mathcal{P}$ to approximate $D^1\mathbb{Y}$. The processes are then sub-sampled over $\mathbb{Q}$ to estimate $\mathbf{A}$. Once $\hat{\mathbb{A}}$ has been estimated we can approximately recover the background driving Lévy process $\mathbb{L}$ and estimate its parameters (b, Σ, F).

CAR PSD is obtained by plugging the parameter estimates into the CAR PSD (3.6.1). We also report the 95% confidence interval for the fitted CAR spectrum accounting for uncertainty in the estimated drift parameters $\hat{\mathbf{A}}$ (but not in $\hat{\sigma}_L^2$). The parametric estimation of the PSD is compared

to the periodogram[4], the classic non-parametric estimator for the PSD. While these two estimates of the spectral density seem in disagreement, Figure 3.8 shows that such apparent inconsistency is a statistical artifact due to the periodogram's biasedness. By bootstrapping the fitted CAR process – i.e. sampling a large set of CAR realizations with parameters $\hat{\mathbf{A}}$ and $(\hat{b}, \hat{\Sigma}, \hat{F})$ along $\mathcal{P}$ – and plotting their corresponding periodograms (in green) we note that the periodogram is a significantly biased estimator for the true PSD (the blue line). The observed periodogram data is well-aligned with the bulk of the simulated periodogram distribution, suggesting there is no evidence to reject the hypothesis of the observed data being an observation of the fitted CAR model. Finally, we plot the measurement error PSD $S_\epsilon(f) \propto \hat{\sigma}_\epsilon^2$ where $\hat{\sigma}_\epsilon^2$ is the variance of measurement residuals, i.e. the difference between actual measurements and smoothed data. The measurement error level gives a rough estimate of the extent to which the smoothing operation contaminates spectral information.

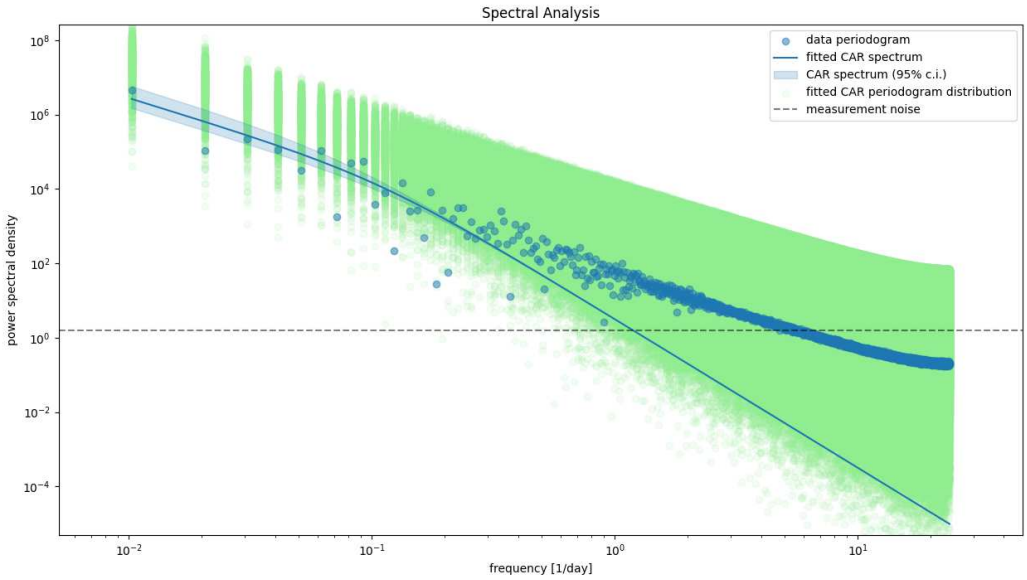


Figure 3.8: Spectral analysis of the Kepler light curve Zw 229-15 in Q9.

3.7 Conclusions and outlook

This chapter introduced novel estimation and inference methods for the drift parameter of a general multivariate Lévy-driven CAR process. The main contributions can be summarized as follows. By working on the path space of the state-space process, we showed the maximum-likelihood estimators based on continuous-time observations are consistent and asymptotically normal. Next, we addressed the more realistic scenario of discrete-time observations. We considered natural

[4]Note that since the curve is irregularly sampled the periodogram's discrete-parameter theoretical properties do not hold but, since the observation partition $\mathcal{P}$ is nearly regular, it still provides reasonable information about the curve's spectral properties.

discretizations of the quantities appearing in the continuous-time setting and proved that the resulting estimators possess the same asymptotic properties. Importantly, the proposed methods also accommodate irregularly spaced observations, which often pose a practical modeling challenge. To prove the desired asymptotic results we carried out a rigorous analysis of the finite difference approximation of the derivatives of the CAR process. To the best of our knowledge, this was the first time such kind of analysis was done for a differentiable stochastic process. Moreover, we addressed what we deem to be inconsistencies in the proofs of the asymptotic properties of the corresponding estimator for the drift parameter of a univariate OU process found in Mai (2014). Under slightly different assumptions we provided modified arguments establishing the desired results in the more general context of multivariate CAR processes. Finally, considering the case where the multivariate observation possesses a known underlying network structure, we introduced the GrCAR architecture, a natural extension of the GrOU process from Courgeau and Veraart (2022), along with the corresponding continuous-time and discrete-time estimators. The reduced parametrization offers an additional modeling advantage when dealing with graphically structured data.

Moving forward, there are several directions for future research that can build upon the findings of this chapter. Firstly, it is important to rigorously explore the practical challenges discussed in Section 3.4.5, for example by developing a joint estimation framework for the CAR drift and the parameters of the driving Lévy process. Investigating methods that simultaneously estimate these components could potentially enhance modeling accuracy and capture the interplay between structure of the CAR process and the underlying Lévy driver. From a practitioner's perspective, exploring different modeling applications of the GrCAR specification represents an exciting area for future empirical work. In such an applied context the GrCAR architecture provides a flexible framework to capture relatively simple relationships in very large networks from (possibly irregular) high-frequency observations. A related interesting problem is to consider the scenario where the adjacency matrix is unknown and needs to be estimated from an observation of the GrCAR process. Developing methodologies that effectively estimate the adjacency matrix can provide insights into the underlying network structure and improve the understanding of complex systems with a graphical topology. Finally, we believe it would be of great research interest to explore the estimation and inference of more general CARMA processes, possibly by extending the results discussed in this chapter for CAR processes. CARMA models incorporate moving average components, allowing for more flexible modeling of time series data. Developing estimators for CARMA processes with strong theoretical underpinnings would provide a more comprehensive toolbox for modeling and analyzing complex (spatio-)temporal data.

Appendix

3.A Preliminaries

3.A.1 Lévy processes

We consider a filtered probability space $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \in \mathbb{R}\}, \mathbb{P}')$ to which all stochastic processes are adapted. Let $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ be a d-dimensional Lévy process, i.e. a stochastic process that is continuous in probability with stationary and independent increments, $\mathbf{L}_0 = \mathbf{0},\ \mathbb{P}' - \text{a.s.}$, which we assume without loss of generality to be càdlàg. Two well-known results about Lévy processes are the Lévy-Khintchine representation and the Lévy-Itô decomposition:

- (Lévy-Khintchine representation, Sato (1999, Theorem 8.1)) For each $t > 0$, the law $\mathcal{L}(\mathbf{L}_t) := (\mathbf{L}_t)_* \mathbb{P}'$ of the random variable $\mathbf{L}_t$ is infinitely divisible with characteristic triplet $(t\mathbf{b}, t\Sigma, tF)$, i.e. $\mathbf{b} \in \mathbb{R}^d$, Σ is a symmetric positive semi-definite $d \times d$ matrix and F is a Lévy measure on $\mathbb{R}^d$, that is $F(\{0\}) = 0$ and $\int_{\mathbb{R}^d} (\|x\|^2 \wedge 1) F(\mathrm{d}x) < \infty$, such that

  $$\mathbb{E}'\left[\exp\{iz^{\mathrm{T}}\mathbf{L}_t\}\right] = \exp\left\{t\left[i\mathbf{b}^{\mathrm{T}}z - \frac{1}{2}z^{\mathrm{T}}\Sigma z + \int_{\mathbb{R}^d}(e^{iy^{\mathrm{T}}z} - 1 - iy^{\mathrm{T}}z\tau(y))F(\mathrm{d}y)\right]\right\}, \quad z \in \mathbb{R}^d,$$

  where $\tau : \mathbb{R}^d \to \mathbb{R}$ is a fixed truncation function, i.e. a bounded measurable function such that $\tau(x) = 1 + o(\|x\|)$ as $\|x\| \to 0$ and $\tau(x) = \mathcal{O}(\|x\|^{-1})$ as $\|x\| \to \infty$. We say that $(\mathbf{b}, \Sigma, F)$ are the characteristics of $\mathbb{L}$.

  Remark 3.A.1. Strictly speaking, the triplet parameter $\mathbf{b}$ depends on the choice of truncation function in the Lévy-Khintchine representation (cf. Sato (1999, Remark 8.4)). In this chapter, we always take the truncation function to be $\tau(x) := \mathbb{1}_{\{\|x\| \leq 1\}}$ but other choices are possible.

- (Lévy-Itô decomposition, Sato (1999, Theorem 19.2)) One has the following representation for the Lévy process $\mathbb{L}$ with characteristics $(\mathbf{b}, \Sigma, F)$: for $t \geq 0$, $\mathbb{P}'-$a.s.

  $$\mathbf{L}_t = \mathbf{b}t + \Sigma^{1/2}\mathbf{W}_t + \int_0^t \int_{\{x:\|x\|>1\}} x\mu(\mathrm{d}t, \mathrm{d}x) + \int_0^t \int_{\{x:\|x\|\leq 1\}} x\{\mu(\mathrm{d}t, \mathrm{d}x) - \mathrm{d}t\, F(\mathrm{d}x)\}, \quad (3.A.1)$$

  where:

  - $\mathbb{W} = \{\mathbf{W}_t,\ t \geq 0\}$ is a standard Brownian motion on $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \in \mathbb{R}\}, \mathbb{P}')$ and $\Sigma^{1/2}$ is a $d \times d$ matrix s.t. $\Sigma = \Sigma^{1/2}\Sigma^{1/2,\mathrm{T}}$;
  - $\{\mu(B) : B \in \mathcal{B}((0,\infty) \times \mathbb{R}^d \setminus \{0\})\}$ is a Poisson random measure independent of $\mathbb{W}$ such

that

$$\mu(\mathrm{d}t, \mathrm{d}x) = \sum_{s \geq 0 : \Delta \mathbf{L}_s \neq 0} \delta_{(s, \Delta \mathbf{L}_s)}(\mathrm{d}t, \mathrm{d}x),$$

and $\Delta\mathbb{L} = \{\Delta \mathbf{L}_t,\ t \geq 0\}$ is the jump process of $\mathbb{L}$, i.e. $\Delta \mathbf{L}_t = \mathbf{L}_t - \mathbf{L}_{t-}$,

– we define the integral

$$\int_0^t \int_{\{x : \|x\| \leq 1\}} x\{\mu(\mathrm{d}t, \mathrm{d}x) - \mathrm{d}t F(\mathrm{d}x)\} := \lim_{\epsilon \downarrow 0} \int_0^t \int_{\{x : \epsilon < \|x\| \leq 1\}} x\{\mu(\mathrm{d}t, \mathrm{d}x) - \mathrm{d}t F(\mathrm{d}x)\}.$$

3.A.2 Stochastic integration

In order to define an MCAR process we require a notion of stochastic differential, or, more formally, that of the stochastic integral

$$(H \cdot \mathbf{L})_t := \int_0^t H_s \mathrm{d}\mathbf{L}_s, \quad t \geq 0,$$

for general (possibly stochastic) $\mathbb{R}^{m \times d}$-valued integrands $\mathbb{H} = \{H_t,\ t \geq 0\}$. When considering a Lévy process $\mathbb{L}$ as integrator we will define the integral above via the definition of Lévy type stochastic integral given in Applebaum (2009, Section 4.3.3) and Applebaum (2009, Example 4.3.2). For any predictable process $\mathbb{H} = \{H_t,\ t \geq 0\}$ s.t. $\int_0^t \|H_s\|^2 \mathrm{d}s < \infty$, $\mathbb{P}'$-a.s. we use the Lévy-Itô decomposition to define for any $t \geq 0$

$$\begin{aligned} \int_0^t H_s \mathrm{d}\mathbf{L}_s = &\int_0^t H_s \mathbf{b}\, \mathrm{d}s + \int_0^t H_s \Sigma^{1/2} \mathrm{d}\mathbf{W}_s \\ &+ \int_0^t \int_{\{x : \|x\| > 1\}} H_s x \mu(\mathrm{d}s, \mathrm{d}x) + \int_0^t \int_{\{x : \|x\| \leq 1\}} H_s x \{\mu(\mathrm{d}s, \mathrm{d}x) - \mathrm{d}s\, F(\mathrm{d}x)\}. \end{aligned} \quad (3.A.2)$$

Here we use the following definitions:

- The first integral $\int_0^t H_s \mathbf{b}\, \mathrm{d}s$ is a (stochastic) Lebesgue-Stieljes integral w.r.t. the Lebesgue measure on $\mathbb{R}$, i.e. for each $\omega \in \Omega', t \geq 0$ one defines $\int_0^t H_s \mathbf{b}\, \mathrm{d}s(\omega) := \int_0^t H_s(\omega) \mathbf{b}\, \mathrm{d}s$ as a Lebsgue-Stieljes integral.

- The second integral $\int_0^t H_s \Sigma^{1/2} \mathrm{d}\mathbf{W}_s$ is a stochastic Itô integral w.r.t. Brownian motion defined via an L_2-limit construction. The original theory was developed by Itô (1944) for Brownian motion and then extended to general square-integrable martingales by Kunita and Watanabe (1967). Expositions of the general theory can be found in Karatzas and Shreve (1996) when the integrator is continuous or in Jacod and Shiryaev (1987, Theorem I.4.40), when the integrator is discontinuous (slightly different classes of integrands need to be considered in the two cases, progressively measurable integrands when the integrator is continuous and predictable integrands, a smaller class, when the integrator is discontinuous). Applebaum (2009, Section 4.3.1) gives a slightly different but equivalent construction by treating $\mathbb{W}$ as an $\mathbb{R}^d$-valued martingale measure on $\mathbb{R}$.

- The third integral is defined as $\int_0^t \int_{\{x:\|x\|>1\}} H_s x \mu(\mathrm{d}s, \mathrm{d}x) = \sum_{0\leq s\leq t:\|\Delta \mathbf{L}_s\|>1} H_s \Delta \mathbf{L}_s$.
- The last integral $\int_0^t \int_{\{x:\|x\|\leq 1\}} H_s x\{\mu(\mathrm{d}s, \mathrm{d}x) - \mathrm{d}s\, F(\mathrm{d}x)\}$ is a stochastic integral w.r.t. the $(2, F)$-martingale valued measure $M(\mathrm{d}t, \mathrm{d}x) := \mu(\mathrm{d}t, \mathrm{d}x) - \mathrm{d}t\, F(\mathrm{d}x)$, cf. Applebaum (2009, Section 4.2.2) where $(2, F)$-martingale valued measures and the corresponding integrals are defined.

Remark 3.A.2. There are less general definitions of the integral which are consistent with this construction when suitable stricter conditions are satisfied:

- When $\mathbb{L}$ is of finite variation this coincides $\mathbb{P}'$-a.s. with a stochastic pathwise Lebesgue-Stieljes integral, see Millar (1972, page 314).

There are also more general definitions that are consistent with the above construction of the integral:

- Every Lévy process is a semimartingale, so any definition of the stochastic integral for semimartingale integrators applies to Lévy integrators:
  - The uniformly on compacts in probability (ucp) limit construction of Protter (1990). In the simplest case the integrand $\mathbb{H}$ is an adapted càglàd process but the construction can be extended to more general predictable processes $\mathbb{H}$ which satisfy joint conditions with $\mathbb{X}$, cf. Protter (1990).
  - The L_2-limit construction of Dellacherie and Meyer (1978) (also treated in Jacod and Shiryaev (1987, Theorem I.4.31)). This approach uses the decomposition of a semimartingale into a predictable finite variation process and a local martingale to define the stochastic integral, as discussed below.
- Every Lévy process can be viewed as an $\mathbb{R}^d$-valued Lévy random measure on $\mathbb{R}$, so definitions of stochastic integral w.r.t. infinitely divisible random measures apply. For example, Rajput and Rosinski (1989) construct the integral using the $\mathbb{P}'$-limit construction of Urbanik and Woyczynski (1967).

In the following section we will also be integrating with respect to general semimartingales, i.e. we will be considering processes $\mathbb{H} \cdot \mathbb{X} = \{(H \cdot \mathbf{X})_t,\ t \geq 0\}$ where

$$(H \cdot \mathbf{X})_t := \int_0^t H_s \mathrm{d}\mathbf{X}_s,$$

for $\mathbb{R}^d$-valued semimartingale $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ and suitable $\mathbb{R}^{m\times d}$-valued integrand $\mathbb{H} = \{H_t,\ t \geq 0\}$. When considering such integrals, we will be using the definition of the stochastic integral obtained by extending the classical L_2-integration theory of square-integrable martingales to semimartingales.

To do so, one decomposes $\mathbb{X} = \mathbb{A} + \mathbb{M}$, where $\mathbb{A} = \{\mathbf{A}_t,\ t \geq 0\}$ is a predictable finite variation process and $\mathbb{M} = \{\mathbf{M}_t,\ t \geq 0\}$ is a local martingale and defines

$$\int_0^t H_s \mathrm{d}\mathbf{X}_s = \int_0^t H_s \mathrm{d}\mathbf{A}_s + \int_0^t H_s \mathrm{d}\mathbf{M}_s,$$

where the first integral is defined as a stochastic Lebesgue-Stieljes integral while the latter is obtained by an extension via localization of the classical L_2-theory for square-integrable martingales by Kunita and Watanabe (1967). For further details see Jacod and Shiryaev (1987, Theorem I.4.31), which follows a similar exposition as that in Dellacherie and Meyer (1978). Note that the integral in the most general form is defined for any locally bounded predictable process $\mathbb{H}$, that is $\mathbb{H} = \{H_t,\ t \geq 0\}$ needs to be predictable, i.e. measurable w.r.t. the σ-algebra generated by all left-continuous adapted processes, and locally bounded, i.e. there exists a localizing sequence of stopping times $(\tau_n)_{n\geq 0}$ increasing to infinity such that the stopped process $\mathbb{H}^{\tau_n}$ is uniformly bounded for all $n \geq 0$. In particular, it holds that $\mathbb{H} \cdot \mathbb{X}$ is also a semimartingale and, whenever $\mathbb{X}$ is a local martingale, then so is $\mathbb{H} \cdot \mathbb{X}$. The way these integrals are defined automatically ensures that when observations of the continuous-time processes are only available on a discrete grid, the left Riemann-Stieltjes sums constructed from the discrete observations yield a natural approximator for the stochastic integral (under suitable continuity conditions for the integrand). This is of fundamental importance when constructing estimators from discrete observations in Section 3.4.

3.A.3 Canonical space

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t,\ t \geq 0\})$ denote the canonical space of $\mathbb{R}^{pd}$-valued paths, i.e.

- $\Omega = D([0,\infty); \mathbb{R}^{pd})$ the space of càdlàg processes $\omega : [0,\infty) \to \mathbb{R}^{pd}$,
- $\{\mathcal{F}_t,\ t \geq 0\}$ is the filtration s.t. $\forall t \geq 0$

$$\mathcal{F}_t = \bigcap_{u>t} \sigma\Big(\omega(s) : 0 \leq s \leq u\Big),$$

- $\mathcal{F}$ is the smallest sigma algebra containing $\{\mathcal{F}_t,\ t \geq 0\}$.

This is the (filtered) probability space on which the probability measures $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ for $\mathbf{x}_0 \in \mathbb{R}^{pd}$ and $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$ are defined.

Remark 3.A.3. Here $\sigma(\omega(s) : s \in \mathcal{I})$ indicates the smallest sigma algebra such that $\omega \in D([0,\infty)) \mapsto \omega(s) \in \mathbb{R}^{pd}$ is measurable for each $s \in \mathcal{I}$. We have that $\mathcal{F} = \sigma\left(\omega(s) : s \geq 0\right)$ and hence $\mathcal{F}$ can be characterized as the restriction to $D([0,\infty))$ of the product σ-algebra $\otimes_{\mathbb{R}_+} \mathcal{B}(\mathbb{R}^{pd})$ defined as the smallest σ-algebra such that $\pi \in \{f \text{ s.t. } f : [0,\infty) \to \mathbb{R}^{pd}\} \mapsto \pi(s) \in \mathbb{R}^{pd}$ is measurable for each $s \in \mathbb{R}_+$. Alternatively, we can define $\mathcal{F}$ as the σ-algebra generated by cylinder sets of the form

$$C = \{\omega \in D([0,\infty)) \text{ s.t. } \omega(t_1) \in A_1, \ldots, \omega(t_n) \in A_N\},$$

for any $n \in \mathbb{N}, t_1 < \ldots < t_n \in \mathbb{R}_+$ and $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R}^{pd})$. Moreover, by Ethier and Kurtz (1986, Proposition III.7.1), this σ-algebra is the same as the Borel σ-algebra generated by the Skorokhod topology on $D([0,\infty))$. It is a well-known property that $(\Omega, \mathcal{F})$ is a Polish space.

For a parameter $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$ consider the pd-dimensional state-space representation $\mathbb{X}_{\mathbf{A},\mathbf{x}_0} = \{\mathbf{X}_{\mathbf{A},\mathbf{x}_0,t},\ t \geq 0\}$ of a MCAR(p) process driven by a Lévy process $\mathbb{L}$ with characteristics $(\mathbf{b}, \Sigma, F)$, parameter $\mathbf{A}$ and deterministic initial condition $\xi = \mathbf{x}_0$ on an arbitrary stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \geq 0\}, \mathbb{P}')$. Under $\mathbb{P}'$, the process $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}$ is a semimartingale with local characteristics $(\mathbf{B}'_{\mathbf{A}}, C', \nu')$ on $(\Omega', \mathcal{F}', \{\mathcal{F}'_t,\ t \geq 0\}, \mathbb{P}')$ given by

$$\begin{aligned} \mathbf{B}'_{\mathbf{A},t}(\omega) &= \int_0^t (\mathcal{A}_{\mathbf{A}} \mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}(\omega) + \mathcal{E}\mathbf{b}) \mathrm{d}s = \mathcal{A}_{\mathbf{A}} \int_0^t \mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}(\omega) \mathrm{d}s + \mathcal{E}\mathbf{b}t, \\ C'_t(\omega) &= \int_0^t \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}} \mathrm{d}s = \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}} t, \\ \nu'(\omega; \mathrm{d}t, \mathrm{d}x) &= \mathrm{d}t \int_{\mathbb{R}^d} \mathbb{1}_{\{\mathrm{d}x\}}(\mathcal{E}z) F(\mathrm{d}z) = \mathrm{d}t\, \delta_0(\mathrm{d}x_{-p})\, F(\mathrm{d}x_p), \end{aligned} \tag{3.A.3}$$

where we write $x = (x_1^{\mathrm{T}}, \ldots, x_p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{pd}$ and $x_{-p} = (x_1^{\mathrm{T}}, \ldots, x_{p-1}^{\mathrm{T}})^{\mathrm{T}}$ for $x_1, \ldots, x_p \in \mathbb{R}^d$.

3.B Properties of the state-space representation of the MCAR(p) process

Note that the state-space representation of an $\mathbb{R}^d$-dimensional MCAR(p) process $\mathbb{Y} = \{\mathbf{Y}_t,\ t \geq 0\}$ driven by $\mathbb{L} = \{\mathbf{L}_t,\ t \geq 0\}$ with coefficient matrices $\mathbf{A} = (A_1, \ldots, A_p) \in (\mathcal{M}_d(\mathbb{R}))^p$ and initial state-space representation $\mathbf{X}_0$ is the $\mathbb{R}^{pd}$-dimensional Ornstein-Uhlenbeck process $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ driven by the (degenerate) Lévy process $\mathcal{E}\mathbb{L} := \{\mathcal{E}\mathbf{L}_t,\ t \geq 0\}$ and dynamics matrix

$$\mathcal{A} = \begin{pmatrix} 0_{d\times d} & I_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} & I_{d\times d} & \cdots & 0_{d\times d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{d\times d} & 0_{d\times d} & 0_{d\times d} & \cdots & I_{d\times d} \\ -A_p & -A_{p-1} & -A_{p-2} & \cdots & -A_1 \end{pmatrix},$$

i.e. it can be written as

$$\mathbf{X}_t = e^{\mathcal{A}t}\mathbf{X}_0 + \int_0^t e^{\mathcal{A}(t-s)} \mathcal{E}\mathrm{d}\mathbf{L}_s, \quad t \geq 0.$$

We can thus apply standard results from the theory of multivariate Ornstein–Uhlenbeck processes driven by a general Lévy process (Masuda, 2004; Sato and Yamazato, 1984). For example, we have that for $t > 0$, the law of $\mathbf{X}_t$ given $\mathbf{X}_0 = \mathbf{x}_0$ is infinitely divisible with characteristics $(\mathbf{b}_{t,\mathbf{x}_0,\mathcal{A}}, \Sigma_{t,\mathcal{A}}, F_{t,\mathcal{A}})$ given by

$$\mathbf{b}_{t,\mathbf{x}_0,\mathcal{A}} = e^{t\mathcal{A}}\mathbf{x}_0 + \int_0^t e^{s\mathcal{A}} \mathcal{E}\mathbf{b}\, \mathrm{d}s$$

$$+\int_{\mathbb{R}^{pd}}\int_0^t e^{s\mathcal{A}}z\left(\mathbb{1}_{\{y:\|y\|\leq 1\}}\left(e^{s\mathcal{A}}z\right)-\mathbb{1}_{\{y:\|y\|\leq 1\}}(z)\right)\mathrm{d}s\,\delta_0(\mathrm{d}z_{-p})\,F(\mathrm{d}z_p),$$

$$\Sigma_{t,\mathcal{A}}=\int_0^t e^{s\mathcal{A}}\mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}e^{s\mathcal{A}^{\mathrm{T}}}\mathrm{d}s,$$

$$F_{t,\mathcal{A}}(E)=\int_{\mathbb{R}^{pd}}\int_0^t \mathbb{1}_E\left(e^{s\mathcal{A}}z\right)\mathrm{d}s\,\delta_0(\mathrm{d}z_{-p})\,F(\mathrm{d}z_p),\quad \forall E\in\mathcal{B}(\mathbb{R}^{pd}),$$

where $(\mathbf{b},\Sigma,F)$ are the characteristics of the driving Lévy process $\mathbb{L}=\{\mathbf{L}_t,\ t\geq 0\}$, cf. Sato and Yamazato (1984, Theorem 3.1). If the driving Lévy process $\mathbb{L}$ is assumed to have second moments then also $\mathbb{X}$ will have second moments and we can explicitly compute

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}_t-\mathbf{X}_0\|^2]&=\mathrm{tr}\,(\Sigma_{t,\mathcal{A}})+\int_{\mathbb{R}^{pd}}\|z\|^2F_{t,\mathcal{A}}(\mathrm{d}z)+\mathbb{E}\left[\left\|\mathbf{X}_0-\left(\mathbf{b}_{t,\mathbf{X}_0,\mathcal{A}}+\int_{\{z:\|z\|>1\}}zF_{t,\mathcal{A}}(\mathrm{d}z)\right)\right\|^2\right]\\&=\mathrm{tr}\left(\int_0^t e^{s\mathcal{A}}\mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}e^{s\mathcal{A}^{\mathrm{T}}}\mathrm{d}s\right)+\int_{\mathbb{R}^d}\int_0^t\|e^{s\mathcal{A}}\mathcal{E}y\|^2F(\mathrm{d}y)\mathrm{d}s+\\&\qquad+\mathbb{E}\left[\left\|(e^{t\mathcal{A}}-I)\mathbf{X}_0+\int_0^t e^{(t-s)\mathcal{A}}\mathcal{E}\mathbb{E}[\mathbf{L}_1]\mathrm{d}s\right\|^2\right].\end{aligned}$$

And one can thus obtain the bound

$$\mathbb{E}[\|\mathbf{X}_t-\mathbf{X}_0\|^2]=\mathcal{O}\left(\int_0^t e^{2s\|\mathcal{A}\|}\mathrm{d}s+t\|\mathcal{A}\|e^{t\|\mathcal{A}\|}\right)=\mathcal{O}\left(t(e^{2t\|\mathcal{A}\|}+e^{t\|\mathcal{A}\|})\right)=\mathcal{O}(t),\quad t\to 0. \tag{3.B.1}$$

3.C Continuous observations: proofs

3.C.1 Proof of Proposition 3.3.3

A brief outline of this proof was given in Section 3.3.1. We shall now provide the complete details. Let $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}=\{\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s},\ t\geq 0\}$ denote a MCAR(p) process driven by a Lévy process $\mathbb{L}$ with characteristics $(\mathbf{b},\Sigma,F)$ on the stochastic basis $(\Omega',\mathcal{F}',\{\mathcal{F}'_s,\ s\geq 0\},\mathbb{P}')$ and parameter $\mathbf{A}$, i.e. such as it satisfies SDE (3.3.1). Recall that $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ represents the unique solution-measure to the SDE (3.3.1) with initial condition $\mathbf{X}_0=\mathbf{x}_0$, i.e. the unique solution to the martingale problem for $(\mathbf{B}_\mathbf{A},C,\nu)$ on the canonical space with initial condition $\eta=\delta_{\mathbf{x}_0}$, where $(\mathbf{B}_\mathbf{A},C,\nu)$ are defined in Equation (3.3.2).

Fix a deterministic time horizon $t>0$, then, under $\mathbb{P}'$, the stopped process $\mathbb{X}^t_{\mathbf{A},\mathbf{x}_0}=\{\mathbf{X}^t_{\mathbf{A},\mathbf{x}_0,s},\ s\geq 0\}$ such that $\mathbf{X}^t_{\mathbf{A},\mathbf{x}_0,s}(\omega)=\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s\wedge t}(\omega)$ satisfies the stopped SDE

$$\begin{aligned}\mathrm{d}\mathbf{X}_s&=(\mathcal{A}_\mathbf{A}\mathbf{X}_s+\mathcal{E}\mathbf{b})\mathbb{1}_{[0,t]}(s)\mathrm{d}s+\mathcal{E}\Sigma^{1/2}\mathbb{1}_{[0,t]}(s)\mathrm{d}\mathbf{W}_s\\&\quad+\int_{\|z\|\leq 1}\mathcal{E}z\{\mu^t(\mathrm{d}s,\mathrm{d}z)-\mathbb{1}_{[0,t]}(s)F(\mathrm{d}z)\mathrm{d}s\}+\int_{\|z\|>1}\mathcal{E}z\mu^t(\mathrm{d}s,\mathrm{d}z),\end{aligned} \tag{3.C.1}$$

where $\mu^t(\mathrm{d}s,\mathrm{d}z)=\mathbb{1}_{[0,t]}(s)\mu^t(\mathrm{d}s,\mathrm{d}z)$ is again a Poisson random measure on $\mathbb{R}_+\times\mathbb{R}^d$ with compen-

sator $\mathbb{1}_{[0,t]}(s)F(\mathrm{d}z)\mathrm{d}s$ (abusing notation slightly). The process $\mathbb{X}^t_{\mathbf{A},\mathbf{x}_0}$ is a semimartingale with local characteristics $(\mathbf{B}'^{,t}_{\mathbf{A}}, C'^{,t}, \nu'^{,t})$ (Jacod and Mémin, 1976, Proposition 2.4) where $\mathbf{B}'^{,t}_{\mathbf{A}}$ and $C'^{,t}$ denote the stopped at $t > 0$ processes $\mathbf{B}'_{\mathbf{A}}$ and C' respectively and

$$\nu'^{,t}(A) = \int_A \mathbb{1}_{[0,t]}(s)\nu'(\mathrm{d}s, \mathrm{d}x), \quad \forall A \in \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd}).$$

We start by showing that for any $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ the martingale problem for $(\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t)$ with deterministic initial condition has a unique solution. The characteristics $(\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t)$ are defined on the canonical space by

$$\begin{aligned} \mathbf{B}^t_{\mathbf{A},s}(\omega) = \mathcal{A}_{\mathbf{A}} \int_0^{s\wedge t} \omega(u)\mathrm{d}u + \mathcal{E}\mathbf{b}(s \wedge t), \quad C^t_s(\omega) = \tilde{\Sigma}(s \wedge t), \\ \nu^t(\omega; \mathrm{d}s, \mathrm{d}x) = \mathbb{1}_{[0,t]}(s)\, \mathrm{d}s\, \delta_0(\mathrm{d}x_{-p})\, F(\mathrm{d}x_p), \end{aligned} \tag{3.C.2}$$

and we write $x = (x_1^{\mathrm{T}}, \ldots, x_p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{pd}$ and $x_{-p} = (x_1^{\mathrm{T}}, \ldots, x_{p-1}^{\mathrm{T}})^{\mathrm{T}}$ for $x_1, \ldots, x_p \in \mathbb{R}^d$.

Lemma 3.C.1. *Let $\mathbb{X}$ denote the canonical process on $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$. For any $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ has a unique solution.*

Proof of Lemma 3.C.1. Uniqueness can be proved as follows. First, we note that there exists a solution-process to the stopped SDE (3.C.1) starting at $\mathbf{x}_0$ on $(\Omega', \mathcal{F}', \{\mathcal{F}'_s,\ s \geq 0\}, \mathbb{P}')$ given by $\mathbb{X}^t_{\mathbf{A},\mathbf{x}_0}$ (and this implies the existence of a solution-measure to SDE (3.C.1)). If $\mathbb{Y}$ is another solution-process of SDE (3.C.1) starting at $\mathbf{x}_0$ then the difference $\mathbb{D}_{\mathbf{A}} = \mathbb{Y} - \mathbb{X}^t_{\mathbf{A},\mathbf{x}_0}$ must satisfy the ODE

$$\mathrm{d}\mathbf{D}_s = \mathcal{A}_{\mathbf{A}}\mathbf{D}_s\mathbb{1}_{[0,t]}\mathrm{d}s, \quad \mathbf{D}_0 = 0,$$

$\mathbb{P}'$-a.s. and hence $\mathbb{D}_{\mathbf{A}} \equiv 0$, $\mathbb{P}'$-a.s. Thus SDE (3.C.1) has a unique solution-process (for any stochastic basis $(\Omega', \mathcal{F}', \{\mathcal{F}'_s,\ s \geq 0\}, \mathbb{P}')$). By Jacod and Shiryaev (1987, Theorem III.2.33) this implies uniqueness of the solution-measure with initial condition $\delta_{\mathbf{x}_0}$. In turn, by Jacod and Shiryaev (1987, Theorem III.2.26), this implies uniqueness of the solution to the martingale problem on the canonical space $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$ for the characteristics $(\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t)$ and initial condition $\eta = \delta_{\mathbf{x}_0}$. The same reasoning holds with $\mathbf{A}$ replaced by $\mathbf{A}^{(0)}$. □

Proof of Proposition 3.3.3. According to Lemma 3.C.1, the martingale problems $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ and $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}^{(0)}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ have unique solutions. Let us denote by $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ such unique measures on canonical path space $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$. Then we clearly have that the restricted measures on $(\Omega, \mathcal{F}_t)$ coincide, i.e.

$$\mathbb{Q}^t_{\mathbf{A},\mathbf{x}_0} = \mathbb{P}^t_{\mathbf{A},\mathbf{x}_0} \quad \text{and} \quad \mathbb{Q}^t_{\mathbf{A}^{(0)},\mathbf{x}_0} = \mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}.$$

Note that for $\omega \in \Omega$, $s \geq 0$

$$\mathbf{B}^t_{\mathbf{A},s}(\omega) = \mathbf{B}^t_{\mathbf{A}^{(0)},s}(\omega) + (\mathcal{A}_{\mathbf{A}} - \mathcal{A}_{\mathbf{A}^{(0)}}) \int_0^s \omega(u)\mathrm{d}u = \mathbf{B}^t_{\mathbf{A}^{(0)},s}(\omega) + \tilde{\mathcal{A}}_{\mathbf{A}} \int_0^s \omega(u-)\mathrm{d}u.$$

Due to the specific structure of $\tilde{\mathcal{A}}_{\mathbf{A}} := \mathcal{A}_{\mathbf{A}} - \mathcal{A}_{\mathbf{A}^{(0)}}$, we can write $\forall \mathbf{x} \in \mathbb{R}^{pd}$

$$\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{x} = \tilde{\Sigma}\tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{x},$$

where we define the pseudo-inverse $\tilde{\Sigma}^{-1}$ for $\tilde{\Sigma}$ as

$$\tilde{\Sigma}^{-1} := \begin{pmatrix} 0_{d\times d} & \cdots & 0_{d\times d} & 0_{d\times d} \\ \vdots & \ddots & \vdots & \vdots \\ 0_{d\times d} & \cdots & 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & \cdots & 0_{d\times d} & \Sigma^{-1} \end{pmatrix}.$$

We can thus define the $\mathbb{R}^{pd}$-valued predictable process $\mathbb{Z}_{\mathbf{A}} = \{\mathbf{Z}_{\mathbf{A},s},\ s \geq 0\}$ on $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$ such that $\mathbf{Z}_{\mathbf{A},s}(\omega) = \tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\omega(s-)$ and for $s \geq 0$, $\omega \in \Omega$

$$\mathbf{B}^t_{\mathbf{A},s}(\omega) = \mathbf{B}^t_{\mathbf{A}^{(0)},s}(\omega) + \int_0^s \mathrm{d}C^t_u \mathbf{Z}_{\mathbf{A},u}(\omega). \tag{3.C.3}$$

The results from Jacod and Mémin (1976) can be immediately generalized to the multivariate setting, cf. Sørensen (1991). We can thus check the hypothesis of Jacod and Mémin (1976, Theorem 4.2.(c)) for $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$:

- For $s \in \mathbb{R}_+$ let

$$A^t_{\mathbf{A},s}(\omega) := \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}(\omega)\mathrm{d}C^t_u \mathbf{Z}_{\mathbf{A},u}(\omega) = \int_0^{s\wedge t} \omega(u)^{\mathrm{T}} \tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}} \tilde{\Sigma}^{-1} \tilde{\mathcal{A}}_{\mathbf{A}} \omega(u)\mathrm{d}u,$$

  and define the $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$-stopping times (note that, by construction, this is a strict stopping time in the sense of Jacod and Shiryaev (1987, Definition III.2.35)):

$$\tau_n(\omega) = \inf\{s \geq 0 : A^t_{\mathbf{A},s}(\omega) \geq n\}.$$

  In order to check τ_n-uniqueness of the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$, in the sense of Jacod and Mémin (1976, Section 4), we show the (stronger) condition of local uniqueness, as defined in Jacod and Shiryaev (1987, Definition III.2.37). Due to the Markovian structure of the martingale problem we can apply Jacod and Shiryaev (1987, Theorem III.2.40) to the characteristics $(\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t)$ to deduce local uniqueness (note that, unlike Sørensen (1991), we cannot apply Jacod and Shiryaev (1987, Corollary III.2.41) directly because Sørensen (1991,

Condition C) is not satisfied). For $s \geq 0, u \geq 0, \omega \in \Omega$, define

$$\begin{aligned}(p_s \mathbf{B}_{\mathbf{A}}^t)_u(\omega) &= \mathbf{B}_{\mathbf{A},u}^{(t-s)^+}(\omega) = \mathcal{A}_{\mathbf{A}} \int_0^{u \wedge (t-s)^+} \omega(r)\mathrm{d}r + \mathcal{E}\mathbf{b}(u \wedge (t-s)^+), \\ (p_s C^t)_u(\omega) &= C_u^{(t-s)^+}(\omega) = \tilde{\Sigma}(u \wedge (t-s)^+), \\ (p_s \nu^t)(\omega; \mathrm{d}u, \mathrm{d}x) &= \nu^{(t-s)^+}(\omega; \mathrm{d}u, \mathrm{d}x) = \mathbb{1}_{[0,(t-s)^+]}(u)\, \mathrm{d}u\, \delta_0(\mathrm{d}x_{-p})\, F(\mathrm{d}x_p).\end{aligned}$$

Note that for fixed $u \geq 0$ and $A \in \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd})$ the maps

$$(s, \omega) \mapsto (p_s \mathbf{B}_{\mathbf{A}}^t)_u(\omega), \ (s, \omega) \mapsto (p_s C^t)_u(\omega), \text{ and } (s, \omega) \mapsto (p_s \nu^t)(\omega; A)$$

are clearly $\mathcal{F} \otimes \mathcal{B}(\mathbb{R}_+)$-measurable. Moreover, letting $\theta : \Omega \to \Omega$ denote the shift operator, it can be easily verified that the following relationships are satisfied $\forall s, u \geq 0$, $\forall \omega \in \Omega$, $\forall A \in \mathcal{B}(\mathbb{R}^{pd})$:

$$\begin{aligned}(p_s \mathbf{B}_{\mathbf{A}}^t)_u(\theta_s \omega) &= \mathbf{B}_{\mathbf{A},s+u}^t(\omega) - \mathbf{B}_{\mathbf{A},s}^t(\omega), \\ (p_s C^t)_u(\theta_s \omega) &= C_{s+u}^t(\omega) - C_s^t(\omega), \\ (p_s \nu^t)(\theta_s \omega; (0, u] \times A) &= \nu^t(\omega; (s, s+u] \times A),\end{aligned}$$

i.e. the mapping defined by p_s satisfies Jacod and Shiryaev (1987, Assumption III.2.39) (moreover Jacod and Shiryaev (1987, Assumption III.2.13) is implicit as we are working on the canonical space). Next, we note that, by Lemma 3.C.1, the martingale problem for

$$(\mathbb{X}, (\mathbf{B}_{\mathbf{A}}^{(t-s)\vee 0}, C^{(t-s)\vee 0}, \nu^{(t-s)\vee 0}), \delta_x),$$

admits a unique solution, $\mathbb{Q}_{\mathbf{A},s,x}$, which, for fixed $\mathbf{A}$, can be shown to define a transition kernel from $(\mathbb{R}_+ \times \mathbb{R}^{pd}, \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd}))$ to $(\Omega, \mathcal{F})$ by

$$\mathcal{Q}_{s,x}(\mathrm{d}\omega) = \mathbb{Q}_{\mathbf{A},s,x}(\mathrm{d}\omega),$$

cf. Lemma 3.C.3. Moreover, by setting $s = 0$, we know that $(\mathbb{X}, (\mathbf{B}_{\mathbf{A}}^t, C^t, \nu^t), \delta_{\mathbf{x}_0})$ admits a unique solution. All hypothesis of Jacod and Shiryaev (1987, Theorem III.2.40) are thus satisfied and we can hence conclude that the martingale problem for $(\mathbb{X}, (\mathbf{B}_{\mathbf{A}}^t, C^t, \nu^t), \delta_{\mathbf{x}_0})$ is locally unique. In particular, we have τ_n-uniqueness in the sense of Jacod and Mémin (1976, Section 4) for $(\mathbb{X}, (\mathbf{B}_{\mathbf{A}}^t, C^t, \nu^t), \delta_{\mathbf{x}_0})$ for all n.

- The initial conditions are both $\delta_{\mathbf{x}_0}$ and thus trivially equivalent.
- Finally we note that

$$A_{\mathbf{A},\infty}^t = \int_0^t \omega(u)^{\mathrm{T}} \tilde{\mathcal{A}}_{\mathbf{A}}^{\mathrm{T}} \tilde{\Sigma}^{-1} \tilde{\mathcal{A}}_{\mathbf{A}} \omega(u) \mathrm{d}u \leq \|\Sigma^{-1}\| \|\tilde{\mathcal{A}}_{\mathbf{A}}\| \int_0^t \|\omega(u)\|^2 \mathrm{d}u < \infty, \tag{3.C.4}$$

$\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ almost surely, where $\|\cdot\|$ denote appropriate arbitrary norms on the underlying (finite dimensional) spaces. Since the driving Lévy process $\mathbb{L}$ is assumed to have finite second moments on $(\Omega', \mathcal{F}', \{\mathcal{F}'_s,\ s \geq 0\}, \mathbb{P}')$, we have that, for $s \leq t$, a solution of the stopped SDE (3.C.1) with initial condition $\mathbf{X}_0 = \mathbf{x}_0$ has law $\mathbb{P}'(\mathbf{X}^t_{\mathbf{A},\mathbf{x}_0,s} \in \cdot) = \mathbb{P}'(\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s} \in \cdot) = P_{s,\mathbf{x}_0,\mathbf{A}}(\cdot)$ which is infinitely divisible with second moments and characteristics $(\mathbf{b}_{s,\mathbf{x}_0,\mathbf{A}}, \Sigma_{s,\mathbf{A}}, F_{s,\mathbf{A}})$ given by

$$\begin{aligned}
\mathbf{b}_{s,\mathbf{x}_0,\mathbf{A}} &= e^{s\mathcal{A}_\mathbf{A}}\mathbf{x}_0 + \int_0^s e^{u\mathcal{A}_\mathbf{A}}\mathcal{E}\mathbf{b}\,\mathrm{d}u \\
&\quad + \int_{\mathbb{R}^{pd}}\int_0^s e^{u\mathcal{A}_\mathbf{A}}z\left(\mathbb{1}_{\{y:\|y\|\leq 1\}}\left(e^{u\mathcal{A}_\mathbf{A}}z\right) - \mathbb{1}_{\{y:\|y\|\leq 1\}}(z)\right)\mathrm{d}u\,\delta_0(\mathrm{d}z_{-p})\,F(\mathrm{d}z_p), \\
\Sigma_{s,\mathbf{A}} &= \int_0^s e^{s\mathcal{A}_\mathbf{A}}\mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}e^{u\mathcal{A}_\mathbf{A}^{\mathrm{T}}}\mathrm{d}u, \\
F_{s,\mathbf{A}}(E) &= \int_{\mathbb{R}^{pd}}\int_0^s \mathbb{1}_E\left(e^{u\mathcal{A}_\mathbf{A}}z\right)\mathrm{d}u\,\delta_0(\mathrm{d}z_{-p})\,F(\mathrm{d}z_p), \quad \forall E \in \mathcal{B}(\mathbb{R}^{pd}),
\end{aligned}$$

where $(\mathbf{b}, \Sigma, F)$ are the characteristics of the driving Lévy process $\mathbb{L} = \{\mathbf{L}_s,\ s \geq 0\}$, cf. Sato and Yamazato (1984, Theorem 3.1). Thus we can bound uniformly in $s \in [0,t]$

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}'}\left[\|\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}\|^2\right] &= \mathrm{tr}(\mathrm{Var}^{\mathbb{P}'}[\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}]) + \|\mathbb{E}^{\mathbb{P}'}[\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}]\|^2 \\
&= \mathrm{tr}(\Sigma_{s,\mathbf{A}}) + \int_{\mathbb{R}^{pd}}\|z\|^2 F_{s,\mathbf{A}}(\mathrm{d}z) + \left\|\mathbf{b}_{s,\mathbf{x}_0,\mathbf{A}} + \int_{\{z:\|z\|>1\}} zF_{s,\mathbf{A}}(\mathrm{d}z)\right\|^2 \\
&\leq \kappa(t,\mathbf{x}_0,\mathcal{A}_\mathbf{A},\mathbf{b},\Sigma,F), \qquad (3.C.5)
\end{aligned}$$

for all $s \in [0,t]$, where $\kappa(t) = \kappa(t,\mathbf{x}_0,\mathcal{A}_\mathbf{A},\mathbf{b},\Sigma,F)$ is a constant which depends on $t, \mathbf{x}_0, \mathcal{A}_\mathbf{A}, \mathbf{b}, \Sigma$, and F. Thus we can show that $A^t_{\mathbf{A},\infty}$ is bounded in $L^1(\Omega,\mathcal{F},\mathbb{Q}_{\mathbf{A},\mathbf{x}_0})$

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}}[A^t_{\mathbf{A},\infty}] &\leq \|\Sigma^{-1}\|\|\tilde{\mathcal{A}}_\mathbf{A}\|\int_0^t \mathbb{E}^{\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}}\left[\|\omega(s)\|^2\right]\mathrm{d}s \\
&= \|\Sigma^{-1}\|\|\tilde{\mathcal{A}}_\mathbf{A}\|\int_0^t \mathbb{E}^{\mathbb{P}'}\left[\|\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}\|^2\right]\mathrm{d}s \leq t\|\Sigma^{-1}\|\|\tilde{\mathcal{A}}_\mathbf{A}\|\kappa(t,\mathbf{A}) < \infty,
\end{aligned}$$

and hence

$$A^t_{\mathbf{A},\infty} < \infty, \quad \mathbb{Q}_{\mathbf{A},\mathbf{x}_0}\text{-a.s.}$$

Similarly, we get

$$\mathbb{E}^{\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}}[A^t_{\mathbf{A},\infty}] \leq t\|\Sigma^{-1}\|\|\tilde{\mathcal{A}}_\mathbf{A}\|\kappa(t,\mathbf{A}^{(0)}) < \infty,$$

and hence

$$A^t_{\mathbf{A},\infty} < \infty, \quad \mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}\text{-a.s.}$$

The assumptions of Jacod and Mémin (1976, Theorem 4.2.(c)) for $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ are satisfied and hence we have

$$\mathbb{Q}_{\mathbf{A},\mathbf{x}_0} \sim \mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$$

which, by construction, implies

$$\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0} \sim \mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}.$$

In order to deduce the form of the likelihood we apply Jacod and Mémin (1976, Theorem 4.5.(b)) to $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$. Note that, by the arguments above, $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ is the (unique) solution to the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}^{(0)}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ on the canonical space $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$. Moreover we have that $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ is the unique solution to the martingale problem $(\mathbb{X}, (\mathbf{B}^t_{\mathbf{A}}, C^t, \nu^t), \delta_{\mathbf{x}_0})$ on $(\Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\})$ and hence satisfies the martingale representation condition (Jacod and Mémin, 1976, Theorem 4.4). Finally we note that by construction $\{A^t_{\mathbf{A},s},\ s \geq 0\}$ has continuous trajectories $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$-a.s. and hence for $\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}$-a.e. $\omega \in \Omega$ and $\forall s \geq 0$

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\left[\frac{\mathrm{d}\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\middle|\mathcal{F}_s\right](\omega) &= \exp\Big\{-\int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A}^{(0)},u}(\omega) - \frac{1}{2}\int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}(\omega)\mathrm{d}C^t_u\mathbf{Z}_{\mathbf{A},u}(\omega)\Big\} \\ &= \exp\Big\{-\int_0^s \mathbf{X}^{\mathrm{T}}_{u-}\tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A}^{(0)},u} - \frac{1}{2}\int_0^{s\wedge t}\mathbf{X}^{\mathrm{T}}_u\tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\tilde{\mathcal{A}}_{\mathbf{A}}\mathbf{X}_u\mathrm{d}u\Big\},\end{aligned}$$

and noting that, by construction, if we set $s = t$, for $\mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}$-a.e. $\omega \in \Omega$

$$\mathbb{E}^{\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\left[\frac{\mathrm{d}\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{Q}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\middle|\mathcal{F}_t\right](\omega) = \frac{\mathrm{d}\mathbb{Q}^t_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{Q}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}}(\omega) = \frac{\mathrm{d}\mathbb{P}^t_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{P}^t_{\mathbf{A}^{(0)},\mathbf{x}_0}}(\omega) = \mathbb{E}^{\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}{\mathrm{d}\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}}\middle|\mathcal{F}_t\right](\omega),$$

which completes the proof. □

Remark 3.C.2. In the general setting of Jacod and Mémin (1976, Theorem 4.2 and Theorem 4.5) the condition $A_{\mathbf{A},t} = A^t_{\mathbf{A},\infty} < \infty$, $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$-a.s. ensures the stochastic integral

$$(\mathbf{Z}^{\mathrm{T}}_{\mathbf{A}} \cdot \mathbf{X}^c_{\mathbf{A}})_s := \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A}^{(0)},u} = \int_0^s \mathbf{X}^{\mathrm{T}}_{u-}\tilde{\mathcal{A}}^{\mathrm{T}}_{\mathbf{A}}\tilde{\Sigma}^{-1}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A}^{(0)},u},$$

is well-defined for $s \leq t$. Note that one can write

$$\mathbf{X}^c_{\mathbf{A}^{(0)},s} = \mathbf{X}^c_{\mathbf{A},s} + (\mathbf{B}^t_{\mathbf{A}^{(0)},s} - \mathbf{B}^t_{\mathbf{A},s}), \quad s \in [0,t],$$

where $\{\mathbf{X}^c_{\mathbf{A},s}, s \in [0,t]\}$ is a continuous local martingale under $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ and $\{\mathbf{B}^t_{\mathbf{A}^{(0)},s} - \mathbf{B}^t_{\mathbf{A},s},\ s \in [0,t]\}$ is a finite variation process. Then one defines

$$\int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A}^{(0)},u} = \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\,\mathrm{d}\mathbf{X}^c_{\mathbf{A},u} + \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u}\,\mathrm{d}(\mathbf{B}^t_{\mathbf{A}^{(0)},u} - \mathbf{B}^t_{\mathbf{A},u}), \quad s \in [0,t],$$

where the first integral is well-defined as the integrand is in

$$\mathcal{P}^*(\mathbf{X}^c_{\mathbf{A}}; \Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\}, \mathbb{Q}_{\mathbf{A},\mathbf{x}_0}) := \left\{\mathbb{H} \text{ prog. meas. s.t. } \mathbb{Q}_{\mathbf{A},\mathbf{x}_0}\left(\int_0^t \mathbf{H}^{\mathrm{T}}_s d\langle \mathbf{X}^c_{\mathbf{A}}\rangle_s \mathbf{H}_s < \infty\right) = 1\right\},$$

cf. Karatzas and Shreve (1996, Theorem 3.2.23), and the second integral is defined for any optional

(and thus any predictable) integrand, cf. Jacod and Shiryaev (1987, Definition 3.4). In our specific setting the processes $\{\mathbf{X}^c_{\mathbf{A},s}, s \in [0,t]\}$ and $\{\mathbf{Z}_{\mathbf{A},s}, s \in [0,t]\}$ satisfy stricter conditions: the first is actually a true square-integrable martingale under $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$ since $\mathbb{E}^{\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}}[\langle \mathbf{X}^c_{\mathbf{A}} \rangle_s] = \tilde{\Sigma} s < \infty$, Protter (1990, Corollary II.3), and the latter is in

$$\mathcal{L}^*(\mathbf{X}^c_{\mathbf{A}}; \Omega, \mathcal{F}, \{\mathcal{F}_s,\ s \geq 0\}, \mathbb{Q}_{\mathbf{A},\mathbf{x}_0}) := \left\{ \mathbb{H} \text{ prog. meas. s.t. } \mathbb{E}_{\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}} \left[\int_0^t \mathbf{H}_s^{\mathrm{T}} d\langle \mathbf{X}^c_{\mathbf{A}} \rangle_s \mathbf{H}_s \right] < \infty \right\},$$

since $A^t_{\mathbf{A},\mathbf{x}_0}$ is bounded in $L^1(\Omega, \mathcal{F}, \mathbb{Q}_{\mathbf{A},\mathbf{x}_0})$. We can hence apply Karatzas and Shreve (1996, Definition 3.2.9) (or from the original work of Kunita and Watanabe (1967)) to deduce that the stochastic integral

$$(\mathbf{Z}^{\mathrm{T}}_{\mathbf{A}} \cdot \mathbf{X}^c_{\mathbf{A}})_s = \int_0^s \mathbf{Z}^{\mathrm{T}}_{\mathbf{A},u} \, \mathrm{d}\mathbf{X}^c_{\mathbf{A},u}, \quad s \in [0,t],$$

is not just well-defined but is actually a continuous square-integrable martingale. This observation is fundamentally linked to Lemma 3.3.8: the process $(\mathbb{Z}^{\mathrm{T}}_{\mathbf{A}} \cdot \mathbb{X}^c_{\mathbf{A}})$ is the sum of the components of the score, i.e. Equation (3.3.14), and is proved to be a $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$-square-integrable martingale (i.e. a $\mathbb{Q}_{\mathbf{A},\mathbf{x}_0}$-square-integrable martingale up to time t).

Lemma 3.C.3. *Fix $\mathbf{A} \in (\mathcal{M}_d(\mathbb{R}))^p$ and let $\mathcal{Q} : \mathbb{R}_+ \times \mathbb{R}^{pd} \times \mathcal{F} \to [0,1]$ be given by*

$$(s, x, A) \mapsto \mathcal{Q}_{s,x}(A) = \mathbb{Q}_{\mathbf{A},s,x}(A),$$

where $\mathbb{Q}_{\mathbf{A},s,x}$ is the unique solution to the martingale problem for

$$(\mathbb{X}, (\mathbf{B}^{(t-s)\vee 0}_{\mathbf{A}}, C^{(t-s)\vee 0}, \nu^{(t-s)\vee 0}), \delta_x).$$

Then $\mathcal{Q}$ defines a transition kernel from $(\mathbb{R}_+ \times \mathbb{R}^{pd}, \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd}))$ to $(\Omega, \mathcal{F})$.

Proof of Lemma 3.C.3. Clearly, for fixed (s,x), $\mathcal{Q}_{s,x}(\cdot)$ defines a probability measure on $(\Omega, \mathcal{F})$. Next, to show that for fixed $A \in \mathcal{F}$, $(s,x) \mapsto Q_{s,x}(A)$ is $\mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd})$-measurable, we use Dynkin's $\pi - \lambda$ theorem (Klenke, 2008, Theorem 1.19). Let $\mathcal{C} \subset \mathcal{F}$ denote the π-system of finite cylinder sets, i.e. $C \in \mathcal{C}$ of the form

$$C = \{\omega \in \Omega \text{ s.t. } \omega(s_1) \in A_1, \ldots, \omega(s_n) \in A_N\} \in \mathcal{F},$$

for any $n \in \mathbb{N}, s_1 < \ldots < s_n \in \mathbb{R}_+$ and $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R}^{pd})$. Define

$$\mathcal{L} = \{L \in \mathcal{F} \text{ s.t. } (s,x) \mapsto \mathcal{Q}_{s,x}(L) \text{ is } \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}^{pd}) - \text{measurable}\}.$$

This is a λ-system, since:

1. $\Omega \in \mathcal{L}$ since $(s,x) \mapsto \mathcal{Q}_{s,x}(\Omega) = \mathbb{Q}_{\mathbf{A},s,x}(\Omega) = 1$ is trivially measurable;

2. if $A, B \in \mathcal{L}$ and $A \subseteq B$ then $B \setminus A \in \mathcal{L}$ since

$$(s,x) \mapsto \mathcal{Q}_{s,x}(B \setminus A) = \mathbb{Q}_{\mathbf{A},s,x}(B \setminus A) = \mathbb{Q}_{\mathbf{A},s,x}(B) - \mathbb{Q}_{\mathbf{A},s,x}(A) = \mathcal{Q}_{s,x}(B) - \mathcal{Q}_{s,x}(A),$$

is the difference of measurable functions and thus measurable;

3. if $A_1 \subseteq A_2 \subseteq \cdots \in \mathcal{L}$ then $\bigcup_{n\geq 1} A_n \in \mathcal{L}$ since

$$(s,x) \mapsto \mathcal{Q}_{s,x}(\bigcup_{n\geq 1} A_n) = \mathbb{Q}_{\mathbf{A},s,x}(\bigcup_{n\geq 1} A_n) = \lim_{n\to\infty} \mathcal{Q}_{s,x}(A_n),$$

is the pointwise limit of measurable functions and thus measurable.

Moreover $\mathcal{C} \subset \mathcal{L}$ since for $C \in \mathcal{C}$ we can write

$$(s,x) \mapsto \mathcal{Q}_{s,x}(C) = \mathbb{Q}_{\mathbf{A},s,x}(C) = \int_{A_1} \cdots \int_{A_n} Q_{\mathbf{A},s}(s_{n-1}, x_{n-1}, s_n, \mathrm{d}x_n) \cdots Q_{\mathbf{A},s}(0, x, s_1, \mathrm{d}x_1),$$

where $Q_{\mathbf{A},s}(\cdot,\cdot,\cdot,\cdot) : \mathbb{R}_+ \times \mathbb{R}^{pd} \times \mathbb{R}_+ \times \mathcal{B}(\mathbb{R}^{pd}) \to [0,1]$ are the time-inhomogenous Markov transition probabilities associated with $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}^{(t-s)\vee 0}$ given by

$$Q_{\mathbf{A},s}(s_1, y, s_2, A) = \begin{cases} P_{\mathbf{A}}(y, s_2 - s_1, A), & \text{if } 0 \leq s_1 \leq s_2 \leq (t-s) \vee 0, \\ P_{\mathbf{A}}(y, (t-s) - s_1, A), & \text{if } 0 \leq s_1 \leq (t-s) \vee 0 \leq s_2, \\ \mathbb{1}_A(y), & \text{if } 0 \leq (t-s) \vee 0 \leq s_1 \leq s_2, \end{cases}$$

for $0 \leq s_1 \leq s_2$, $y \in \mathbb{R}^{pd}$, $A \in \mathcal{B}(\mathbb{R}^{pd})$ and $P_{\mathbf{A}}(\cdot,\cdot,\cdot) : \mathbb{R}^{pd} \times \mathbb{R}_+ \times \mathcal{B}(\mathbb{R}^{pd}) \to [0,1]$ denoting the time-homogeneous Markov transition probabilities associated with $\mathbb{X}_{\mathbf{A},\mathbf{x}_0}$ given in Masuda (2004, Proposition 2.1), or Sato and Yamazato (1984, Theorem 3.1). The map $(s,x) \mapsto \mathcal{Q}_{s,x}(C)$ is:

- measurable in $x \in \mathbb{R}^{pd}$, using measurability of $P_{\mathbf{A}}(x, s_1, A), P_{\mathbf{A}}(x, t-s, A), \mathbb{1}_A(x)$ to obtain measurability of $Q_{\mathbf{A},s}(0, x, s_1, A)$ for general $A \in \mathcal{B}(\mathbb{R}^{pd})$ and then extending from simple functions to obtain measurability of general integrals w.r.t. $Q_{\mathbf{A},s}(0, x, s_1, \mathrm{d}x_1)$,
- right-continuous in $s \in \mathbb{R}_+$, using continuity in s of $P_{\mathbf{A}}$ and treating the cases $s < t$ and $s \geq t$ separately,

and thus jointly measurable in (s,x). Note that $\mathcal{F} = \sigma(\mathcal{C})$ and by the $\pi - \lambda$ theorem we have $\sigma(\mathcal{C}) \subseteq \mathcal{L}$. We can hence conclude that measurability in (s,x) holds for any $A \in \mathcal{F}$. □

3.C.2 Proof of Lemma 3.3.8

Proof of Lemma 3.3.8. Under $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$, the process $\mathbb{H}$ is a continuous local martingale as it is given by the integral of a locally bounded predictable process against the $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$ Brownian motion $D^{p-1}\mathbb{Y}^{\mathrm{c}}_{\mathbf{A}^{(0)}}$, i.e. $\mathbb{H}$ is a semimartingale with characteristics $(0, [\mathbf{H}], 0)$ under $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$. Hence

by Küchler and Sørensen (1997, Corollary 11.3.2), $\mathbb{H}$ is a semimartingale with characteristics $([\mathbf{H}]\mathrm{vec}(\mathbf{A}), [\mathbf{H}], 0)$ under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ and thus the score $\{\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}),\ t \geq 0\}$ is a mean-zero continuous local martingale with quadratic variation $[\mathbf{H}]_t$. Since, moreover, the driving Lévy process $\mathbb{L}$ is assumed to be square-integrable, the score has a $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$-integrable quadratic variation (here $\|\cdot\|$ denotes the Frobenius norm for matrices and $\|\cdot\|_\Sigma$ the norm induced by $\langle\cdot,\cdot\rangle_\Sigma$ for vectors):

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\|[\mathbf{H}]_t\|\right] &\leq \mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\left(\sum_{l,k=0}^{p-1}\sum_{i,j=1}^{d}\sum_{m,n=1}^{d}\left(\int_0^t D^l Y_s^{(i)}\Sigma^{-1}_{m,n}D^k Y_s^{(j)}\,\mathrm{d}s\right)^2\right)^{1/2}\right] \\
&\leq d\|\Sigma^{-1}\|\mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\left(\int_0^t\int_0^t\sum_{l,k=0}^{p-1}\sum_{i,j=1}^{d} D^l Y_s^{(i)}D^k Y_s^{(j)}D^l Y_r^{(i)}D^k Y_r^{(j)}\,\mathrm{d}s\mathrm{d}r\right)^{1/2}\right] \\
&\leq d\|\Sigma^{-1}\|\mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\left(\int_0^t\int_0^t\left(\sum_{l=0}^{p-1}\sum_{i=1}^{d} D^l Y_s^{(i)}D^l Y_r^{(i)}\right)^2\mathrm{d}s\mathrm{d}r\right)^{1/2}\right] \\
&\leq d\|\Sigma^{-1}\|\mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\left(\int_0^t\int_0^t\sum_{l=0}^{p-1}\sum_{i=1}^{d}\left(D^l Y_s^{(i)}D^l Y_r^{(i)}\right)^2\mathrm{d}s\mathrm{d}r\right)^{1/2}\right] \\
&\leq d\|\Sigma^{-1}\|\mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}\left[\int_0^t\sum_{l=0}^{p-1}\sum_{i=1}^{d}\left(D^l Y_s^{(i)}\right)^2\mathrm{d}s\right] \\
&\leq d\|\Sigma^{-1}\|\int_0^t \mathbb{E}^{\mathbb{P}_{\mathbf{A},\mathbf{x}_0}}[\|\mathbf{X}_s\|^2]\mathrm{d}s = d\|\Sigma^{-1}\|\int_0^t \mathbb{E}^{\mathbb{P}'}[\|\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s}\|^2]\,\mathrm{d}s < \infty,
\end{aligned}$$

by noting in the last step that $\{\mathbf{X}_s,\ s \in [0,t]\}$ and $\{\mathbf{X}_{\mathbf{A},\mathbf{x}_0,s},\ s \in [0,t]\}$ have the same law under $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ and $\mathbb{P}'$ respectively and using the bound in Equation (3.C.5). Thus the $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ continuous local martingale $\{\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}),\ t \geq 0\}$ is indeed a $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$ continuous square-integrable martingale by Protter (1990, Corollary II.3). In particular, $\mathbf{H}_t$ and $[\mathbf{H}]_t$ are finite $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$-a.s. □

Remark 3.C.4. Alternatively one can explicitly substitute

$$D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},t} = D^{p-1}\mathbf{Y}^c_{\mathbf{A},t} - \sum_{j=1}^{p}\int_0^t A_j D^{p-j}\mathbf{Y}_s\,\mathrm{d}s$$

into the score. This yields an integral with respect to the $\mathbb{P}_{\mathbf{A},\mathbf{x}_0}$-Brownian motion $D^{p-1}\mathbb{Y}^c_{\mathbf{A}}$, which one can then check to have quadratic variation $[\mathbf{H}]_t$.

3.C.3 Proof of Theorem 3.3.9

Proof of Theorem 3.3.9. For $\phi \in \mathbb{R}^{pd^2}$ the log-likelihood is a quadratic form given by:

$$l_t(\phi) := \log\mathcal{L}(\mathrm{vec}^{-1}(\phi); \mathbb{Y}_{[0,t]}) = \phi^{\mathrm{T}}\mathbf{H}_t - \frac{1}{2}\phi^{\mathrm{T}}[\mathbf{H}]_t\phi,$$

i.e. it has gradient

$$\nabla_\phi l_t(\phi) = \mathbf{H}_t - [\mathbf{H}]_t \phi,$$

and Hessian

$$\mathrm{Hess}_\phi l_t(\phi) = -[\mathbf{H}]_t.$$

First, note that, for any $t > 0$, the (log)likelihood has finite coefficients $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$-a.s. (see discussion at the end of the proof of Lemma 3.3.8). Moreover the matrix $[\mathbf{H}]_t$ is strictly positive definite $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$-a.s. Let $\phi \in \mathbb{R}^{pd^2}$ with $\phi = (\phi_{1,1}^{\mathrm{T}}, \ldots, \phi_{d,p}^{\mathrm{T}})^{\mathrm{T}}$ for $\phi_{1,1}, \ldots, \phi_{d,p} \in \mathbb{R}^d$. Then

$$\phi^{\mathrm{T}}[\mathbf{H}]_t \phi = \int_0^t \| \sum_{i=1}^{d} \sum_{l=0}^{p-1} \phi_{i,l} D^l Y_s^{(i)} \|_\Sigma \, \mathrm{d}s \geq 0, \quad \mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}\text{−a.s.}$$

with equality if and only if

$$\sum_{i=1}^{d} \sum_{l=0}^{p-1} \phi_{i,l} D^l Y_s^{(i)} = 0, \quad \mathrm{Leb} \otimes \mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}\text{−a.s.}$$

We define a collection of random variables $\{Z^i,\ i \in \mathcal{I}\}$ on $(\mathbb{P}', \Omega', \mathcal{F}')$ to be affinely independent under $\mathbb{P}'$ if, for any $\xi \in \mathbb{R}^{\mathcal{I}}, \xi_0 \in \mathbb{R}$

$$\sum_{i \in \mathcal{I}} \xi_i Z^i = \xi_0 \quad \mathbb{P}'\text{−a.s.} \implies \xi, \xi_0 = 0.$$

Thus, showing that under $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$, $\{D^l Y_s^{(i)},\ i = 1, \ldots, d, l = 0, \ldots, p-1\}$ is an affinely independent collection of random variables for any $s \in (0, t]$ would imply $\phi \equiv 0$ and thus that $[\mathbf{H}]_t$ is strictly positive definite. Under $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$, with $\mathbf{A}^{(0)} = (0_{d\times d}, \ldots, 0_{d\times d})$, we have that

$$D^{p-1} Y_t^{(i)} = L_t^{(i)},\ D^l Y_t^{(i)} = \int_0^t D^{l+1} Y_s^{(i)} \mathrm{d}s \quad l = p-2, \ldots, 0,$$

and one can check that the (independent) components of a non-degenerate Lévy process and their first $p-1$ time integrals are affinely independent under $\mathbb{P}_{\mathbf{A}^{(0)},\mathbf{x}_0}$, and hence – by equivalence of the restricted measures up to t – also under $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$. Thus we have that the (log)likelihood is a strictly concave continuous function on $\mathbb{R}^{pd^2}$ with unique maximizer

$$\hat{\phi}_t := [\mathbf{H}]_t^{-1} \mathbf{H}_t,$$

and thus

$$\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]}) = \mathrm{vec}^{-1}([\mathbf{H}]_t^{-1} \mathbf{H}_t).$$

In the remaining part of the proof, we assume that $\mathbf{A}^* \in \mathfrak{A}$. Under $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$, the MCAR($p$)

process $\mathbb{Y}$ and its state-space representation are ergodic. Thus, $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$ almost surely

$$\frac{1}{t}[\mathbf{H}]_t \to \mathcal{H}_\infty, \tag{3.C.6}$$

where $\mathcal{H}_\infty$ is a strictly positive definite matrix (by similar reasoning as for $[\mathbf{H}]_t$) given by

$$\mathcal{H}_\infty = \begin{pmatrix} \mathbb{E}_{\mathbf{A}^*}[D^{p-1}Y_\infty^{(1)}\Sigma^{-1}D^{p-1}Y_\infty^{(1)}] & \cdots & \mathbb{E}_{\mathbf{A}^*}[D^{p-1}Y_\infty^{(1)}\Sigma^{-1}Y_\infty^{(d)}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{\mathbf{A}^*}[Y_\infty^{(d)}\Sigma^{-1}D^{p-1}Y_\infty^{(1)}] & \cdots & \mathbb{E}_{\mathbf{A}^*}[Y_\infty^{(d)}\Sigma^{-1}Y_\infty^{(d)}] \end{pmatrix}, \tag{3.C.7}$$

and $\mathcal{L}([\mathbf{Y}_\infty^{\mathrm{T}},\ldots,D^{p-1}\mathbf{Y}_\infty^{\mathrm{T}}]^{\mathrm{T}}) = \mathcal{L}(\int_0^\infty e^{s\mathcal{A}_{\mathbf{A}^*}}\mathcal{E}\,\mathrm{d}\mathbf{L}_s)$ under $\mathbb{P}_{\mathbf{A}^*}$. Note that

$$\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})) = [\mathbf{H}]_t^{-1}\mathbf{H}_t = \mathrm{vec}(\mathbf{A}^*) + [\mathbf{H}]_t^{-1}(\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*)),$$

and, by Lemma 3.3.8, the score $\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*)$ is a mean-zero square-integrable martingale under $\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}$. By the central limit theorem for martingales, i.e. Küchler and Sørensen (1997, Theorem A.7.7 and Corollary A.7.8), applied to the score, we have that

$$[\mathbf{H}]_t^{-1}(\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*)) \xrightarrow{\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}} 0,$$

and hence

$$\mathrm{vec}(\hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})) \xrightarrow{\mathbb{P}_{\mathbf{A}^*,\mathbf{x}_0}} \mathrm{vec}(\mathbf{A}^*),\ t\to\infty.$$

Again by Küchler and Sørensen (1997, Theorem A.7.7) we have

$$\frac{1}{\sqrt{t}}\mathcal{H}_\infty^{-1/2}(\mathbf{H}_t - [\mathbf{H}]_t\mathrm{vec}(\mathbf{A}^*)) \xrightarrow{\mathcal{L}} \mathbf{W} \sim N(\mathbf{0}, I_{pd^2}),\ t\to\infty.$$

Combining this with Equation (3.C.6) via Slutsky's lemma yields

$$\sqrt{t}\left([\mathbf{H}]_t^{-1}\mathbf{H}_t - \mathrm{vec}(\mathbf{A}^*)\right) \xrightarrow{\mathcal{L}} \mathcal{H}_\infty^{-1/2}\mathbf{W} =: \mathbf{Z} \sim N(\mathbf{0}, \mathcal{H}_\infty^{-1}),\ t\to\infty.$$

□

3.D Discrete observations: proofs

In this section, we will be working heavily with inequalities. We introduce the notation $f(t) \lesssim g(t)$ to denote $f(t) \le Cg(t)$ for some constant C independent of t.

3.D.1 Proof of Theorem 3.4.5

Proof of Theorem 3.4.5. To show Equation (3.4.8) with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{[0,t]})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t},\ldots,D^{p-1}\mathbb{Y}_{\mathcal{P}_t},\mathbb{J}_{\mathcal{P}_t},\mathbb{M}_{\mathcal{P}_t})$, we work in $L^1(\Omega,\mathcal{F},\mathbb{P}_{\mathbf{A}^*})$ and note that for each d-entry batch

with $l \in \{0, \dots, p-1\}$ and $i \in \{1, \dots, d\}$

$$
\begin{aligned}
t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\Big[\Big\|\sum_{n=0}^{N_t-1} D^l Y^{(i)}_{s_n}\Sigma^{-1}(D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s_{n+1}} - D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s_n}) - \int_0^t D^l Y^{(i)}_{s-}\Sigma^{-1}\,\mathrm{d}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s}\Big\|\Big] \\
\overset{(i)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\mathbb{E}_{\mathbf{A}^*}\Big[\Big\|\sum_{n=0}^{N_t-1}\int_{s_n}^{s_{n+1}}\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_{s-}\Big)\Sigma^{1/2}\,\mathrm{d}\mathbf{W}_s \\
\qquad - \sum_{n=0}^{N_t-1}\int_{s_n}^{s_{n+1}}\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_{s-}\Big)\sum_{j=1}^{p} A_j^* D^{p-j}\mathbf{Y}_s\,\mathrm{d}s\Big\|\Big] \\
\overset{(ii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\Bigg\{\mathbb{E}_{\mathbf{A}^*}\Big[\Big\|\int_0^t \sum_{n=0}^{N_t-1}\mathbb{1}_{[s_n,s_{n+1}]}(s)(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_{s-})\Sigma^{1/2}\,\mathrm{d}\mathbf{W}_s\Big\|^2\Big]^{1/2} \\
\qquad + \sum_{n=0}^{N_t-1}\mathbb{E}_{\mathbf{A}^*}\Big[\Big\|\int_{s_n}^{s_{n+1}}(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_{s-})\sum_{j=1}^{p} A_j^* D^{p-j}\mathbf{Y}_s\,\mathrm{d}s\Big\|\Big]\Bigg\} \\
\overset{(iii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\Bigg\{\mathrm{tr}(\Sigma)^{1/2}\Big(\int_0^t \sum_{n=0}^{N_t-1}\mathbb{1}_{[s_n,s_{n+1}]}(s)\,\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)^2\Big]\,\mathrm{d}s\Big)^{1/2} \\
\qquad + \sum_{n=0}^{N_t-1}\sum_{j=1}^{p}\int_{s_n}^{s_{n+1}}\mathbb{E}_{\mathbf{A}^*}\Big[\Big\|\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)A_j^* D^{p-j}\mathbf{Y}_s\Big\|\Big]\,\mathrm{d}s\Bigg\} \\
\overset{(iv)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\Bigg\{\mathrm{tr}(\Sigma)^{1/2}\Big(\sum_{n=0}^{N_t-1}\int_{s_n}^{s_{n+1}}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)^2\Big]\,\mathrm{d}s\Big)^{1/2} \\
\qquad + \sum_{n=0}^{N_t-1}\sum_{j=1}^{p}\int_{s_n}^{s_{n+1}}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)^2\Big]^{1/2}\mathbb{E}_{\mathbf{A}^*}\Big[\|A_j^* D^{p-j}\mathbf{Y}_s\|^2\Big]^{1/2}\,\mathrm{d}s\Bigg\} \\
\overset{(v)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\Bigg\{\mathrm{tr}(\Sigma)^{1/2}\Big(\sum_{n=0}^{N_t-1}\sup_{s\in[s_n,s_{n+1}]}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)^2\Big](s_{n+1}-s_n)\Big)^{1/2} \\
\quad + p\max_{j=1,\dots,p}\|A_j^*\|\,\mathbb{E}_{\mathbf{A}^*}\Big[\|\mathbf{X}_\infty\|^2\Big]^{1/2}\sum_{n=0}^{N_t-1}\sup_{s\in[s_n,s_{n+1}]}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_{s_n} - D^l Y^{(i)}_s\Big)^2\Big]^{1/2}(s_{n+1}-s_n)\Bigg\} \\
\overset{(vi)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\Bigg\{\mathrm{tr}(\Sigma)^{1/2}\Big(t\sup_{s\in[0,\Delta_{\mathcal{P}_t}]}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_0 - D^l Y^{(i)}_s\Big)^2\Big]\Big)^{1/2} \\
\quad + p\max_{j=1,\dots,p}\|A_j^*\|\,\mathbb{E}_{\mathbf{A}^*}\Big[\|\mathbf{X}_\infty\|^2\Big]^{1/2} t\sup_{s\in[0,\Delta_{\mathcal{P}_t}]}\mathbb{E}_{\mathbf{A}^*}\Big[\Big(D^l Y^{(i)}_0 - D^l Y^{(i)}_s\Big)^2\Big]^{1/2}\Bigg\} \\
\overset{(vii)}{\lesssim} \Delta_{\mathcal{P}_t}^{1/2} + t^{1/2}\Delta_{\mathcal{P}_t}^{1/2} \to 0, \quad t\to\infty,
\end{aligned}
$$

by using (i) the decomposition (3.4.16) of the integrator $\{D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s},\ s \geq 0\}$, (ii) triangle inequality and domination of $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ norm by $L^2(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ norm, (iii) Itô isometry, triangle inequality and Tonelli's theorem, (iv) Cauchy-Schwarz inequality, (v) sup bounds on integral and stationarity of the state-space representation, (vi) stationarity of the state-space representation again, (vii) Equation (3.B.1) and the high-frequency sampling condition, i.e. Assumption 3.4.4.

Next, to show Equation (3.4.9), we work again in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ which then implies the corresponding limit in probability. For each $d \times d$ sub-matrix with $l, k \in \{0, \dots, p-1\}$ and $i, j \in \{1, \dots, d\}$

we note that

$$
\begin{aligned}
&t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{n=0}^{N_t-1} D^l Y_{s_n}^{(i)}\Sigma^{-1}D^k Y_{s_n}^{(j)}(s_{n+1}-s_n) - \int_0^t D^l Y_s^{(i)}\Sigma^{-1}D^k Y_s^{(j)}\,\mathrm{d}s\right\|\right] \\
&\overset{(i)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{n=0}^{N_t-1}\int_{s_n}^{s_{n+1}} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{s_n}^{(i)}D^k Y_{s_n}^{(j)} - D^l Y_s^{(i)}D^k Y_s^{(j)}\right|\right]\mathrm{d}s \\
&\overset{(ii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{n=0}^{N_t-1}\sup_{s\in[s_n,s_{n+1}]} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{s_n}^{(i)}D^k Y_{s_n}^{(j)} - D^l Y_s^{(i)}D^k Y_s^{(j)}\right|\right](s_{n+1}-s_n) \\
&\overset{(iii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sup_{s\in[0,\Delta_{\mathcal{P}_t}]} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_0^{(i)}D^k Y_0^{(j)} - D^l Y_s^{(i)}D^k Y_s^{(j)}\right|\right]\sum_{n=0}^{N_t-1}(s_{n+1}-s_n) \\
&\overset{(iv)}{\leq} t^{1/2}\|\Sigma^{-1}\|\sup_{s\in[0,\Delta_{\mathcal{P}_t}]}\left\{\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_0^{(i)}(D^k Y_0^{(j)} - D^k Y_s^{(j)})\right|\right] + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^k Y_s^{(j)}(D^l Y_0^{(i)} - D^l Y_s^{(i)})\right|\right]\right\} \\
&\overset{(v)}{\leq} t^{1/2}\|\Sigma^{-1}\|\,2\,\mathbb{E}_{\mathbf{A}^*}\left[\|\mathbf{X}_\infty\|^2\right]^{1/2}\sup_{s\in[0,\Delta_{\mathcal{P}_t}]}\mathbb{E}_{\mathbf{A}^*}\left[\|\mathbf{X}_0-\mathbf{X}_s\|^2\right]^{1/2} \\
&\overset{(vi)}{\lesssim} t^{1/2}\Delta_{\mathcal{P}_t}^{1/2} \to 0, \quad t\to\infty,
\end{aligned}
$$

where we have used (i) triangle inequality and Tonelli's theorem, (ii) sup bound on the integral, (iii) stationarity of the state-space representation

$$
\mathbf{X}_s = \begin{pmatrix} \mathbf{Y}_s \\ \vdots \\ D^{p-1}\mathbf{Y}_s \end{pmatrix}, \quad s \geq 0,
$$

of the MCAR(p) process, (iv) triangle inequality, (v) Cauchy-Schwarz inequality and stationarity of state-space representation, (vi) Equation (3.B.1) and the high-frequency sampling condition, i.e. Assumption 3.4.4. □

3.D.2 Proof of Lemma 3.4.8

Proof of Lemma 3.4.8. Let $k \in \{1,\dots,p-1\}$. First note that for $n \in \{0,\dots,N_t-k\}$ by Lemma 3.F.1 and Lemma 3.F.2 we can express the forward difference as

$$
\hat{D}^k\mathbf{Y}_{s_n} = \sum_{i=1}^{k}(-1)^{k+i}\sum_{\substack{n_1,\dots,n_i\geq 1\\ n_1+\cdots+n_i=k}}(s_{n+1}-s_n)^{-n_1}\cdots(s_{n+i}-s_{n+i-1})^{-n_i}(\mathbf{Y}_{s_{n+i}}-\mathbf{Y}_{s_{n+i-1}}),
$$

and, for each $i \in \{1,\dots,k\}$, the increments as

$$
\begin{aligned}
&\mathbf{Y}_{s_{n+i}} - \mathbf{Y}_{s_{n+i-1}} \\
&\quad = \sum_{j=1}^{k-1} D^j\mathbf{Y}_{s_n}\frac{(s_{n+i}-s_n)^j - (s_{n+i-1}-s_n)^j}{j!} + \int_{s_{n+i-1}}^{s_{n+i}}\int_{s_n}^{u_1}\cdots\int_{s_n}^{u_{k-1}} D^k\mathbf{Y}_{u_k}\,\mathrm{d}u_k\dots\mathrm{d}u_1.
\end{aligned}
$$

These can be combined to write

$$\hat{D}^k \mathbf{Y}_{s_n} = \sum_{i=1}^{k} (-1)^{k+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=k}} (s_{n+1}-s_n)^{-n_1} \cdots (s_{n+i}-s_{n+i-1})^{-n_i} \times \Big[\sum_{j=1}^{k-1} D^j \mathbf{Y}_{s_n} \frac{(s_{n+i}-s_n)^j - (s_{n+i-1}-s_n)^j}{j!} + \int_{s_{n+i-1}}^{s_{n+i}} \int_{s_n}^{u_1} \cdots \int_{s_n}^{u_{k-1}} D^k \mathbf{Y}_{u_k} \, \mathrm{d}u_k \dots \mathrm{d}u_1 \Big].$$

Thus for any $k \in \{1 \dots, p-1\}$ we can write

$$\begin{aligned} &\hat{D}^k \mathbf{Y}_{s_n} - D^k \mathbf{Y}_{s_n} \\ &\quad = \sum_{j=1}^{k-1} D^j \mathbf{Y}_{s_n} C(s_n, \dots, s_{n+k}; j) + D^k \mathbf{Y}_{s_n} \left[C(s_n, \dots, s_{n+k}; k) - 1 \right] \\ &\qquad + \Bigg[\sum_{i=1}^{k} (-1)^{k+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=k}} (s_{n+1}-s_n)^{-n_1} \cdots (s_{n+i}-s_{n+i-1})^{-n_i} \\ &\qquad\quad \times \int_{s_{n+i-1}}^{s_{n+i}} \int_{s_n}^{u_1} \cdots \int_{s_n}^{u_{k-1}} (D^k \mathbf{Y}_{u_k} - D^k \mathbf{Y}_{s_n}) \, \mathrm{d}u_k \dots \mathrm{d}u_1 \Bigg], \end{aligned} \tag{3.D.1}$$

where we define, for $j \in \{1, \dots, k\}$,

$$\begin{aligned} C(s_n, \dots, s_{n+k}; j) := \sum_{i=1}^{k} (-1)^{k+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=k}} &(s_{n+1}-s_n)^{-n_1} \cdots (s_{n+i}-s_{n+i-1})^{-n_i} \\ &\times \frac{(s_{n+i}-s_n)^j - (s_{n+i-1}-s_n)^j}{j!}. \end{aligned} \tag{3.D.2}$$

We note that, if $\mathcal{P}_t$ is almost evenly spaced, then

$$C(s_n, \dots, s_{n+k}; j) \approx \begin{cases} 0, & j = 0, \dots, k-1, \\ 1, & j = k. \end{cases}$$

More rigorously, we can write (here we consider k even, the case where k is odd is covered by multiplying both inequalities through by -1):

$$\begin{aligned} &C(s_n, \dots, s_{n+k}; j) \\ &= \Big(\sum_{\substack{i=1 \\ i \text{ even}}}^{k} - \sum_{\substack{i=1 \\ i \text{ odd}}}^{k} \Big) \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=k}} (s_{n+1}-s_n)^{-n_1} \cdots (s_{n+i}-s_{n+i-1})^{-n_i} \frac{(s_{n+i}-s_n)^j - (s_{n+i-1}-s_n)^j}{j!} \\ &\leq \sum_{\substack{i=1 \\ i \text{ even}}}^{k} \binom{k-1}{i-1} c_{\mathcal{P}_t}^{-k} \Delta_{\mathcal{P}_t}^{-k} \frac{(i\Delta_{\mathcal{P}_t})^j - c_{\mathcal{P}_t}^j ((i-1)\Delta_{\mathcal{P}_t})^j}{j!} - \sum_{\substack{i=1 \\ i \text{ odd}}}^{k} \binom{k-1}{i-1} \Delta_{\mathcal{P}_t}^{-k} \frac{c_{\mathcal{P}_t}^j (i\Delta_{\mathcal{P}_t})^j - ((i-1)\Delta_{\mathcal{P}_t})^j}{j!} \end{aligned}$$

$$\leq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!}\Big\{ \sum_{\substack{i=1 \\ i \text{ even}}}^{k} \binom{k-1}{i-1} c_{\mathcal{P}_t}^{-k}[i^j - c_{\mathcal{P}_t}^{j}(i-1)^j] - \sum_{\substack{i=1 \\ i \text{ odd}}}^{k} \binom{k-1}{i-1}[c_{\mathcal{P}_t}^{j} i^j - (i-1)^j]\Big\}$$
$$\leq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!}\Big\{ \sum_{i=1}^{k}(-1)^i \binom{k-1}{i-1}[i^j - (i-1)^j] + \sum_{i=1}^{k} \binom{k-1}{i-1}\Big[(|c_{\mathcal{P}_t}^{-k} - 1| + |c_{\mathcal{P}_t}^{j} - 1|)i^j + |1 - c_{\mathcal{P}_t}^{j-k}|(i-1)^j\Big]\Big\},$$

and

$$C(s_n, \ldots, s_{n+k}; j)$$
$$\geq \sum_{\substack{i=1 \\ i \text{ even}}}^{k} \binom{k-1}{i-1} \Delta_{\mathcal{P}_t}^{-k} \frac{c_{\mathcal{P}_t}^{j}(i\Delta_{\mathcal{P}_t})^j - ((i-1)\Delta_{\mathcal{P}_t})^j}{j!} - \sum_{\substack{i=1 \\ i \text{ odd}}}^{k} \binom{k-1}{i-1} c_{\mathcal{P}_t}^{-k}\Delta_{\mathcal{P}_t}^{-k} \frac{(i\Delta_{\mathcal{P}_t})^j - c_{\mathcal{P}_t}^{j}((i-1)\Delta_{\mathcal{P}_t})^j}{j!}$$
$$\geq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!}\Big\{ \sum_{\substack{i=1 \\ i \text{ even}}}^{k} \binom{k-1}{i-1}[c_{\mathcal{P}_t}^{j} i^j - (i-1)^j] - \sum_{\substack{i=1 \\ i \text{ odd}}}^{k} \binom{k-1}{i-1} c_{\mathcal{P}_t}^{-k}[i^j - c_{\mathcal{P}_t}^{j}(i-1)^j]\Big\}$$
$$\geq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!}\Big\{ \sum_{i=1}^{k}(-1)^i \binom{k-1}{i-1}[i^j - (i-1)^j] - \sum_{i=1}^{k} \binom{k-1}{i-1}\Big[(|c_{\mathcal{P}_t}^{j} - 1| + |1 - c_{\mathcal{P}_t}^{-k}|)i^j + |c_{\mathcal{P}_t}^{j-k} - 1|(i-1)^j\Big]\Big\},$$

where we use that for any $m \geq 1$ the relationship $c_{\mathcal{P}_t}\Delta_{\mathcal{P}_t} \leq |s_{n+1} - s_n| \leq \Delta_{\mathcal{P}_t}$ implies

$$c_{\mathcal{P}_t}^{m}\Delta_{\mathcal{P}_t}^{m} \leq |s_{n+1} - s_n|^m \leq \Delta_{\mathcal{P}_t}^{m}, \ -\Delta_{\mathcal{P}_t}^{m} \leq -|s_{n+1} - s_n|^m \leq -c_{\mathcal{P}_t}^{m}\Delta_{\mathcal{P}_t}^{m},$$
$$\Delta_{\mathcal{P}_t}^{-m} \leq |s_{n+1} - s_n|^{-m} \leq c_{\mathcal{P}_t}^{-m}\Delta_{\mathcal{P}_t}^{-m}, \ -c_{\mathcal{P}_t}^{-m}\Delta_{\mathcal{P}_t}^{-m} \leq -|s_{n+1} - s_n|^{-m} \leq -\Delta_{\mathcal{P}_t}^{-m}.$$

and for any $k \geq 1$, $1 \leq i \leq k$

$$\#\{(n_1, \ldots, n_i) : n_1, \ldots, n_i \geq 1, n_1 + \cdots + n_i = k\} = \binom{k-1}{i-1}.$$

We can rearrange and combine the upper and lower bounds for $C(s_n, \ldots, s_{n+k}; j)$ to obtain

$$\begin{aligned} &|C(s_n, \ldots, s_{n+k}; j) - F(j,k)| \\ &\qquad \leq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!} \sum_{i=1}^{k} \binom{k-1}{i-1} \Big[(|c_{\mathcal{P}_t}^{j} - 1| + |1 - c_{\mathcal{P}_t}^{-k}|)i^j + |c_{\mathcal{P}_t}^{j-k} - 1|(i-1)^j\Big], \end{aligned} \tag{3.D.3}$$

where, by Lemma 3.F.3, we have

$$F(j,k) := (-1)^k \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!} \sum_{i=1}^{k}(-1)^i \binom{k-1}{i-1}[i^j - (i-1)^j] = \begin{cases} 0, & j = 0, \ldots, k-1, \\ 1, & j = k. \end{cases}.$$

Next, we can use $c_{\mathcal{P}_t} \in (0,1)$ and Assumption 3.4.6 to deduce that, for $1 \leq j \leq k$,

$$|c_{\mathcal{P}_t}^{j} - 1| \leq |c_{\mathcal{P}_t}^{k} - 1|, \quad |c_{\mathcal{P}_t}^{-k} - 1| = \frac{|c_{\mathcal{P}_t}^{k} - 1|}{|c_{\mathcal{P}_t}^{k}|} \leq 2^k |c_{\mathcal{P}_t}^{k} - 1|, \quad |c_{\mathcal{P}_t}^{j-k} - 1| = \frac{|c_{\mathcal{P}_t}^{k-j} - 1|}{|c_{\mathcal{P}_t}^{k-j}|} \leq 2^{k-j} |c_{\mathcal{P}_t}^{k} - 1|,$$

for $t \in \mathcal{T}$ sufficiently large (i.e. such that $c_{\mathcal{P}_t} \geq 1/2$). Due to the "convergence to uniform spacing" condition on the sequence of partitions, i.e. Assumption 3.4.6, we can bound Equation (3.D.3) by

$$\begin{aligned} &|C(s_n, \dots, s_{n+k}; j) - F(j,k)| \\ &\qquad \leq \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!} |c_{\mathcal{P}_t}^k - 1| \sum_{i=1}^{k} \binom{k-1}{i-1} \left[(1+2^k) i^j + 2^{k-j} (i-1)^j \right] \lesssim \Delta_{\mathcal{P}_t}^{j-1/2}. \end{aligned} \tag{3.D.4}$$

This implies

$$\begin{aligned} &\mathbb{E}_{\mathbf{A}^*} \left[\left\| \hat{D}^k \mathbf{Y}_{s_n} - D^k \mathbf{Y}_{s_n} \right\|^2 \right] \\ &\overset{(i)}{\leq} \sum_{j=1}^{k} \sum_{l=1}^{k} \mathbb{E}_{\mathbf{A}^*} \left[D^j \mathbf{Y}_{s_n}^{\mathrm{T}} D^l \mathbf{Y}_{s_n} \right] |C(s_n, \dots, s_{n+k}; j) - F(j,k)| |C(s_n, \dots, s_{n+k}; l) - F(l,k)| \\ &\quad + 2 \sum_{j=1}^{k} \sup_{u \in [s_n, s_{n+k}]} \mathbb{E}_{\mathbf{A}^*} \left[D^j \mathbf{Y}_{s_n}^{\mathrm{T}} (D^k \mathbf{Y}_u - D^k \mathbf{Y}_{s_n}) \right] |C(s_n, \dots, s_{n+k}; j) - F(j,k)| \, \bar{C}(s_n, \dots, s_{n+k}; k) \\ &\quad + \sup_{u \in [s_n, s_{n+k}]} \mathbb{E}_{\mathbf{A}^*} \left[\| D^k \mathbf{Y}_u - D^k \mathbf{Y}_{s_n} \|^2 \right] \bar{C}(s_n, \dots, s_{n+k}; k)^2 \\ &\overset{(ii)}{\leq} \sum_{j=1}^{k} \sum_{l=1}^{k} \mathbb{E}_{\mathbf{A}^*} \left[D^j \mathbf{Y}_{\infty}^{\mathrm{T}} D^l \mathbf{Y}_{\infty} \right] |C(s_n, \dots, s_{n+k}; j) - F(j,k)| |C(s_n, \dots, s_{n+k}; l) - F(l,k)| \\ &\quad + 2 \sum_{j=1}^{k} \mathbb{E}_{\mathbf{A}^*} \left[\| D^j \mathbf{Y}_{\infty} \|^2 \right]^{1/2} \sup_{u \in [0, k\Delta_{\mathcal{P}_t}]} \mathbb{E}_{\mathbf{A}^*} \left[\| D^k \mathbf{Y}_u - D^k \mathbf{Y}_0 \|^2 \right]^{1/2} \\ &\qquad\qquad \times |C(s_n, \dots, s_{n+k}; j) - F(j,k)| \, \bar{C}(s_n, \dots, s_{n+k}; k) \\ &\quad + \sup_{u \in [0, k\Delta_{\mathcal{P}_t}]} \mathbb{E}_{\mathbf{A}^*} \left[\| D^k \mathbf{Y}_u - D^k \mathbf{Y}_0 \|^2 \right] \bar{C}(s_n, \dots, s_{n+k}; k)^2 \\ &\overset{(iii)}{\lesssim} \left[\sum_{j=1}^{k} \sum_{l=1}^{k} \Delta_{\mathcal{P}_t}^{j-1/2} \Delta_{\mathcal{P}_t}^{l-1/2} + 2 \sum_{j=1}^{k} (k\Delta_{\mathcal{P}_t})^{1/2} \Delta_{\mathcal{P}_t}^{j-1/2} + k\Delta_{\mathcal{P}_t} \right] \lesssim \Delta_{\mathcal{P}_t}, \end{aligned}$$

where

$$\begin{aligned} \bar{C}(s_n, \dots, s_{n+k}; k) &:= \sum_{i=1}^{k} \sum_{\substack{n_1, \dots, n_i \geq 1 \\ n_1 + \cdots + n_i = k}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} \frac{(s_{n+i} - s_n)^k - (s_{n+i-1} - s_n)^k}{k!} \\ &\leq \sum_{i=1}^{k} \binom{k-1}{i-1} \Delta_{\mathcal{P}_t}^{-k} c_{\mathcal{P}_t}^{-k} \frac{(i\Delta_{\mathcal{P}_t})^k - c_{\mathcal{P}_t}^k ((i-1)\Delta_{\mathcal{P}_t})^k}{k!} = \sum_{i=1}^{k} \binom{k-1}{i-1} \frac{2^k i^k - (i-1)^k}{k!}, \end{aligned}$$

by applying (i) Equation (3.D.1), sup-bounds on the integrals and triangle inequality, (ii) stationarity of the state-space representation and Cauchy-Schwarz inequality and (iii) Equation (3.D.4), Equation (3.B.1) and $\mathbb{E}_{\mathbf{A}^*}[\mathbf{X}_\infty \mathbf{X}_\infty^{\mathrm{T}}] < \infty$. □

3.D.3 Proof of Theorem 3.4.14

Proof of Theorem 3.4.14. In order to apply Lemma 3.4.3 we need to show Equation (3.4.8) and Equation (3.4.9) hold for $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{Q}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{Q}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$.

For ease of notation let us write the increments of the partition $\mathcal{Q}_t$ with mesh $\Delta_{\mathcal{Q}_t}$ as

$$\Delta^m_{\mathcal{Q}_t} := u_{m+1} - u_m, \quad m = 0, \dots, M_t - 1,$$

and the increment of a process $\mathbb{Z}$ over the interval $[u_m, u_{m+1}] \in \mathcal{Q}_t$ as

$$\Delta^m_{\mathcal{Q}_t} \mathbf{Z} := (\mathbf{Z}_{u_{m+1}} - \mathbf{Z}_{u_m}), \quad m = 0, \dots, M_t - 1.$$

We stress that, throughout the whole proof, the finite difference approximations are computed on the finer partition $\mathcal{P}_t$. We will be repeatedly applying the bound (3.4.19) from Lemma 3.4.8 evaluated at points $s_n = u_m \in \mathcal{Q}_t \subset \mathcal{P}_t$. In this context, the notation $\hat{D}^l \mathbf{Y}_{u_m}$ denotes the $\mathcal{P}_t$-finite difference approximation (3.4.17) at the point $u_m \in \mathcal{Q}_t$ which is, by construction, also in $\mathcal{P}_t$.

We start by showing Equation (3.4.8). We work in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ which then implies the corresponding limit in probability. We note that for each d-entry batch with $l \in \{0, \dots, p-1\}$ and $i \in \{1, \dots, d\}$

$$\begin{aligned}
&t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \Sigma^{-1} \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} - \sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \Sigma^{-1} \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}}\right\|\right] \\
&\quad \overset{(i)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m}\left[\hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},u_{m+1}} - \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},u_m} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}\right]\right.\right. \\
&\qquad\qquad \left.\left. + \sum_{m=0}^{M_t-1} (\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m}) \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}\right\|\right] \\
&\quad \overset{(ii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\left\{\sum_{m=0}^{M_t-1} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^l Y^{(i)}_{u_m}\left[(\hat{D}^{p-1}\mathbf{Y}_{u_{m+1}} - D^{p-1}\mathbf{Y}_{u_{m+1}}) - (\hat{D}^{p-1}\mathbf{Y}_{u_m} - D^{p-1}\mathbf{Y}_{u_m})\right]\right\|\right]\right. \\
&\qquad\qquad \left. + \mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} (\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m}) \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}\right\|\right]\right\} \\
&\quad \overset{(iii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\left\{\sum_{m=0}^{M_t-1} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^l Y^{(i)}_{u_m}\right\|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^{p-1}\mathbf{Y}_{u_{m+1}} - D^{p-1}\mathbf{Y}_{u_{m+1}}\right\|^2\right]^{1/2}\right. \\
&\qquad + \mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^l Y^{(i)}_{u_m}\right\|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^{p-1}\mathbf{Y}_{u_m} - D^{p-1}\mathbf{Y}_{u_m}\right\|^2\right]^{1/2} \\
&\qquad \left. + \mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} (\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m})[\Sigma^{1/2}\Delta^m_{\mathcal{Q}_t}\mathbf{W} - \sum_{j=1}^{p}\int_{u_m}^{u_{m+1}} A^*_j D^{p-j}\mathbf{Y}_u \, \mathrm{d}u]\right\|\right]\right\} \\
&\quad \overset{(iv)}{\lesssim} t^{-1/2}\left\{\sum_{m=0}^{M_t-1} \Delta^{1/2}_{\mathcal{P}_t} + \mathbb{E}_{\mathbf{A}^*}\left[\left\|\int_0^t \sum_{m=0}^{M_t-1} \mathbb{1}_{[u_m,u_{m+1}]}(s)(\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m})\Sigma^{1/2}\mathrm{d}\mathbf{W}_s\right\|\right]\right. \\
&\qquad\qquad \left. + \mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1}\sum_{j=1}^{p}\int_{u_m}^{u_{m+1}} A^*_j(\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m}) D^{p-j}\mathbf{Y}_u \, \mathrm{d}u\right\|\right]\right\} \\
&\quad \overset{(v)}{\lesssim} t^{-1/2}\left\{M_t \Delta^{1/2}_{\mathcal{P}_t} + \mathrm{tr}(\Sigma)^{1/2}\left(\int_0^t \sum_{m=0}^{M_t-1} \mathbb{1}_{[u_m,u_{m+1}]}(u)\mathbb{E}_{\mathbf{A}^*}\left[\left(\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m}\right)^2\right]\mathrm{d}u\right)^{1/2}\right. \\
&\qquad\qquad \left. + \sum_{m=0}^{M_t-1}\sum_{j=1}^{p}\int_{u_m}^{u_{m+1}} \|A^*_j\|\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\|D^{p-j}\mathbf{Y}_\infty\|^2\right]^{1/2} \mathrm{d}u\right\}
\end{aligned}$$

$$\overset{(vi)}{\lesssim} t^{-1/2} M_t \Delta_{\mathcal{P}_t}^{1/2} + \Delta_{\mathcal{P}_t}^{1/2} + t^{1/2}\Delta_{\mathcal{P}_t}^{1/2}$$
$$\lesssim t^{-1/2} N_t' \Delta_{\mathcal{P}_t}^{1/2} = \left(t^{-1} N_t' \Delta_{\mathcal{Q}_t}\right)\left(t\Delta_{\mathcal{Q}_t}^{-2}\Delta_{\mathcal{P}_t}\right)^{1/2} \to 0, \quad t \to \infty,$$

where we use (i) simple rearranging, (ii) triangle inequality, (iii) triangle inequality, Cauchy-Schwarz inequality and representation (3.4.16), (iv) bound (3.4.19), triangle inequality and $\mathbb{E}_{\mathbf{A}^*}[\mathbf{X}_\infty \mathbf{X}_\infty^{\mathrm{T}}] < \infty$, (v) Itô isometry and triangle inequlity, Tonelli theorem, Cauchy-Schwarz and stationarity of state-space, (vi) bound (3.4.19) again and, finally, $M_t \leq N_t'$ by definition, Assumption 3.4.10 and Assumption 3.4.12.

Next, to show Equation (3.4.8) holds, we again work in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ and note that for each $d \times d$ sub-matrix with $l, k \in \{0, \dots, p-1\}$ and $i, j \in \{1, \dots, d\}$ we note that

$$t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} \hat{D}^l Y_{u_m}^{(i)} \Sigma^{-1} \hat{D}^k Y_{u_m}^{(j)} \Delta_{\mathcal{Q}_t}^m - \sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)} \Sigma^{-1} D^k Y_{u_m}^{(j)} \Delta_{\mathcal{Q}_t}^m\right\|\right]$$
$$\overset{(i)}{\leq} t^{-1/2}\|\Sigma^{-1}\| \sum_{m=0}^{M_t-1} \left\{\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)}(\hat{D}^k Y_{u_m}^{(j)} - D^k Y_{u_m}^{(j)})\right|\right] + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^k Y_{u_m}^{(j)}(\hat{D}^l Y_{u_m}^{(i)} - D^l Y_{u_m}^{(i)})\right|\right]\right\}\Delta_{\mathcal{Q}_t}^m$$
$$\overset{(ii)}{\leq} t^{-1/2}\|\Sigma^{-1}\| \sum_{m=0}^{M_t-1} \left\{\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^k Y_{u_m}^{(j)} - D^k Y_{u_m}^{(j)}\right|^2\right]^{1/2}\right.$$
$$\left. + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^k Y_{u_m}^{(j)}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)} - D^l Y_{u_m}^{(i)}\right|^2\right]^{1/2}\right\}\Delta_{\mathcal{Q}_t}^m$$
$$\overset{(iii)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1} \left[\Delta_{\mathcal{P}_t}^{1/2} + \Delta_{\mathcal{P}_t}^{1/2}\right]\Delta_{\mathcal{Q}_t}^m$$
$$\overset{(iv)}{\lesssim} (t\Delta_{\mathcal{P}_t})^{1/2} \to 0, \quad t \to \infty,$$

where we have used (i) the triangle inequality, (ii) Cauchy-Schwarz inequality, (iii) bound (3.4.19), the fact that $\hat{D}^l \mathbf{Y}_{s_n}$ is a linear combination of stationary $\mathbf{Y}_{s_n}, \dots, \mathbf{Y}_{s_{n+l}}$ with finite second moments, stationarity of the state-space representation and $\mathbb{E}_{\mathbf{A}^*}[\mathbf{X}_\infty \mathbf{X}_\infty^{\mathrm{T}}] < \infty$, and (iv) the high-frequency sampling condition, i.e. Assumption 3.4.4. □

Remark 3.D.1. Note that the estimator $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{Q}_t}, \dots, D^{p-1}\mathbb{Y}_{\mathcal{Q}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ is defined with the summations up to $N_t' - 1$ and not $M_t - 1$. In order to prove Equation (3.4.8) and Equation (3.4.9) we are thus missing

$$\sum_{m=M_t}^{N_t'-1} D^l Y_{u_m}^{(i)} \Sigma^{-1} D^k Y_{u_m}^{(j)} \Delta_{\mathcal{Q}_t}^m \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0, \text{ and } \sum_{m=M_t}^{N_t'-1} D^l Y_{u_m}^{(i)} \Sigma^{-1} \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0.$$

as $t \to \infty$ for $0 \leq l, k \leq p-1$ and $1 \leq i, j \leq d$. The first limit can be shown in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ by

$$\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=M_t}^{N_t'-1} D^l Y_{u_m}^{(i)} \Sigma^{-1} D^k Y_{u_m}^{(j)} \Delta_{\mathcal{Q}_t}^m\right\|\right] \leq \|\Sigma^{-1}\| \sum_{m=M_t}^{N_t'-1} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_\infty^{(i)} D^k Y_\infty^{(j)}\right|\right]\Delta_{\mathcal{Q}_t}^m$$
$$\lesssim (N_t' - M_t - 1)\Delta_{\mathcal{Q}_t} \lesssim (p-1)\Delta_{\mathcal{Q}_t} \to 0, \quad t \to \infty,$$

and similarly the second limit follows by

$$\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=M_t}^{N_t'-1} D^l Y_{u_m}^{(i)} \Sigma^{-1} \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c\right\|\right] \leq \|\Sigma^{-1}\| \sum_{m=M_t}^{N_t'-1} \mathbb{E}_{\mathbf{A}^*}\left[\left\|D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c\right\|\right]$$
$$\lesssim (N_t' - M_t - 1)\Delta_{\mathcal{Q}_t} \lesssim (p-1)\Delta_{\mathcal{Q}_t} \to 0, \quad t \to \infty.$$

3.D.4 Proof of Theorem 3.4.19

Proof of Theorem 3.4.19. Recall that by Lemma 3.4.3 it suffices to show Equation (3.4.8) with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$. To show this we work with a slightly different approach than in the proofs in Section 3.D.1 and Section 3.D.3. We first introduce the event A_t, representing whether the thresholding technique is succesful, and show this event has probability approaching one as $t \to \infty$, i.e. Equation (3.D.8). Next, we show that on the set A_t, $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ converges to $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$, i.e. Equation (3.D.11). Combining these results implies the desired condition, i.e. Equation (3.4.8).

For $t \in \mathcal{T}$, $m \in \{0, \ldots, M_t - 1\}$ and $i \in \{1, \ldots, d\}$ let us introduce the event that the thresholding technique is successful on the m-th interval of the t-th partition in the i-th component, i.e.

$$\begin{aligned} A_t^{m,i} :&= \left\{\omega \in \Omega : \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y^{(i)} - \tilde{b}^{(i)} \Delta_{\mathcal{Q}_t}^m\right| \leq \nu_t^{(i),m}\right\}}(\omega) = \mathbb{1}_{\left\{\Delta_{\mathcal{Q}_t}^m N^{(i)} = 0\right\}}(\omega)\right\}, \\ &= (K_t^{m,i} \cap O_t^{m,i}) \cup (K_t^{m,i,c} \cap O_t^{m,i,c}) = (K_t^{m,i} \setminus O_t^{m,i})^c \cap (O_t^{m,i} \setminus K_t^{m,i})^c, \end{aligned} \tag{3.D.5}$$

where

$$\begin{aligned} K_t^{m,i} &:= \left\{\omega \in \Omega : \left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y^{(i)}(\omega) - \tilde{b}^{(i)} \Delta_{\mathcal{Q}_t}^m\right| \leq \nu_t^{(i),m}\right\}, \\ O_t^{m,i} &:= \left\{\omega \in \Omega : \Delta_{\mathcal{Q}_t}^m N^{(i)}(\omega) = 0\right\}. \end{aligned} \tag{3.D.6}$$

On the set $A_t^{m,i}$ either:

- no jumps have occurred and the *approximated* de-trended increment is kept, i.e. we perfectly recover $\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)}$ on $[u_m, u_{m+1}]$; or
- jumps have occurred and the *approximated* de-trended increment is discarded, i.e. $\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c$ is set to 0 on $[u_m, u_{m+1}]$.

Recall that by definition of $\hat{D}^{p-1} \mathbb{Y}_{\mathbf{A}^{(0)}}^c$, i.e. Equation (3.4.23), we have

$$\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c - \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c = \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} \mathbf{Y} - \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y},$$

and thus, since the finite difference approximations live on $\mathcal{P}_t$, by Equation (3.4.19) we have the

bound

$$\begin{aligned}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} - \Delta_{\mathcal{Q}_t}^m D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}\right\|^2\right] \\ = \mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} - \Delta_{\mathcal{Q}_t}^m D^{p-1}\mathbf{Y}\right\|^2\right] \lesssim \Delta_{\mathcal{P}_t}.\end{aligned} \tag{3.D.7}$$

Define the event that the thresholding technique is successful over the whole partition $\mathcal{Q}_t$

$$A_t := \bigcap_{i=1}^{d}\bigcap_{m=0}^{M_t-1} A_t^{m,i}.$$

We start by showing the sequence of events $\{A_t,\ t \in \mathcal{T}\}$ has probability approaching one as $t \to \infty$, i.e.

$$\mathbb{P}_{\mathbf{A}^*}(A_t) \to 1, \quad t \to \infty. \tag{3.D.8}$$

Note

$$\begin{aligned}&\mathbb{P}_{\mathbf{A}^*}\left(A_t^c\right) \qquad (3.D.9)\\
&= \mathbb{P}_{\mathbf{A}^*}\left(\bigcup_{i=1}^{d}\bigcup_{m=0}^{M_t-1} A_t^{m,i,c}\right)\\
&\overset{(i)}{\leq} \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\left[\mathbb{P}_{\mathbf{A}^*}\left(K_t^{m,i} \setminus O_t^{m,i}\right) + \mathbb{P}_{\mathbf{A}^*}\left(O_t^{m,i} \setminus K_t^{m,i}\right)\right]\\
&\overset{(ii)}{\leq} \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\Big[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{(i)} - \tilde{b}^{(i)}\Delta_{\mathcal{Q}_t}^m\right| \leq \nu_t^{(i),m},\ \Delta_{\mathcal{Q}_t}^m N^{(i)} > 0\right)\\
&\qquad + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{(i)} - \tilde{b}^{(i)}\Delta_{\mathcal{Q}_t}^m\right| > \nu_t^{(i),m},\ \Delta_{\mathcal{Q}_t}^m N^{(i)} = 0\right)\Big]\\
&\overset{(iii)}{\leq} \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\Big[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}} + \Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| \leq \nu_t^{(i),m},\ \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\\
&\qquad + \mathbb{P}_{\mathbf{A}*}\left(\Delta_{\mathcal{Q}_t}^m N^{(i)} \geq 2\right) + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(i),m}\right)\Big]\\
&\overset{(iv)}{\leq} \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\Big[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}} + \Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| \leq \nu_t^{(i),m},\ \left|\Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| > 2\nu_t^{(i),m},\ \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\\
&\qquad + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}} + \Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| \leq \nu_t^{(i),m},\ \left|\Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| \leq 2\nu_t^{(i),m},\ \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\\
&\qquad + \left(\lambda^{(i)}\Delta_{\mathcal{Q}_t}^m\right)^2 + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(i),m}\right)\Big]\\
&\overset{(v)}{\leq} \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\Big[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(i),m}, \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\\
&\qquad + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)}\right| \leq 2\nu_t^{(i),m} \,\middle|\, \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\mathbb{P}_{\mathbf{A}^*}\left(\Delta_{\mathcal{Q}_t}^m N^{(i)} = 1\right)\\
&\qquad + \left(\lambda^{(i)}\Delta_{\mathcal{Q}_t}^m\right)^2 + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(i)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(i),m}\right)\Big]\end{aligned}$$

$$\overset{(vi)}{\leq} \sum_{i=1}^{d} \sum_{m=0}^{M_t-1} \left[2\mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)} \right| > \nu_t^{(i),m} \right) + \tilde{F}^{(i)} \left(\left(-2\nu_t^{(i),m}, 2\nu_t^{(i),m} \right) \right) \lambda^{(i)} \Delta_{\mathcal{Q}_t}^m + \left(\lambda^{(i)} \Delta_{\mathcal{Q}_t}^m \right)^2 \right]$$

where we use in (i) Equation (3.D.5), in (ii) Equation (3.D.6), in (iii) Equation (3.4.27), in (iv) theorem of total probability and $\Delta_{\mathcal{Q}_t}^m N^{(i)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx}$ Poisson $\left(\lambda^{(i)} \Delta_{\mathcal{Q}_t}^m \right)$, in (v) triangle inequality and definition of conditional probability, in (vi) the jump distribution $\Delta_{\mathcal{Q}_t}^m \tilde{J}^{(i)} \,|\, \Delta_{\mathcal{Q}_t}^m N^{(i)} = 1 \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \tilde{F}^{(i)}(\cdot)$. Next, we can bound

$$\begin{aligned}
&\mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)} \right| > \nu_t^{(i),m} \right) \\
&\quad \overset{(i)}{\leq} \mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)} - \Delta_{\mathcal{Q}_t}^m D^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)} \right| > \frac{\nu_t^{(i),m}}{2} \right) + \mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m D^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(i)} \right| > \frac{\nu_t^{(i),m}}{2} \right) \\
&\quad \overset{(ii)}{\leq} \frac{4}{\left(\nu_t^{(i),m} \right)^2} \mathbb{E}_{\mathbf{A}^*} \left[\left\| \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c - \Delta_{\mathcal{Q}_t}^m D^{p-1} \mathbf{Y}_{\mathbf{A}^{(0)}}^c \right\|^2 \right] \\
&\qquad\qquad + \mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m (\Sigma^{1/2} \mathbf{W})^{(i)} - \int_{u_m}^{u_{m+1}} \sum_{j=1}^{p} (A_j^* D^{p-j} \mathbf{Y}_u)^{(i)} \, \mathrm{d}u \right| > \frac{\nu_t^{(i),m}}{2} \right) \\
&\quad \overset{(iii)}{\lesssim} \frac{\Delta_{\mathcal{P}_t}}{\left(\nu_t^{(i),m} \right)^2} + \mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta_{\mathcal{Q}_t}^m (\Sigma^{1/2} \mathbf{W})^{(i)} \right| > \frac{\nu_t^{(i),m}}{4} \right) + \mathbb{P}_{\mathbf{A}^*} \left(\left| \int_{u_m}^{u_{m+1}} \sum_{j=1}^{p} (A_j^* D^{p-j} \mathbf{Y}_u)^{(i)} \, \mathrm{d}u \right| > \frac{\nu_t^{(i),m}}{4} \right) \\
&\quad \overset{(iv)}{\lesssim} \frac{\Delta_{\mathcal{P}_t}}{\left(\nu_t^{(i),m} \right)^2} + \frac{\left(\Delta_{\mathcal{Q}_t}^m \right)^{1/2}}{\nu_t^{(i),m}} \exp \left\{ -\frac{1}{32\varsigma_i^2} \left(\Delta_{\mathcal{Q}_t}^m \right)^{-1} \left(\nu_t^{(i),m} \right)^2 \right\} \\
&\qquad\qquad + \frac{16}{\left(\nu_t^{(i),m} \right)^2} \mathbb{E}_{\mathbf{A}^*} \left[\left| \int_{u_m}^{u_{m+1}} \sum_{j=1}^{p} (A_j^* D^{p-j} \mathbf{Y}_u)^{(i)} \, \mathrm{d}u \right|^2 \right] \\
&\quad \overset{(v)}{\lesssim} \frac{\Delta_{\mathcal{P}_t}}{\left(\nu_t^{(i),m} \right)^2} + \frac{\left(\Delta_{\mathcal{Q}_t}^m \right)^{1/2}}{\nu_t^{(i),m}} \exp \left\{ -\frac{1}{32\varsigma_i^2} \left(\Delta_{\mathcal{Q}_t}^m \right)^{-1} \left(\nu_t^{(i),m} \right)^2 \right\} \\
&\qquad\qquad + \frac{\Delta_{\mathcal{Q}_t}^m}{\left(\nu_t^{(i),m} \right)^2} \int_{u_m}^{u_{m+1}} \mathbb{E}_{\mathbf{A}^*} \left[\left| \sum_{j=1}^{p} (A_j^* D^{p-j} \mathbf{Y}_\infty)^{(i)} \right|^2 \right] \mathrm{d}u \\
&\quad \overset{(vi)}{\lesssim} \frac{\Delta_{\mathcal{P}_t}}{\left(\nu_t^{(i),m} \right)^2} + \frac{\left(\Delta_{\mathcal{Q}_t}^m \right)^{1/2}}{\nu_t^{(i),m}} \exp \left\{ -\frac{1}{32\varsigma_i^2} \left(\Delta_{\mathcal{Q}_t}^m \right)^{-1} \left(\nu_t^{(i),m} \right)^2 \right\} + \frac{\left(\Delta_{\mathcal{Q}_t}^m \right)^2}{\left(\nu_t^{(i),m} \right)^2} \\
&\quad \overset{(vii)}{\lesssim} \left(\Delta_{\mathcal{P}_t} + \left(\Delta_{\mathcal{Q}_t}^m \right)^2 \right) \left(\Delta_{\mathcal{Q}_t}^m \right)^{-2\beta^{(i)}}, \text{ for } t \in \mathcal{T} \text{ sufficiently large,}
\end{aligned} \tag{3.D.10}$$

where we use in (i) triangle inequality, in (ii) Markov's inequality and Equation (3.4.16), in (iii) Equation (3.D.7) and triangle inequality, in (iv) Klenke (2008, Lemma 22.2) with $\varsigma_i^2 \Delta_{\mathcal{Q}_t}^m = \mathrm{Var}[\Delta_{\mathcal{Q}_t}^m (\Sigma^{1/2} \mathbf{W})^{(i)}]$ and Markov's inequality, in (v) Jensen's inequality and stationarity, in (vi) $\mathbb{E}_{\mathbf{A}^*}[\mathbf{X}_\infty \mathbf{X}_\infty^{\mathrm{T}}] < \infty$ and in (vii) Assumption 3.4.17.(i), $\exp\{-x^2/(32\varsigma_i^2)\} \lesssim x^{-3}$ for $x = (\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}-1/2}$ sufficiently large and $\Delta_{\mathcal{Q}_t}^m \lesssim 1$ for t sufficiently large. Plugging this into the previous inequality and using $\Delta_{\mathcal{Q}_t}^m \geq c_{\mathcal{Q}_t} \Delta_{\mathcal{Q}_t}$, for $t \in \mathcal{T}$ sufficiently large we have

$$\mathbb{P}_{\mathbf{A}^*} (A_t^c) \lesssim \sum_{i=1}^{d} \left[c_{\mathcal{Q}_t}^{-2\beta^{(i)}} M_t \Delta_{\mathcal{P}_t} \Delta_{\mathcal{Q}_t}^{-2\beta^{(i)}} + t \Delta_{\mathcal{Q}_t}^{1-2\beta^{(i)}} + t \tilde{F}^{(i)} \left(\left(-2(\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}, \, 2(\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}} \right) \right) + t \Delta_{\mathcal{Q}_t} \right]$$

$$\lesssim \sum_{i=1}^{d} \left[c_{\mathcal{Q}_t}^{-2\beta^{(i)}} (t^{-1} N_t' \Delta_{\mathcal{Q}_t})(t\Delta_{\mathcal{Q}_t}^{-2}\Delta_{\mathcal{P}_t}) \Delta_{\mathcal{Q}_t}^{1-2\beta^{(i)}} + t\Delta_{\mathcal{Q}_t}^{1-2\beta^{(i)}} + t\tilde{F}^{(i)}\left(\left(-2\Delta_{\mathcal{Q}_t}^{\beta^{(i)}},\ 2\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}\right)\right)\right] \to 0,$$

as $t \to \infty$, by Assumption 3.4.10, $M_t \leq N_t'$ by definition, Assumption 3.4.12, Assumption 3.4.17.(*ii*) and Assumption 3.4.17.(*iii*). We note that it now suffices to show the $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ convergence

$$\mathbb{E}_{\mathbf{A}^*}\left[\left\| t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\mathbb{1}_{A_t}\right\|\right] \to 0, \quad t \to \infty, \tag{3.D.11}$$

as this immediately implies convergence in probability

$$t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\mathbb{1}_{A_t} \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0, \quad t \to \infty, \tag{3.D.12}$$

and hence $\forall \epsilon > 0$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{A}^*}\left(\left\| t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\right\| > \epsilon\right) &= \mathbb{P}_{\mathbf{A}^*}\left(\left\{\left\| t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\right\| > \epsilon\right\} \cap A_t\right) \\
&\quad + \mathbb{P}_{\mathbf{A}^*}\left(\left\{\left\| t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\right\| > \epsilon\right\} \cap A_t^c\right) \\
&\leq \mathbb{P}_{\mathbf{A}^*}\left(\left\| t^{-1/2}\left(\mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t,\boldsymbol{\nu}_t} - \mathbf{H}_{\mathcal{P}_t,\mathcal{Q}_t}\right)\mathbb{1}_{A_t}\right\| > \epsilon\right) + \mathbb{P}_{\mathbf{A}^*}\left(A_t^c\right) \\
&\to 0, \quad t \to \infty,
\end{aligned}$$

by combining Equation (3.D.12) and Equation (3.D.8), i.e. this implies the sufficient condition (3.4.8) for the required asymptotic properties of $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t^m)$.

To show Equation (3.D.11) note that for each d-entry batch

$$\begin{aligned}
&t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\left(\sum_{m=0}^{M_t-1}\hat{D}^l Y_{u_m}^{(i)}\Sigma^{-1}\Delta_{\mathcal{Q}_t,\boldsymbol{\nu}_t}^m \hat{D}^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)}}^c - \sum_{m=0}^{M_t-1}\hat{D}^l Y_{u_m}^{(i)}\Sigma^{-1}\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)}}^c\right)\mathbb{1}_{A_t}\right\|\right] \\
&\overset{(i)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{m=0}^{M_t-1}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\hat{D}^l Y_{u_m}^{(i)}\left(\Delta_{\mathcal{Q}_t,\boldsymbol{\nu}_t}^m \hat{D}^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)}}^c - \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)}}^c\right)\mathbb{1}_{A_t}\right\|\right] \\
&\overset{(ii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)}\ \Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\ \mathbb{1}_{A_t\cap O_t^{m,i,c}}\right|\right] \\
&\overset{(iii)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\left\{\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)}\ \Delta_{\mathcal{Q}_t}^m\left(\hat{D}^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)} - D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right)\right|\right]\right. \\
&\qquad + \mathbb{E}_{\mathbf{A}^*}\left[\left|\left(\hat{D}^l Y_{u_m}^{(i)} - D^l Y_{u_m}^{(i)}\right)\Delta_{\mathcal{Q}_t}^m D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right|\right] \\
&\qquad \left. + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)}\ \Delta_{\mathcal{Q}_t}^m D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\ \mathbb{1}_{O_t^{m,i,c}}\right|\right]\right\} \\
&\overset{(iv)}{\leq} t^{-1/2}\|\Sigma^{-1}\|\sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\left\{\mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{\infty}^{(i)}\right|^2\right]^{1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta_{\mathcal{Q}_t}^m\left(\hat{D}^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)} - D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right)\right|^2\right]^{1/2}\right. \\
&\qquad + \mathbb{E}_{\mathbf{A}^*}\left[\left|\hat{D}^l Y_{u_m}^{(i)} - D^l Y_{u_m}^{(i)}\right|^2\right]^{1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right|^2\right]^{1/2} \\
&\qquad \left. + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)}\ \Delta_{\mathcal{Q}_t}^m(\Sigma^{1/2}\mathbf{W})^{(i)}\ \mathbb{1}_{O_t^{m,i,c}}\right|\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
&\qquad + \mathbb{E}_{\mathbf{A}^*}\Big[\Big|D^l Y^{(i)}_{u_m} \int_{u_m}^{u_{m+1}} \sum_{j=1}^{p} (A_j^* D^{p-j}\mathbf{Y}_u)^{(i)}\,\mathrm{d}u\; \mathbb{1}_{O_t^{m,i,c}}\Big|\Big]\Big\} \\
&\overset{(v)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\Big\{\Delta_{\mathcal{P}_t}^{1/2} + \Delta_{\mathcal{P}_t}^{1/2}\Delta_{\mathcal{Q}_t}^{1/2} + \mathbb{E}_{\mathbf{A}^*}\Big[\Big|D^l Y^{(i)}_{\infty}\Big|\Big]\mathbb{E}_{\mathbf{A}^*}\Big[\Big|\Delta^m_{\mathcal{Q}_t}(\Sigma^{1/2}\mathbf{W})^{(i)}\Big|\Big]\mathbb{P}_{\mathbf{A}^*}\left(O_t^{m,i,c}\right) \\
&\qquad + \sum_{j=1}^{p}\mathbb{E}_{\mathbf{A}^*}\Big[\Big|D^l Y^{(i)}_{u_m}\,\mathbb{1}_{O_t^{m,i,c}}\Big|^2\Big]^{1/2}\mathbb{E}_{\mathbf{A}^*}\Big[\Big|\int_{u_m}^{u_{m+1}}(A_j^* D^{p-j}\mathbf{Y}_u)^{(i)}\,\mathrm{d}u\Big|^2\Big]^{1/2}\Big\} \\
&\overset{(vi)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\Big\{\Delta_{\mathcal{P}_t}^{1/2} + \Delta_{\mathcal{P}_t}^{1/2}\Delta_{\mathcal{Q}_t}^{1/2} + (\Delta^m_{\mathcal{Q}_t})^{1/2}\Delta^m_{\mathcal{Q}_t} + \sum_{j=1}^{p}\mathbb{E}_{\mathbf{A}^*}\Big[\Big|D^l Y^{(i)}_{\infty}\Big|^2\Big]^{1/2}\mathbb{P}_{\mathbf{A}^*}\left(O_t^{m,i,c}\right)^{1/2}\Delta^m_{\mathcal{Q}_t}\Big\} \\
&\overset{(vii)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1}\sum_{i=1}^{d}\Big\{\Delta_{\mathcal{P}_t}^{1/2} + \Delta_{\mathcal{P}_t}^{1/2}\Delta_{\mathcal{Q}_t}^{1/2} + (\Delta^m_{\mathcal{Q}_t})^{3/2} + (\Delta^m_{\mathcal{Q}_t})^{3/2}\Big\} \\
&\overset{(viii)}{\lesssim} \left(t^{-1}N_t'\Delta_{\mathcal{Q}_t}\right)\left(t\Delta_{\mathcal{Q}_t}^{-2}\Delta_{\mathcal{P}_t}\right)^{1/2}\Delta_{\mathcal{P}_t}^{1/2} + t^{1/2}\Delta_{\mathcal{Q}_t}^{1/2} \to 0, \quad t\to\infty,
\end{aligned}
\tag{3.D.13}
$$

where we use in (i) the triangle inequality, in (ii) the domination $\|\cdot\|_2 \le \|\cdot\|_1$ and the fact that on A_t either the increment is perfectly recovered, on $O_t^{m,i}$, or set to zero by the thresholding, on $O_t^{m,i,c}$, in (iii) the triangle inequality and the fact that $\mathbb{1}_{A\cap O} \le \mathbb{1}_O \le 1$ for sets $A, O \subset \Omega$, in (iv) Cauchy-Schwarz inequality (and stationarity) on the first two terms, Equation (3.4.16) and triangle inequality on the third term, in (v) Equation (3.D.7) and boundedness of the second moments on the first term, the bounds (3.4.19) and (3.D.14) on the second term, independence of $D^l\mathbf{Y}_{u_m}$, $\Delta^m_{\mathcal{Q}_t}\mathbf{W}$ and $\Delta^m_{\mathcal{Q}_t}\mathbf{N}$ on the third term, triangle inequality and Cauchy-Schwarz on the fourth term, in (vi) boundedness of second moments of $\mathbf{X}_t$, second moment of Brownian increments and $\mathbb{P}_{\mathbf{A}^*}(O_t^{m,i,c}) \le \lambda^{(i)}\Delta^m_{\mathcal{Q}_t}$ on the third term, independence of $D^l\mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t}\mathbf{N}$, Jensen's inequality and uniform boundedness of second moments in the fourth term, and finally $(viii)$ Assumption 3.4.10, Assumption 3.4.12 and Assumption 3.4.17.(ii).

In the above, we applied the following bound on the squared increment of the continuous martingale part

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}}\right\|^2\right]^{1/2} \\
&\overset{(i)}{\le} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta^m_{\mathcal{Q}_t}\Sigma^{1/2}\mathbf{W}\right\|^2\right]^{1/2} + \sum_{j=1}^{p}\|A_j^*\|\,\mathbb{E}_{\mathbf{A}^*}\left[\left\|\int_{u_m}^{u_{m+1}} D^{p-j}\mathbf{Y}_u\,\mathrm{d}u\right\|^2\right]^{1/2} \\
&\overset{(ii)}{\le} \mathbb{E}_{\mathbf{A}^*}\left[\left\|\Delta^m_{\mathcal{Q}_t}\Sigma^{1/2}\mathbf{W}\right\|^2\right]^{1/2} + \sum_{j=1}^{p}\|A_j^*\|\left(\Delta^m_{\mathcal{Q}_t}\int_{u_m}^{u_{m+1}}\mathbb{E}_{\mathbf{A}^*}\left[\left\|D^{p-j}\mathbf{Y}_u\right\|^2\right]\mathrm{d}u\right)^{1/2} \\
&\overset{(iii)}{\lesssim} (\Delta^m_{\mathcal{Q}_t})^{1/2},
\end{aligned}
\tag{3.D.14}
$$

obtained by using (i) the representation (3.4.16) and the triangle inequality, (ii) Jensen's inequality and (iii) the second moment of Brownian increments and stationarity and boundedness of second moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$. □

3.D.5 Proof of Theorem 3.4.23

Proof of Theorem 3.4.23. Recall that by Lemma 3.4.3 it suffices to show Equation (3.4.8) with $\hat{\mathbf{A}}_{1,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}, \mathbb{J}_{\mathcal{Q}_t}, \mathbb{M}_{\mathcal{Q}_t})$ and $\hat{\mathbf{A}}_{2,t} = \hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$. We start by splitting the limiting random variable in (3.4.8) into three components. For $i \in \{1, \dots, d\}$ and $l \in \{0, \dots, p-1\}$, after factoring out the Σ^{-1} term, we decompose each d-batch entry as

$$
\begin{aligned}
&t^{-1/2}\Big(\sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t,\boldsymbol{\nu}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} - \sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \Big) \\
&\quad = t^{-1/2} \left(\sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \Sigma^{-1} \left(\Delta^m_{\mathcal{Q}_t,\boldsymbol{\nu}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right) \\
&\quad = t^{-1/2} \left(\sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \left(\left(\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right) \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right) \\
&\quad \overset{(*)}{\approx} t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \left(\left(\Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right) \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right) \\
&\quad = t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right) \\
&\qquad\qquad + t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} \mathbf{M} \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} \right) \\
&\quad = t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{\mathbf{Y}} \right| \le 2\boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right) \\
&\qquad + t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{\mathbf{Y}} \odot \left(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} - \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{\mathbf{Y}} \right| \le 2\boldsymbol{\nu}^m_t \right\}} \right) \right) \\
&\qquad\qquad + t^{-1/2} \left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} \mathbf{M} \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} \mathbf{Y} - \mathbf{b} \Delta^m_{\mathcal{Q}_t} \right| \le \boldsymbol{\nu}^m_t \right\}} \right) \\
&\quad =: \mathbf{S}^1_t(i,l) + \mathbf{S}^2_t(i,l) + \mathbf{S}^3_t(i,l),
\end{aligned}
$$

setting $D^{p-1}\tilde{\mathbb{Y}} = \{D^{p-1}\tilde{\mathbf{Y}}_s,\ s \ge 0\}$ with

$$
D^{p-1}\tilde{\mathbf{Y}}_s := D^{p-1}\mathbf{Y}_0 + D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} + \mathbf{J}_s = D^{p-1}\mathbf{Y}_s - \mathbf{b}s - \mathbf{M}_s, \tag{3.D.15}
$$

and understanding the approximation step $(*)$ in the $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$ limit (and hence in the $\mathbb{P}_{\mathbf{A}^*}$ limit) as $t \to \infty$. Note we directly define the auxiliary process $D^{p-1}\tilde{\mathbb{Y}}$ but no $\tilde{\mathbb{Y}}$. From Equation (3.4.28) recall $\mathbb{J}$ is the finite activity pure jump Lévy process with jumps greater than one given in Equation (3.4.3) and $\mathbb{M}$ is the pure jump martingale with jumps smaller or equal to one given in Equation (3.4.4). Note that $\mathbb{J}$ is a compound Poisson process with elementwise Poisson counting processes $\{N^{(i)}_t,\ t \ge 0\}$ for $i = 1, \dots, d$. We denote by $\lambda^{(i)}$ the jump rate of the process $\{N^{(i)}_t,\ t \ge 0\}$ and by $\tilde{F}^{(i)}(\cdot)$ the probability distribution of jump sizes in the i-th component of $\mathbb{J}$, i.e. the i-th marginal of $\tilde{F} := F|_{\{\mathbf{x}:\|\mathbf{x}\|>1\}}/F(\{\mathbf{x} : \|\mathbf{x}\| > 1\})$.

The validity of the approximation $(*)$ can be shown as follows for $t \in \mathcal{T}$ sufficiently large:

$$
\begin{aligned}
t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\Bigg[\Bigg\| \sum_{m=0}^{M_t-1} \hat{D}^l Y^{(i)}_{u_m} \left(\left(\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right) \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right| \leq \boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \\
- \sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m} \left(\left(\Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right) \odot \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right| \leq \boldsymbol{\nu}^m_t \right\}} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \Bigg\| \Bigg] \\
\overset{(i)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \Bigg\{ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right) - D^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right) \right\| \right] \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} - D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right\| \right] \Bigg\} \\
\overset{(ii)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \Bigg\{ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} \right) \right\| \right] \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \left(\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m} \right) \left(\Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right) \right\| \right] \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \left(\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right) \right\| \right] \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \left(\hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m} \right) \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right\| \right] \Bigg\} \\
\overset{(iii)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1} \Bigg\{ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \right\|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} \right\|^2 \right]^{1/2} \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m} \right\|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\| \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y} - \mathbf{b}\Delta^m_{\mathcal{Q}_t} \right\|^2 \right]^{1/2} \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} \right\|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right\|^2 \right]^{1/2} \\
+ \mathbb{E}_{\mathbf{A}^*}\left[\left\| \hat{D}^l Y^{(i)}_{u_m} - D^l Y^{(i)}_{u_m} \right\|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left\| \Delta^m_{\mathcal{Q}_t} D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \right\|^2 \right]^{1/2} \Bigg\} \\
\overset{(iv)}{\lesssim} t^{-1/2} \sum_{m=0}^{M_t-1} \{ \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{1/2}_{\mathcal{P}_t}\Delta^{1/2}_{\mathcal{Q}_t} + \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{1/2}_{\mathcal{P}_t}\Delta^{1/2}_{\mathcal{Q}_t} \} \\
\overset{(v)}{\lesssim} t^{-1/2} M_t \Delta^{1/2}_{\mathcal{P}_t} \lesssim (t^{-1}N'_t\Delta_{\mathcal{Q}_t})(t\Delta^{-2}_{\mathcal{Q}_t}\Delta_{\mathcal{P}_t})^{1/2} \to 0, \quad t \to \infty,
\end{aligned}
$$

where we use in (i) triangle inequality and $\|\mathbf{x} \odot \mathbb{1}_{\{\mathbf{y}\in A\}}\| \leq \|\mathbf{x}\|$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, in (ii) adding and subtracting the same quantity and applying the triangle inequality multiple times, in (iii) Cauchy-Schwarz inequality on all four terms, in (iv) stationarity and boundedness of second moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$, Equation (3.D.7), Equation (3.4.19), Equation (3.B.1) and Equation (3.D.14), and finally in (v) Assumption 3.4.10 and Assumption 3.4.12. It now remains to prove $\mathbf{S}^1_t(i,l), \mathbf{S}^2_t(i,l), \mathbf{S}^3_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$ for $i \in \{1, \ldots, d\}, l \in \{0, \ldots, p-1\}$.

To show $\mathbf{S}^1_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$ note that $D^{p-1}\tilde{\mathbb{Y}}$ is a finite jump activity process with $\tilde{\mathbf{b}} = 0$ and $\tilde{F}^{(i)}$ for $i \in \{1, \ldots, d\}$ satisfies Assumption 3.4.20.(iii), i.e. a similar condition to Assumption 3.4.17.(iii). We can thus proceed with a similar strategy as in the proof of Theorem 3.4.19 presented in Appendix 3.D.4. We do not repeat the full argument but summarize the main steps.

Importantly we note that here we can work with the "exact" derivatives $D^{p-1}\tilde{\mathbf{Y}}$ directly while in Appendix 3.D.4 we were working with the finite difference approximations $\hat{D}^{p-1}\mathbf{Y}$. We consider the events

$$\tilde{A}_t := \bigcap_{i=1}^{d}\bigcap_{m=0}^{M_t-1} \tilde{A}_t^{m,i}, \quad \text{where } \tilde{A}_t^{m,i} := \left\{\omega \in \Omega : \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}\tilde{Y}^{(i)}\right| \le 2\nu_t^{(i),m}\right\}}(\omega) = \mathbb{1}_{\left\{\Delta_{\mathcal{Q}_t}^m N^{(i)}=0\right\}}(\omega)\right\},$$

for $t \in \mathcal{T}$, $m \in \{0, \dots, M_t - 1\}$ and $i \in \{1, \dots, d\}$. We note $\tilde{A}_t$ has probability approaching one:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{A}^*}\left(\tilde{A}_t^c\right) \\
&\le \sum_{i=1}^{d}\sum_{m=0}^{M_t-1}\left[2\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right| > 2\nu_t^{(i),m}\right) + \tilde{F}^{(i)}\left(\left(-4\nu_t^{(i),m}, 4\nu_t^{(i),m}\right)\right)\lambda^{(i)}\Delta_{\mathcal{Q}_t}^m + \left(\lambda^{(i)}\Delta_{\mathcal{Q}_t}^m\right)^2\right] \\
&\lesssim \sum_{i=1}^{d}\left[t\Delta_{\mathcal{Q}_t}^{1-2\beta^{(i)}} + t\tilde{F}^{(i)}\left(\left(-4(\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}},\ 4(\Delta_{\mathcal{Q}_t}^m)^{\beta^{(i)}}\right)\right)\right] \to 0, \quad t \to \infty,
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}Y_{\mathbf{A}^{(0)}}^{c,(i)}\right| > 2\nu_t^{(i),m}\right) \\
&\quad\le \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta_{\mathcal{Q}_t}^m(\Sigma^{1/2}\mathbf{W})^{(i)}\right| > \nu_t^{(i),m}\right) + \mathbb{P}_{\mathbf{A}^*}\left(\left|\int_{u_m}^{u_{m+1}}\sum_{j=1}^{p}(A_j^* D^{p-j}\mathbf{Y}_u)^{(i)}\,\mathrm{d}u\right| > \nu_t^{(i),m}\right) \\
&\quad\lesssim (\Delta_{\mathcal{Q}_t}^m)^{2-2\beta^{(i)}}, \text{ for } t \in \mathcal{T} \text{ sufficiently large.}
\end{aligned} \tag{3.D.16}$$

The first inequality is obtained by proceeding exactly as in Equation (3.D.9) with $\tilde{b}^{(i)} = 0$ and the finite difference approximation $\hat{D}^{p-1}\mathbf{Y}$ replaced by the exact derivative $D^{p-1}\mathbf{Y}$. The second inequality follows from the bound on the second term in Equation (3.D.10).(i) with $\nu_t^{(i),m}/2$ replaced by $2\nu_t^{(i),m}$.

Moreover, on $\tilde{A}_t$, the term $\mathbf{S}_t^1(i,l)$ vanishes in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$. This follows by noting that the same approach used in Equation (3.D.13) with the finite difference approximations replaced by exact derivatives, i.e. the first two terms in step (3.D.13).(iii) set to zero, yields

$$\begin{aligned}
&t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)}\Delta_{\mathcal{Q}_t}^m D^{p-1}\tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}\tilde{\mathbf{Y}}\right| \le 2\boldsymbol{\nu}_t^m\right\}} - \sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)}\Delta_{\mathcal{Q}_t}^m D^{p-1}\mathbf{Y}_{\mathbf{A}^{(0)}}^c\right)\mathbb{1}_{\tilde{A}_t}\right\|\right] \\
&\quad\lesssim t^{1/2}\Delta_{\mathcal{Q}_t}^{1/2} \to 0, \quad t \to \infty.
\end{aligned}$$

We can thus conclude that $\mathbf{S}_t^1(i,l) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0$ as $t \to \infty$ by the same discussion as in Appendix 3.D.4 which ensures Equation (3.D.8) and Equation (3.D.11) imply the desired convergence in probability.

Next, to show $\mathbf{S}_t^2(i,l) \xrightarrow{\mathbb{P}_{\mathbf{A}^*}} 0$ as $t \to \infty$ let us write

$$\mathbf{S}_t^2(i,l) = t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)}\Delta_{\mathcal{Q}_t}^m D^{p-1}\tilde{\mathbf{Y}} \odot \left(\mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta_{\mathcal{Q}_t}^m\right| \le \boldsymbol{\nu}_t^m\right\}} - \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m D^{p-1}\tilde{\mathbf{Y}}\right| \le 2\boldsymbol{\nu}_t^m\right\}}\right)\right)$$

$$
\begin{aligned}
&= t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|\le \boldsymbol{\nu}^m_t, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{\mathbf{Y}}\right|>2\boldsymbol{\nu}^m_t\right\}}\right)\\
&\quad - t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|> \boldsymbol{\nu}^m_t, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{\mathbf{Y}}\right|\le 2\boldsymbol{\nu}^m_t\right\}}\right)\\
&=: \mathbf{S}^{2,1}_t(i,l) - \mathbf{S}^{2,2}_t(i,l).
\end{aligned}
$$

We have $\mathbf{S}^{2,1}_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t\to\infty$ since

$$
\begin{aligned}
\mathbb{P}_{\mathbf{A}^*}\left(\|\mathbf{S}^{2,1}_t(i,l)\|>0\right) &\overset{(i)}{\le} \mathbb{P}_{\mathbf{A}^*}\left(\bigcup_{m=0}^{M_t-1}\bigcup_{j=1}^{d}\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t}\right| \le \nu^{(j),m}_t, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| > 2\nu^{(j),m}_t\right\}\right)\\
&\overset{(ii)}{\le} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t}\right| \le \nu^{(j),m}_t, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| > 2\nu^{(j),m}_t\right)\\
&\overset{(iii)}{\le} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\left\{\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{(j)}\right| > \frac{\nu^{(j),m}_t}{2}\right)\right.\\
&\qquad \left. + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t}\right| \le \frac{3\nu^{(j),m}_t}{2}, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| > 2\nu^{(j),m}_t\right)\right\}\\
&\overset{(iv)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\left\{\Delta_{\mathcal{P}_t}\left(\nu^{(j),m}_t\right)^{-2} + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| > \frac{\nu^{(j),m}_t}{2}, \Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right)\right.\\
&\qquad \left. + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| > 2\nu^{(j),m}_t, \Delta^m_{\mathcal{Q}_t} J^{(j)} = 0\right)\right\}\\
&\overset{(v)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\left\{\Delta_{\mathcal{P}_t}\left(\nu^{(j),m}_t\right)^{-2} + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| > \frac{\nu^{(j),m}_t}{2}\right)\mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right)\right.\\
&\qquad \left. + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > 2\nu^{(j),m}_t\right)\right\}\\
&\overset{(vi)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\left\{\Delta_{\mathcal{P}_t}\left(\nu^{(j),m}_t\right)^{-2} + \Delta^m_{\mathcal{Q}_t}\left(\nu^{(j),m}_t\right)^{-2}\Delta^m_{\mathcal{Q}_t} + \left(\Delta^m_{\mathcal{Q}_t}\right)^2\left(\nu^{(j),m}_t\right)^{-2}\right\}\\
&\overset{(vii)}{\lesssim} \sum_{j=1}^{d}\left\{c^{-2\beta^{(j)}}_{\mathcal{Q}_t} M_t\Delta_{\mathcal{P}_t}\Delta^{-2\beta^{(j)}}_{\mathcal{Q}_t} + t\Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t}\right\}\\
&\overset{(viii)}{\lesssim} \sum_{j=1}^{d}\left\{c^{-2\beta^{(j)}}_{\mathcal{Q}_t}(t^{-1}N'_t\Delta_{\mathcal{Q}_t})(t\Delta^{-2}_{\mathcal{Q}_t}\Delta_{\mathcal{P}_t})\Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t} + t\Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t}\right\} \to 0, \quad t\to\infty,
\end{aligned}
$$

by using in (i) the fact that if the norm is non-zero then at least one component of one of the summands is non-zero and thus the corresponding indicator is non-zero, in (ii) sub-additivity of probability, in (iii) theorem of total probability and $|\hat{y}-y|\le\nu/2, |\hat{y}|\le\nu, |\tilde{y}|>2\nu \implies |y|\le 3/2\nu, |\tilde{y}|>2\nu$ for $y,\hat{y},\tilde{y},\nu\in\mathbb{R}$, in (iv) Markov's inequality and Equation (3.D.7) on the first term, theorem of total probability,

$$
\Delta_{\mathcal{Q}_t} D^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t} = \Delta_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)} + \Delta^m_{\mathcal{Q}_t} M^{(j)},
$$

by definition of $D^{p-1}\tilde{\mathbb{Y}}$ in Equation (3.D.15), and $|y| \leq 3/2\nu, |y-m| > 2\nu \implies |m| > \nu/2$ for $y, m, \nu \in \mathbb{R}$ on the second term, in (v) independence of $\mathbb{J}$ and $\mathbb{M}$ on the second term, in (vi) Markov's inequality,

$$\mathbb{E}_{\mathbf{A}^*}\left[\|\Delta^m_{\mathcal{Q}_t}\mathbf{M}\|^2\right], \mathbb{E}_{\mathbf{A}^*}\left[\|\Delta^m_{\mathcal{Q}_t}\mathbf{J}\|^2\right] \lesssim \Delta^m_{Q_t} \tag{3.D.17}$$

since $\mathbb{M}$ and $\mathbb{J}$ are Lévy processes with bounded second moments and $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\sim}$ Poisson$(\lambda^{(j)}\Delta^m_{\mathcal{Q}_t})$ on the second term, Equation (3.D.16) on the third term, in (vii) Assumption 3.4.20.(i) and $\Delta^m_{\mathcal{Q}_t} \geq c_{\mathcal{Q}_t}\Delta_{\mathcal{Q}_t}$, in $(viii)$ Assumption 3.4.10, Assumption 3.4.12 and Assumption 3.4.20.(ii).

To show $\mathbf{S}^{2,2}_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$ we write using Equation (3.4.16)

$$D^{p-1}\tilde{\mathbf{Y}}_s - D^{p-1}\tilde{\mathbf{Y}}_0 = D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)},s} + \mathbf{J}_s = \Sigma^{1/2}\mathbf{W}_s - \sum_{j=1}^{p}\int_0^s A^*_j D^{p-j}\mathbf{Y}_u \, \mathrm{d}u + \mathbf{J}_s,$$

and we can hence further decompose $\mathbf{S}^{2,2}_t(i,l)$ as

$$\begin{aligned}
\mathbf{S}^{2,2}_t(i,l) &= t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}\tilde{\mathbf{Y}} \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|>\boldsymbol{\nu}^m_t, \left|\Delta^m_{\mathcal{Q}_t}D^{p-1}\tilde{\mathbf{Y}}\right|\leq 2\boldsymbol{\nu}^m_t\right\}}\right) \\
&= t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}\mathbf{Y}^c_{\mathbf{A}^{(0)}} \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|>\boldsymbol{\nu}^m_t, \left|\Delta^m_{\mathcal{Q}_t}D^{p-1}\tilde{\mathbf{Y}}\right|\leq 2\boldsymbol{\nu}^m_t\right\}}\right) \\
&\qquad + t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}\mathbf{J} \odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|>\boldsymbol{\nu}^m_t, \left|\Delta^m_{\mathcal{Q}_t}D^{p-1}\tilde{\mathbf{Y}}\right|\leq 2\boldsymbol{\nu}^m_t\right\}}\right) \\
&=: \mathbf{S}^{2,2,1}_t(i,l) + \mathbf{S}^{2,2,2}_t(i,l).
\end{aligned}$$

We start by showing $\mathbf{S}^{2,2,1}_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$. We work in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\|\mathbf{S}^{2,2,1}_t(i,l)\|\right] \\
&\overset{(i)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)}-b^{(j)}\Delta^m_{\mathcal{Q}_t}\right|>\nu^{(j),m}_t\right\}}\right] \\
&\overset{(ii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| \left(\mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)}-\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{(j)}\right|>\nu^{(j),m}_t/2\right\}}\right.\right. \\
&\qquad\qquad \left.\left. + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{(j)}-b^{(j)}\Delta^m_{\mathcal{Q}_t}\right|>\nu^{(j),m}_t/2\right\}}\right)\right] \\
&\overset{(iii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| \left(\mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)}-\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{(j)}\right|>\nu^{(j),m}_t/2\right\}}\right.\right. \\
&\qquad\qquad \left.\left. + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}}+\Delta^m_{\mathcal{Q}_t}M^{(j)}\right|>\nu^{(j),m}_t/2\right\}} + \mathbb{1}_{\left\{\Delta^m_{\mathcal{Q}_t}N^{(j)}>0\right\}}\right)\right] \\
&\overset{(iv)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| \left(\mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)}-\Delta^m_{\mathcal{Q}_t}D^{p-1}Y^{(j)}\right|>\nu^{(j),m}_t/2\right\}}\right.\right.
\end{aligned}$$

$$+ \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(j),m}/4\right\}} + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| > \nu_t^{(j),m}/4\right\}} + \mathbb{1}_{\left\{\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right\}}\Big)\Big]$$

$$\overset{(v)}{\lesssim} t^{-1/2} \sum_{j=1}^{d} \sum_{m=0}^{M_t-1} \left[\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} (\Sigma^{1/2}\mathbf{W})^{(j)}\right|^2\right]^{1/2} + \sum_{k=1}^{p} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m} \int_{u_m}^{u_{m+1}} (A_k^* D^{p-k} \mathbf{Y}_u)^{(j)} \, \mathrm{d}u\right|^2\right]^{1/2}\right]$$

$$\times \left[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{(j)}\right| > \nu_t^{(j),m}/2\right)^{1/2} + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(j),m}/4\right)^{1/2}\right.$$

$$\left. + \mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| > \nu_t^{(j),m}/4\right) + \mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right)\right]$$

$$\overset{(vi)}{\lesssim} t^{-1/2} \sum_{j=1}^{d} \sum_{m=0}^{M_t-1} \left[\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} (\Sigma^{1/2}\mathbf{W})^{(j)}\right|^2\right]^{1/2}\right.$$

$$\left. + \sum_{k=1}^{p} \left(\Delta^m_{\mathcal{Q}_t} \int_{u_m}^{u_{m+1}} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m} (A_k^* D^{p-k} \mathbf{Y}_u)^{(j)}\right|^2\right] \mathrm{d}u\right)^{1/2}\right]$$

$$\times \left[\Delta_{\mathcal{P}_t}^{1/2} \left(\nu_t^{(j),m}\right)^{-1} + \left(\Delta^m_{\mathcal{Q}_t}\right)^{1-\beta^{(j)}} + \Delta^m_{\mathcal{Q}_t} \left(\nu_t^{(j),m}\right)^{-2} + \Delta^m_{\mathcal{Q}_t}\right]$$

$$\overset{(vii)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left(\Delta^m_{\mathcal{Q}_t}\right)^{1/2} \left[\Delta_{\mathcal{P}_t}^{1/2} \left(\Delta^m_{\mathcal{Q}_t}\right)^{-\beta^{(j)}} + \left(\Delta^m_{\mathcal{Q}_t}\right)^{1-\beta^{(j)}} + \left(\Delta^m_{\mathcal{Q}_t}\right)^{1-2\beta^{(j)}} + \Delta^m_{\mathcal{Q}_t}\right]$$

$$\overset{(viii)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left[\Delta_{\mathcal{P}_t}^{1/2} \Delta_{\mathcal{Q}_t}^{1/2-\beta^{(j)}} + \Delta_{\mathcal{Q}_t}^{3/2-2\beta^{(j)}}\right]$$

$$\overset{(ix)}{\lesssim} \sum_{j=1}^{d} \left[(t^{-1} N'_t \Delta_{\mathcal{Q}_t})(t \Delta_{\mathcal{Q}_t}^{-2} \Delta_{\mathcal{P}_t})^{1/2} \Delta_{\mathcal{Q}_t}^{1/2-\beta^{(j)}} + (t^{-1} N'_t \Delta_{\mathcal{Q}_t}) \left(t \Delta_{\mathcal{Q}_t}^{1-4\beta^{(j)}}\right)^{1/2}\right] \to 0, \quad t \to \infty,$$

where in (i) we use the triangle inequality, in (ii) $|\hat{y}| > \nu \implies |\hat{y} - y| > \nu/2 \vee |y| > \nu/2$ for $y, \hat{y}, \nu \in \mathbb{R}$ and $\mathbb{1}_A \leq \mathbb{1}_{B \cup C} \leq \mathbb{1}_B + \mathbb{1}_C$ for sets $A, B, C \subset \Omega$ such that $A \subset B \cup C$, in (iii) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$ and $\mathbb{1}_{A \cap B} \leq \mathbb{1}_A$ for sets $A, B \subset \Omega$, in (iv) $|y + m| > \nu/2 \implies |y| > \nu/4 \vee |m| > \nu/4$ for $y, m, \nu \in \mathbb{R}$ $\mathbb{1}_A \leq \mathbb{1}_{B \cup C} \leq \mathbb{1}_B + \mathbb{1}_C$ for sets $A, B, C \subset \Omega$ such that $A \subset B \cup C$, in (v) Cauchy-Schwarz inequality and triangle inequality in the first two terms, independence of $D^l \mathbf{Y}_{u_m}, \Delta^m_{\mathcal{Q}_t} D^{p-1} \mathbf{Y}^c_{\mathbf{A}^{(0)}}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{M}, \Delta^m_{\mathcal{Q}_t} \mathbf{J}$ in the last two terms, in (vi) independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{W}$ on the first term, Jensen's inequality on the second term, Markov's inequality and Equation (3.D.7) on the third term, a slight modification of Equation (3.D.16) on the fourth term, Markov inequality and Equation (3.D.17) on the fifth term, $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \text{Poisson}(\lambda^{(j)} \Delta^m_{\mathcal{Q}_t})$ on the sixth term, in (vii) boundedness of fourth moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1} \mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$, standard deviation of Brownian increments and Assumption 3.4.20.(i), in $(viii)$ $\Delta^m_{\mathcal{Q}_t} \leq \Delta_{\mathcal{Q}_t}$, in (ix) Assumption 3.4.10, Assumption 3.4.12 and Assumption 3.4.20.(ii).

To show $\mathbf{S}_t^{2,2,2}(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$ we note

$$\mathbb{P}_{\mathbf{A}^*}\left(\|\mathbf{S}_t^{2,2,2}(i,l)\| > 0\right)$$

$$\overset{(i)}{\leq} \mathbb{P}_{\mathbf{A}^*}\left(\bigcup_{m=0}^{M_t-1} \bigcup_{j=1}^{d} \left\{\Delta^m_{\mathcal{Q}_t} J^{(j)} \neq 0, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1} \tilde{Y}^{(j)}\right| \leq 2\nu_t^{(j),m}\right\}\right)$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} J^{(j)} \neq 0, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| \leq 2\nu_t^{(j),m}\right)\\
&\overset{(iii)}{\leq} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \left[\mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} = 1, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{(j)}\right| \leq 2\nu_t^{(j),m}\right) + \mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 1\right)\right]\\
&\overset{(iv)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \left[\mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} = 1, \left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{c,(j)}_{\mathbf{A}^{(0)}} + \Delta^m_{\mathcal{Q}_t} J^{(j)}\right| \leq 2\nu_t^{(j),m}\right) + (\Delta^m_{\mathcal{Q}_t})^2\right]\\
&\overset{(v)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \left[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{c,(j)}_{\mathbf{A}^{(0)}} + \Delta^m_{\mathcal{Q}_t} J^{(j)}\right| \leq 2\nu_t^{(j),m}, \left|\Delta^m_{\mathcal{Q}_t} J^{(j)}\right| \geq 4\nu_t^{(j),m}\right)\right.\\
&\qquad\qquad \left. + \mathbb{P}_{\mathbf{A}^*}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} = 1, \left|\Delta^m_{\mathcal{Q}_t} J^{(j)}\right| < 4\nu_t^{(j),m}\right) + (\Delta^m_{\mathcal{Q}_t})^2\right]\\
&\overset{(vi)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \left[\mathbb{P}_{\mathbf{A}^*}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}\tilde{Y}^{c,(j)}_{\mathbf{A}^{(0)}}\right| > 2\nu_t^{(j),m}\right) + \Delta^m_{\mathcal{Q}_t}\tilde{F}^{(j)}\left(\left(-4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}, 4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}\right)\right) + (\Delta^m_{\mathcal{Q}_t})^2\right]\\
&\overset{(vii)}{\lesssim} \sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \left[(\Delta^m_{\mathcal{Q}_t})^{2-2\beta^{(j)}} + \Delta^m_{\mathcal{Q}_t}\tilde{F}^{(j)}\left(\left(-4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}, 4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}\right)\right) + (\Delta^m_{\mathcal{Q}_t})^2\right]\\
&\overset{(viii)}{\lesssim} \sum_{j=1}^{d} \left[t\Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t} + t\tilde{F}^{(j)}\left(\left(-4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}, 4\Delta^{\beta^{(j)}}_{\mathcal{Q}_t}\right)\right)\right] \to 0, \quad t\to\infty,
\end{aligned}
$$

where we use in (i) the fact that if the norm is non-zero then at least one component of one of the summands is non-zero and thus the corresponding indicator is non-zero, in (ii) sub-additivity of probability, in (iii) theorem of total probability, in (iv) $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \text{Poisson}(\lambda^{(j)}\Delta^m_{\mathcal{Q}_t})$, in (v) theorem of total probability, in (vi) $|y+j|<2\nu, |j|\geq 4\nu \implies |y|>2\nu$ for $y, j, \nu \in \mathbb{R}$ on the first term, $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \text{Poisson}(\lambda^{(j)}\Delta^m_{\mathcal{Q}_t})$ and $\Delta^m_{\mathcal{Q}_t}\tilde{J}^{(i)} \,|\, \Delta^m_{\mathcal{Q}_t} N^{(i)} = 1 \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \tilde{F}^{(i)}(\cdot)$ on the second term, in (vii) Equation (3.D.16), in $(viii)$ Assumption 3.4.20.(ii) and Assumption 3.4.20.(iii).

It remains to show $\mathbf{S}^3_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t\to\infty$. We work in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\left\|\mathbf{S}^3_t(i,l)\right\|\right]\\
&\quad = t^{-1/2}\mathbb{E}_{\mathbf{A}^*}\left[\left\|\sum_{m=0}^{M_t-1} D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t}\mathbf{M}\odot \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}\mathbf{Y}-\mathbf{b}\Delta^m_{\mathcal{Q}_t}\right|\leq \boldsymbol{\nu}^m_t\right\}}\right\|\right]\\
&\quad \overset{(i)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t}\right|\leq \nu_t^{(j),m}\right\}}\right]\\
&\quad \overset{(ii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| \left(\mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{(j)}\right|>\nu_t^{(j),m}/2\right\}}\right.\right.\\
&\qquad\qquad \left.\left. + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{(j)} - b^{(j)}\Delta^m_{\mathcal{Q}_t}\right|\leq 3\nu_t^{(j),m}/2\right\}}\right)\right]\\
&\quad \overset{(iii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| \left(\mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t}\hat{D}^{p-1}Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{(j)}\right|>\nu_t^{(j),m}/2\right\}}\right.\right.\\
&\qquad\qquad \left.\left. + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} D^{p-1}Y^{c,(j)}_{\mathbf{A}^{(0)}} + \Delta^m_{\mathcal{Q}_t} M^{(j)}\right|\leq 3\nu_t^{(j),m}/2\right\}} + \mathbb{1}_{\left\{\Delta^m_{\mathcal{Q}_t} N^{(j)}>0\right\}}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(iv)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \Bigg[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \Bigg(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{(j)} \right| > \nu_t^{(j),m}/2 \right\}} \\
&\qquad + \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right\}} + \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \leq 2\nu_t^{(j),m} \right\}} + \mathbb{1}_{\left\{ \Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right\}} \Bigg) \Bigg] \\
&\overset{(v)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \Bigg\{ \mathbb{E}_{\mathbf{A}^*} \left[\left| D^l Y^{(i)}_{u_m} \right|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*} \left[\left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right|^2 \right]^{1/2} \\
&\qquad \times \Bigg[\mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(j)} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{(j)} \right| > \nu_t^{(j),m}/2 \right)^{1/2} \\
&\qquad\quad + \mathbb{P}_{\mathbf{A}^*} \left(\left| \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right)^{1/2} + \mathbb{P}_{\mathbf{A}^*} \left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right) \Bigg] \\
&\qquad + \mathbb{E}_{\mathbf{A}^*} \left[\left| D^l Y^{(i)}_{u_m} \right| \right] \mathbb{E}_{\mathbf{A}^*} \left[\left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \leq 2\nu_t^{(j),m} \right\}} \right] \Bigg\} \\
&\overset{(vi)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left\{ \left(\Delta^m_{\mathcal{Q}_t} \right)^{1/2} \left[\left(\nu_t^{(j),m} \right)^{-1} \Delta^{1/2}_{\mathcal{P}_t} + \left(\Delta^m_{\mathcal{Q}_t} \right)^{1-\beta^{(j)}} + \Delta^m_{\mathcal{Q}_t} \right] + \Delta^{1+\epsilon^{(j)}}_{\mathcal{Q}_t} \right\} \\
&\overset{(vii)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left\{ \left(\Delta^m_{\mathcal{Q}_t} \right)^{1/2-\beta^{(j)}} \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{3/2-\beta^{(j)}}_{\mathcal{Q}_t} + \Delta^{1+\epsilon^{(j)}}_{\mathcal{Q}_t} \right\} \\
&\overset{(viii)}{\lesssim} \sum_{j=1}^{d} (t^{-1} N'_t \Delta_{\mathcal{Q}_t}) \left[\left(t \Delta^{-2}_{\mathcal{Q}_t} \Delta_{\mathcal{P}_t} \right)^{1/2} \Delta^{1/2-\beta^{(j)}}_{\mathcal{Q}_t} + \left(t \Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t} \right)^{1/2} + \left(t \Delta^{2\epsilon^{(j)}}_{\mathcal{Q}_t} \right)^{1/2} \right] \to 0, \quad t \to \infty,
\end{aligned}
$$

where we use in (i) triangle inequality, in (ii) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$, $|\hat{y}| \leq \nu \wedge |\hat{y} - y| \leq \nu/2 \implies |y| \leq 3\nu/2$, $\mathbb{1}_A \leq \mathbb{1}_B$ and $\mathbb{1}_{A \cap C} \leq \mathbb{1}_C$ for $y, \hat{y}, \nu \in \mathbb{R}$ and sets $A, B, C \subset \Omega$ with $A \subset B$, in (iii) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$ and $\mathbb{1}_{A \cap B} \leq \mathbb{1}_A$ for sets $A, B \subset \Omega$, in (iv) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$, $|y + m| \leq 3\nu/2 \wedge |m| > 2\nu \implies |y| > \nu/2$, $\mathbb{1}_A \leq \mathbb{1}_B$ and $\mathbb{1}_{A \cap C} \leq \mathbb{1}_C$ for $y, m, \nu \in \mathbb{R}$ and sets $A, B, C \subset \Omega$ with $A \subset B$, in (v) Cauchy-Schwarz inequality and independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{M}$ on the first two terms, independence of $D^l \mathbf{Y}_{u_m}$, $\Delta^m_{\mathcal{Q}_t} \mathbf{M}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{N}$ on the third term, independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{M}$ on the fourth term, in (vi) boundedness of second moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1} \mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$, Equation (3.D.17), Markov inequality, Equation (3.4.19), Equation (3.D.16), $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\sim} \text{Poisson}(\lambda^{(j)} \Delta^m_{\mathcal{Q}_t})$ and Assumption 3.4.20.(iii), in (vii) Assumption 3.4.20.(i), in $(viii)$ Assumption 3.4.10, Assumption 3.4.12 and Assumption 3.4.20.(ii) and Assumption 3.4.20.(iii). □

Infinite jump activity: the symmetric driver case

In this section we give a slightly different version of Theorem 3.4.23 with Assumption 3.4.20 replaced by the following joint condition on $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$.

Assumption 3.D.2. The sequence of partitions $\{\mathcal{Q}_t,\ t \in \mathcal{T}\}$ and the sequence of jump thresholds $\{(\boldsymbol{\nu}_t^m)_{m=0}^{M_t-1},\ t \in \mathcal{T}\}$ satisfy for $i \in \{1, \ldots, d\}$

(i) $\nu_t^{(i),m} = (\Delta^m_{\mathcal{Q}_t})^{\beta^{(i)}}$ with $\beta^{(i)} \in (0,\ 1/4)$;

(ii) $t \Delta^{1-4\beta^{(i)}}_{\mathcal{Q}_t} \to 0,\ t \to \infty$;

(iii) $t\tilde{F}^{(i)}\left(\left(-4\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}, 4\Delta_{\mathcal{Q}_t}^{\beta^{(i)}}\right)\right) \to 0,\ t\to\infty$ and $\exists \epsilon^{(i)} > 2\beta^{(i)}$ such that $t\Delta_{\mathcal{Q}_t}^{\epsilon^{(i)}-2\beta^{(i)}} \to 0,\ t\to\infty$ and

$$\mathbb{E}_{\mathbf{A}^*}\left[|M_s^{(i)}|^2 \mathbb{1}_{\left\{|M_s^{(i)}|\leq s^{\beta^{(i)}}\right\}}\right] = \mathcal{O}\left(s^{1+\epsilon^{(i)}}\right) \text{ as } s\downarrow 0;$$

(iv) $\exists \epsilon_0 > 0, t_0 \in \mathcal{T}$ s.t. $\forall \epsilon \leq \epsilon_0, t \geq t_0$,

$$\mathbb{E}_{\mathbf{A}^*}\left[\Delta_{\mathcal{Q}_t}^m M^{(i)} \mathbb{1}_{\{|\Delta_{\mathcal{Q}_t}^m M^{(i)}|\leq\epsilon\}}\right] = 0 \quad \text{for } i\in\{1,\dots,d\},\ m\in\{0,\dots,M_t-1\}.$$

Remark 3.D.3. Assumption 3.D.2.(*iv*) was introduced in Mai (2014) and is a symmetry condition on $\mathbb{M}$. In Mai (2014, Lemma 4.4) shows that a sufficient condition for Assumption 3.D.2.(*iv*) is that $F|_{[-1,1]}$ is symmetric. We note that introducing this condition allows us to modify Assumption 3.4.20.(*iii*): we now control the squared norm of small increments of $\mathbb{M}$, a "weaker" condition.

Theorem 3.D.4. *Theorem 3.4.23 holds ceteris paribus with Assumption 3.4.20 replaced by Assumption 3.D.2.*

Proof of Theorem 3.D.4. The proof proceeds exactly as in the proof of Theorem 3.4.23, cf. Appendix 3.D.5 for details, with the only difference being in the arguments used to show $\mathbf{S}_t^3(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t\to\infty$. In this setting, we rewrite $\mathbf{S}_t^3(i,l)$ as

$$\begin{aligned}
\mathbf{S}_t^3(i,l) &= t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m \mathbf{M} \odot \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta_{\mathcal{Q}_t}^m\right| \leq \boldsymbol{\nu}_t^m\right\}}\right) \\
&= t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m \mathbf{M} \odot \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta_{\mathcal{Q}_t}^m\right| \leq \boldsymbol{\nu}_t^m, |\Delta_{\mathcal{Q}_t}^m \mathbf{M}| > 2\boldsymbol{\nu}_t^m\right\}}\right) \\
&\quad + t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m \mathbf{M} \odot \left(\mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}\mathbf{Y} - \mathbf{b}\Delta_{\mathcal{Q}_t}^m\right| \leq \boldsymbol{\nu}_t^m, |\Delta_{\mathcal{Q}_t}^m \mathbf{M}| \leq 2\boldsymbol{\nu}_t^m\right\}} - \mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^m \mathbf{M}| \leq 2\boldsymbol{\nu}_t^m\right\}}\right)\right) \\
&\quad + t^{-1/2}\left(\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m \mathbf{M} \odot \mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^m \mathbf{M}| \leq 2\boldsymbol{\nu}_t^m\right\}}\right) \\
&=: \mathbf{S}_t^{3,1}(i,l) + \mathbf{S}_t^{3,2}(i,l) + \mathbf{S}_t^{3,3}(i,l),
\end{aligned}$$

For $\mathbf{S}_t^{3,1}(i,l)$ note that in $L^1(\Omega,\mathcal{F},\mathbb{P}_{\mathbf{A}^*})$ we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\left\|\mathbf{S}_t^{3,1}(i,l)\right\|\right] \\
&\overset{(i)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m M^{(j)}\right| \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{(j)} - b^{(j)}\Delta_{\mathcal{Q}_t}^m\right| \leq \nu_t^{(j),m}, |\Delta_{\mathcal{Q}_t}^m M^{(j)}| > 2\nu_t^{(j),m}\right\}}\right] \\
&\overset{(ii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m M^{(j)}\right| \mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1}Y^{c,(j)} + \Delta_{\mathcal{Q}_t}^m J^{(j)}\right| > \nu_t^{(j),m}\right\}}\right] \\
&\overset{(iii)}{\leq} t^{-1/2}\sum_{m=0}^{M_t-1}\sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)} \Delta_{\mathcal{Q}_t}^m M^{(j)}\right| \left(\mathbb{1}_{\left\{\left|\Delta_{\mathcal{Q}_t}^m \hat{D}^{p-1} Y_{\mathbf{A}^{(0)}}^{c,(j)}\right| > \nu_t^{(j),m}\right\}} + \mathbb{1}_{\left\{\Delta_{\mathcal{Q}_t}^m N^{(j)} > 0\right\}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \Bigg[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \Bigg(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right\}} \\
&\qquad\qquad + \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right\}} + \mathbb{1}_{\left\{ \Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right\}} \Bigg) \Bigg] \\
&\overset{(v)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \left[\left| D^l Y^{(i)}_{u_m} \right|^2 \right]^{1/2} \mathbb{E}_{\mathbf{A}^*} \left[\left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right|^2 \right]^{1/2} \\
&\qquad \times \Bigg[\mathbb{P} \left(\left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right)^{1/2} \\
&\qquad\qquad + \mathbb{P} \left(\left| \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right)^{1/2} + \mathbb{P} \left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right) \Bigg] \\
&\overset{(vi)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left(\Delta^m_{\mathcal{Q}_t} \right)^{1/2} \left[\left(\nu_t^{(j),m} \right)^{-1} \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{1-\beta^{(j)}}_{\mathcal{Q}_t} + \Delta_{\mathcal{Q}_t} \right] \\
&\overset{(vii)}{\lesssim} \sum_{j=1}^{d} (t^{-1} N'_t \Delta_{\mathcal{Q}_t}) \left[\left(t \Delta_{\mathcal{P}_t} \Delta^{-2}_{\mathcal{Q}_t} \right)^{1/2} \Delta^{1/2-\beta^{(j)}}_{\mathcal{Q}_t} + \left(t \Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t} \right)^{1/2} \right] \to 0, \quad t \to \infty,
\end{aligned}
$$

where we use in (i) triangle inequality, in (ii) $|\hat{y}| \leq \nu, |m| > 2\nu \implies |\hat{y} - m| > \nu$ for $\hat{y}, m, \nu \in \mathbb{R}$ and $\mathbb{1}_A \leq \mathbb{1}_B$ for sets $A \subset B \subset \Omega$, in (iii) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$ and $\mathbb{1}_{A \cap B} \leq \mathbb{1}_A$ for sets $A, B \subset \Omega$, in (iv) $|\hat{y}| > \nu \implies |y| > \nu/2 \vee |y - \hat{y}| > \nu/2$ for $y, \hat{y}, \nu \in \mathbb{R}$ and $\mathbb{1}_A \leq \mathbb{1}_B + \mathbb{1}_C$ for sets $A, B, C \subset \Omega$ such that $A \subset B \cup C$, in (v) Cauchy-Schwarz inequality and independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{M}$ on the first two terms, independence of $D^l \mathbf{Y}_{u_m}$, $\Delta^m_{\mathcal{Q}_t} \mathbf{M}$ and $\Delta^m_{\mathcal{Q}_t} \mathbf{N}$ on the third term, in (vi) boundedness of second moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1} \mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$, Equation (3.D.17), Markov inequality, Equation (3.4.19), Equation (3.D.16), $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\approx} \text{Poisson}(\lambda^{(j)} \Delta^m_{\mathcal{Q}_t})$, in (vii) Assumption 3.4.10, Assumption 3.D.2.(i) and $\Delta^m_{\mathcal{Q}_t} \leq \Delta_{\mathcal{Q}_t}$, Assumption 3.4.12 and Assumption 3.D.2.(ii).

Next, we consider $\mathbf{S}^{3,2}_t(i,l)$. Again we work in $L^1(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}^*})$.

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*} \left[\left\| \mathbf{S}^{3,2}_t(i,l) \right\| \right] \\
&\overset{(i)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \left[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(j)} - b^{(j)} \Delta^m_{\mathcal{Q}_t} \right| > \nu_t^{(j),m}, |\Delta^m_{\mathcal{Q}_t} M^{(j)}| \leq 2\nu_t^{(j),m} \right\}} \right] \\
&\overset{(ii)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \Bigg[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \Bigg(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} + \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| > \nu_t^{(j),m}, |\Delta^m_{\mathcal{Q}_t} M^{(j)}| \leq 2\nu_t^{(j),m} \right\}} \\
&\qquad\qquad + \mathbb{1}_{\left\{ \Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right\}} \Bigg) \Bigg] \\
&\overset{(iii)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \Bigg[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \Bigg(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/2 \right\}} \\
&\qquad\qquad + \mathbb{1}_{\left\{ \nu_t^{(j),m}/2 < \left| \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| < 2\nu_t^{(j),m} \right\}} + \mathbb{1}_{\left\{ \Delta^m_{\mathcal{Q}_t} N^{(j)} > 0 \right\}} \Bigg) \Bigg] \\
&\overset{(iv)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*} \Bigg[\left| D^l Y^{(i)}_{u_m} \Delta^m_{\mathcal{Q}_t} M^{(j)} \right| \Bigg(\mathbb{1}_{\left\{ \left| \Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} \right| > \nu_t^{(j),m}/4 \right\}}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(j),m}/4\right\}} + \mathbb{1}_{\left\{\nu_t^{(j),m}/2 < \left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| < 2\nu_t^{(j),m}\right\}} + \mathbb{1}_{\left\{\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right\}}\Bigg)\Bigg] \\
&\overset{(v)}{\leq} t^{-1/2} \sum_{m=0}^{M_t-1} \Bigg\{ \sum_{j=1}^{d} \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right|^2\right]^{1/2} \\
&\qquad \times \Bigg[\mathbb{P}\left(\left|\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}} - \Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(j),m}/4\right)^{1/2} \\
&\qquad\qquad + \mathbb{P}\left(\left|\Delta^m_{\mathcal{Q}_t} D^{p-1} Y^{c,(j)}_{\mathbf{A}^{(0)}}\right| > \nu_t^{(j),m}/4\right)^{1/2} + \mathbb{P}\left(\Delta^m_{\mathcal{Q}_t} N^{(j)} > 0\right) \Bigg] \\
&\qquad + \mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y^{(i)}_{u_m}\right|\right] \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| \mathbb{1}_{\left\{\nu_t^{(j),m}/2 < \left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| < 2\nu_t^{(j),m}\right\}}\right] \Bigg\} \\
&\overset{(vi)}{\lesssim} \sum_{j=1}^{d} \left\{ t^{-1/2} M_t \left(\Delta^m_{\mathcal{Q}_t}\right)^{1/2} \left[\left(\nu_t^{(j),m}\right)^{-1} \Delta^{1/2}_{\mathcal{P}_t} + \left(\Delta^m_{\mathcal{Q}_t}\right)^{1-\beta^{(j)}} + \Delta^m_{\mathcal{Q}_t}\right] + t^{-1/2} M_t \Delta^{1-\beta^{(j)}+\epsilon^{(j)}/2}_{\mathcal{Q}_t} \right\} \\
&\overset{(vii)}{\lesssim} \sum_{j=1}^{d} t^{-1/2} M_t \left\{ \left(\Delta^m_{\mathcal{Q}_t}\right)^{1/2-\beta^{(j)}} \Delta^{1/2}_{\mathcal{P}_t} + \Delta^{3/2-\beta^{(j)}}_{\mathcal{Q}_t} + \Delta^{1-\beta^{(j)}+\epsilon^{(j)}/2}_{\mathcal{Q}_t} \right\} \\
&\overset{(viii)}{\lesssim} \sum_{j=1}^{d} (t^{-1} N'_t \Delta_{\mathcal{Q}_t}) \left[\left(t\Delta^{-2}_{\mathcal{Q}_t} \Delta_{\mathcal{P}_t}\right)^{1/2} \Delta^{1/2-\beta^{(j)}}_{\mathcal{Q}_t} + \left(t\Delta^{1-2\beta^{(j)}}_{\mathcal{Q}_t}\right)^{1/2} + \left(t\Delta^{\epsilon^{(j)}-2\beta^{(j)}}_{\mathcal{Q}_t}\right)^{1/2} \right] \to 0, \quad t \to \infty,
\end{aligned}
$$

where we use in (i) triangle inequality, in (ii) $1 = \mathbb{1}_A + \mathbb{1}_{A^c}$ and $\mathbb{1}_{A\cap B} \leq \mathbb{1}_A$ for sets $A, B \subset \Omega$, in (iii) $|\hat{y}+m| > \nu \implies |\hat{y}| > \nu/2 \vee |m| > \nu/2$ for $\hat{y}, m, \nu, \in \mathbb{R}$ and $\mathbb{1}_A \leq \mathbb{1}_B + \mathbb{1}_C$ for sets $A, B, C \subset \Omega$ such that $A \subset B \cup C$, in (iv) $|\hat{y}| > \nu/2 \implies |y| > \nu/4 \vee |y - \hat{y}| > \nu/4$ for $y, \hat{y}, \nu \in \mathbb{R}$ and $\mathbb{1}_A \leq \mathbb{1}_B + \mathbb{1}_C$ for sets $A, B, C \subset \Omega$ such that $A \subset B \cup C$, in (v) Cauchy-Schwarz inequality and independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t}\mathbf{M}$ on the first two terms, independence of $D^l \mathbf{Y}_{u_m}$, $\Delta^m_{\mathcal{Q}_t}\mathbf{M}$ and $\Delta^m_{\mathcal{Q}_t}\mathbf{N}$ on the third term, independence of $D^l \mathbf{Y}_{u_m}$ and $\Delta^m_{\mathcal{Q}_t}\mathbf{M}$ on the fourth term, in (vi) boundedness of second moments of $\mathbf{X}_t = (\mathbf{Y}_t^{\mathrm{T}}, \ldots, D^{p-1}\mathbf{Y}_t^{\mathrm{T}})^{\mathrm{T}}$, Equation (3.D.17), Markov's inequality, Equation (3.4.19), Equation (3.D.16), $\Delta^m_{\mathcal{Q}_t} N^{(j)} \overset{\mathbb{P}_{\mathbf{A}^*}}{\sim} \text{Poisson}(\lambda^{(j)} \Delta^m_{\mathcal{Q}_t})$ and

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| \mathbb{1}_{\left\{\nu_t^{(j),m}/2 < \left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| < 2\nu_t^{(j),m}\right\}}\right] \\
&\leq \mathbb{P}_{\mathbf{A}^*}\left(\nu_t^{(j),m}/2 < \left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right|\right)^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right|^2 \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| < 2\nu_t^{(j),m}\right\}}\right]^{1/2} \\
&\leq \left(\nu_t^{(j),m}\right)^{-1} \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right|^2\right]^{1/2} \mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right|^2 \mathbb{1}_{\left\{\left|\Delta^m_{\mathcal{Q}_t} M^{(j)}\right| < 2\nu_t^{(j),m}\right\}}\right]^{1/2} \\
&\leq (\Delta^m_{\mathcal{Q}_t})^{-\beta^{(j)}} (\Delta^m_{\mathcal{Q}_t})^{1/2} (\Delta^m_{\mathcal{Q}_t})^{1/2+\epsilon^{(j)}/2} = (\Delta^m_{\mathcal{Q}_t})^{1-\beta^{(j)}+\epsilon^{(j)}/2},
\end{aligned} \tag{3.D.18}
$$

in (vii) Assumption 3.D.2.(i), in $(viii)$ Assumption 3.4.10, Assumption 3.4.12 and Assumption 3.D.2.(iii).

Finally we show $\mathbf{S}^{3,3}_t(i,l) \overset{\mathbb{P}_{\mathbf{A}^*}}{\to} 0$ as $t \to \infty$. In this case we work in $L^2(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{A}*})$ and note that,

for t sufficiently large, $\mathbf{S}_t^{3,3}(i,l)$ is mean-zero by Assumption 3.D.2.(iv):

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}^*}\left[\left\|\mathbf{S}_t^{3,3}(i,l)\right\|^2\right] \\
&\quad= t^{-1}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{A}^*}\left[\left|\sum_{m=0}^{M_t-1} D^l Y_{u_m}^{(i)}\Delta_{\mathcal{Q}_t}^m M^{(j)}\mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^m M^{(j)}|\le 2\nu_t^{(j),m}\right\}}\right|^2\right] \\
&\quad= t^{-1}\sum_{j=1}^{d}\sum_{m=0}^{M_t-1}\sum_{m'=0}^{M_t-1}\mathbb{E}_{\mathbf{A}^*}\left[D^l Y_{u_m}^{(i)} D^l Y_{u_{m'}}^{(i)}\Delta_{\mathcal{Q}_t}^m M^{(j)}\Delta_{\mathcal{Q}_t}^{m'} M^{(j)}\mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^m M^{(j)}|\le 2\nu_t^{(j),m}\right\}}\mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^{m'} M^{(j)}|\le 2\nu_t^{(j),m}\right\}}\right] \\
&\quad= t^{-1}\sum_{j=1}^{d}\sum_{m=0}^{M_t-1}\mathbb{E}_{\mathbf{A}^*}\left[\left|D^l Y_{u_m}^{(i)}\right|^2\right]\mathbb{E}_{\mathbf{A}^*}\left[\left|\Delta_{\mathcal{Q}_t}^m M^{(j)}\right|^2\mathbb{1}_{\left\{|\Delta_{\mathcal{Q}_t}^m M^{(j)}|\le 2\nu_t^{(j),m}\right\}}\right] \\
&\quad\lesssim \sum_{j=1}^{d} t^{-1}\sum_{m=0}^{M_t-1}(\Delta_{\mathcal{Q}_t}^m)^{1+\epsilon^{(j)}} \lesssim \sum_{j=1}^{d}\Delta_{\mathcal{Q}_t}^{\epsilon^{(j)}} \to 0, \quad t\to\infty,
\end{aligned}$$

using independence of $\Delta_{\mathcal{Q}_t}^m M^{(j)}$, $\Delta_{\mathcal{Q}_t}^{m'} M^{(j)}$, $D^l Y_{u_m}^{(i)}$, $D^l Y_{u_{m'}}^{(i)}$ for $m \neq m'$ and Assumption 3.D.2.(iv) to show that off-diagonal elements are zero and Assumption 3.D.2.(iii). □

3.E GrCAR: proofs

3.E.1 Proof of Theorem 3.5.3

Proof of Theorem 3.5.3. We proceed exactly as in Section 3.3 replacing the dependencies on the parameter $\mathbf{A} \in (\mathcal{M}(\mathbb{R}^d))^p$ with dependencies on $\theta \in \mathbb{R}^{p\times 2}$. Expanding the form of the likelihood in Proposition 3.3.3 and grouping in terms of $\theta^{(i,j)}$ for $i = 1,\dots,p$ and $j = 1,2$ we can rewrite the likelihood of the GrCAR(p) process as

$$\mathcal{L}(\theta; \mathbb{Y}_{[0,t]}) = \exp\left\{\mathrm{vec}(\theta)^{\mathrm{T}}\mathbf{K}_t - \frac{1}{2}\mathrm{vec}(\theta)^{\mathrm{T}}[\mathbf{K}]_t\mathrm{vec}(\theta)\right\}, \quad \theta \in \mathbb{R}^{p\times 2},$$

with $\mathbb{K} = \{\mathbf{K}_t,\ t \ge 0\}$ as in the statement of Theorem 3.5.3. One can then proceed as in Lemma 3.3.8 to check that for any $\theta \in \mathbb{R}^{p\times 2}$ the score

$$\nabla_{\mathrm{vec}(\theta)} \log \mathcal{L}(\theta; \mathbb{Y}_{[0,t]}) = \mathbf{K}_t - [\mathbf{K}]_t\mathrm{vec}(\theta),$$

defines a continuous square-integrable martingale under $\mathbb{P}_{\theta,\mathbf{x}_0}$. This is ensured by showing the quadratic variation is integrable, i.e.

$$\begin{aligned}
\mathbb{E}_{\theta,\mathbf{x}_0}\left[\|[\mathbf{K}]_t\|\right] &\le (\|I\| + \|\bar{A}^{\mathrm{T}}\|)\mathbb{E}_{\theta,\mathbf{x}_0}\left[\int_0^t \sum_{i=0}^{p-1}\|D^i\mathbf{Y}_s\|_\Sigma^2\, ds\right] \\
&\le (\|I\| + \|\bar{A}^{\mathrm{T}}\|)\|\Sigma^{-1}\|\int_0^t \mathbb{E}_{\theta,\mathbf{x}_0}[\|\mathbf{X}_s\|^2]\, ds < \infty.
\end{aligned}$$

Finally, as in Theorem 3.3.9, we check that $\forall t > 0$ the matrix $[\mathbf{K}]_t$ is strictly positive definite $\mathbb{P}_{\theta^*,\mathbf{x}_0}$-a.s. By carrying out the full computation one can show this boils down to proving that

$$\{D^l Y_s^{(i)},\ i = 1,\dots,d, l = 0,\dots,p-1\},$$

is an affinely independent collection of random variables for any $s \in (0,t]$ under $\mathbb{P}_{\theta^{(0)},\mathbf{x}_0}$, which follows by the same arguments as before. The rest of the results carry over ceteris paribus for the MLE $\hat{\theta}(\mathbb{Y}_{[0,t]})$ with

$$\mathcal{K}_\infty = \begin{pmatrix} \mathbb{E}_{\theta^*}[\langle D^{p-1}\mathbf{Y}_\infty, D^{p-1}\mathbf{Y}_\infty\rangle_\Sigma] & \cdots & \mathbb{E}_{\theta^*}[\langle D^{p-1}\mathbf{Y}_\infty, \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty\rangle_\Sigma] \\ \mathbb{E}_{\theta^*}[\langle \bar{A}^{\mathrm{T}} D^{p-1}\mathbf{Y}_\infty, D^{p-1}\mathbf{Y}_\infty\rangle_\Sigma] & \cdots & \mathbb{E}_{\theta^*}[\langle \bar{A}^{\mathrm{T}} D^{p-1}\mathbf{Y}_\infty, \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty\rangle_\Sigma] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{\theta^*}[\langle \mathbf{Y}_\infty, D^{p-1}\mathbf{Y}_\infty\rangle_\Sigma] & \cdots & \mathbb{E}_{\theta^*}[\langle \mathbf{Y}_\infty, \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty\rangle_\Sigma] \\ \mathbb{E}_{\theta^*}[\langle \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty, D^{p-1}\mathbf{Y}_\infty\rangle_\Sigma] & \cdots & \mathbb{E}_{\theta^*}[\langle \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty, \bar{A}^{\mathrm{T}}\mathbf{Y}_\infty\rangle_\Sigma] \end{pmatrix}. \tag{3.E.1}$$

The inference results for the discretized estimator follow by the same subsequent approximations as in Section 3.4 (and the corresponding proofs in Appendix 3.D), noting that one can simply factor out the $\bar{A}^{\mathrm{T}}$ term. □

3.F Finite difference approximations of derivatives

For $k \in \{1,\dots,p-1\}$ let $D^k\mathbb{Y}_{[0,t]} = \{D^k\mathbf{Y}_s,\ s \in [0,t]\}$ denote the k-th time derivative of the process $\mathbb{Y}_{[0,t]} = \{\mathbf{Y}_s,\ s \in [0,t]\}$. We define the forward finite difference approximation of the derivatives over the partition

$$\mathcal{P}_t = \{0 = s_0 < s_1 < \ldots < s_{N_t} = t\},$$

iteratively by

$$\hat{D}^k\mathbb{Y}_{\mathcal{P}_t} := \{\hat{D}^k\mathbf{Y}_{s_n},\ n = 0,\dots,N_t - k\} \text{ s.t. } \hat{D}^j\mathbf{Y}_{s_n} = \frac{\hat{D}^{k-1}\mathbf{Y}_{s_{n+1}} - \hat{D}^{k-1}\mathbf{Y}_{s_n}}{s_{n+1} - s_n},$$

for $j = k,\dots,p-1$ and $\hat{D}^0\mathbb{Y}_{\mathcal{P}_t} = \mathbb{Y}_{\mathcal{P}_t}$. We make use of the following results.

Lemma 3.F.1. *Let $k \in \{1,\dots,p-1\}$. The forward differences can be written as follows for $n \in \{0,\dots,N_t - k\}$*

$$\hat{D}^k\mathbf{Y}_{s_n} = \sum_{i=1}^{k}(-1)^{k+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=k}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i}(\mathbf{Y}_{s_{n+i}} - \mathbf{Y}_{s_{n+i-1}}).$$

Proof of Lemma 3.F.1. For $m = 1, \dots, k$,

$$\hat{D}^k \mathbf{Y}_{s_n} = \sum_{i=1}^{m} (-1)^{m+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=m}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} (\hat{D}^{k-m} \mathbf{Y}_{s_{n+i}} - \hat{D}^{k-m} \mathbf{Y}_{s_{n+i-1}}).$$

This can be proved inductively:

- for the case $m = 1$ this is the iterative definition of the forward finite difference;
- given the case $m < k$

$$\begin{aligned}
\hat{D}^k \mathbf{Y}_{s_n} &= \sum_{i=1}^{m} (-1)^{m+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=m}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} \\
&\qquad \times (\hat{D}^{k-m} \mathbf{Y}_{s_{n+i}} - \hat{D}^{k-m} \mathbf{Y}_{s_{n+i-1}}) \\
&= \sum_{i=1}^{m} (-1)^{m+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=m}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} \\
&\qquad \times \Big[(s_{n+i+1} - s_{n+i})^{-1} (\hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i+1}} - \hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i}}) \\
&\qquad\quad - (s_{n+i} - s_{n+i-1})^{-1} (\hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i}} - \hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i-1}}) \Big] \\
&= \sum_{i=1}^{m+1} (-1)^{m+1+i} \left(\sum_{\substack{n_1,\dots,n_{i-1} \geq 1, n_i = 1 \\ n_1+\cdots+n_i=m+1}} + \sum_{\substack{n_1,\dots,n_{i-1} \geq 1, n_i \geq 2 \\ n_1+\cdots+n_i=m+1}} \right) (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} \\
&\qquad \times (\hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i}} - \hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i-1}}) \\
&= \sum_{i=1}^{m+1} (-1)^{m+1+i} \sum_{\substack{n_1,\dots,n_i \geq 1 \\ n_1+\cdots+n_i=m+1}} (s_{n+1} - s_n)^{-n_1} \cdots (s_{n+i} - s_{n+i-1})^{-n_i} \\
&\qquad \times (\hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i}} - \hat{D}^{k-m-1} \mathbf{Y}_{s_{n+i-1}}),
\end{aligned}$$

where we use that

$$\begin{aligned}
\{(n_1, \dots, n_i) : n_1, \dots, n_i \geq 1, n_1 + \cdots + n_i = m + 1\} = \\
\{(n_1, \dots, n_{i-1}, 1) : n_1, \dots, n_{i-1} \geq 1, n_1 + \cdots + n_{i-1} = m\} + \\
\{(n_1, \dots, n_i) : n_1, \dots, n_{i-1} \geq 1, n_i \geq 2, n_1 + \cdots + n_i = m + 1\}.
\end{aligned}$$

□

Lemma 3.F.2. *Let $k \in \{1, \dots, p-1\}$. We can write for $n \in \{0, \dots, N_t - k\}$ and $i \in \{1, \dots, k\}$*

$$\begin{aligned}
\mathbf{Y}_{s_{n+i}} - \mathbf{Y}_{s_{n+i-1}} = \sum_{j=1}^{k-1} D^j \mathbf{Y}_{s_n} \frac{(s_{n+i} - s_n)^j - (s_{n+i-1} - s_n)^j}{j!} \\
+ \int_{s_{n+i-1}}^{s_{n+i}} \int_{s_n}^{u_1} \cdots \int_{s_n}^{u_{k-1}} D^k \mathbf{Y}_{u_k} \, \mathrm{d}u_k \dots \mathrm{d}u_1.
\end{aligned}$$

Proof of Lemma 3.F.2. This follows by writing

$$D^j \mathbf{Y}_s = D^j \mathbf{Y}_{s_n} + \int_{s_n}^{s} D^{j+1} \mathbf{Y}_u \, \mathrm{d}u,$$

and noting that for $j = 1, \ldots, k-1$

$$\int_{s_{n+i-1}}^{s_{n+i}} \int_{s_n}^{u_1} \cdots \int_{s_n}^{u_{j-1}} \mathrm{d}u_j \ldots \mathrm{d}u_1 = \frac{(s_{n+i} - s_n)^j - (s_{n+i-1} - s_n)^j}{j!}.$$

□

Lemma 3.F.3. *The following holds for $k \geq 1$ and $j \in \{0, \ldots, k\}$,*

$$F(j,k) := (-1)^k \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!} \sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} [i^j - (i-1)^j] = \begin{cases} 0, & j = 0, \ldots, k-1, \\ 1, & j = k. \end{cases} \tag{3.F.1}$$

Proof of Lemma 3.F.3. Note that we can write

$$F(j,k) = \frac{\Delta_{\mathcal{P}_t}^{j-k}}{j!} f(j,k)$$

where

$$\begin{aligned}
f(j,k) &= (-1)^k \sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} [i^j - (i-1)^j] \\
&= (-1)^k \left[\sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} i^j - \sum_{i=2}^{k} (-1)^i \frac{(k-1)!}{(i-1)!(k-i)!} (i-1)^j \right] \\
&= (-1)^k \left[\sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} i^j - \sum_{i=2}^{k} (-1)^i \frac{(k-1)!}{(i-2)!(k-i)!} (i-1)^{j-1} \right] \\
&= (-1)^k \left[\sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} i^j + \sum_{i=1}^{k-1} (-1)^i \frac{(k-1)!}{(i-1)!(k-i-1)!} i^{j-1} \right] \\
&= (-1)^k \left[(-1)^k k^j + \sum_{i=1}^{k-1} (-1)^i \left(\frac{(k-1)!}{(i-1)!(k-i)!} i^j + \frac{(k-1)!}{(i-1)!(k-i-1)!} i^{j-1} \right) \right] \\
&= (-1)^k \left[(-1)^k k^j + \sum_{i=1}^{k-1} (-1)^i \frac{(k-1)!}{(i-1)!(k-i-1)!} i^{j-1} \left(\frac{i}{k-i} + 1 \right) \right] \\
&= (-1)^k \left[(-1)^k k^j + \sum_{i=1}^{k-1} (-1)^i \frac{k!}{(i-1)!(k-i)!} i^{j-1} \right] \\
&= (-1)^k k \sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} i^{j-1}.
\end{aligned}$$

We note that $f(j,k)$ satisfies the relationship

$$\begin{aligned}
f(j,k) &= (-1)^k k \sum_{i=1}^{k} (-1)^i \binom{k-1}{i-1} \sum_{m=0}^{j-1} \binom{j-1}{m} (i-1)^m \\
&= (-1)^k k \left\{ \sum_{i=0}^{k-1} (-1)^{i+1} \binom{k-1}{i} + \sum_{m=1}^{j-1} \binom{j-1}{m} \left[\sum_{i=1}^{k} (-1)^i \frac{(k-1)!}{(i-1)!(k-i)!} (i-1)^m \right] \right\} \\
&= (-1)^k k \left\{ (1-1)^{k-1} + \sum_{m=1}^{j-1} \binom{j-1}{m} \left[(k-1) \sum_{i=2}^{k} (-1)^i \frac{(k-2)!}{(i-2)!(k-i)!} (i-1)^{m-1} \right] \right\} \\
&= (-1)^k k \sum_{m=1}^{j-1} \binom{j-1}{m} (k-1) \sum_{i=2}^{k} (-1)^i \binom{k-2}{i-2} (i-1)^{m-1} \\
&= k \sum_{m=1}^{j-1} \binom{j-1}{m} (-1)^{k-1} (k-1) \sum_{i=1}^{k-1} (-1)^i \binom{k-2}{i-1} i^{m-1} \\
&= k \sum_{m=1}^{j-1} \binom{j-1}{m} f(m, k-1).
\end{aligned}$$

Proceeding by induction on $k \geq 1$ one can easily show that

$$f(j,k) = \begin{cases} 0, & j = 0, \dots, k-1, \\ k!, & j = k. \end{cases}$$

□

3.G Simulating MCAR(p) processes

In this section, we will discuss numerical methods for simulating a discrete-time realization $\mathbb{Y}_{\mathcal{P}_t}$ of the MCAR(p) processes $\mathbb{Y}$ defined in Definition 3.2.1. We note that the state-space representation of the MCAR(p) process $\mathbb{X} = \{\mathbf{X}_t,\ t \geq 0\}$ satisfies the Ornstein-Uhlenbeck SDE (3.2.3) with drift matrix $\mathcal{A}_{\mathbf{A}}$ and degenerate Lévy driver $\mathcal{E}\mathbb{L} = \{\mathcal{E}\mathbf{L}_t,\ t \geq 0\}$. The task of simulating a realization of $\mathbb{Y}_{\mathcal{P}_t}$ can thus be tackled by simulating the OU process $\mathbb{X}_{\mathcal{P}_t}$ and then keeping only the first d entries. We first discuss an approximate scheme and then consider some special cases where exact simulation is possible.

3.G.1 An approximate simulation method: Euler-Maruyama scheme

We note SDE (3.2.3) can be approximated over the partition $\mathcal{P}_t := \{0 = t_0 < t_1 < \ldots < t_{N_t} = t\}$ by

$$\mathbf{X}_0^{\mathcal{P}_t} = \mathbf{X}_0, \quad \mathbf{X}_{t_{n+1}}^{\mathcal{P}_t} = \mathbf{X}_{t_n}^{\mathcal{P}_t} + \mathcal{A}_{\mathbf{A}} \mathbf{X}_{t_n}^{\mathcal{P}_t} \Delta_{\mathcal{P}_t}^n + \mathcal{E} \Delta_{\mathcal{P}_t}^n \mathbf{L}, \quad n = 0, \dots, N_t - 1, \tag{3.G.1}$$

where we simulate Lévy increments $\Delta_{\mathcal{P}_t}^n \mathbf{L}$ using an exact simulation method. The Brownian component and the jump component of $\Delta_{\mathcal{P}_t}^n \mathbf{L} = \Delta_{\mathcal{P}_t}^n \mathbf{W}^{\Sigma} + \Delta_{\mathcal{P}_t}^n \mathbf{J}$ can be simulated independently:

- the continuous increment is given by $\Delta^n_{\mathcal{P}_t}\mathbf{W}^\Sigma \sim N(\mathbf{0}, \Delta^n_{\mathcal{P}_t}\Sigma)$;
- the jump component depends on whether:
  - $\mathbb{L}$ has finite jump activity, i.e. $F = \lambda\tilde{F}$ where $\tilde{F}$ is the jump size distribution and λ is the jump rate, in which case we can simulate the number of jumps in $[t_n, t_{n+1}]$ by a Poisson random variable $K_{\Delta^n_{\mathcal{P}_t}} \sim \text{Poisson}(\Delta^n_{\mathcal{P}_t}\lambda)$ and i.i.d. jump sizes from $\tilde{F}$;
  - $\mathbb{L}$ has infinite jump activity, in which case one simulates directly from the prescribed increment distribution, e.g. assuming the d jump components follow independent Gamma processes we can sample each increment as $\Delta^n_{\mathcal{P}_t}J^{(j)} \sim \Gamma(\gamma^j\Delta^n_{\mathcal{P}_t}, \lambda^j)$, $j =, \ldots, d$.

Remark 3.G.1. We can vectorize the operations to speed up computations. For example, if $\mathbf{L} = \mathbf{W}^\Sigma + \tilde{\mathbf{J}}$ has finite jump activity, one can generate

- $\{\Delta^n_{\mathcal{P}_t}\mathbf{W}^\Sigma, n = 0, \ldots, N_t - 1\}$ by simulating N_t i.i.d. $N(\mathbf{0}, \Sigma)$ variables and scaling them by $\sqrt{\Delta^n_{\mathcal{P}_t}}$;
- $\{\Delta^n_{\mathcal{P}_t}\tilde{\mathbf{J}}, n = 0, \ldots, N_t - 1\}$ by
  - simulating $K_t \sim \text{Poisson}(t\lambda)$ the total number of jumps over $[0, t]$,
  - generating the jump times $U^{(1)} \leq \ldots \leq U^{(K_t)}$ as uniform order statistics over $[0, t]$ and the corresponding i.i.d. jump sizes $\Delta\tilde{\mathbf{J}}_1, \ldots, \Delta\tilde{\mathbf{J}}_{K_t} \sim \tilde{F}$,
  - aggregate the jumps over the partition bins $[t_n, t_{n+1}]$ by $\Delta^n_{\mathcal{P}_t}\tilde{\mathbf{J}} = \sum_{k:U^{(k)}\in[t_n,t_{n+1})} \Delta\tilde{\mathbf{J}}_k$.

The strong convergence rate of this Euler-Maruyama scheme is studied in Kühn and Schilling (2019, Theorem 2.1). Kühn and Schilling (2019, Remark 2.3.(i)) covers the case where $\mathbb{L}$ has a non-vanishing (possibly degenerate) diffusion part, as in SDE (3.2.3). For the OU process $\mathbb{X}$, under sufficient regularity conditions for the transition density of $\mathcal{E}\mathbb{L}$, they prove the strong convergence rate is

$$\mathbb{E}_{\mathbf{A}}\left[\sup_{s\in[0,t]} \|\mathbf{X}_s - \mathbf{X}^{\mathcal{P}_t}_s\|^2\right] \lesssim \Delta_{\mathcal{P}_t}.$$

3.G.2 An exact simulation method: the finite jump activity case

When the driving Lévy process has finite jump activity we can simulate the OU process $\mathbb{X}$ exactly. Recall that when $\mathbb{L}$ has finite jump activity we can write the Lévy measure as $F = \lambda\tilde{F}$ where λ is the jump arrival rate and $\tilde{F}$ is the distribution of jump sizes. The process $\mathbb{X}$ satisfies the integral equation

$$\mathbf{X}_t = e^{\mathcal{A}_{\mathbf{A}}(t-s)}\mathbf{X}_s + \int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\mathrm{d}\mathbf{L}_u, \quad t \geq 0. \tag{3.G.2}$$

By plugging-in the Lévy-Itô decomposition of $\mathbb{L}$ when the truncation $\tau(z) \equiv 0$ is used, i.e. $\mathbf{L}_t = \tilde{\mathbf{b}}t + \mathbf{W}^\Sigma_t + \tilde{\mathbf{J}}_t$, we can write

$$\mathbf{X}_t = e^{\mathcal{A}_{\mathbf{A}}(t-s)}\mathbf{X}_s + \int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\tilde{\mathbf{b}}\mathrm{d}u + \int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\mathrm{d}\mathbf{W}^\Sigma_u + \int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}d\tilde{\mathbf{J}}_u,$$

where

- $\int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\tilde{\mathbf{b}}\mathrm{d}u = \left(e^{\mathcal{A}_{\mathbf{A}}(t-s)} - I\right)\mathcal{A}_{\mathbf{A}}^{-1}\mathcal{E}\tilde{\mathbf{b}}$;

- $\int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\mathrm{d}\mathbf{W}_u^{\Sigma} \sim N\left(\mathbf{0}, F(t-s)e^{\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}(t-s)}\right)$, where we use the Itô isometry to compute

$$\begin{aligned}\mathbb{E}_{\mathbf{A}}\left[\left(\int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\mathrm{d}\mathbf{W}_u^{\Sigma}\right)\left(\int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\mathrm{d}\mathbf{W}_u^{\Sigma}\right)^{\mathrm{T}}\right] &= \int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}e^{\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}(t-u)}\mathrm{d}u \\ &= \int_0^{t-s} e^{\mathcal{A}_{\mathbf{A}}(t-s-u)}\mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}}e^{\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}(t-s-u)}\mathrm{d}u \\ &= F(t-s)e^{\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}(t-s)},\end{aligned}$$

  and define

$$M = \begin{pmatrix}\mathcal{A}_{\mathbf{A}} & \mathcal{E}\Sigma\mathcal{E}^{\mathrm{T}} \\ 0 & -\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}\end{pmatrix},$$

  to compute

$$e^{M(t-s)} = \begin{pmatrix}e^{\mathcal{A}_{\mathbf{A}}(t-s)} & F(t-s) \\ 0 & e^{-\mathcal{A}_{\mathbf{A}}^{\mathrm{T}}(t-s)}\end{pmatrix};$$

- $\int_s^t e^{\mathcal{A}_{\mathbf{A}}(t-u)}\mathcal{E}d\tilde{\mathbf{J}}_u = \sum_{\tau_k\in[s,t]} e^{\mathcal{A}_{\mathbf{A}}(t-\tau_k)}\mathcal{E}\Delta\tilde{\mathbf{J}}_{\tau_k}$, where

  - $\{\tau_k,\ k \geq 1\}$ denote the jump times of $\tilde{\mathbb{J}}$, i.e. Poisson arrival times with rate λ, and
  - $\{\Delta\tilde{\mathbf{J}}_{\tau_k},\ k \geq 1\}$ denote the jump sizes of $\tilde{\mathbb{J}}$, i.e. i.i.d. $\tilde{F}$-distributed random variables.

Starting from $\mathbf{X}_0 = \mathbf{x}_0$ we can thus evolve the process $\mathbb{X}$ along $\mathcal{P}_{t'} = \{0 = t_0 < t_1 < \ldots < t_{N_{t'}} = t'\}$ by setting $s = t_n$ and $t = t_{n+1}$ in Equation (3.G.2) and simulating the increments due to $\mathbf{W}^{\Sigma}$ and $\tilde{\mathbf{J}}$ independently. Similarly, we can simulate the initial condition from the stationary distribution by writing

$$\begin{aligned}\mathbf{X}_0 &= \int_0^{\infty} e^{\mathcal{A}_{\mathbf{A}}u}\mathcal{E}\tilde{\mathbf{b}}\mathrm{d}u + \int_0^{\infty} e^{\mathcal{A}_{\mathbf{A}}u}\mathcal{E}\mathrm{d}\mathbf{W}_u^{\Sigma,\prime} + \int_0^{\infty} e^{\mathcal{A}_{\mathbf{A}}u}\mathcal{E}d\tilde{\mathbf{J}}'_u \\ &\approx -\mathcal{A}_{\mathbf{A}}^{-1}\mathcal{E}\tilde{\mathbf{b}} + \int_0^{T} e^{\mathcal{A}_{\mathbf{A}}u}\mathcal{E}\mathrm{d}\mathbf{W}_u^{\Sigma,\prime} + \sum_{\tau_k\in[0,T]} e^{\mathcal{A}_{\mathbf{A}}\tau_k}\mathcal{E}\Delta\tilde{\mathbf{J}}'_{\tau_k},\end{aligned}$$

where $T > 0$ is a big enough cutoff threshold (which can be chosen depending on $\|\mathcal{A}_{\mathbf{A}}\|$) and $\mathbb{W}^{\Sigma,\prime}$ and $\tilde{\mathbb{J}}'$ are independent copies of $\mathbb{W}^{\Sigma}$ and $\tilde{\mathbb{J}}$.

3.H Practical considerations: details and simulation study

3.H.1 Disentangling Brownian and jump component of a Lévy process

In this section, we focus on how to disentangle the continuous and jump components of a discretely observed d-dimensional Lévy process $\mathbb{L}_{\mathcal{Q}_t} = \{\mathbf{L}_u,\ u \in \mathcal{Q}_t\}$ with $\mathcal{Q}_t = \{0 = t_0 < t_1 < \ldots < t_{M_t} = t\}$. In our setting, this is particularly useful when estimating the covariance matrix of the Brownian component Σ and choosing the thresholding powers $\beta^{(i)}$, cf. Section 3.4.5. To do so we follow the approach outlined in Gegler (2011) based on a critical region B_t^γ but generalize it to irregularly spaced data. The idea is quite simple and is essentially a multidimensional version of thresholding. Let us fix the hyperparameter $\gamma > 1$ and define the true critical regions depending on the unknown parameter Σ

$$B_{t,m}^\gamma(\Sigma) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} \leq 2\gamma\Delta_{\mathcal{Q}_t}^m \log(M_t)\}. \tag{3.H.1}$$

We can then classify each Lévy increment in $\{\Delta_{\mathcal{Q}_t}^m\mathbf{L},\ m = 0, \ldots, M_t - 1\}$ as

- a Brownian increment if $\Delta_{\mathcal{Q}_t}^m\mathbf{L} \in B_{t,m}^\gamma(\Sigma)$;
- a jump increment if $\Delta_{\mathcal{Q}_t}^m\mathbf{L} \notin B_{t,m}^\gamma(\Sigma)$.

The choice of such critical regions, inherently tied to the elliptical shape of multivariate normal level curves, ensures that asymptotically only Brownian increments are retained. This procedure leads to a consistent (as $\Delta_{\mathcal{Q}_t} \to 0$ with t fixed) estimator for Σ

$$\hat{\Sigma}(\mathbb{L}_{\mathcal{Q}_t}; \gamma, \Sigma) = \frac{1}{M_t}\sum_{m=0}^{M_t-1}\frac{1}{\Delta_{\mathcal{Q}_t}^m}(\Delta_{\mathcal{Q}_t}^m\mathbf{L})(\Delta_{\mathcal{Q}_t}^m\mathbf{L})^{\mathrm{T}}\mathbb{1}_{\{\Delta_{\mathcal{Q}_t}^m\mathbf{L} \in B_{t,m}^\gamma(\Sigma)\}}. \tag{3.H.2}$$

Clearly, this estimator is infeasible in practice since the true critical regions depend on the unknown covariance matrix Σ. To use this in practice we apply the following iterative scheme with tolerance level $\epsilon > 0$:

1. initialize $\hat{\Sigma}^{(0)} = \frac{1}{M_t}\sum_{m=0}^{M_t-1}\frac{1}{\Delta_{\mathcal{Q}_t}^m}(\Delta_{\mathcal{Q}_t}^m\mathbf{L})(\Delta_{\mathcal{Q}_t}^m\mathbf{L})^{\mathrm{T}}$;
2. for $n \geq 0$ set $\hat{\Sigma}^{(n+1)} = \hat{\Sigma}(\mathbb{L}_{\mathcal{Q}_t}; \gamma, \hat{\Sigma}^{(n)})$ until $\|\hat{\Sigma}^{(n+1)} - \hat{\Sigma}^{(n)}\| \leq \epsilon$.

Gegler (2011) shows that the limiting critical regions $B_{t,m}^\gamma(\hat{\Sigma}^*)$ where $\hat{\Sigma}^* = \lim_{n\to\infty}\hat{\Sigma}^{(n)}$ converge in probability to the true critical regions $B_{t,m}^\gamma(\Sigma)$. From the iterative procedure above we thus obtain an estimator $\hat{\Sigma} = \hat{\Sigma}(\mathbb{L}_{\mathcal{Q}_t}; \gamma, \epsilon)$ for Σ and feasible critical regions $\{B_{t,m}^\gamma(\hat{\Sigma}),\ m = 0, \ldots, M_t - 1\}$ which can be used to disentangle the Brownian and jump components of $\mathbb{L}_{\mathcal{Q}_t}$.

The obtained critical regions can also be used to estimate integrals of the Lévy measure F of the form $F(f|_K) := \int_K f(\mathbf{z})F(\mathrm{d}\mathbf{z})$ for sets $K \subseteq \mathbb{R}^d$ and functions $f : K \mapsto \mathbb{R}^k$. In this case, the critical region-based estimators are defined as

$$\hat{F}(f|_K) = \frac{1}{t}\sum_{m=0}^{M_t-1} f(\Delta_{\mathcal{Q}_t}^m\mathbf{L})\mathbb{1}_{\{\Delta_{\mathcal{Q}_t}^m\mathbf{L} \in B^c \cap K\}},$$

where $B = B^{\gamma}_{t,m}(\Sigma)$ or $B^{\gamma}_{t,m}(\hat{\Sigma})$ depending on whether Σ is known or has to be estimated. Gegler (2011) provides theoretical guarantees of convergence for functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ of the form $f(\mathbf{z}) = (\mathbf{z}^{\mathrm{T}} M \mathbf{z})^l$ with $M \in \mathbb{R}^{d\times d}$ and $l \in \mathbb{N}$ and sets $K \subseteq \mathbb{R}^d$ of the form $K = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| \leq c\}$ for $c \in (0, \infty) \cup \{\infty\}$.

3.H.2 Joint estimation

As discussed in Section 3.4.5 we do not require to know the parameters $(\mathbf{b}, \Sigma, F)$ of the driving Lévy process to estimate the MCAR drift parameter $\mathbf{A}$ since we can make some pretty good approximations and data-driven choices. More precisely, the choice of $\mathbf{b}$ should have little impact on the value of $\hat{\mathbf{A}}$, Σ is not required and the jump thresholds (for which one would require knowledge of F) can be chosen empirically via the critical region procedure described in Appendix 3.H.1. In practice though, one might be interested in estimating such parameters for inference procedures, for example, to produce the asymptotic confidence regions described in Section 3.4.5. We hence propose the following iterative scheme given the observations $\mathbb{Y}_{\mathcal{P}_t} = \{\mathbf{Y}_s,\ s \in \mathcal{P}_t\}$ and refinement $\mathcal{Q}_t \subseteq \mathcal{P}_t$:

1. Set $\hat{\mathbf{b}} = \mathbf{0}$.

2. Until convergence criterion (in terms of $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, $\hat{\Sigma}$ and $\hat{F}(\mathrm{id}|_{\|\mathbf{z}\|\geq 1})$) is met:

   (a) estimate $\hat{\mathbf{A}}$ using Equation (3.4.25) with $\mathbf{b} = \hat{\mathbf{b}}$ and $\boldsymbol{\nu}_t$ chosen as described in Section 3.4.5, i.e. for each $i = 1, \ldots, d$ treat the increments $\Delta^m_{\mathcal{Q}_t} \hat{D}^{p-1} Y^{(i)}$, $m = 0, \ldots, M_t - 1$ as univariate Lévy increments and identify the critical interval $B^{\gamma,(i)}_{t,m}(\hat{\Sigma}) = [-\nu_t^{m,(i)}, \nu_t^{m,(i)}]$ as in Appendix 3.H.1;

   (b) recover the increments of the background driving Levy process $\{\Delta \mathbf{L}_u(\hat{\mathbf{A}}),\ u \in \mathcal{Q}_t\}$ using Equation (3.4.30) and estimate:

   - $\hat{\boldsymbol{\mu}} = t^{-1}\mathbf{L}_t(\hat{\mathbf{A}})$ or $\hat{A}_p\left(\frac{1}{|\mathcal{P}_t|}\sum_{s\in\mathcal{P}_t} \mathbf{Y}_s\right)$;
   - $\hat{\Sigma}$ and $\hat{F}(\mathrm{id}|_{\|\mathbf{z}\|\geq 1})$ by applying the critical region-based estimators described in Appendix 3.H.1;
   - $\hat{\mathbf{b}} = \hat{\boldsymbol{\mu}} - \hat{F}(\mathrm{id}|_{\|\mathbf{z}\|\geq 1})$.

In the above procedure, we can estimate the auxiliary quantity $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{A}}[\mathbf{L}_1]$ by using either of the two proposed estimators. The first is a "law of large numbers" estimator while the second is a "method of moments" estimator obtained by inverting $\mathbb{E}_{\mathbf{A}}[\mathbf{Y}_1] = A_p^{-1}\boldsymbol{\mu}$.

Once the convergence criterion is met, we can estimate other quantities related to $F(f|_K)$ by applying the critical region-based estimators described in Appendix 3.H.1 to the recovered Lévy increments.

3.H.3 Simulation study with unknown b, Σ and F

In this section, we repeat the simulation study carried out in Section 3.4.4 but now assume the Lévy process parameters $\mathbf{b}$, Σ and F are unknown. We use the joint estimation procedure described in Appendix 3.H.2.

The experiment is set up exactly as in Section 3.4.4, i.e. we consider a $d = 1$ dimensional CAR process of order $p = 2$ such that the driving Lévy process has triplet $b = (0, \Sigma = 1, F)$ and F follows the three jump regimes (BM), (CP) and (Γ). To evaluate the quality of estimation of the Lévy measure F we consider the statistics $F(\mathrm{id}) = \int_{\mathbb{R}} xF(\mathrm{d}x)$ and $F(\mathrm{id}^2) = \int_{\mathbb{R}} x^2F(\mathrm{d}x)$. Note that the choices of F are such that:

(BM) $F(\mathrm{id}) = 0$, $F(\mathrm{id}^2) = 0$;

(CP) $F(\mathrm{id}) = 0$, $F(\mathrm{id}^2) = 1$;

(Γ) $F(\mathrm{id}) = 0$, $F(\mathrm{id}^2) = 2$;

As before $\mathcal{P}_t$ and $\mathcal{Q}_t$ are chosen to be uniform with mesh sizes $\Delta_{\mathcal{P}_t} = t^{-6}$ and $\Delta_{\mathcal{Q}_t} = t^{-2}$. The hyperparameters of the iterative scheme are set to $\gamma = 1.01$ and $\epsilon = 0.01$ respectively.

The results are reported in Figure 3.H.1, Figure 3.H.2, Figure 3.H.3, 3.H.4, Figure 3.H.5 and Figure 3.H.6 which empirically suggest consistency and asymptotic normality of $\hat{\mathbf{A}}$ is retained and the estimators $\hat{b}, \hat{\Sigma}$, $\hat{F}(\mathrm{id})$ and $\hat{F}(\mathrm{id}^2)$ are consistent. It is worth noting that in the low data regime (at $t = 2$ with $N_{\mathcal{P}_t} = 128$ observations) the estimators for $\mathbf{A}$ perform considerably worse than in the case where the Lévy process parameters are known. When more data is available though the performance of the estimators appears comparable.

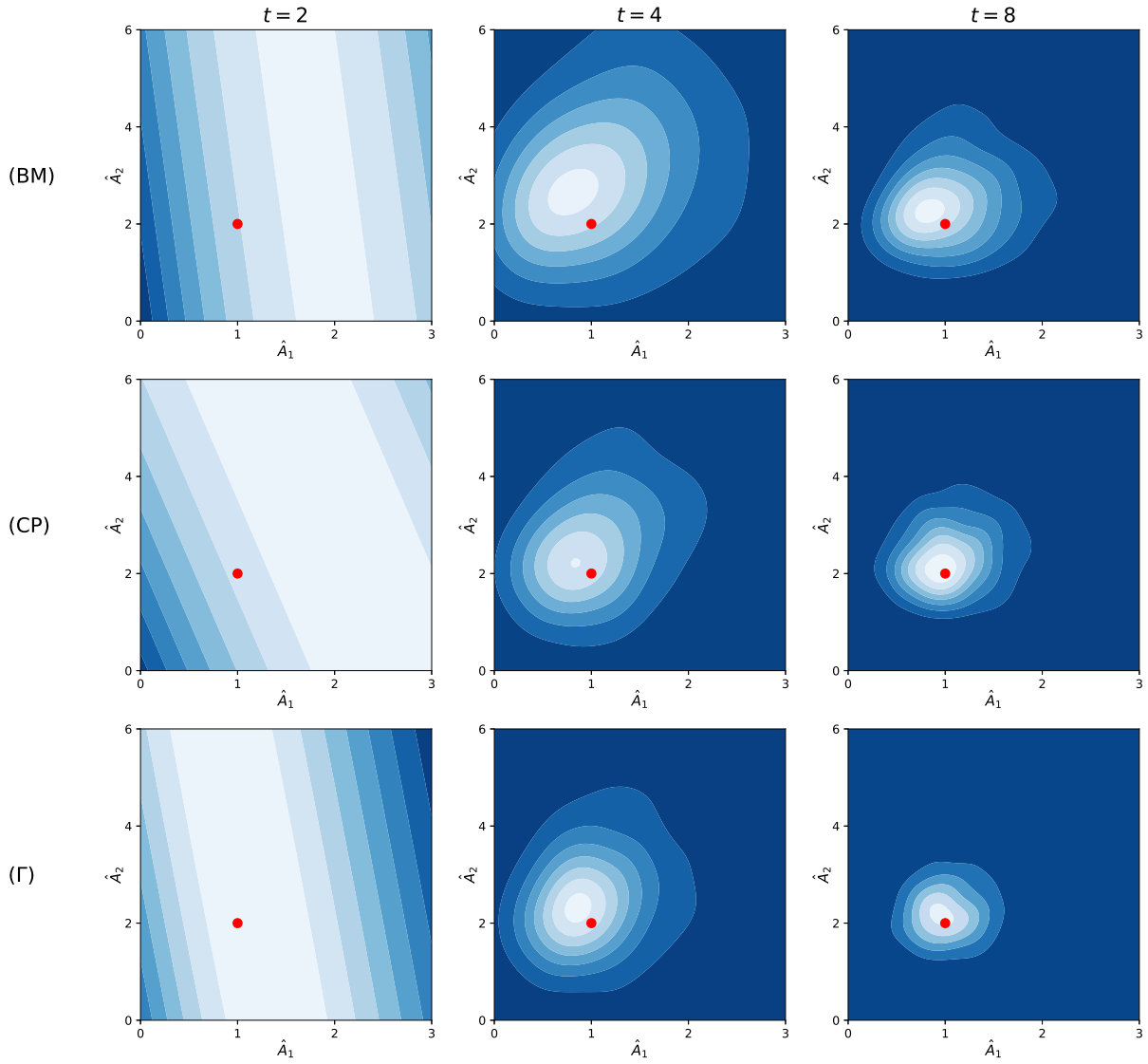


Figure 3.H.1: Empirical distribution of the estimator $\hat{\mathbf{A}}(\mathbb{Y}_{\mathcal{P}_t}; \mathcal{Q}_t, \boldsymbol{\nu}_t)$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

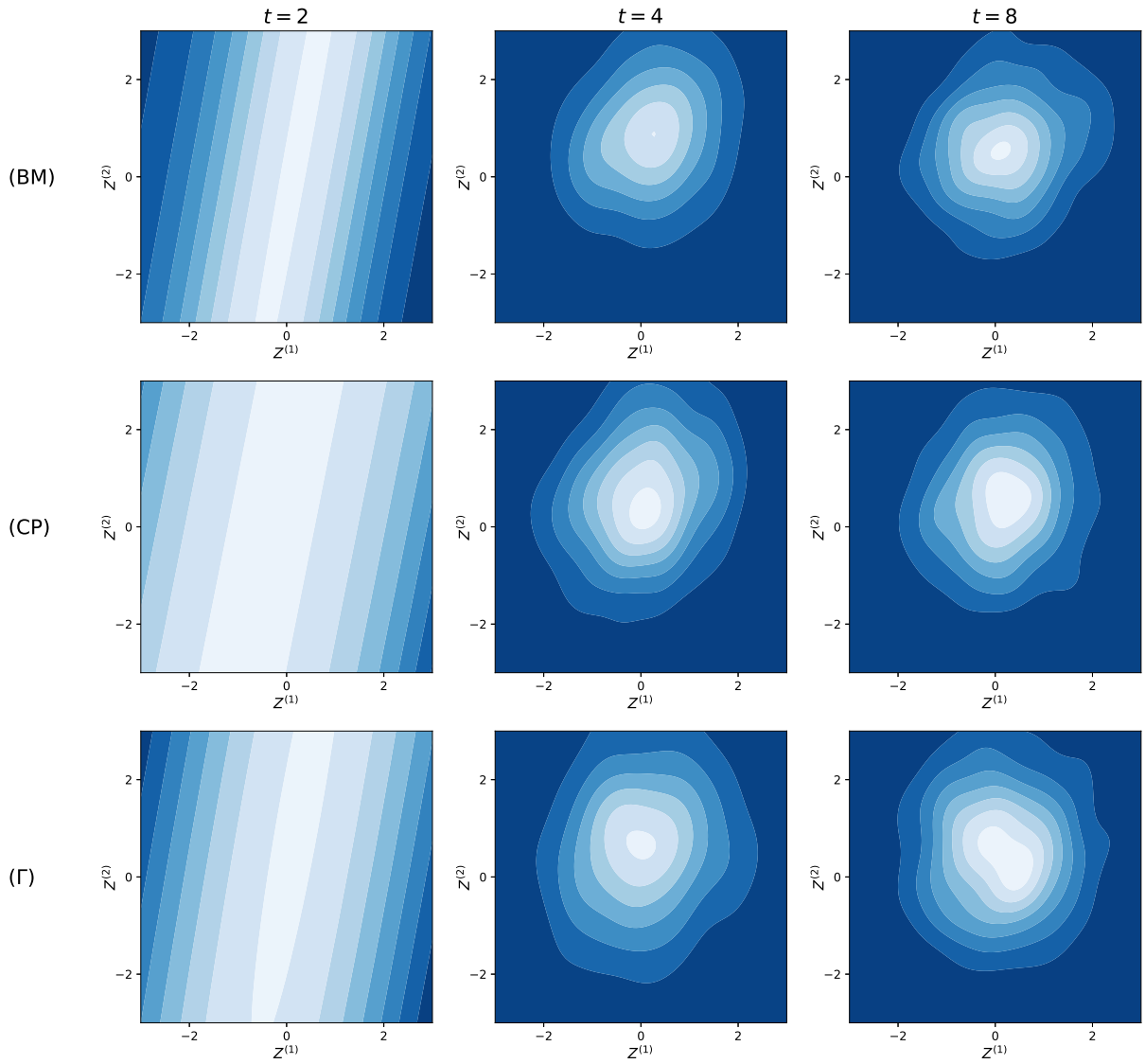


Figure 3.H.2: Empirical distribution of the statistic $\mathbf{Z}_t(\hat{\Sigma})$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

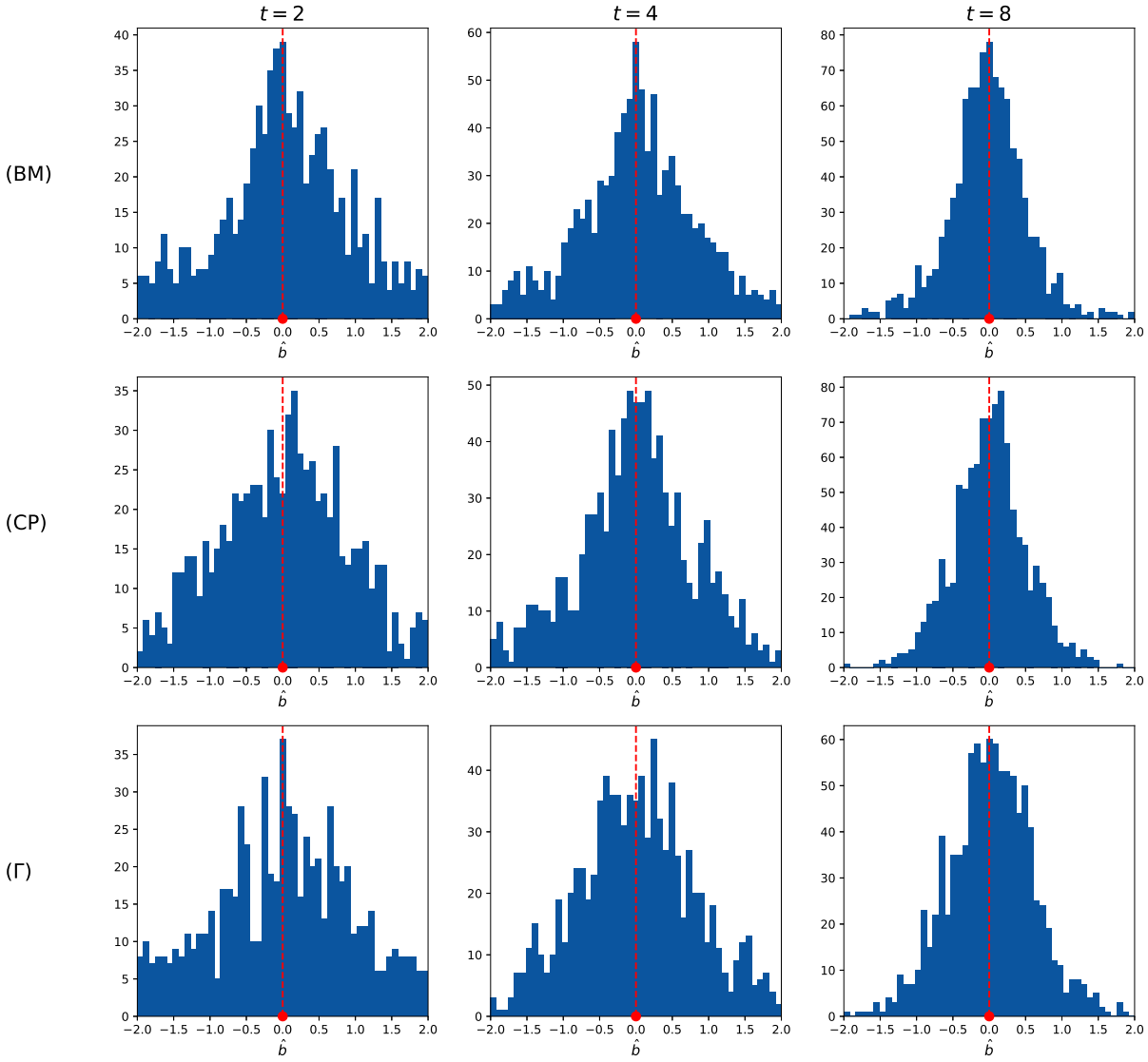


Figure 3.H.3: Empirical distribution of the estimator $\hat{b}$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

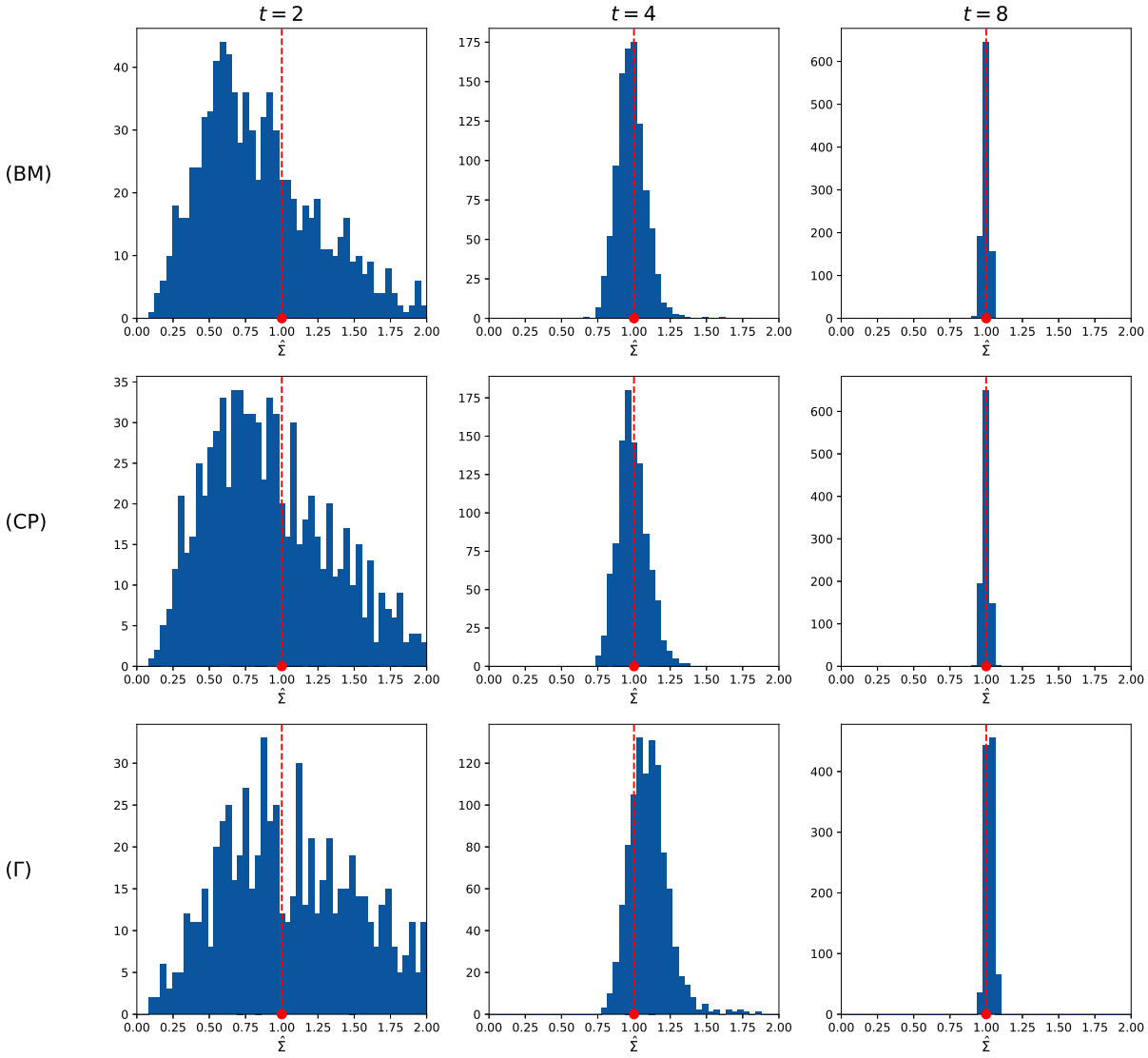


Figure 3.H.4: Empirical distribution of the estimator $\hat{\Sigma}$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

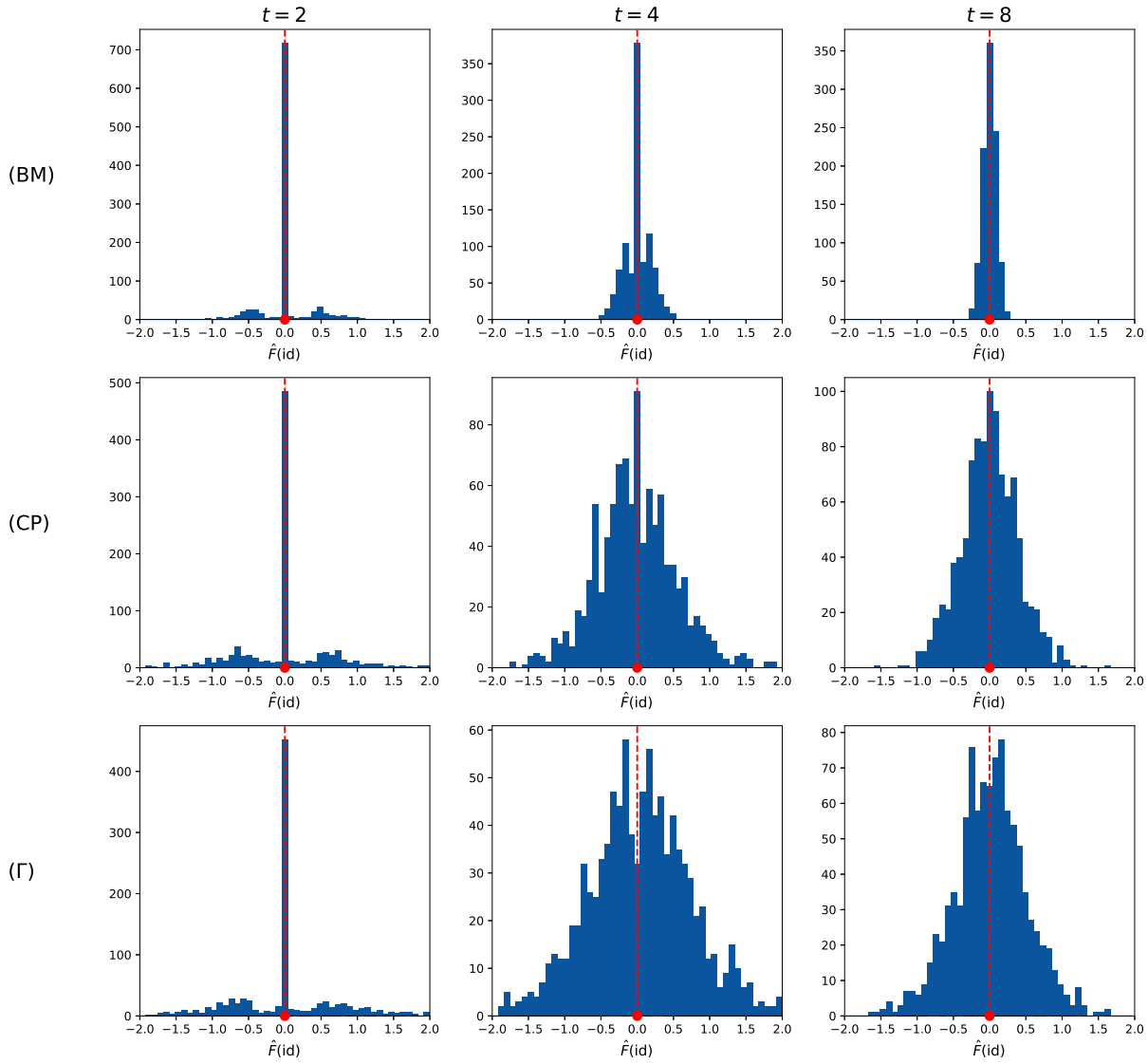


Figure 3.H.5: Empirical distribution of the estimator $\hat{F}(\mathrm{id})$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

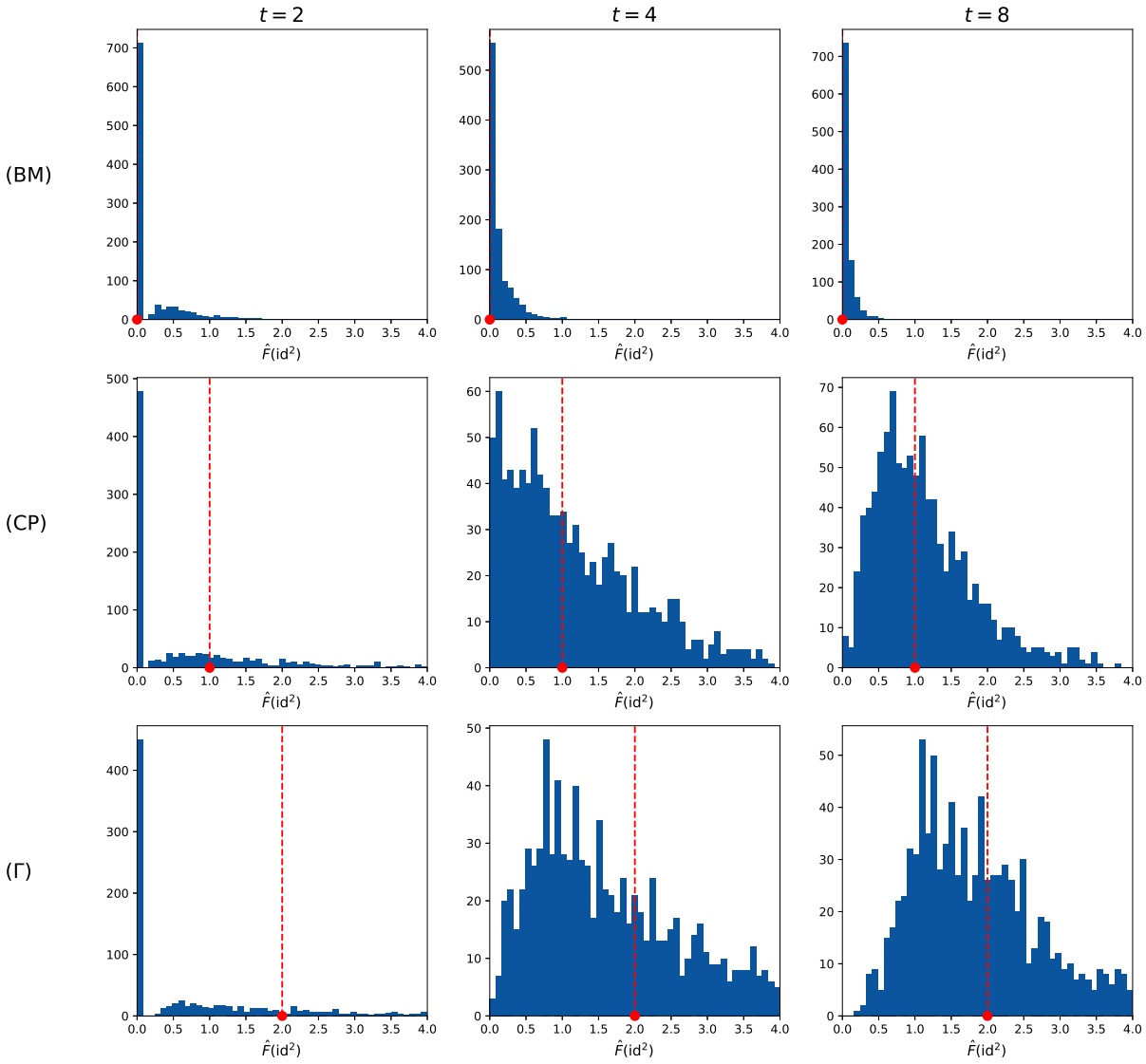


Figure 3.H.6: Empirical distribution of the estimator $\hat{F}(\mathrm{id}^2)$ for 1000 Monte Carlo samples under the three jump regimes for $t = 2, 4, 8$ when b, Σ and F are unknown.

Chapter 4

Expected Signature Estimation

In the third and final chapter of this thesis, we shift from parametric to non-parametric statistics. In the previous chapter, the dynamics of CAR processes were specified through a parametric stochastic differential equation, interpreted rigorously via classical Itô calculus. More recently, it has been shown that such stochastic integrals can also be constructed in the broader framework of rough path theory. For a given stochastic process, this pathwise construction uniquely defines the expected signature: the expected value, under the law of the process, of a sequence of its iterated integrals.

This infinite but countable object is exceptionally rich and – when the integral is constructed "geometrically" – uniquely characterizes the law of the process. Leveraging this property, the expected signature has been widely applied in practical settings to design various machine learning algorithms. In this chapter, we provide asymptotic results for the expected signature of a finite collection of discrete paths, demonstrating that this non-parametric statistic can consistently estimate the target continuous-time expected signature under realistic sampling schemes. Additionally, we propose a modification to this statistic with better finite sample properties when the underlying process is a martingale. Finally, we showcase the practical advantages of this modified statistic across several machine learning tasks.

4.1 Introduction

Originally introduced by Chen (1954) and serving as a fundamental object of rough path analysis (Lyons, Caruana, and Lévy, 2007), the signature

$$\mathbb{S} = \{S(\mathbb{X})_{[0,t]} \in T((\mathbb{R}^d)),\ t \in [0,T]\},$$

of a path $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\} \in C([0,T], \mathbb{R}^d)$ is a lift (in the sense that it embeds $\mathbb{X}$) to the space of continuous functions over the tensor algebra $T((\mathbb{R}^d)) := \bigoplus_{k\geq 0}(\mathbb{R}^d)^{\otimes k}$ possessing some nice algebraic and geometric properties. When the path is of bounded variation, i.e. $\mathbb{X} \in \mathrm{BV}([0,T], \mathbb{R}^d)$, the signature is defined as the sequence of iterated Riemann-Stieltjes integrals of $\mathbb{X}$,

$$S^k(\mathbb{X})_{[0,t]} = \int\limits_{0\leq s_1\leq \ldots \leq s_k\leq t} \cdots \int \mathrm{d}\mathbf{X}_{s_1} \otimes \cdots \otimes \mathrm{d}\mathbf{X}_{s_k}, \quad t \in [0,T],\ k \geq 0.$$

In many practical applications the path $\mathbb{X}$ is taken to be the piecewise linear interpolation of a discrete-time stream of data, which is of bounded variation by construction. Signature-based

machine learning (ML) approaches (Lemercier et al., 2021; Ni et al., 2022; Liao et al., 2024) thus often restrict the theoretical framework to the bounded variation case. In this setting, two fundamental properties of the signature are the characterization result of Hambly and Lyons (2005) and the universality approximation theorem of Levin, Lyons, and Ni (2016). Moreover, when the path $\mathbb{X}$ is understood as a (realization of a) random process with distribution $\mathbb{P}$ over $\mathrm{BV}([0,T],\mathbb{R}^d)$, the shuffle property of the signature implies that all moments of the random variable $S(\mathbb{X})_{[0,T]} \in T((\mathbb{R}^d))$ are determined by the expected signature $\phi(T) = \mathbb{E}[S(\mathbb{X})_{[0,T]}] \in T((\mathbb{R}^d))$. A natural question, known as the Hamburger moment problem (Fawcett, 2003), is thus whether the expectation of the signature characterizes its law (and thus the law of the path). When imposing a probability distribution $\mathbb{P}$ on $\mathbb{X}$ the assumption of bounded variation paths becomes quite restrictive: Brownian motion, the basic building block of many stochastic models, has paths of infinite variation almost surely. Even if we observe a discrete-time stream of data, we often still would like to define the process $\mathbb{X}$ as a latent stochastic process of which we observe the linear interpolation over some partition π of $[0,T]$, hereafter denoted by $\mathbb{X}^\pi$. We hence wish to make sense of the signature of a stochastic process $\mathbb{X}$ with paths of unbounded variation. For a given path $\mathbb{X} \in C([0,T],\mathbb{R}^d)$ of finite p-variation, once we "lift" the process to a p-rough path (Lyons, Caruana, and Lévy, 2007, Definition 3.11) then the signature $\mathbb{S}$ of $\mathbb{X}$ is uniquely defined – this is the first fundamental theorem in the theory of rough paths (Lyons, Caruana, and Lévy, 2007, Theorem 3.7). Without delving into the details of rough path theory, for our purposes it suffices to interpret the choice of lift as fixing a notion of integration with respect to $\mathbb{X}$: the higher order signatures terms are then understood as iterated integrals of the path $\mathbb{X}$ defined in this sense[1]. To help intuitive understanding of this statement, let us consider a couple of examples. When $\mathbb{X}$ is a semimartingale it has finite p-variation for any $p > 2$. It hence suffices to lift $\mathbb{X}$ to a 2-rough path to fix the definition of signature. It turns out that, for a given notion of stochastic integral (e.g. Stratonovich or Itô), the lifted process

$$t \mapsto \left(1,\ \mathbf{X}_t,\ \int_0^t \mathbf{X}_s \otimes \mathrm{d}\mathbf{X}_s\right) \in T^2((\mathbb{R}^d)), \quad t \in [0,T],$$

is a p-rough path for all $p > 2$ and hence uniquely defines a full signature for $\mathbb{X}$. All higher-order signature terms can then be shown to be the iterated stochastic integrals in the sense fixed by the lift. Similarly, a fractional Brownian motion with Hurst parameter $H \in (1/4, 1/3]$ has finite p variation for all $p > 1/H$. In this case, we need to lift $\mathbb{X}$ to a 3-rough path. For a large class of Gaussian processes, including fractional Brownian motion, a canonical lift is defined in Coutin and Qian (2002). For a complete treatment of stochastic processes as rough paths we refer to the detailed monograph Friz and Victoir (2010).

Motivated by the fact that we can only ever observe the process $\mathbb{X}$ over a discrete partition π of $[0,T]$ we restrict our attention to the class of stochastic processes whose lift (and hence signature) can be approximated by the lift (and hence signature) of the bounded variation path $\mathbb{X}^\pi$. Following

[1]Note that, when $\mathbb{X}$ is a bounded variation path, it does not require to be lifted as all notions of integral coincide and we can define all higher-order signature terms as iterated Riemann-Stieltjes integrals.

the rough path literature we take such approximation in the p-variation metric to define the notion of *canonical geometric stochastic process*, cf. Definition 4.3.1. This definition ensures the resulting lift of $\mathbb{X}$ is almost surely a geometric p-rough path in the sense of Lyons, Caruana, and Lévy (2007, Definition 3.13) and Friz and Victoir (2010, Definition 9.16). When $\mathbb{X}$ is a semimartingale, the Stratonovich lift ensures the resulting 2-rough path is geometric and is given by the p-variation limit of the lifts of piecewise linear approximations $\mathbb{X}^\pi$ along *any* sequence of partitions with $|\pi| \to 0$ (Friz and Victoir, 2010, Theorem 14.16)[2]. For more general stochastic processes the choice of sequence of partitions π along which we approximate the lift of $\mathbb{X}$ plays a fundamental role, in which case we call the sequence of partitions signature defining. In Chevyrev and Lyons (2016) and Chevyrev and Oberhauser (2018) the authors provide characterization results[3] for the expected signature of canonical geometric stochastic processes, i.e. conditions under which the map $\mathbb{P} \mapsto \mathbb{E}[S(\mathbb{X})_{[0,T]}]$ is injective. The characterizing property of the expected signature has found practical use in a wide range of applications, ranging from pure machine learning tasks (Lemercier et al., 2021; Triggiano and Romito, 2024; Ni et al., 2022; Liao et al., 2024) to mathematical finance (Lyons, Nejad, and Pérez Arribas, 2021; Futter, Horvath, and Wiese, 2023).

The expected signature

$$\phi(T) := \mathbb{E}[S(\mathbb{X})_{[0,T]}] \in T((\mathbb{R}^d)),$$

is thus a highly informative quantity and, consequently, methods for computing $\phi(T)$ have received considerable research interest. Such methods can be broadly categorized into two classes: those employing an analytical approach and those following a statistical one. Analytical methods aim to develop exact formulas for specific classes of models. A first step in this direction was taken in Ni (2012, Section 4) showing that the expected signature of an Itô diffusion satisfies an explicit partial differential equation (PDE). This result was subsequently generalized in Cuchiero, Svaluto-Ferro, and Teichmann (2023) to the class of signature-SDEs and in Friz, Hager, and Tapia (2022) and Friz, Hager, and Tapia (2024) to (discontinuous) semimartingales. On the other hand, the statistical approach aims to estimate $\phi(T)$ directly from observed data, preserving the model-free nature of the expected signature. For a given set of observations $\mathbb{X}^{\pi,1}, \dots, \mathbb{X}^{\pi,n}$ one can form the estimator

$$\hat{\phi}^{N,\pi}(T) := \frac{1}{N} \sum_{n=1}^{N} S\left(\mathbb{X}^{n,\pi}\right)_{[0,T]},$$

and study its in-fill $|\pi| \to 0$ and long-span $N \to \infty$ asymptotics. This line of work includes the explicit results of Ni (2012, Section 3.2) for Brownian motion and of Passeggeri (2020) for fractional Brownian motion with Hurst parameter $H > 1/2$ as well as the preliminary results in Friz and Victoir (2010) for more general semimartingales. In this chapter we aim to provide a unifying set of general conditions under which the expected signature estimator $\hat{\phi}^{N,\pi}(T)$ displays desirable

[2] In Theorem 4.3.8 we give a weaker version of this result sufficient for our purposes.

[3] In Chevyrev and Oberhauser (2018) the authors prove universality and characterization of a robustified version of the signature. Nevertheless, the authors suggest the use of the classical expected signature is still justified in certain situations (Chevyrev and Oberhauser, 2018, Remark 4).

asymptotic statistical properties, namely consistency and asymptotic normality. Our results allow for irregular observation partitions[4] π – possibly different across samples – and for dependency across the samples $\mathbb{X}^1, \dots, \mathbb{X}^n$. We consider the practically relevant case where the samples are obtained by chopping-and-shifting a single long observation of the process $\{\mathbf{X}_t,\ t \geq 0\}$.

Remark 4.1.1. In Bonnier and Oberhauser (2020) the authors introduce the log expected signature, i.e. $\kappa(T) := \log \mathbb{E}[S(\mathbb{X})_{[0,T]}]$ where log denotes the logarithm on the space of formal power series. Much like the expected signature $\phi(T)$ yields a natural generalization of the moments of the process $\mathbb{X}$, the log expected signature $\kappa(T)$ provides a generalized notion of the cumulants of $\mathbb{X}$. We refer to Schell and Oberhauser (2023) for an application of the log expected signature to nonlinear independent component analysis. In this work we focus on estimating terms in the expected signature $\phi(T)$ but the consistency results immediately carry over to the estimation of terms in $\kappa(T)$ by a simple application of the continuous mapping theorem.

The rest of this chapter is structured as follows. In Section 4.2 we start by considering the simple but insightful case where the process $\mathbb{X}$ is taken to be a standard Brownian motion. Section 4.3 then sets the scene by defining canonical geometric stochastic processes and their signatures. Section 4.4 contains the main theoretical results of the chapter. This section is quite naturally divided into two parts treating the in-fill and the long-span asymptotics of the expected signature estimator. Piecing together these two components yields the complete asymptotic theorems, namely Theorem 4.4.19, Theorem 4.4.29 and Theorem 4.4.34. Having established the theoretical asymptotic properties of the estimator, Section 4.5 is devoted to empirically assessing its finite samples properties. In Section 4.6 we then focus on the case where $\mathbb{X}$ is a martingale and introduce a modification of the expected signature estimator with better finite sample properties. A review of machine learning applications employing the expected signature estimator is given in Section 4.6.2, with an emphasis on how the modified expected signature estimator can be used to improve empirical performance. Some final remarks and potential future research directions are discussed in Section 4.7.

4.2 A simple motivating example: Brownian motion

In this section, we will focus on the task of estimating the signature of Brownian motion from a finite collection of discretely sampled Brownian paths. From an epistemic perspective, this is somewhat unnecessary, as $\mathbb{E}\left[S(\mathbb{B})_{[0,T]}\right]$ is known explicitly. Nevertheless, this serves as an important motivating example for the results developed in the rest of this work. In the first part of this section, we follow quite closely Ni's doctoral dissertation (Ni, 2012). The results on asymptotic normality and variance reduction via control variates are new.

In the following we fix $T > 0$ and let $\Omega = C([0,T]; \mathbb{R}^d)$ be the set of continuous paths from $[0,T]$ to $\mathbb{R}^d$. We equip Ω with the Borel sigma-algebra $\mathcal{F}$, i.e. the σ-algebra generated by the open sets

[4] Clearly, for the estimation problem to be well-posed, the sequence of partitions needs to be signature defining in the sense of Definition 4.3.6.

with respect to the sup-norm $\|\cdot\|_\infty$ on Ω, and the Wiener measure $\mathbb{P}$, i.e. the measure on $(\Omega, \mathcal{F})$ such that the canonical process $\mathbb{B} = \{\mathbf{B}_t,\ t \in [0,T]\}$ with $\mathbf{B}_t(\omega) = \omega(t)$ for all $t \in [0,T]$, $\omega \in \Omega$ is a standard d-dimensional Brownian motion. Fixing the standard basis for $\mathbb{R}^d$ we write

$$\mathbf{B}_t = \sum_{i=1}^{d} B_t^{(i)} e_i,\ t \in [0,T].$$

We will be interested in estimating the expected (Stratonovich) signature of Brownian motion:

$$\phi(T) := \mathbb{E}\left[S(\mathbb{B})_{[0,T]}\right],$$

where

$$S(\mathbb{B})_{[0,T]} = \left(1,\ S^1(\mathbb{B})_{[0,T]},\ S^2(\mathbb{B})_{[0,T]},\ \ldots\right) \in T((\mathbb{R}^d)),$$

with

$$S^k(\mathbb{B})_{[0,T]} = \int\cdots\int_{0<u_1<\ldots<u_k<T} \circ\, \mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\, \mathrm{d}\mathbf{B}_{u_k}, \quad k \geq 1.$$

The expected signature of Brownian motion can be explicitly computed, its analytic form is given by the following theorem.

Theorem 4.2.1. *The expected signature of Brownian motion up to time $t \in [0,T]$ is given by*

$$\phi(t) = \mathbb{E}\left[S(\mathbb{B})_{[0,t]}\right] = \exp_\otimes\left\{\frac{t}{2}\sum_{i=1}^{d} e_i \otimes e_i\right\} := \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{t}{2}\right)^k \left(\sum_{i=1}^{d} e_i \otimes e_i\right)^{\otimes k}.$$

Remark 4.2.2. This result can be found in Fawcett (2003).

Proof. See Appendix 4.A.1. □

In the general setting, when $\mathbb{X}$ is a process whose expected signature is not available analytically (either because the model is not fully specified or simply because no closed form formula is known for the fully specified model), we are interested in estimating $\phi(T)$ from samples of the discretely observed process. In the Brownian case let

$$\mathbb{B}^\pi = \{\mathbf{B}_t,\ t \in \pi\} \quad \text{where} \quad \pi = \{0 = t_0 < \ldots < t_M = T\}.$$

We define the (continuous-time) linear approximation of $\mathbb{B} = \mathbb{B}_{[0,T]}$ on π by

$$\hat{\mathbb{B}}^\pi = \hat{\mathbb{B}}^\pi_{[0,T]} = \{\hat{\mathbf{B}}^\pi_t,\ t \in [0,T]\},$$

where

$$\hat{\mathbf{B}}^\pi_t = \mathbf{B}_{t_m} + \frac{t - t_m}{t_{m+1} - t_m}\left(\mathbf{B}_{t_{m+1}} - \mathbf{B}_{t_m}\right), \quad t \in [t_m, t_{m+1}].$$

For simplicity, in this section we assume evenly spaced observations of Brownian motion, i.e. $t_m = m\Delta$ where $\Delta = \frac{T}{M}$, and thus make the dependence on the number of steps M explicit by writing $\hat{\mathbb{B}}^M = \hat{\mathbb{B}}^\pi$. We start by showing that the expected signature of the linearly interpolated process $\hat{\mathbb{B}}^M$ is close to the expected signature of Brownian motion. This assumes we have access to "infinitely" many observations of $\hat{\mathbb{B}}^M$ and thus can perfectly compute its expectation. In the following, the results rely on being able to compute the expected signatures of $\mathbb{B}$ and $\hat{\mathbb{B}}^M$ explicitly but in the more general setting of Section 4.4 we will develop approximations results which will not require an analytical form of the limiting quantity $\phi(T)$.

We start by computing the expected signature of the linearly approximated Brownian motion explicitly.

Theorem 4.2.3. *The expected signature of the linearly approximated Brownian motion up to time $m\Delta \in [0,T]$ for $m = 0, \dots, M$ is given by*

$$\phi^M(m\Delta) = \mathbb{E}\left[S(\hat{\mathbb{B}}^M)_{[0,T]}\right] = \left(\sum_{k=0}^{\infty} \sum_{I \in \mathcal{K}_{2k}} \frac{C_I}{k!} \left(\frac{\Delta}{2}\right)^k \otimes_{i \in I} e_i\right)^{\otimes m},$$

where

- *$\mathcal{K}_{2k}$ is the set of even words of length $2k$, i.e. $n_j(I) = |\{i \in I : i = j\}|$ is even $\forall j \in \{1, \dots, d\}$;*
- *$C_I = \binom{k}{\frac{n_1(I)}{2}, \dots, \frac{n_d(I)}{2}} \Big/ \binom{2k}{n_1(I), \dots, n_d(I)}$ for $I \in \mathcal{K}_{2k}$.*

Remark 4.2.4. This result is from Ni (2012). Here and in what follows we use the convention that when $I = ()$ (the empty word) we write $\otimes_{i \in I} e_i = 1$.

Proof. See Appendix 4.A.2. □

Corollary 4.2.5. *The k-th level of the expected signature of the linearly approximated Brownian motion up to time $m\Delta \in [0,T]$ for $m = 0, \dots, M$ is given by*

$$\phi_k^M(m\Delta) = \begin{cases} \displaystyle\sum_{\substack{j_1, \dots, j_m \geq 0 \\ 2j_1 + \dots + 2j_m = k}} \bigotimes_{l=1}^{m} \sum_{I \in \mathcal{K}_{2j_l}} \frac{C_I}{j_l!} \left(\frac{\Delta}{2}\right)^{j_l} \otimes_{i \in I} e_i, & \text{if } k \text{ is even,} \\ 0, & \text{if } k \text{ is odd.} \end{cases}$$

Proof. See Appendix 4.A.3. □

Given the exact forms of ϕ^M and ϕ, we can compute the following error bound explicitly, which shows that ϕ^M converges to ϕ with rate $\mathcal{O}\left(M^{-1}\right)$ as $M \to \infty$.

Theorem 4.2.6. *For every even integer $k \geq 4$ it holds that*

$$\left\|\phi_k^M(T) - \phi_k(T)\right\|_1 \leq \frac{d^{k/2}\Delta}{(k/2-2)!}\left(\frac{T}{2}\right)^{k/2-1} = \mathcal{O}\left(\frac{1}{M}\right), \tag{4.2.1}$$

while for all other integers, i.e $k \in \{0,2\}$ or k odd, we have

$$\phi_k^M(T) = \phi_k(T).$$

Remark 4.2.7. This result is from Ni (2012), with the only minor correction being the explicit dependence on the dimension d.

Proof. See Appendix 4.A.4. □

We can now consider the more realistic setting where we do not observe the full law of $\hat{\mathbb{B}}^M$ but instead have access to N i.i.d. copies of $\hat{\mathbb{B}}^M$. In this case we cannot compute $\phi^M(T)$ exactly but need to estimate it. In view of Theorem 4.2.6 such estimator will also be a (biased) estimator for $\phi(T)$. Let $\{\hat{\mathbb{B}}^{M,n}\}_{n=1}^N$ denote N i.i.d. linear approximations of Brownian motion over the evenly space partition π. By the independent increments property of Brownian motion, one may equivalently assume that such copies are obtained by chopping up and shifting (so that each subsample starts from zero) a single long discrete observation over $[0, NT]$. We then define the estimator for the expected signature of Brownian motion as

$$\hat{\phi}^{M,N}(T) := \frac{1}{N}\sum_{n=1}^{N} S(\hat{\mathbb{B}}^{M,n})_{[0,T]}. \tag{4.2.2}$$

Note that, for every word I and fixed M we can apply the strong law of large numbers and the classic central limit theorem to the i.i.d. random variables $\{S^I(\hat{\mathbb{B}}^{M,n})_{[0,T]}\}_{n=1}^N$ with

$$\mathbb{E}\left[S^I(\hat{\mathbb{B}}^M)_{[0,T]}\right] = \phi_I^M(T),$$

and variance

$$\begin{aligned}\sigma_{M,I}^2 &:= \mathrm{Var}\left(S^I(\hat{\mathbb{B}}^M)_{[0,T]}\right)\\ &= \mathbb{E}\left[\left(S^I(\hat{\mathbb{B}}^M)_{[0,T]}\right)^2\right] - \mathbb{E}\left[S^I(\hat{\mathbb{B}}^M)_{[0,T]}\right]^2\\ &= \sum_{K \in I \sqcup\!\sqcup I} \phi_K^M(T) - \phi_I^M(T)^2 < \infty,\end{aligned}$$

to deduce that

$$\hat{\phi}_I^{M,N}(T) - \phi_I^M(T) \overset{\mathbb{P}\text{-a.s.}}{\to} 0, \quad N \to \infty,$$

and

$$\frac{\sqrt{N}}{\sigma_{M,I}}\left(\hat{\phi}_I^{M,N}(T) - \phi_I^M(T)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad N \to \infty.$$

These results only describe the relationship between $\hat{\phi}^{N,M}(T)$ and $\phi^M(T)$ for each *fixed* M. The following theorem is of more practical relevance as it gives the asymptotic properties of $\hat{\phi}^{N,M}(T)$ as an estimator for $\phi(T)$ as both N *and* M increase to ∞. Note that, due to the interplay between the in-fill and the large sample limits, only an L^2 consistency result can be obtained.

Theorem 4.2.8. *Fix a word $I \neq ()$, then the estimator $\hat{\phi}_I^{M,N}(T)$ for $\phi_I(T)$ is*

1. *consistent, i.e.*

$$\hat{\phi}_I^{M,N}(T) \xrightarrow{L^2} \phi_I(T),$$

   for any sequence (M,N) such that $M \to \infty$, $N \to \infty$;

2. *asymptotically normal, i.e.*

$$\frac{\sqrt{N}}{\sigma_I}\left(\hat{\phi}_I^{M,N}(T) - \phi_I(T)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

   where

$$\sigma_I^2 := \mathrm{Var}\left(S^I(\mathbb{B})_{[0,T]}\right) = \sum_{K \in I \sqcup\!\sqcup I} \phi_K(T) - \phi_I(T)^2 \in (0,\infty),$$

   for any sequence (M,N) such that $M,N \to \infty$ and, when $|I| \geq 4$ and even, $\sqrt{N}/M \to 0$.

Proof. See Appendix 4.A.5. □

Remark 4.2.9. The proof of Theorem 4.2.8 immediately yields a quantitative version of the CLT, i.e. denoting by $\Phi_{N,M,I}$ the c.d.f. of $\frac{\sqrt{N}}{\sigma_I}\left(\hat{\phi}_I^{M,N}(T) - \phi_I(T)\right)$ one has for $N, M \geq 1$ and $x \in \mathbb{R}$

$$\left|\Phi_{N,M,I}\left(\frac{\sigma_{M,I}}{\sigma_I}x + \mu_{N,M,I}\right) - \Phi(x)\right| \leq \frac{C\rho_{M,I}}{\sigma_{M,I}^3\sqrt{N}},$$

with "CLT bias"

$$\mu_{N,M,I} = \frac{\sqrt{N}}{\sigma_I}\left(\phi_I^M(T) - \phi_I(T)\right).$$

Finally, we extend the previous result to the joint estimation of multiple signature terms by a simple application of the Cramér-Wold theorem.

Theorem 4.2.10. *Fix a collection of words $\mathbf{I}$ such that $I \neq ()$ for all $I \in \mathbf{I}$, then the estimator $\hat{\phi}_{\mathbf{I}}^{M,N}(T)$ for $\phi_{\mathbf{I}}(T)$ is*

1. *consistent, i.e.*

$$\hat{\phi}_{\mathbf{I}}^{M,N}(T) \xrightarrow{L^2} \phi_{\mathbf{I}}(T),$$

   for any sequence (M,N) such that $M \to \infty$, $N \to \infty$;

2. asymptotically normal, i.e.

$$\sqrt{N}\left(\hat{\phi}_{\mathbf{I}}^{M,N}(T) - \phi_{\mathbf{I}}(T)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{\mathbf{I}}),$$

where

$$\Sigma^{\mathbf{I}}_{I,J} := \mathrm{Cov}\left(S^I(\mathbb{B})_{[0,T]}, S^J(\mathbb{B})_{[0,T]}\right) = \sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T), \quad I, J \in \mathbf{I},$$

for any sequence (M, N) such that $M, N \to \infty$ and, when $|I| \geq 4$ and even for some $I \in \mathbf{I}$, $\sqrt{N}/M \to 0$.

Proof. See Appendix 4.A.6. □

Remark 4.2.11. The CLT result of this theorem is feasible since we can explicitly compute $\Sigma^{\mathbf{I}}$ using Theorem 4.2.1. In the more general setting when $\phi(T)$ is unknown, to make the CLT feasible we will need to find a consistent estimator for $\Sigma^{\mathbf{I}}$. Note that this is a function of higher order signature terms and hence it will be sufficient to apply the consistency part of the estimation result, cf. Remark 4.4.22 and Remark 4.4.30.

Note that in the proofs of the central limit theorem, we use $\phi^M(T)$ as "intermediate" approximating quantity between $\hat{\phi}^{M,N}(T)$ and $\phi(T)$. One could alternatively consider as "intermediate" quantity

$$\hat{\phi}^N(T) := \frac{1}{N}\sum_{n=1}^{N} S(\mathbb{B}^n)_{[0,T]},$$

where $\{\mathbb{B}^n\}_{n=1}^N$ are the i.i.d. Brownian motions from which $\{\hat{\mathbb{B}}^{M,n}\}_{n=1}^N$ are discretely sampled and decompose

$$\begin{aligned}\hat{\phi}^{M,N}(T) - \phi(T) &= \left(\hat{\phi}^{M,N}(T) - \hat{\phi}^N(T)\right) + \left(\hat{\phi}^N(T) - \phi(T)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(S(\hat{\mathbb{B}}^{M,n})_{[0,T]} - S(\mathbb{B}^n)_{[0,T]}\right) + \left(\frac{1}{N}\sum_{n=1}^{N} S(\mathbb{B}^n)_{[0,T]} - \mathbb{E}\left[S(\mathbb{B})_{[0,T]}\right]\right).\end{aligned}$$

In this case, the proof of Theorem 4.2.8 would simplify considerably once we obtain a result on the convergence of $S(\hat{\mathbb{B}}^M)_{[0,T]}$ to $S(\mathbb{B})_{[0,T]}$ in an appropriate probabilistic sense, a stronger result than the convergence of the expectations given in Theorem 4.2.6. With the results we develop in Section 4.4.1 we can easily show such convergence indeed holds in L^m for any $m \geq 2$. This follows by applying Theorem 4.4.7 to the process $\mathbb{X} = \mathbb{B}$, which clearly satisfies Assumption (Aα) and Assumption (Aδ) with $\alpha = 1/2$, $\delta = 1$ and $p > 2$ and to the uniform sequence of partitions π with $|\pi| = M^{-1}$, which by Theorem 4.3.8 is signature-defining for the Stratonovich signature.

4.2.1 Variance reduction via control variates

In this section, we explore whether we can find a better estimator for $\phi(T)$ than $\hat{\phi}^{N,M}(T)$ given the observations $\{\mathbb{B}^{n,\pi}\}_{n=1}^N$, i.e. an alternative estimator with the same bias as $\hat{\phi}^{N,M}(T)$ but smaller variance. We will focus on the one-dimensional case ($d = 1$) to keep notation as light as possible.

To provide some initial intuition we start by considering the unrealistic setting where N i.i.d. continuous-time observations, i.e. $\{\mathbb{B}^n_{[0,T]}\}_{n=1}^N$, are available. We will then move on to the more realistic scenario where the processes are discretely observed and the signature needs to be approximated.

Continuous-time observations

Motivation: level $k = 2$. To provide some motivation for the choice of control variate we will be using we start by considering the estimation of the level $k = 2$ expected signature of one-dimensional ($d = 1$) Brownian motion. Assuming N i.i.d. observations of the Brownian motion one has the "naive" estimator

$$\hat{\phi}_2^N(T) := \frac{1}{N}\sum_{n=1}^N S^2(\mathbb{B}^n)_{[0,T]},$$

where

$$S^2(\mathbb{B})_{[0,T]} = \int_0^T B_u \circ \mathrm{d}B_u = \frac{B_T^2}{2}.$$

At least formally, we can understand $\hat{\phi}^N(T)$ as $\hat{\phi}^{N,M}(T)$ when $M = \infty$. One can easily check that the estimator is unbiased

$$\mathbb{E}[\hat{\phi}_2^N(T)] = \phi_2(T),$$

and has variance

$$\mathrm{Var}(\hat{\phi}_2^N(T)) = \frac{1}{N}\mathrm{Var}\left(\frac{B_T^2}{2}\right) = \frac{T}{4N} > 0.$$

If we instead consider

$$\hat{\phi}_2^{N,c}(T) := \frac{1}{N}\sum_{n=1}^N \left\{S^2(\mathbb{B}^n)_{[0,T]} - S_c^2(\mathbb{B}^n)_{[0,T]}\right\},$$

where

$$S_c^2(\mathbb{B})_{[0,T]} = \int_0^T B_u \mathrm{d}B_u = \frac{B_T^2}{2} - \frac{T}{2},$$

is the level $k = 2$ signature term in the Itô sense, it can be easily checked this is a "perfect" estimator since

$$\hat{\phi}_2^{N,c}(T) = \frac{T}{2}, \quad \mathbb{P} - a.s.$$

This simple but powerful example motivates us to "control" the naive estimator based on iterated Stratonovich integrals via a similar object obtained through Itô integration. Such control will be, by construction of the Itô integral, mean zero and hence won't affect the (un)biasness of the original

estimator. As we will see later, the variance reduction effect of the control depends directly on its correlation with the original iterated Stratonovich integral, and thus, intuitively aiming to maximize correlation, we will consider as control the signature-like object obtained by interpreting only the outermost integral in the Itô sense.

General signature level k. We continue assuming access to N independent Brownian motion observations over $[0,T]$ in dimension $d=1$, i.e. $\{\mathbb{B}^n\}_{n=1}^N$, and consider the (unbiased) estimator

$$\hat{\phi}_k^N(T) = \frac{1}{N}\sum_{n=1}^{N} S^k(\mathbb{B}^n)_{[0,T]},$$

for the target parameter

$$\phi_k(T) = \mathbb{E}\left[S^k(\mathbb{B})_{[0,T]}\right],$$

where

$$S^k(\mathbb{B})_{[0,T]} := \int\cdots\int_{0<u_1<\cdots<u_k<T} \circ \mathrm{d}B_{u_1}\cdots\circ \mathrm{d}B_{u_{k-1}} \circ \mathrm{d}B_{u_k}.$$

As above, at least formally, we can understand $\hat{\phi}^N(T)$ as $\hat{\phi}^{N,M}(T)$ when $M=\infty$.

We then define the control variate estimator

$$\hat{\phi}_k^{N,c}(T) = \frac{1}{N}\sum_{n=1}^{N}\left\{S^k(\mathbb{B}^n)_{[0,T]} - cS_c^k(\mathbb{B}^n)_{[0,T]}\right\},$$

where the controls are defined as

$$S_c^k(\mathbb{B})_{[0,T]} := \int\cdots\int_{0<u_1<\cdots<u_k<T} \circ \mathrm{d}B_{u_1}\cdots\circ \mathrm{d}B_{u_{k-1}} \mathrm{d}B_{u_k}.$$

One can verify that the resulting estimator is unbiased

$$\mathbb{E}\left[\hat{\phi}_k^{N,c}(T)\right] = \mathbb{E}\left[\hat{\phi}_k^{N}(T)\right] = \phi_k(T),$$

and, setting

$$c_* = \frac{\mathrm{Cov}\left(S^k(\mathbb{B})_{[0,T]}, S_c^k(\mathbb{B})_{[0,T]}\right)}{\mathrm{Var}\left(S_c^k(\mathbb{B})_{[0,T]}\right)},$$

it has variance

$$\mathrm{Var}\left(\hat{\phi}_k^{N,c_*}(T)\right) = (1-\rho_k^2)\mathrm{Var}\left(\hat{\phi}_k^{N}(T)\right),$$

where

$$\rho_k = \mathrm{Corr}\left(S^k(\mathbb{B})_{[0,T]}, S_c^k(\mathbb{B})_{[0,T]}\right)$$

$$= \operatorname{Corr}\left(\int_{0<u_1<\cdots<u_k<T} \cdots \int \circ \mathrm{d}B_{u_1} \cdots \circ \mathrm{d}B_{u_{k-1}} \circ \mathrm{d}B_{u_k}, \int_{0<u_1<\cdots<u_k<T} \cdots \int \circ \mathrm{d}B_{u_1} \cdots \circ \mathrm{d}B_{u_{k-1}} \mathrm{d}B_{u_k} \right).$$

As discussed in the previous motivational paragraph we choose $S_c^k(\mathbb{B})_{[0,T]}$ with only the outermost integral interpreted in the Itô sense to preserve the "mean-zero" property of the control while aiming to maximize ρ_k and, thus, minimize the resulting estimator's variance.

Proposition 4.2.12. *The correlation is ρ_k is given by*

$$\rho_k = \operatorname{Corr}\left(S^k(\mathbb{B})_{[0,T]}, S_c^k(\mathbb{B})_{[0,T]}\right) = \delta_k \sqrt{2 - \frac{1}{k}} \left\{ 1 - \frac{k!^2}{(2k)!} \sum_{\substack{i=0 \\ i+k\text{even}}}^{k} \binom{k+i-2}{i} \right\},$$

where

$$\delta_k = \begin{cases} \left(1 - \dfrac{k!^3}{(2k)!(k/2)!^2}\right)^{-1/2}, & \text{if } k \text{ even}, \\ 1, & \text{if } k \text{ odd}. \end{cases}$$

Proof. See Appendix 4.A.7. □

In Figure 4.1 we plot the correlation ρ_k for $k \in \{2, \ldots, 50\}$ and note that using the control variate $S_c^k(\mathbb{B})_{[0,T]}$ leads to a decrease in the variance of the estimator by a factor ≥ 5 and asymptotically ≈ 9.

Discrete-time path observations

We now consider the more realistic setting of discrete-time observations, i.e. we assume access to N independent discrete-time observations of one-dimensional Brownian motion over the evenly-spaced partition $\pi = \pi_M$ in $d = 1$ and denote them by $\{\mathbb{B}^{n,\pi}\}_{n=1}^N$. We consider the biased but consistent estimator

$$\hat{\phi}_k^{N,M}(T) = \frac{1}{N} \sum_{n=1}^{N} S^k(\hat{\mathbb{B}}^{n,M})_{[0,T]},$$

for the target parameter

$$\phi_k(T) = \mathbb{E}\left[S^k(\mathbb{B})_{[0,T]}\right],$$

where

$$S^k(\hat{\mathbb{B}}^M)_{[0,T]} := \int_{0<u_1<\cdots<u_k<T} \cdots \int \circ \mathrm{d}\hat{B}_{u_1}^M \cdots \circ \mathrm{d}\hat{B}_{u_k}^M = \int_{0<u_1<\cdots<u_k<T} \cdots \int \mathrm{d}\hat{B}_{u_1}^M \cdots \mathrm{d}\hat{B}_{u_k}^M,$$

since the linear approximation of Brownian motion $\hat{\mathbb{B}}^M$ has finite variation and thus we have a unique notion of integral (which can be defined in the Riemannian sense).

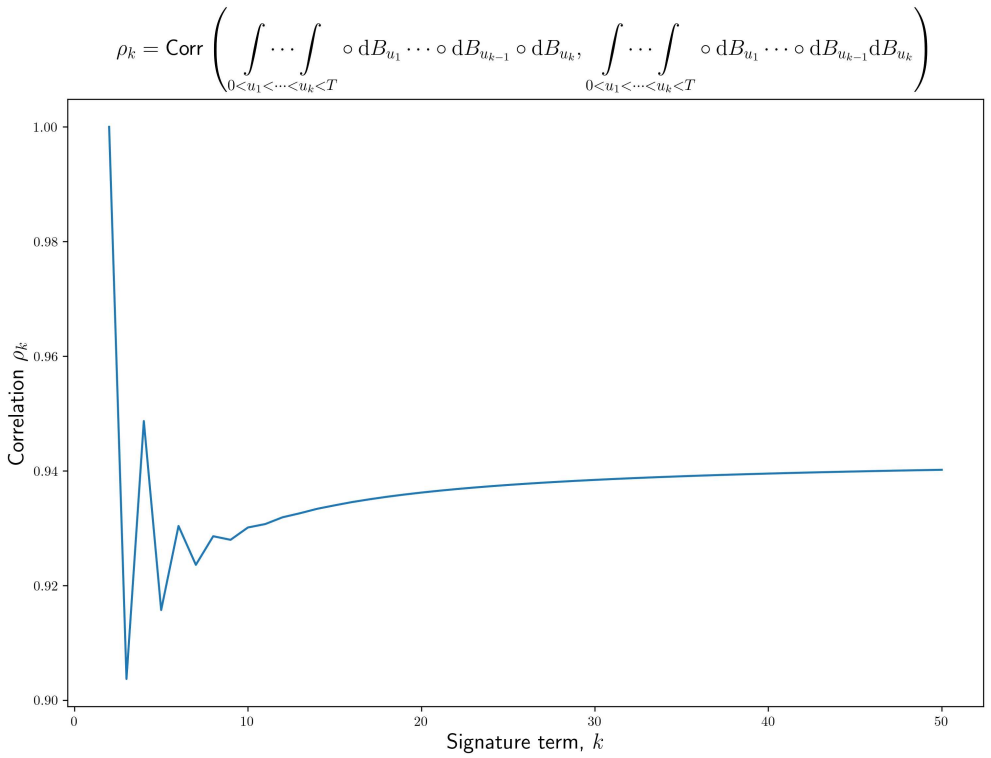


Figure 4.1: Control variates correlation, continuous-time observations.

As in the previous setting, we define the controlled estimator

$$\hat{\phi}_k^{N,M,c}(T) = \frac{1}{N}\sum_{n=1}^{N}\left\{S^k(\hat{\mathbb{B}}^{n,M})_{[0,T]} - cS_c^k(\hat{\mathbb{B}}^{n,M})_{[0,T]}\right\},$$

where the discrete-time analog of the Itô control is given by

$$S_c^k(\hat{\mathbb{B}}^M)_{[0,T]} := \sum_{m=0}^{M-1}\int\cdots\int_{0<u_1<\cdots<u_{k-1}<t_m} \mathrm{d}\hat{B}_{u_1}^M\cdots\mathrm{d}\hat{B}_{u_{k-1}}^M(B_{t_{m+1}} - B_{t_m}).$$

One can verify that the resulting controlled estimator is also consistent, with the same bias as $\hat{\phi}_k^{N,M}(T)$ since

$$\mathbb{E}\left[\hat{\phi}_k^{N,M,c}(T)\right] = \mathbb{E}\left[\hat{\phi}_k^{N,M}(T)\right],$$

and, when

$$c_* = \frac{\mathrm{Cov}\left(S^k(\hat{\mathbb{B}}^M)_{[0,T]}, S_c^k(\hat{\mathbb{B}}^M)_{[0,T]}\right)}{\mathrm{Var}\left(S_c^k(\hat{\mathbb{B}}^M)_{[0,T]}\right)},$$

has variance

$$\mathrm{Var}\left(\hat{\phi}_k^{N,M,c_*}(T)\right) = (1-\rho_{k,M}^2)\mathrm{Var}\left(\hat{\phi}_k^{N,M}(T)\right),$$

where

$$\begin{aligned}\rho_{k,M} &= \mathrm{Corr}\left(S^k(\hat{\mathbb{B}}^M)_{[0,T]}, S_c^k(\hat{\mathbb{B}}^M)_{[0,T]}\right) \\ &= \mathrm{Corr}\left(\int\cdots\int_{0<u_1<\cdots<u_k<T} \mathrm{d}\hat{B}^M_{u_1}\cdots\mathrm{d}\hat{B}^M_{u_k}, \sum_{m=0}^{M-1}\int\cdots\int_{0<u_1<\cdots<u_{k-1}<t_m} \mathrm{d}\hat{B}^M_{u_1}\cdots\mathrm{d}\hat{B}^M_{u_{k-1}}(B_{t_{m+1}}-B_{t_m})\right).\end{aligned}$$

As before, we can explicitly compute the correlation $\rho_{k,M}$.

Proposition 4.2.13. *The correlation $\rho_{k,M}$ is given by*

$$\rho_{k,M} = \frac{\frac{1}{2^{k-1}(k-1)!}\left(\frac{1}{M}\right)^k \sum_{m=0}^{M-1}\sum_{j=0}^{\lfloor (k-1)/2\rfloor}\binom{2k-2j-2}{j,k-j-1,k-2j-1} m^{k-j-1}(M-m)^j}{\sqrt{\left(\frac{(2k-1)!!}{k!^2} - \mathbb{1}_{2|k}\frac{(k-1)!!^2}{k!^2}\right)\left(\frac{(2k-3)!!}{(k-1)!^2}\left(\frac{1}{M}\right)^k\sum_{m=0}^{M-1} m^k\right)}}.$$

Proof. See Appendix 4.A.8. □

In Figure 4.2 we revisit Figure 4.1 by comparing the signature control variate correlation $\rho_k,\ k\in\{2,\dots,50\}$ obtained from continuous-time observations (in blue) with the signature control variate correlation $\rho_{k,M},\ k\in\{2,\dots,50\}$ computed from discrete-time observations (in green, with color gradient increasing in the number of discretization steps $M\in\{5,10,20,50,100\}$). We note that, for fixed $M<\infty$, as k increases the benefit of using the Itô control decreases, while for fixed signature level k the finer the observational grid, i.e. the bigger the M, the higher the variance reduction obtained by using the control. In general, we note that for reasonable values of k (and M) the controlled estimator $\hat{\phi}^{N,M,c_*}(T)$ has a significantly lower variance than $\hat{\phi}^{N,M}(T)$.

4.3 Canonical geometric stochastic processes

In this section, we introduce the class of stochastic processes for which we will develop expected signature estimators. Recall from rough path theory that the signature of a stochastic process is generally not unique, as it depends on the choice of lift (for example, we may lift a Brownian motion to a 2-rough path using either Itô or Stratonovich lifts). Given the discussion in the introduction, we will focus on processes that can be lifted to geometric rough paths in a canonical way, i.e. from limits of piecewise linear interpolations of discrete observations (this will correspond to the Stratonovich lift in the aforementioned Brownian setting), and will be interested in estimating the corresponding signatures. In the following let $\mathbb{X}=\{\mathbf{X}_t,\ t\in[0,T]\}$ denote a d-dimensional stochastic process over a probability space $(\Omega,\mathcal{F},\mathbb{P})$.

Definition 4.3.1. We say $\mathbb{X}=\{\mathbf{X}_t,\ t\in[0,T]\}$ is a canonical geometric stochastic process of rough order p if there exists a deterministic sequence of partitions ρ with $|\rho|\to 0$ such that the limit in

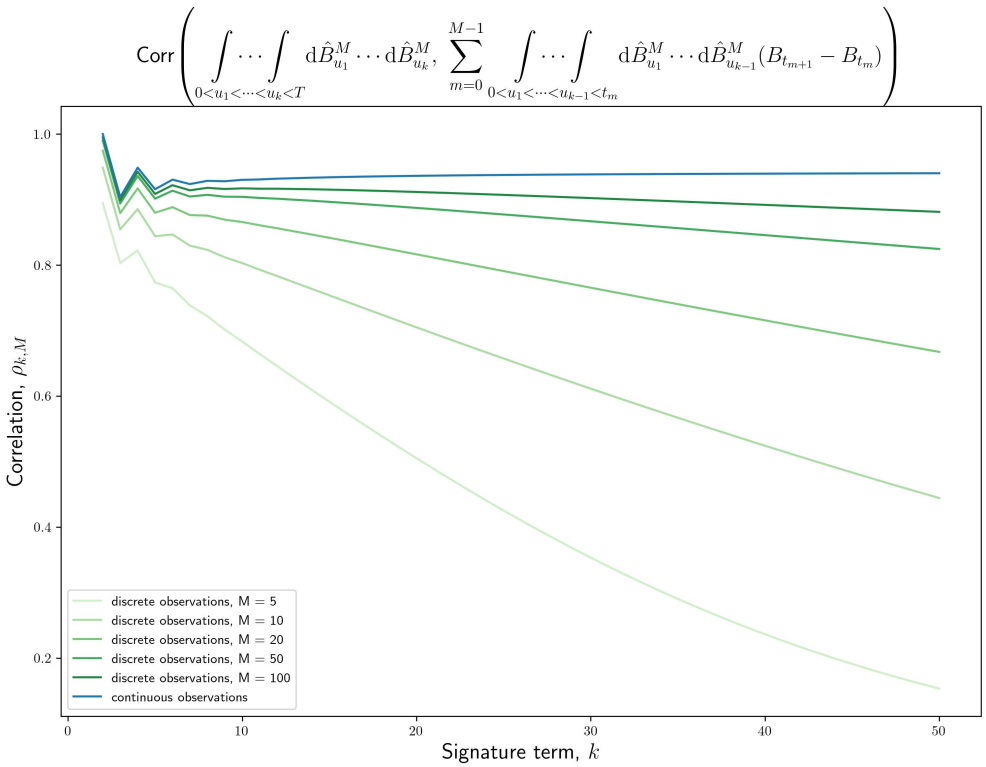


Figure 4.2: Control variates correlation, discrete-time observations.

the p-variation metric of the canonically lifted linearly interpolated process $\mathbb{X}^\rho$ exists in probability. Convergence in probability implies almost sure convergence (along a subsequence) and hence we can almost surely define the lift of $\mathbb{X}$ as such limit.

Remark 4.3.2. Any canonical geometric stochastic process of rough order p can thus be almost surely lifted to a geometric p-rough path. A question for the rough path community: is this a strict inclusion?

If we replace geometric by *weakly* geometric in the previous question then the answer is affirmative. For example, fractional Brownian motion with $H \leq 1/4$ can be lifted to a weakly geometric p-rough path (Nualart and Tindel, 2011) but for any sequence of partitions ρ the limit in the p-variation metric of the canonically lifted linearly interpolated process $\mathbb{X}^\rho$ does not converge (Coutin and Qian, 2002).

Remark 4.3.3. The definition of lift suggests this might depend on the choice of the sequence of partitions ρ. In any case, for a wide range of stochastic processes there exist *canonical* lifts that satisfy our definition of canonical geometric rough path. These include:

- Semimartingales: for $p \in (2, 3)$ any semimartingale can be lifted to a geometric p-rough path by defining the lift via Stratonovich integration; the signature of $\mathbb{X}$ then coincides with iterated Stratonovich integrals. For any sequence of partitions ρ the lifts of the linear interpolations $\{S_2(\mathbb{X}^\rho)_{[0,t]},\ t \in [0, T]\}$ converge in p-variation metric to the Stratonovich lift (Friz and Victoir,

2010, Chapter 14) and hence $\mathbb{X}$ is a canonical geometric stochastic process in the sense of Definition 4.3.1. In Section 4.3.1 we discuss a slightly weaker result, but sufficient for our purposes, implying that for any sequence of partitions ρ, $S^2(\mathbb{X}^\rho)_{[0,t]}$ converges uniformly over compacts in probability to the Stratonovich lift.

- Gaussian processes: many Gaussian processes admit canonical lifts to geometric p-rough paths (Friz and Victoir, 2010; Coutin and Qian, 2002, Theorem 15.34, Definition 15.35) with the existence criterion for such canonical lifts easily stated in terms of the covariance function. The definition of the lift implicitly yields ρ to be any sequence such that $\mathbb{X}^\rho$ converges uniformly to $\mathbb{X}$ almost surely. For example, fractional Brownian motion with Hurst parameter $H > 1/4$ can be canonically lifted to a geometric p-rough path with $p > 1/H$ by choosing ρ to be the sequence of dyadic partitions.

In what follows, we will assume the canonical geometric stochastic process $\mathbb{X}$ has a *canonical* lift (i.e. a canonical sequence of partitions ρ along which the lift is defined) and unambiguously refer to it as *the* lift of $\mathbb{X}$.

More precisely, let ρ denote a partition of $[0,T]$ with mesh $|\rho| = \max_{[u,v]\in\rho} |v-u|$ and let $\mathbb{X}^\rho = \{\mathbf{X}^\rho_t,\ t \in [0,T]\}$ denote the linear approximation of $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ over ρ, i.e.

$$\mathbf{X}^\rho_t = \mathbf{X}_u + \frac{t-u}{v-u}\mathbf{X}_{u,v}, \quad \text{for } t \in [u,v].$$

The signature of $\mathbb{X}^\rho$ up to time $t \in [0,T]$ is defined as

$$S(\mathbb{X}^\rho)_{[0,t]} = \left(1,\ S^1(\mathbb{X}^\rho)_{[0,t]},\ S^2(\mathbb{X}^\rho)_{[0,t]},\ \ldots\right) \in T((\mathbb{R}^d)), \tag{4.3.1}$$

with

$$S^k(\mathbb{X}^\rho)_{[0,t]} = \int\limits_{0<u_1<\ldots<u_k<t} \cdots \int \mathrm{d}\mathbf{X}^\rho_{u_1} \otimes \cdots \otimes \mathrm{d}\mathbf{X}^\rho_{u_k}, \quad k \geq 1,$$

where the integrals are defined in the classic Riemann-Stieltjes sense along the piecewise-linear path $\mathbb{X}^\rho$. The canonical lift of $\mathbb{X}^\rho$ to a (geometric) p-rough path is

$$\left(1,\ S^1(\mathbb{X}^\rho)_{[0,t]},\ \ldots,\ S^{\lfloor p\rfloor}(\mathbb{X}^\rho)_{[0,t]}\right) \in T^{\lfloor p\rfloor}\left(\left(\mathbb{R}^d\right)\right), \quad t \in [0,T]. \tag{4.3.2}$$

Definition 4.3.1 requires that there exists a sequence of partitions ρ for which this sequence of geometric p-rough paths converges in probability in the p-variation metric.

A key result from rough path theory is that a (geometric) p-rough path has a full signature. Fixing the lift of $\mathbb{X}$ via Definition 4.3.1 we thus have a uniquely specified signature for $\mathbb{X}$.

Definition 4.3.4. The signature of a canonical geometric stochastic process $\mathbb{X}$,

$$\mathbb{S} = \{S(\mathbb{X})_{[0,t]} \in T((\mathbb{R}^d)),\ t \in [0,T]\},$$

is defined pathwise (on a set of full measure) as the unique extension of the lift of $\mathbb{X}$ to a multiplicative functional of arbitrary order in the sense of Lyons, Caruana, and Lévy (2007, Theorem 3.7). The elements of the signature are the rough iterated integrals of $\mathbb{X}$.

Remark 4.3.5. Taking ρ to be a sequence such that Definition 4.3.1 holds, by continuity of the extension map (Lyons, Caruana, and Lévy, 2007, Theorem 3.10), it immediately follows that the signature of $\mathbb{X}^\rho$ (truncated at level $K \geq \lfloor p \rfloor$) converges in probability to the signature of $\mathbb{X}$ (up to level K) in the p-variation topology. In particular, this implies that for any finite collection of words $\mathbf{I}$,

$$S^{\mathbf{I}}(\mathbb{X}^\rho)_{[0,t]} \xrightarrow{\mathbb{P}} S^{\mathbf{I}}(\mathbb{X})_{[0,t]}. \tag{4.3.3}$$

Similar arguments imply that when convergence to the lift along ρ holds almost surely in the p-variation metric then also the higher order signature terms converge almost surely, and in particular Equation (4.3.3) holds in the almost sure limit.

In this work, we will be interested in estimating the expected signature at the fixed time horizon $T > 0$. To develop the properties of these estimators it will be thus sufficient to work with pointwise limits like Equation (4.3.3) without having to deal with the stronger pathwise p-variation convergence used to define canonical geometric stochastic processes. This mode of convergence will thus be sufficient to consider a sequence of partitions as *signature-defining*.

Definition 4.3.6. Let $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ be a canonical geometric stochastic process, we say that a determinstic sequence of partitions π of the interval $[s,t] \subseteq [0,T]$ with $|\pi| \to 0$ is signature-defining if for any collection of words $\mathbf{I}$,

$$S^{\mathbf{I}}(\mathbb{X}^\pi)_{[s,t]} \xrightarrow{\mathbb{P}} S^{\mathbf{I}}(\mathbb{X})_{[s,t]}, \quad |\pi| \to 0. \tag{4.3.4}$$

Remark 4.3.7. By Remark 4.3.5 every canonical geometric stochastic process has at least one signature-defining sequence of partitions over any $[s,t] \subseteq [0,T]$ given by the sequence ρ in Definition 4.3.1. Moreover, by passing to a subsequence if necessary, this also guarantees the existence of a sequence of partitions ρ_* along which Equation (4.3.4) holds almost surely.

In the next section, we show that when $\mathbb{X}$ is a semimartingale, every sequence of partitions π with $|\pi| \to 0$ is signature-defining.

4.3.1 Semimartingales: convergence in probability of signatures

In this section, we consider the case where $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a d-dimensional continuous semimartingale, i.e. we can decompose it as $\mathbb{X} = \mathbb{M} + \mathbb{A}$ where $\mathbb{M}$ is a continuous local martingale and $\mathbb{A}$ is a continuous finite variation process (requiring $\mathbf{A}_0 = 0$ ensures the decomposition is

unique). The Stratonovich[5] lift of $\mathbb{X}$ to a p-rough path for $p \geq 2$ is

$$\left(1,\, \mathbf{X}_t,\, \int_0^t \mathbf{X}_u \otimes \circ\, \mathrm{d}\mathbf{X}_u\right) \in T^2\left(\left(\mathbb{R}^d\right)\right),$$

and the corresponding signature

$$S(\mathbb{X})_{[0,T]} = \left(1,\, S^1(\mathbb{X})_{[0,T]},\, S^2(\mathbb{X})_{[0,T]},\, \ldots\right) \in T((\mathbb{R}^d)),$$

with

$$S^k(\mathbb{X})_{[0,T]} = \int\cdots\int_{0<u_1<\ldots<u_k<T} \circ\, \mathrm{d}\mathbf{X}_{u_1} \otimes \cdots \otimes \circ\, \mathrm{d}\mathbf{X}_{u_k}, \quad k \geq 1.$$

We denote by $\{S^{\leq K}(\mathbb{X})_{[0,t]},\ t \in [0,T]\}$ the signature truncated at level $K \geq 0$, i.e.

$$S^{\leq K}(\mathbb{X})_{[0,t]} = \left(1,\, S^1(\mathbb{X})_{[0,t]},\, \ldots,\, S^K(\mathbb{X})_{[0,t]}\right) \in T^K((\mathbb{R}^d)).$$

Theorem 4.3.8. *If $\mathbb{X}$ is a continuous d-dimensional semimartingale then for $t \in [0,T]$ and any (deterministic) sequence of partitions π with $|\pi| \to 0$*

$$\sup_{s\in[0,t]} \left\| S^{\leq K}(\mathbb{X}^\pi)_{[0,s]} - S^{\leq K}(\mathbb{X})_{[0,s]} \right\| \xrightarrow{\mathbb{P}} 0, \quad |\pi| \to 0,$$

which immediately implies

$$S^{\leq K}(\mathbb{X}^\pi)_{[0,t]} \xrightarrow{\mathbb{P}} S^{\leq K}(\mathbb{X})_{[0,t]}, \quad |\pi| \to 0.$$

Proof. See Appendix 4.B.1 □

Remark 4.3.9. We abuse notation slightly be writing π for the restriction of π to $[0,t]$.

4.4 Estimating the expected signatures of a canonical geometric stochastic process

In the following, let $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ be a d-dimensional canonical geometric stochastic process, as defined in Definition 4.3.1. Let π denote a signature-defining sequence of partitions of $[0,T]$ with

[5] Recall that, for a (one-dimensional) semimartingale $\mathbb{X} = \{X_t,\ t \geq 0\}$ and suitable integrand $\mathbb{Y} = \{Y_t,\ t \geq 0\}$, the Stratonovich integral is defined (Protter, 1990, Chapter II.7) in terms of the Itô integral as

$$\int_0^t Y_{s-} \circ \mathrm{d}X_s := \int_0^t Y_{s-}\mathrm{d}X_s + \frac{1}{2}[Y,X]^c_t.$$

In the case of continuous $\mathbb{X}$ and $\mathbb{Y}$ this simplifies to

$$\int_0^t Y_s \circ \mathrm{d}X_s = \int_0^t Y_s \mathrm{d}X_s + \frac{1}{2}[Y,X]_t = \int_0^t Y_s \mathrm{d}X_s + \frac{1}{2}\langle Y,X\rangle_t.$$

For a multi-dimensional semimartingale, the integral can thence be defined component-wise.

mesh $|\pi| \to 0$, cf. Definition 4.3.6, and let $\mathbb{X}^\pi = \{\mathbf{X}_t^\pi,\ t \in [0,T]\}$ denote the linear approximation of $\mathbb{X}$ over π. Assuming we have access to N "copies" of $\mathbb{X}^\pi$ denoted by $\{\mathbb{X}^{\pi,n}\}_{n=1}^N$ we consider the estimator

$$\hat{\phi}_{\mathbf{I}}^{\pi,n}(T) := \frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]}, \tag{4.4.1}$$

where $\mathbf{I}$ is a finite collection of words. Assuming the target expected signature

$$\phi_{\mathbf{I}}(T) := \mathbb{E}[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}], \tag{4.4.2}$$

exists, i.e. $\phi_{\mathbf{I}}(T) < \infty$. We can decompose

$$\begin{aligned}\hat{\phi}_{\mathbf{I}}^{\pi,n}(T) - \phi_{\mathbf{I}}(T) = \frac{1}{N}\sum_{n=1}^{N}\left(S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]} - S^{\mathbf{I}}(\mathbb{X}^{n})_{[0,T]}\right) \\ + \left(\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^{n})_{[0,T]} - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right]\right).\end{aligned} \tag{4.4.3}$$

Under suitable conditions, we will prove the estimator $\hat{\phi}_{\mathbf{I}}^{\pi,n}(T)$ is consistent and asymptotically normal for $\phi_{\mathbf{I}}(T)$ by showing that

1. each summand in the first term converges to zero in L^m in the in-fill asymptotics $|\pi| \to 0$ (in Section 4.4.1);
2. the second term, when inflated by $\sqrt{N}$, converges in distribution to a normal random variable in the large sample asymptotics $N \to \infty$ (in Section 4.4.2).

Remark 4.4.1. Note that if $\{\mathbf{X}_t,\ t \geq 0\}$ has stationary increments of length T, in the sense discussed in Section 4.4.2, and each observation partition over $[0, NT]$ can be decomposed as

$$\Pi(N) = \pi \cup (T + \pi) \ldots \cup ((N-1)T + \pi),$$

then the set $\{\mathbb{X}^{\pi,n}\}_{n=1}^N$ can be obtained by chopping up and shifting (so that each $\mathbb{X}^{\pi,n}$ starts from zero) a single long observation over $\Pi(N)$. This is why we equivalently refer to the large sample asymptotics $N \to \infty$ as "long-span" asymptotics.

4.4.1 In-fill asymptotics

In this section we explore conditions on the process $\mathbb{X}$ under which the summands in the first term of Equation (4.4.3) converge to zero in L^m as $|\pi| \to 0$, i.e.

$$S^{\mathbf{I}}(\mathbb{X}^\pi)_{[0,T]} \overset{L^m}{\to} S^{\mathbf{I}}(\mathbb{X})_{[0,T]}. \tag{4.4.4}$$

Recall that, by Definition 4.3.6, along the sequence of partitions π this convergence holds in probability, i.e. Equation (4.3.4). We first discuss why convergence in probability is not strong

enough to obtain the desired asymptotic properties for the estimator $\hat{\phi}^{\pi,n}_{\mathbf{I}}(T)$.

We then give conditions on $\mathbb{X}$ which ensure convergence in L^m of the signature terms. This yields the desired in-fill asymptotics, Equation (4.4.4), along any signature-defining sequence of partitions π. Through various examples, we provide verifiable sufficient conditions when $\mathbb{X}$ is assumed to be a martingale, an Itô processes, an Itô diffusions, or a fractional Brownian motion, effectively showcasing the broad applicability of our results.

Why is convergence in probability not sufficient?

We give two motivating examples illustrating why the convergence in probability (4.3.4) might not be strong enough to estimate $\phi_{\mathbf{I}}(T)$. In the first example, $S^{\mathbf{I}}(\mathbb{X}^{\pi})_{[0,T]}$ converges to $S^{\mathbf{I}}(\mathbb{X})_{[0,T]}$ in probability but the limiting signature is not integrable, i.e.

$$\mathbb{E}[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}] = \infty,$$

and hence the estimation problem is ill-posed. This example naturally pushes us to look for conditions on $\mathbb{X}$ that ensure the approximators $S^{\mathbf{I}}(\mathbb{X}^{\pi})_{[0,T]}$ are (square-)integrable and converge in L^m, ensuring the limiting quantity is also (square-)integrable, cf. Theorem 4.4.4 and Theorem 4.4.7. The additional requirement of square integrability will be fundamental for the CLT results of Theorem 4.4.19 and Theorem 4.4.29.

The second example illustrates another shortcoming of working with convergence in probability: even if the limiting quantity is integrable, convergence in probability of each summand does not necessarily imply convergence in probability of the sum. As this example illustrates, one would need explicit rates of convergence in terms of $|\pi|$ for Equation (4.3.4) to get conditions on the relative speed at which $|\pi| \to 0$ and $N \to \infty$ ensuring the first term in the decomposition (4.4.3) vanishes. Moreover, when considering this mode of convergence and moving beyond the i.i.d. setting, one would need to take care of the dependence structure of the summands. On the other hand, L^m convergence (4.4.4) of each summand immediately implies L^m convergence of the sum, regardless of the dependence structure and the relative rate at which $|\pi| \to 0$ and $N \to \infty$. This explains why the L^m convergence result given in Theorem 4.4.7 naturally leads to consistency results, Theorem 4.4.19 and Theorem 4.4.29, which do not impose any joint conditions of $|\pi|$ and N[6].

Example 4.4.2. Let $\mathbb{X} = \{\mathbf{X}_t = \omega_t \in \mathbb{R},\ t \in [0,1]\}$ be defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{P}(\omega^m) = 2^{-m}, \quad m \geq 1,$$

where

$$\omega^m_t = \left(t,\ 4^m \phi(2^m t)\, \mathbb{1}_{[0,2^{-m}]}(t)\right) \in \mathbb{R}^2, \quad t \in [0,1],$$

[6]The asymptotic normality results of Theorem 4.4.19 and Theorem 4.4.29 instead requires a joint condition on $|\pi|$ and N, to ensure the first term in (4.4.3) vanishes faster than $\frac{1}{\sqrt{N}}$.

and $\phi : [0,1] \to \mathbb{R}$ is the tent map

$$\phi(t) = 4t\,\mathbb{1}_{[0,1/2)}(t) + (4-4t)\,\mathbb{1}_{[1/2,1]}(t), \quad t \in [0,1].$$

Let $\{\pi_k,\ k \geq 1\}$ denote a sequence of dyadic partitions of $[0,1]$ with $|\pi_k| = 2^{-k}$. Then one has

$$S^{(2,1)}(\mathbb{X})_{[0,1]}(\omega^m) = 2^m \int_0^1 \phi(s)\mathrm{d}s = 2^m, \quad m \geq 1,$$

and

$$S^{(2,1)}(\mathbb{X}^{\pi_k})_{[0,1]}(\omega^m) = \begin{cases} 0, & k \leq m, \\ 2^m, & k > m. \end{cases}$$

Hence

$$\mathbb{P}\left(\left|S^{(2,1)}(\mathbb{X}^{\pi_k})_{[0,1]} - S^{(2,1)}(\mathbb{X})_{[0,1]}\right| > \epsilon\right) = \sum_{m \geq k} 2^{-m} = 2^{1-k} \to 0, \quad k \to \infty.$$

While

$$\mathbb{E}[S^{(2,1)}(\mathbb{X})_{[0,1]}] = \sum_{m \geq 1} 2^m 2^{-m} = \infty.$$

Figure 4.3 gives a graphic visualization of the process $\mathbb{X}$ and its approximations $\mathbb{X}^{\pi_k}$.

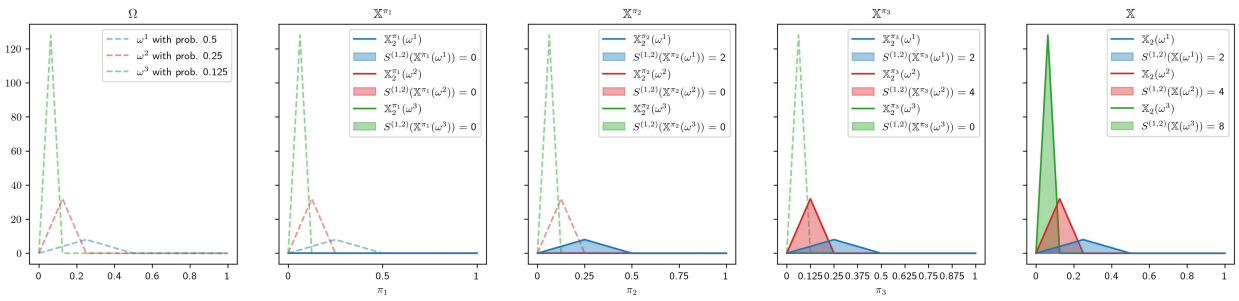


Figure 4.3: Some realizations (corresponding to $\omega^m \in \Omega$ for $m = 1,2,3$) of the process $\mathbb{X}$ and its approximations $\mathbb{X}^{\pi_k}$ for $k = 1,2,3$ as defined in Example 4.4.2.

Example 4.4.3. Assume $\mathbb{X}^n$'s are independent and identically distributed and assume

$$Z_n^\pi := S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} - S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]} \geq 0, \quad \forall \pi, n.$$

Then

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^N Z_n^\pi\right| > \epsilon\right) = \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^N |Z_n^\pi| > \epsilon\right) &= 1 - \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^N |Z_n^\pi| \leq \epsilon\right) \\ &\geq 1 - \mathbb{P}\left(\max_n |Z_n^\pi| \leq N\epsilon\right) = 1 - \mathbb{P}\left(|Z^\pi| \leq N\epsilon\right)^N.\end{aligned}$$

Assume that for all $\epsilon > 0$,

$$\kappa(|\pi|, \epsilon) := \mathbb{P}\left(|Z^\pi| \leq \epsilon\right) \uparrow 1, \quad |\pi| \to 0.$$

Let $\{\pi_k,\ k \geq 1\}$ be such that $|\pi_k| \to 0$ and $\{N_k,\ k \geq 1\}$ be such that $N_k \to \infty$, $k \to \infty$. Then $\forall n \geq 1$,

$$Z_n^{\pi_k} \xrightarrow{\mathbb{P}} 0, \quad k \to \infty,$$

but if

$$\kappa\left(|\pi_k|, N_k\epsilon\right)^{N_k} \not\to 1, \quad N \to \infty,$$

then

$$\frac{1}{N_k}\sum_{n=1}^{N_k} Z_n^{\pi_k} \not\xrightarrow{\mathbb{P}} 0, \quad k \to \infty.$$

We can construct a concrete example as follows: let $\mathbb{X} = \{\mathbf{X}_t = \omega_t \in \mathbb{R},\ t \in [0,1]\}$ be defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{P}(\omega^m) = 2^{-m}, \quad m \geq 1,$$

where

$$\omega_t^m = \left(t,\ 2^m\phi(2^m t)\,\mathbb{1}_{[0,2^{-m}]}(t)\right) \in \mathbb{R}^2, \quad t \in [0,1],$$

and $\phi : [0,1] \to \mathbb{R}$ is the tent map. Let $\{\pi_k,\ k \geq 1\}$ denote a sequence of dyadic partitions of $[0,1]$ with $|\pi_k| = 2^{-k}$. Note this is a very similar setting to Example 4.4.2 (and displayed in Figure 4.3) but the different scaling of the ω^m's ensures

$$\mathbb{E}[S^{(2,1)}(\mathbb{X})_{[0,1]}] = \sum_{m\geq 1} 2^{-m} = 1 < \infty,$$

since

$$S^{(2,1)}(\mathbb{X})_{[0,1]}(\omega^m) = \int_0^1 \phi(s)\mathrm{d}s = 1, \quad \forall m \geq 1.$$

Similar computations as in Example 4.4.2 yield

$$\mathbb{P}\left(\left|S^{(2,1)}(\mathbb{X}^{\pi_k})_{[0,1]} - S^{(2,1)}(\mathbb{X})_{[0,1]}\right| > \epsilon\right) = 2^{1-k} \to 0, \quad k \to \infty,$$

and indeed

$$Z^{\pi_k} := S^{(2,1)}(\mathbb{X})_{[0,1]} - S^{(2,1)}(\mathbb{X}^{\pi_k})_{[0,1]} \geq 0, \quad \forall k \geq 1,$$

with

$$\kappa(|\pi|, \epsilon) = 1 - 2|\pi|.$$

Then, letting $\{N_k, k \geq 1\}$ increase as $N_k = 2^k$ and recalling that $\{\pi_k,\ k \geq 1\}$ was chosen with $|\pi_k| = 2^{-k}$ we have

$$\kappa\left(|\pi_k|, N_k\epsilon\right)^{N_k} = \left(1 - 2^{1-k}\right)^{2^k} \to e^{-2} < 1,$$

and hence

$$Z_n^{\pi_k} \xrightarrow{\mathbb{P}} 0, \quad \text{but} \quad \frac{1}{N_k}\sum_{n=1}^{N_k} Z_n^{\pi_k} \not\xrightarrow{\mathbb{P}} 0, \quad k \to \infty.$$

Under the convergence in probability (4.3.4) what we *can* show is that we can get arbitrarily close in probability to $\phi_{\mathbf{I}}(T)$, defined in Equation (4.4.2), via an appropriate choice of π and N. Fixing $\epsilon > 0$, suppose a weak law of large numbers holds for the long-span asymptotics (e.g. assuming i.i.d. observations of $\mathbb{X}$), i.e. we have that the second term in the decomposition (4.4.3) satisfies

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right]\right\| > \frac{\epsilon}{2}\right) \to 0, \quad N \to \infty.$$

Then by (4.3.4) we can also find π_N such that

$$\mathbb{P}\left(\left\|S^{\mathbf{I}}(\mathbb{X}^{\pi_N})_{[0,T]} - S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right\| > \frac{\epsilon}{2}\right) = \mathcal{O}(N^{-2}),$$

and hence

$$\begin{aligned}
&\mathbb{P}\left(\left\|\hat{\phi}_{\mathbf{I}}^{\pi_N,N}(T) - \phi_{\mathbf{I}}(T)\right\| > \epsilon\right) \\
&\quad \leq N \times \mathbb{P}\left(\left\|S^{\mathbf{I}}(\mathbb{X}^{\pi_N})_{[0,T]} - S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right\| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left\|\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right]\right\| > \frac{\epsilon}{2}\right) \\
&\quad \to 0, \quad N \to \infty,
\end{aligned}$$

by the pigeonhole principle (applied twice) and identical distribution of the $(\mathbb{X}^{\pi,n}, \mathbb{X}^n)$'s. There are two clear problems with this approach:

- The dependence of π_N on N is not given explicitly, hence in applications we cannot know how fine our partition has to be relative to the sample size;
- The choice of π_N also depends on the required precision level ϵ, therefore this statement is weaker than "there exists a sequence (π_N, N) such that $\hat{\phi}_{\mathbf{I}}^{\pi_N,N}(T)$ is consistent".

Motivated by the discussion at the beginning of this section, we hence look for conditions on $\mathbb{X}$ and π under which $S^{\mathbf{I}}(\mathbb{X}^{\pi})_{[0,T]}$ converges to $S^{\mathbf{I}}(\mathbb{X})_{[0,T]}$ in L^m.

From probability to L^m convergence

We work in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ over which the process $\mathbb{X} := \{\mathbf{X}_t,\ t \in [0,T]\}$ is defined. We denote by $L^k = L^k(\Omega, \mathcal{F}, \mathbb{P})$ the space of random variables $Y : \Omega \mapsto E$ such that

$$\|Y\|_{L^k} := \mathbb{E}[\|Y\|^k]^{1/k} < \infty,$$

where $\|\cdot\|$ is a norm over the Banach space E. When E is a finite dimensional vector space the choice of norm is irrelevant (since all norms are equivalent). Unless explicitly stated otherwise, we

select $\|\cdot\|$ to be the ℓ_2-norm over finite dimensional E. In the following, to simplify notation, we write $\mathbf{X}_{s,t} := \mathbf{X}_t - \mathbf{X}_s$ for any $s, t \in [0,T]$ with $s \le t$.

In the following theorem let $\{\mathcal{F}_{s,t},\ [s,t] \subseteq [0,T]\}$ be a family of sigma-algebras such that, for all $[u,v] \subseteq [s,t] \subseteq [0,T]$, $\mathcal{F}_{u,v} \subseteq \mathcal{F}_{s,t}$ and, for each $[s,t] \subseteq [0,T]$, $\mathbf{X}_{s,u}$ is $\mathcal{F}_{s,t}$-measurable for all $u \in [s,t]$. Note that, by contraction property of conditional expectation, the strongest form of Assumption (Aβ) and Assumption (Aγ) is obtained by setting $\mathcal{F}_{s,t} = \sigma(\mathbf{X}_{s,u},\ u \in [s,t])$.

Theorem 4.4.4. *Let $k, m \ge 2$, $p = mk$. Assume $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ satisfies for some $\alpha > 1/4$,*

(Aα) $\|\mathbf{X}_{s,t}\|_{L^p} \lesssim |t-s|^\alpha$ *for all* $0 \le s \le t \le T$.

If $\alpha \in (1/4, 1/2]$ assume further that for some $\beta > 1$,

(Aβ) $\|\mathbb{E}_{\mathcal{F}_{0,s} \vee \mathcal{F}_{t,T}}[\mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t}]\|_{L^{p/2}} \lesssim |t-s|^\beta$ *for all* $0 \le s < u < t \le T$.

If $\alpha \in (1/4, 1/3]$ and $k \ge 3$ assume further that for some $\gamma > 1$,

(Aγ) $\|\mathbb{E}_{\mathcal{F}_{0,s} \vee \mathcal{F}_{t,T}}[\mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t}^{\otimes 2}]\|_{L^{p/3}}, \|\mathbb{E}_{\mathcal{F}_{0,s} \vee \mathcal{F}_{t,T}}[\mathbf{X}_{s,u}^{\otimes 2} \otimes \mathbf{X}_{u,t}]\|_{L^{p/3}} \lesssim |t-s|^\gamma$ *for all* $0 \le s < u < t \le T$.

Let

$$\epsilon = \epsilon(\alpha, \beta, \gamma) = \begin{cases} 2\alpha - 1, & \text{if } \alpha \in (1/2, 1], \\ 3\alpha \wedge \beta - 1, & \text{if } \alpha \in (1/3, 1/2], \\ 4\alpha \wedge \beta \wedge \gamma - 1, & \text{if } \alpha \in (1/4, 1/3], \end{cases} \tag{4.4.5}$$

and consider a signature-defining, cf. Definition 4.3.6, sequence of refining partitions $\{\pi_n,\ n \ge 1\}$ of the interval $[0,T]$ such that

$$\sum_{n \ge 1} |\pi_n|^\epsilon < \infty,$$

then the stronger convergence holds

$$S^k(\mathbb{X}^{\pi_n})_{[0,T]} \overset{L^m}{\to} S^k(\mathbb{X})_{[0,T]}, \quad n \to \infty, \tag{4.4.6}$$

with rate $\mathcal{O}(\sum_{n' \ge n} |\pi_{n'}|^\epsilon)$.

Remark 4.4.5. The requirement

$$\sum_{n \ge 1} |\pi_n|^\epsilon < \infty,$$

restricts the set of signature-defining partitions for which this result holds. If $\{\pi_n,\ n \ge 1\}$ is a sequence of dyadic partitions with $|\pi_n| = 2^{-n}T$ then for any $\epsilon > 0$,

$$\sum_{n \ge 1} |\pi_n|^\epsilon = \sum_{n \ge 1} 2^{-n\epsilon} T^\epsilon = \frac{T^\epsilon}{1 - 2^{-\epsilon}} < \infty,$$

and the rate of convergence is $\mathcal{O}(2^{-n\epsilon})$.

Proof. See Appendix 4.C.1. □

The conditions of the previous theorem are quite difficult to check when $\alpha \leq 1/2$ as they require to compute (or at least bound) the conditional expectation $\mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}}[\mathbf{X}_{s,u}\otimes\mathbf{X}_{u,t}]$. This would require an understanding of the conditional behavior of the stochastic process $\mathbf{X}_u$ for $u \in (s,t)$ given both past information $\mathcal{F}_{0,s}$ and future information $\mathcal{F}_{t,T}$. For models of practical interest, it is often quite difficult to characterize $\mathbf{X}_u$ given its future values. We hence provide an alternative result for the case when $\alpha = 1/2$, which is often the regularity of processes considered in practice, see Example 4.4.12 and Example 4.4.13.

We first prove a useful lemma that allows us to write the k-th level signature of a piecewise linear path as a "causal" sum of lower order signature terms, i.e. preserving time order. This will serve as the basis to derive an in-fill result with assumptions on the regularity of $\mathbb{E}_{\mathcal{F}_{0,s}}[\mathbf{X}_{s,t}]$, a more natural object than $\mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}}[\mathbf{X}_{s,u}\otimes\mathbf{X}_{u,t}]$. As we shall see we will need to sacrifice some generality on the unconditional regularity of $\mathbb{X}$, i.e. we will require $\mathbb{X}$ to satisfy Assumption (Aα) with $\alpha \geq 1/2$.

Lemma 4.4.6. *Let π be a partition of $[0,T]$ and let $\tau \in \pi$. Then for $k \geq 0$ we can write*

$$S^{k+1}(\mathbb{X}^\pi)_{[0,\tau]} = \sum_{i=0}^{k} \frac{1}{(1+i)!} \sum_{[u,v]\in\pi_{[0,\tau]}} S^{k-i}(\mathbb{X}^\pi)_{[0,u]} \otimes \mathbf{X}_{u,v}^{\otimes(i+1)}.$$

Proof. See Appendix 4.C.2. □

In the following theorem let $\{\mathcal{F}_t,\ t\in[0,T]\}$ be a filtration such that $\mathbb{X}$ is $\mathcal{F}_t$-adapted. We denote by $\mathbb{E}_t$ the conditional expectation given $\mathcal{F}_t$. Note that by contraction property of conditional expectation the strongest form of Assumption (Aδ) is obtained by setting $\mathcal{F}_t = \sigma(\mathbf{X}_s,\ s\in[0,t])$.

Theorem 4.4.7. *Let $k,m \geq 2$, $p > mk$ and $\{\pi_n,\ n\geq 1\}$ be a signature-defining sequence of partitions for the process $\mathbb{X} = \{\mathbf{X}_t,\ t\in[0,T]\}$, cf. Definition 4.3.6. Assume $\mathbb{X}$ satisfies Assumption (Aα) for some $\alpha \geq 1/2$ and*

(Aδ) $\|\mathbb{E}_s[\mathbf{X}_{s,t}]\|_{L^p} \lesssim |t-s|^\delta$ *for all $0\leq s\leq t\leq T$,*

for some $\delta \geq 1$. Then the stronger convergence holds

$$S^k(\mathbb{X}^{\pi_n})_{[0,T]} \overset{L^m}{\to} S^k(\mathbb{X})_{[0,T]}, \quad n\to\infty. \tag{4.4.7}$$

If, moreover, $\{\pi_n,\ n\geq 1\}$ is a sequence of refining partitions with $\sum_{n\geq 1}|\pi_n|^\epsilon < \infty$ for

$$\epsilon = \epsilon(\alpha,\delta) = (\alpha+\delta-1)\wedge(2\alpha-1/2), \tag{4.4.8}$$

it is sufficient for $\mathbb{X}$ to satisfy Assumption (Aα) and Assumption (Aδ) with $p = mk$. In this case, we can explicitly characterize the rate of convergence as $\mathcal{O}(\sum_{n'\geq n}|\pi_{n'}|^\epsilon)$.

Remark 4.4.8. Note that if $\{\pi_n,\ n \geq 1\}$ is a sequence of dyadic partitions with $|\pi_n| = 2^{-n}T$ then

$$\sum_{n\geq 1} |\pi_n|^\epsilon = \sum_{n\geq 1} 2^{-n\epsilon} T^\epsilon = \frac{T^\epsilon}{1-2^{-\epsilon}} < \infty,$$

and the rate of convergence is $\mathcal{O}(2^{-n\epsilon})$.

Remark 4.4.9. When $\{\pi_n,\ n \geq 1\}$ is a sequence of refining partitions with $\sum_{n\geq 1} |\pi_n|^\epsilon < \infty$ the proof actually yields the following (stronger) result

$$\sup_{\tau\in\pi_n} \|S^i(\mathbb{X}^{\pi_n})_{[0,\tau]} - S^i(\mathbb{X})_{[0,\tau]}\|_{L^{mk/i}} \lesssim \sum_{n'\geq n} |\pi_{n'}|^\epsilon \to 0, \quad n \to \infty.$$

for all $i \in \{1, \dots, k\}$. In particular, this implies

$$\|S^i(\mathbb{X}^{\pi_n})_{[0,T]} - S^i(\mathbb{X})_{[0,T]}\|_{L^{mk/i}} \lesssim \sum_{n'\geq n} |\pi_{n'}|^\epsilon \to 0, \quad n \to \infty,$$

which shows that the convergence is stronger for lower-order signature terms. Fixing $m \geq 2$, let $k \geq 1$ be the largest value for which $\mathbb{X}$ satisfies Assumption (Aα) for some $\alpha \geq 1/2$ and Assumption (Aδ) for some $\delta \geq 2$ with $p = mk$. Then, k represents the highest order of the signature that can be approximated in L^m by applying Theorem 4.4.7.

Proof. See Appendix 4.C.3. □

We now consider a few important examples of stochastic processes $\mathbb{X}$ and provide sufficient conditions for Assumption (Aα), Assumption (Aβ), Assumption (Aγ), Assumption (Aδ) to be satisfied.

Example 4.4.10 (Martingales). If $\mathbb{X}$ is a (local) martingale[7] starting at zero then by the Burkholder-Davis-Gundy (BDG) inequality we can write a stronger version of assumptions Assumption (Aα) in terms of the quadratic variation of $\mathbb{X}$,

(Aα.M) $\|\langle \mathbf{X}\rangle_{s,t}\|_{L^{p/2}} \lesssim |t-s|^{2\alpha}$ for all $0 \leq s \leq t \leq T$.

Note that for many martingales[8] Assumption (Aα.M) holds with $\alpha = 1/2$ and hence, since Assumption (Aδ) holds trivially, we can usually apply Theorem 4.4.7. Some non-trivial exceptions exist: for example, consider a one-dimensional mean-zero Gaussian martingale over $[0,1]$ with covariance function $C(s \wedge t)$ where C is the Cantor function. This process has quadratic variation $\langle X\rangle_t = C(t)$ and hence – since the Cantor function is Hölder continuous with Hölder exponent

[7] Under Assumption (Aα.M) the local martingale $\mathbb{X}$ is necessarily a true martingale. This follows by noting that applying the BDG inequality, the inclusion of L^p norms and Assumption (Aα.M) yields

$$\mathbb{E}\left[\sup_{t\in[0,T]} \|\mathbf{M}_{0,t}\|\right] \lesssim \mathbb{E}\left[\langle \mathbf{M}\rangle_{0,T}^{1/2}\right] \leq \|\langle \mathbf{M}\rangle_{0,T}\|_{L^{p/2}}^2 \lesssim |T|^{\alpha p} < \infty,$$

and hence by Protter (1990, Theorem I.47) $\mathbb{X}$ is a true martingale.

[8] For martingales that can be written as Brownian integrals see Example 4.4.12 with $\mathbf{b} \equiv 0$.

$\log_3(2)$ – Assumption (Aα.M) is satisfied with $\alpha = \frac{1}{2}\log_3 2 \in (1/4, 1/3]$. In this case, one can easily check that Assumption (Aβ) and Assumption (Aγ) hold by combining the independent increments and martingale property of $\mathbb{X}$.

Example 4.4.11 (Gaussian processes). Let $\mathbb{X}$ be a Gaussian process with mean function $\mu : [0,T] \to \mathbb{R}^d$. If $\mathbb{X}$ satisfies Assumption (Aα) with $p = 2$ and $\alpha > 1/4$ and μ is α-Hölder continuous then it satisfies Assumption (Aα) for any $p \geq 2$. Note that, by the inclusion of norms in L^p spaces, it suffices to show this holds for arbitrarily large p. Choosing $p = 2q$ we can write

$$\begin{aligned}
\|\mathbf{X}_{s,t}\|_{L^{2q}} &\leq \|\mathbf{X}_{s,t} - \mu_{s,t}\|_{L^{2q}} + \|\mu_{s,t}\| \\
&\lesssim \left(\mathbb{E}\left[\left(\sum_{i=1}^{d}\left|X_{s,t}^{(i)} - \mu_{s,t}^{(i)}\right|^2\right)^q\right]\right)^{1/2q} + |t-s|^\alpha \\
&= \left(\sum_{\substack{q_1+\ldots+q_d=q \\ q_1,\ldots,q_d\geq 0}} \binom{q}{q_1,\ldots,q_d} \mathbb{E}\left[\left|X_{s,t}^{(1)} - \mu_{s,t}^{(1)}\right|^{2q_1} \cdots \left|X_{s,t}^{(d)} - \mu_{s,t}^{(d)}\right|^{2q_d}\right]\right)^{1/2q} + |t-s|^\alpha \\
&\overset{(i)}{=} \left(\sum_{\substack{q_1+\ldots+q_d=q \\ q_1,\ldots,q_d\geq 0}} \binom{q}{q_1,\ldots,q_d} \sum_{p\in P^2_{2q_1,\ldots,2q_d}} \prod_{\{i,j\}\in p} \mathrm{Cov}\left(X_{s,t}^{(i)}, X_{s,t}^{(j)}\right)\right)^{1/2q} + |t-s|^\alpha \\
&\overset{(ii)}{\lesssim} \left(\sum_{\substack{q_1+\ldots+q_d=q \\ q_1,\ldots,q_d\geq 0}} \binom{q}{q_1,\ldots,q_d} \sum_{p\in P^2_{2q_1,\ldots,2m_k}} \prod_{\{i,j\}\in p} |t-s|^{2\alpha}\right)^{1/2q} \lesssim |t-s|^\alpha,
\end{aligned}$$

where in (i) we apply Isserlis' Theorem[9] denoting by $P^2_{2q_1,\ldots,2q_d}$ the set of all the pairings of $S = \{1\}^{2q_1} \cup \{2\}^{2q_2} \cup \cdots \cup \{d\}^{2q_d}$, i.e. all distinct ways of partitioning S into $q_1 + \ldots + q_d = q$ pairs, and in (ii) the fact that Assumption (Aα) with $p = 2$ and $\alpha > 1/4$ and α-Hölder continuity of μ implies for all $i, j = 1, \ldots, d$,

$$\begin{aligned}
\mathrm{Cov}(X_{s,t}^{(i)}, X_{s,t}^{(j)}) &\leq |\mathrm{Cov}(X_{s,t}^{(i)}, X_{s,t}^{(j)})| \\
&\leq |\mathbb{E}[X_{s,t}^{(i)} X_{s,t}^{(j)}]| + |\mu_{s,t}^{(i)}\mu_{s,t}^{(j)}| \\
&\leq \mathbb{E}[|X_{s,t}^{(i)} X_{s,t}^{(j)}|] + |\mu_{s,t}^{(i)}||\mu_{s,t}^{(j)}| \\
&\leq \mathbb{E}[|X_{s,t}^{(i)}|^2]^{1/2}\mathbb{E}[|X_{s,t}^{(j)}|^2]^{1/2} + |\mu_{s,t}^{(i)}||\mu_{s,t}^{(j)}| \\
&\leq \|\mathbf{X}_{s,t}\|^2 + \|\mu_{s,t}\|^2 \lesssim |t-s|^{2\alpha}.
\end{aligned}$$

[9] If $\{Z_i,\ i \in \mathcal{I}\}$ is a collection of jointly normal mean-zero random variables, then

$$\mathbb{E}[\prod_{i\in\mathcal{I}} Z_i] = \sum_{p\in P^2_{\mathcal{I}}} \prod_{\{i,j\}\in p} \mathrm{Cov}(Z_i, Z_j),$$

where $P^2_{\mathcal{I}}$ denotes the set of all pairings of $\mathcal{I}$, i.e. all distinct ways of partitioning $\mathcal{I}$ into pairs (note that if $|\mathcal{I}|$ is odd then $P^2_{\mathcal{I}}$ is the empty set).

Example 4.4.12 (Itô processes). We consider the case where $\mathbb{X}$ is an Itô process, i.e. satisfies

$$\mathbf{X}_t = \mathbf{X}_0 + \int_0^t \mathbf{b}_s \, \mathrm{d}s + \int_0^t V_s \, \mathrm{d}\mathbf{W}_s, \quad t \in [0,T],$$

where $\mathbf{b} = \{\mathbf{b}_t,\ t \in [0,T]\}$ and $V = \{V_t,\ t \in [0,T]\}$ are progressively measurable d- and $d \times q$-dimensional processes such that

$$\sup_{s\in[0,T]} \|\mathbf{b}_s\|_{L^p}, \ \sup_{s\in[0,T]} \|V_s\|_{L^p} < \infty, \tag{4.4.9}$$

$\mathbb{W} = \{\mathbf{W}_t,\ t \geq 0\}$ is a q-dimensional Brownian motion and $\mathbf{X}_0 \in L^p$. The assumptions on $\mathbf{b}$ and V imply that for all $0 \leq s \leq t \leq T$,

$$\left\| \int_s^t \mathbf{b}_u \mathrm{d}u \right\|_{L^p} \leq \int_s^t \|\mathbf{b}_u\|_{L^p} \, \mathrm{d}u \tag{4.4.10}$$

$$\leq \left(\sup_{u\in[0,T]} \|\mathbf{b}_u\|_{L^p} \right) |t-s|, \tag{4.4.11}$$

by Minkowski's integral inequality, and

$$\left\| \int_s^t V_u \mathrm{d}\mathbf{W}_u \right\|_{L^p} \lesssim \mathbb{E}\left[\left(\mathrm{tr} \int_s^t V_u V_u^{\mathrm{T}} \mathrm{d}u \right)^{p/2} \right]^{1/p} = \left\| \int_s^t \|V_u\|^2 \mathrm{d}u \right\|_{L^{p/2}}^{1/2}$$

$$\leq \left(\int_s^t \|V_u\|_{L^p}^2 \, \mathrm{d}u \right)^{1/2} \tag{4.4.12}$$

$$\leq \left(\sup_{u\in[0,T]} \|V_u\|_{L^p} \right) |t-s|^{1/2}, \tag{4.4.13}$$

by Burkholder-Davis-Gundy (BDG) inequality, the formula for the quadratic variation of the Itô integral, and Minkowski integral inequality.

We can show Assumption (Aα) holds with $\alpha = 1/2$ by noting that for all $0 \leq s \leq t \leq T$,

$$\|\mathbf{X}_{s,t}\|_{L^p} \leq \left\| \int_s^t \mathbf{b}_u \mathrm{d}u \right\|_{L^p} + \left\| \int_s^t V_u \mathrm{d}\mathbf{W}_u \right\|_{L^p} \lesssim |t-s| + |t-s|^{1/2} \lesssim |t-s|^{1/2},$$

by combining bounds (4.4.11) and (4.4.13).

Next, we can show Assumption (Aδ) holds with $\delta = 1$ by noting that for all $0 \leq s \leq t \leq T$,

$$\|\mathbb{E}_s[\mathbf{X}_{s,t}]\|_{L^p} = \left\| \mathbb{E}_s \left[\int_s^t \mathbf{b}_u \mathrm{d}u \right] \right\|_{L^p} \leq \left\| \int_s^t \mathbf{b}_u \mathrm{d}u \right\|_{L^p} \lesssim |t-s|,$$

where we use the martingale property of Itô integrals, contractive property of conditional expectation and the bound (4.4.11).

Example 4.4.13 (Itô diffusions). We consider the case where $\mathbb{X}$ is a (possibly time-inhomogeneous)

Itô diffusion, i.e. satisfies

$$d\mathbf{X}_t = f(t, \mathbf{X}_t)\mathrm{d}t + \sigma(t, \mathbf{X}_t)\mathrm{d}\mathbf{W}_t, \quad t \in [0, T],$$

where $\mathbb{W} = \{\mathbf{W}_t,\ t \in [0,T]\}$ is a q-dimensional Brownian motion, $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d, \sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times q}$ and $\mathbf{X}_0 \in L^p$. Itô diffusions form a subclass of Itô processes, which we already covered in Example 4.4.12. Here, we give conditions specific to Itô diffusions – i.e. in terms of f and σ – which imply condition (4.4.9).

- If f and σ are uniformly bounded on $[0,T] \times \mathbb{R}^d$ then condition (4.4.9) immediately holds.
- Assume f and σ are Lipschitz continuous, i.e. for all $s, t \in [0,T]$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\begin{aligned}\|f(t,\mathbf{x}) - f(s,\mathbf{y})\| &\le K_f\|(t,\mathbf{x}) - (s,\mathbf{y})\| \le K_f(|t-s| + \|\mathbf{x}-\mathbf{y}\|),\\ \|\sigma(t,\mathbf{x}) - \sigma(s,\mathbf{y})\| &\le K_\sigma\|(t,\mathbf{x}) - (s,\mathbf{y})\| \le K_\sigma(|t-s| + \|\mathbf{x}-\mathbf{y}\|).\end{aligned}$$

  Then, for all $0 \le s \le t \le T$, we can bound

$$\begin{aligned}
&\|\mathbf{X}_{s,t}\|_{L^p}\\
&= \left\|\int_s^t f(u,\mathbf{X}_u)\mathrm{d}u + \int_s^t \sigma(u,\mathbf{X}_u)\mathrm{d}\mathbf{W}_u\right\|_{L^p}\\
&\overset{(i)}{\lesssim} \int_s^t \|f(u,\mathbf{X}_u)\|_{L^p}\,\mathrm{d}u + \left(\int_s^t \|\sigma(u,\mathbf{X}_u)\|_{L^p}^2\,\mathrm{d}u\right)^{1/2}\\
&\overset{(ii)}{\lesssim} \|f(s,\mathbf{X}_s)\|_{L^p}(t-s) + \int_s^t \|f(u,\mathbf{X}_u) - f(s,\mathbf{X}_s)\|_{L^p}\,\mathrm{d}u\\
&\qquad + \|\sigma(s,\mathbf{X}_s)\|_{L^p}(t-s)^{1/2} + \left(\int_s^t \|\sigma(u,\mathbf{X}_u) - \sigma(s,\mathbf{X}_s)\|_{L^p}^2\,\mathrm{d}u\right)^{1/2}\\
&\overset{(iii)}{\lesssim} \Big[\|f(0,\mathbf{0})\| + K_f(s + \|\mathbf{X}_s\|_{L^p})\Big](t-s) + K_f\frac{(t-s)^2}{2} + \int_s^t K_f\|\mathbf{X}_{s,u}\|_{L^p}\,\mathrm{d}u\\
&\quad + \Big[\|\sigma(0,\mathbf{0})\| + K_\sigma(s + \|\mathbf{X}_s\|_{L^p})\Big](t-s)^{1/2} + K_\sigma\left(\int_s^t \|(u,\mathbf{X}_u) - (s,\mathbf{X}_s)\|_{L^p}^2\,\mathrm{d}u\right)^{1/2}\\
&\overset{(iv)}{\lesssim} \Big[\|f(0,\mathbf{0})\| + K_f(s + \|\mathbf{X}_s\|_{L^p})\Big](t-s) + K_f\frac{(t-s)^2}{2} + \int_s^t K_f\|\mathbf{X}_{s,u}\|_{L^p}\,\mathrm{d}u\\
&\qquad + \Big[\|\sigma(0,\mathbf{0})\| + K_\sigma(s + \|\mathbf{X}_s\|_{L^p})\Big](t-s)^{1/2} + K_\sigma\left(\int_s^t |u-s|^2\mathrm{d}u\right)^{1/2}\\
&\qquad + K_\sigma\left(\int_s^t \|\mathbf{X}_{s,u}\|_{L^p}^2\,\mathrm{d}u\right)^{1/2}\\
&\overset{(v)}{\lesssim} (1 \vee \|\mathbf{X}_s\|_{L^p})(t-s)^{1/2} + \int_s^t \|\mathbf{X}_{s,u}\|_{L^p}\,\mathrm{d}u + \left(\int_s^t \|\mathbf{X}_{s,u}\|_{L^p}^2\,\mathrm{d}u\right)^{1/2}\\
&\overset{(vi)}{\lesssim} (1 \vee \|\mathbf{X}_s\|_{L^p})(t-s)^{1/2} + \left(\int_s^t \|\mathbf{X}_{s,u}\|_{L^p}^2\,\mathrm{d}u\right)^{1/2},
\end{aligned}$$

  by using in (i) triangle inequality and Equations (4.4.10) and (4.4.12), in (ii) triangle inequality,

in (iii) Lipschitzianity of f and σ, in (iv) triangle inequality, in (v) for all $0 \leq s \leq t \leq T$, $s \leq T \lesssim 1$ and $(t-s)^{1/2+\epsilon} \leq T^{\epsilon}(t-s)^{1/2} \lesssim (t-s)^{1/2}$ for $\epsilon \geq 0$, and in (vi) Jensen's inequality. Setting $s = 0$, since we assume $\mathbf{X}_0 \in L^p$, this is

$$\|\mathbf{X}_{0,t}\|_{L^p} \lesssim t^{1/2} + \left(\int_0^t \|\mathbf{X}_{0,u}\|_{L^p}^2 \,\mathrm{d}u\right)^{1/2},$$

and we can apply Willett (1964, Lemma 2.2), a nonlinear generalization of the Gronwall inequality, to deduce for all $0 \leq t \leq T$,

$$\begin{aligned}\|\mathbf{X}_{0,t}\|_{L^p} &\lesssim t^{1/2} + \frac{\left(\int_0^t \exp\{-Cs\} s \,\mathrm{d}s\right)^{1/2}}{1 - \sqrt{1 - \exp\{-Ct\}}} \\ &\lesssim T^{1/2} + \frac{T}{1 - \sqrt{1 - \exp\{-CT\}}} < \infty.\end{aligned}$$

We can hence show the condition for Itô processes (4.4.9) holds by noting that $\mathbf{X}_0 \in L^p$ and Lipschitzianity of f and σ imply

$$\begin{aligned}\|f(t, \mathbf{X}_t)\|_{L^p} &\leq \|f(0, \mathbf{0})\| + K_f(|t| + \|\mathbf{X}_0\|_{L^p} + \|\mathbf{X}_{0,t}\|_{L^p}) < \infty, \\ \|\sigma(t, \mathbf{X}_t)\|_{L^p} &\leq \|\sigma(0, \mathbf{0})\| + K_\sigma(|t| + \|\mathbf{X}_0\|_{L^p} + \|\mathbf{X}_{0,t}\|_{L^p}) < \infty,\end{aligned}$$

uniformly in $t \in [0, T]$, and hence Assumption (Aα) and Assumption (Aδ) hold with $\alpha = 1/2$ and $\delta = 1$ respectively.

- Assume f and σ are time-homogeneous such that f is Lipschitz continuous and σ is $1/2$-Hölder continuous, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\begin{aligned}\|f(\mathbf{x}) - f(\mathbf{y})\| &\leq K_f \|\mathbf{x} - \mathbf{y}\|, \\ \|\sigma(\mathbf{x}) - \sigma(\mathbf{y})\| &\leq K_\sigma \|\mathbf{x} - \mathbf{y}\|^{1/2}.\end{aligned}$$

Then, for all $0 \leq s \leq t \leq T$, we can bound

$$\begin{aligned}\|\mathbf{X}_{s,t}\|_{L^p} &= \left\| \int_s^t f(\mathbf{X}_u)\mathrm{d}u + \int_s^t \sigma(\mathbf{X}_u)\mathrm{d}\mathbf{W}_u \right\|_{L^p} \\ &\overset{(i)}{\lesssim} \int_s^t \|f(\mathbf{X}_u)\|_{L^p} \,\mathrm{d}u + \left(\int_s^t \|\sigma(\mathbf{X}_u)\|_{L^p}^2 \,\mathrm{d}u\right)^{1/2} \\ &\overset{(ii)}{\lesssim} \|f(\mathbf{X}_s)\|_{L^p} (t-s) + \int_s^t \|f(\mathbf{X}_u) - f(\mathbf{X}_s)\|_{L^p} \,\mathrm{d}u \\ &\qquad + \|\sigma(\mathbf{X}_s)\|_{L^p} (t-s)^{1/2} + \left(\int_s^t \|\sigma(\mathbf{X}_u) - \sigma(\mathbf{X}_s)\|_{L^p}^2 \,\mathrm{d}u\right)^{1/2} \\ &\overset{(iii)}{\lesssim} \left[\|f(\mathbf{0})\| + K_f \|\mathbf{X}_s\|_{L^p}\right](t-s) + \int_s^t K_f \|\mathbf{X}_{s,u}\|_{L^p} \,\mathrm{d}u\end{aligned}$$

$$+\left[\|\sigma(\mathbf{0})\| + K_\sigma \|\mathbf{X}_s\|_{L^p}^{1/2}\right](t-s)^{1/2} + K_\sigma^{1/2}\left(\int_s^t \|\mathbf{X}_{s,u}\|_{L^p}\mathrm{d}u\right)^{1/2}$$
$$\overset{(iv)}{\lesssim} \left(1 \vee \|\mathbf{X}_s\|_{L^p} \vee \|\mathbf{X}_s\|_{L^p}^{1/2}\right)(t-s)^{1/2} + \int_s^t \|\mathbf{X}_{s,u}\|_{L^p}^2 \,\mathrm{d}u + \left(\int_s^t \|\mathbf{X}_{s,u}\|_{L^p}\mathrm{d}u\right)^{1/2},$$

by proceeding as in the previous case. Setting $s = 0$, since we assume $\mathbf{X}_0 \in L^p$, this is

$$\|\mathbf{X}_{0,t}\|_{L^p} \lesssim t^{1/2} + \int_0^t \|\mathbf{X}_{0,u}\|_{L^p}\,\mathrm{d}u + \left(\int_0^t \|\mathbf{X}_{0,u}\|_{L^p}\,\mathrm{d}u\right)^{1/2},$$

and we can apply Dragomir (2003, Theorem 41), another nonlinear generalization of the Gronwall inequality, to deduce for all $0 \le t \le T$,

$$\|\mathbf{X}_{0,t}\|_{L^p} \lesssim \left(\sqrt{t} + e^{Ct} + \int_0^t \sqrt{u}e^{Cu}\mathrm{d}u\right)e^{Ct} \lesssim \left(\sqrt{T} + e^{CT} + \int_0^T \sqrt{u}e^{Cu}\mathrm{d}u\right)e^{CT} < \infty.$$

We can hence show the condition for Itô processes (4.4.9) holds by noting that $\mathbf{X}_0 \in L^p$ and the conditions on f and σ imply

$$\|f(\mathbf{X}_t)\|_{L^p} \le \|f(\mathbf{0})\| + K_f(\|\mathbf{X}_0\|_{L^p} + \|\mathbf{X}_{0,t}\|_{L^p}) < \infty,$$
$$\|\sigma(\mathbf{X}_t)\|_{L^p} \le \|\sigma(\mathbf{0})\| + K_\sigma(\|\mathbf{X}_0\|_{L^p} + \|\mathbf{X}_{0,t}\|_{L^p})^{1/2} < \infty,$$

uniformly in $t \in [0,T]$, and hence Assumption (Aα) and Assumption (Aδ) hold with $\alpha = 1/2$ and $\delta = 1$ respectively.

Example 4.4.14 (Semimartingales). Assume $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a continuous semimartingale with decomposition

$$\mathbf{X}_t = \mathbf{X}_0 + \mathbf{M}_t + \mathbf{A}_t, \quad t \in [0,T],$$

where $\mathbb{M} = \{\mathbf{M}_t,\ t \in [0,T]\}$ is a continuous local martingale and $\mathbb{A} = \{\mathbf{A}_t,\ t \in [0,T]\}$ is a continuous adapted process of locally bounded variation with $\mathbf{M}_0 = \mathbf{A}_0 = 0$. Assume Assumption (A$\alpha$.M) holds for $\mathbb{X}$ (or, equivalently, for $\mathbb{M}$) for some $\alpha' \ge 1/2$ and

(Aδ.S) $\|\mathbf{A}_{s,t}\|_{L^p} \lesssim |t-s|^\delta$ for all $0 \le s \le t \le T$,

holds for some $\delta \ge 1$. Then $\mathbb{X}$ satisfies Assumption (Aα) with $\alpha = \alpha' \wedge \delta \ge 1/2$ since

$$\|\mathbf{X}_{s,t}\|_{L^p} \le \|\mathbf{M}_{s,t}\|_{L^p} + \|\mathbf{A}_{s,t}\|_{L^p} \lesssim \|\langle \mathbf{M}_{s,t}\rangle\|_{L^{p/2}} + |t-s|^\delta \lesssim |t-s|^{\alpha\wedge\delta}, \quad 0 \le s \le t \le T,$$

and Assumption (Aδ) for $\delta \ge 1$ since

$$\|\mathbb{E}_s[\mathbf{X}_{s,t}]\|_{L^p} = \|\mathbb{E}_s[\mathbf{M}_{s,t}] + \mathbb{E}_s[\mathbf{A}_{s,t}]\|_{L^p} = \|\mathbb{E}_s[\mathbf{A}_{s,t}]\|_{L^p} \lesssim |t-s|^\delta, \quad 0 \le s \le t \le T,$$

since, by Example 4.4.10, $\mathbb{M}$ is necessarily a true martingale.

Example 4.4.15 (Fractional Brownian motion). Let $\mathbb{X}^H = \{X_t^H,\ t \in [0,T]\}$ be a (one-dimensional) fractional Brownian motion (fBm) with Hurst parameter $H > 1/4$. $\mathbb{X}^H$ is a Gaussian process with stationary increments $X_{s,t}^H \sim \mathcal{N}(0, |t-s|^{2H})$ and hence one can easily check Assumption (Aα) holds with $\alpha = H$ since by Example 4.4.11 it is sufficient to note that

$$\|X_{s,t}^H\|_{L^2} = |t-s|^H, \quad 0 \le s \le t \le T.$$

When $H \in (1/2, 1)$ then all the assumptions of Theorem 4.4.4 are satisfied.

When $H = 1/2$ the process reduces to a standard Brownian motion and hence the strongest form of Assumption (Aβ), or Assumption (Aδ), can be easily proved to hold by using the martingale and independent increments properties.

When $H \in (1/4, 1/2)$ to apply Theorem 4.4.4 we would further require $\mathbb{X}^H$ to satisfy Assumption (Aβ) with respect to a family of sigma algebras $\{\mathcal{F}_{s,t},\ [s,t] \subseteq [0,T]\}$. We can show that Assumption (Aβ) is not satisfied when choosing $\mathcal{F}_{s,t} = \sigma(B_u,\ u \le t)$ for $0 \le s \le t \le T$ where $\mathbb{B} = \{B_t,\ t \in \mathbb{R}\}$ is the driving Brownian motion in the fBm representation[10] of Mandelbrot and Van Ness (1968), i.e.

$$X_t^H = \kappa_H \int_{-\infty}^{+\infty} \left((t-u)_+^{H-1/2} - (-u)_+^{H-1/2}\right) \mathrm{d}B_u.$$

We can write the increments of the process as

$$\begin{aligned} X_{s,t}^H &= \kappa_H \int_{-\infty}^{+\infty} \left((t-u)_+^{H-1/2} - (s-u)_+^{H-1/2}\right) \mathrm{d}B_u \\ &= \kappa_H \int_{-\infty}^{s} \left((t-u)^{H-1/2} - (s-u)^{H-1/2}\right) \mathrm{d}B_u + \kappa_H \int_s^t (t-u)^{H-1/2} \mathrm{d}B_u. \end{aligned}$$

Since $\mathcal{F}_{0,s} \perp \sigma(B_u - B_s, u \ge s)$ we have that for $0 \le s \le t \le T$,

$$\mathbb{E}_{\mathcal{F}_{0,s}}[X_{s,t}^H] = \kappa_H \int_{-\infty}^{s} \left((t-u)^{H-1/2} - (s-u)^{H-1/2}\right) \mathrm{d}B_u,$$

and for $0 \le s \le u \le t \le T$,

$$\begin{aligned} &\mathbb{E}_{\mathcal{F}_{0,s}}[X_{s,u}^H \mathbb{E}_{\mathcal{F}_{0,u}}[X_{u,t}^H]] \\ &\quad = \kappa_H^2 \mathbb{E}_{\mathcal{F}_{0,s}}\left[\left(\int_{-\infty}^{s} (u-v)^{H-1/2} - (s-v)^{H-1/2} \mathrm{d}B_v + \int_s^u (u-v)^{H-1/2} \mathrm{d}B_v\right)\right. \\ &\qquad\qquad \left. \times \left(\int_{-\infty}^{u} (t-v)^{H-1/2} - (u-v)^{H-1/2} \mathrm{d}B_v\right)\right] \\ &\quad = \kappa_H^2 \left(\int_{-\infty}^{s} (u-v)^{H-1/2} - (s-v)^{H-1/2} \mathrm{d}B_v\right)\left(\int_{-\infty}^{s} (t-v)^{H-1/2} - (u-v)^{H-1/2} \mathrm{d}B_v\right) \end{aligned}$$

[10] An alternative representation of fBm is can be obtained with the Volterra kernel, as in Nualart and Tindel (2011). In this case, the filtrations generated by $\mathbb{B} = \{B_t,\ t \ge 0\}$ and $\mathbb{X}^H = \{X_t^H,\ t \ge 0\}$ coincide; however, deriving explicit bounds for the kernel presents a significantly more complex challenge.

$$+ \kappa_H^2 \mathbb{E}\Big[\Big(\int_s^u (u-v)^{H-1/2} \mathrm{d}B_v\Big)\Big(\int_s^u (t-v)^{H-1/2} - (u-v)^{H-1/2} \mathrm{d}B_v\Big)\Big]$$
$$= \kappa_H^2 \Big(\int_{-\infty}^s (u-v)^{H-1/2} - (s-v)^{H-1/2} \mathrm{d}B_v\Big)\Big(\int_{-\infty}^s (t-v)^{H-1/2} - (u-v)^{H-1/2} \mathrm{d}B_v\Big)$$
$$+ \kappa_H^2 \int_s^u (u-v)^{H-1/2} \Big((t-v)^{H-1/2} - (u-v)^{H-1/2}\Big) \mathrm{d}v.$$

And hence

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}_{\mathcal{F}_{0,s}}[X^H_{s,u}\mathbb{E}_{\mathcal{F}_{0,u}}[X^H_{u,t}]]\right|\right] \\
&\geq \kappa_H^2 \Big|\mathbb{E}\Big[\Big(\int_{-\infty}^s (u-v)^{H-1/2} - (s-v)^{H-1/2} \mathrm{d}B_v\Big)\Big(\int_{-\infty}^s (t-v)^{H-1/2} - (u-v)^{H-1/2} \mathrm{d}B_v\Big)\Big] \\
&\qquad + \int_s^u (u-v)^{H-1/2} \Big((t-v)^{H-1/2} - (u-v)^{H-1/2}\Big) \mathrm{d}v\Big| \\
&\geq \kappa_H^2 \Big|\int_{-\infty}^s \big((u-v)^{H-1/2} - (s-v)^{H-1/2}\big)\big((t-v)^{H-1/2} - (u-v)^{H-1/2}\big)\mathrm{d}v \\
&\qquad + \int_s^u (u-v)^{H-1/2} \Big((t-v)^{H-1/2} - (u-v)^{H-1/2}\Big) \mathrm{d}v\Big|.
\end{aligned}$$

Note that $s \leq u \leq t$ implies $(s-v)^{H-1/2} \geq (u-v)^{H-1/2} \geq (t-v)^{H-1/2}$ for all $v \leq s$ and hence the first term in the absolute value is positive while the second is negative. We can show that the second term is strictly larger in magnitude than the first:

$$\begin{aligned}
&\Big|\int_{-\infty}^s \big((u-v)^{H-1/2} - (s-v)^{H-1/2}\big)\big((t-v)^{H-1/2} - (u-v)^{H-1/2}\big)\mathrm{d}v\Big| \\
&\qquad = \int_{-\infty}^s \big((u-v)^{H-1/2} - (s-v)^{H-1/2}\big)\big((t-v)^{H-1/2} - (u-v)^{H-1/2}\big)\mathrm{d}v \\
&\qquad \leq \int_{-\infty}^s \big((u-v)^{H-1/2} - (s-v)^{H-1/2}\big)^2 \mathrm{d}v \\
&\qquad = \left(\frac{2^{-2H}\Gamma(H)\Gamma(1/2+H)}{H\sqrt{\pi}} - \frac{1}{2H}\right)(u-s)^{2H} =: C_{1,H}(u-s)^{2H},
\end{aligned}$$

while

$$\begin{aligned}
&\Big|\int_s^u (u-v)^{H-1/2} \Big((t-v)^{H-1/2} - (u-v)^{H-1/2}\Big) \mathrm{d}v\Big| \\
&\qquad = \int_s^u (u-v)^{H-1/2} \Big((u-v)^{H-1/2} - (t-v)^{H-1/2}\Big) \mathrm{d}v \\
&\qquad \geq \int_s^u \Big((u-v)^{H-1/2} - (t-v)^{H-1/2}\Big)^2 \mathrm{d}v \\
&\qquad \geq \frac{1}{u-s}\Big(\int_s^u (u-v)^{H-1/2} - (t-v)^{H-1/2} \mathrm{d}v\Big)^2 \\
&\qquad \geq \frac{1}{u-s}\frac{1}{(H+1/2)^2}\Big(-(u-s)^{H+1/2} + (t-s)^{H+1/2} - (t-u)^{H+1/2}\Big)^2 \\
&\qquad \geq \frac{((\lambda+1)^{H+1/2} - \lambda^{H+1/2} - 1)^2}{(H+1/2)^2}(u-s)^{2H} =: C_{2,H,\lambda}(u-s)^{2H},
\end{aligned}$$

when $(t-u) = \lambda(u-s)$ (only used in the last step). Choosing λ sufficiently large[11] so that $C_{1,H} < C_{2,H,\lambda}$ we can hence conclude that for $H \in (1/4, 1/2)$ for all $0 \leq s < u < t \leq T$ with $(t-u) = \lambda(u-s)$ we have

$$\begin{aligned}\|\mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}}[X^H_{s,u}X^H_{u,t}]\|_{L^{p/2}} &\geq \|\mathbb{E}_{\mathcal{F}_{0,s}}[X^H_{s,u}\mathbb{E}_{\mathcal{F}_{0,u}}[X^H_{u,t}]]\|_{L^{p/2}} \\ &\geq \|\mathbb{E}_{\mathcal{F}_{0,s}}[X^H_{s,u}\mathbb{E}_{\mathcal{F}_{0,u}}[X^H_{u,t}]]\|_{L^1} > c_H(t-s)^{2H},\end{aligned}$$

for some constant $c_H = \kappa_H^2(C_{2,H,\lambda} - C_{1,H}) > 0$ and hence Assumption (Aβ) does not hold. This serves as an example where the assumptions of Theorem 4.4.4 cannot be easily checked to hold. One might still try to show Assumption (Aβ) (and Assumption (Aγ)) hold under a different family of sigma-algebras $\{\mathcal{F}_{s,t},\ [s,t] \subseteq [0,T]\}$ or establish the L^m-convergence of the signatures in an alternative way.

Example 4.4.16 (Add-time transform). In many practical applications the add-time transformation is applied to the data, i.e. the d-dimensional martingale $\mathbb{X}$ is augmented with a strictly monotonic deterministic function $t \mapsto \psi(t)$ to form the $(d+1)$-dimensional process $\hat{\mathbb{X}} = \{(\psi(t), \mathbf{X}_t),\ t \in [0,T]\}$. Introducing the auxiliary time variable $\psi(t)$ "resolves" tree-like equivalences, ensuring (bounded variation) paths are identifiable from their signature (Hambly and Lyons, 2005). If the function $\psi(t)$ is Lipschitz continuous over $[0,T]$ (as is usually the case) then it can be easily shown that

- $\hat{\mathbb{X}}$ satisfies Assumption (Aα), Assumption (Aβ) and/or Assumption (Aδ) if $\mathbb{X}$ satisfies the corresponding condition(s); and
- any sequence of partitions that is signature defining for $\mathbb{X}$ is also signature defining for $\hat{\mathbb{X}}$ (since any sequence of partitions π is signature-defining for ψ).

In most cases the auxiliary time variable $\psi(t)$ is taken to be $\psi(t) = t$ for $t \in [0,T]$.

4.4.2 Long-span asymptotics

Having established results for the in-fill asymptotics, we now investigate the large sample properties of the second term in the decomposition (4.4.3). We will focus on two ways in which one might obtain "copies" of $\mathbb{X}^\pi = \{\mathbf{X}^\pi_t,\ t \in [0,T]\}$ to estimate the expected signature $\phi_{\mathbf{I}}(T)$.

(ind) Repeatedly observe $\mathbb{X}$ through N independent experiments over partitions $\pi_{N,1}, \ldots, \pi_{N,N}$ of the interval $[0,T]$, in which case the "underlying" signatures $S(\mathbb{X}^n)_{[0,T]}$ for $n = 1, \ldots, N$ are independent and identically distributed.

(chop) Chop-up (and shift in time) the process $\mathbb{X}^{\Pi(N)} = \{\mathbf{X}^{\Pi(N)}_t,\ t \in [0,NT]\}$ into $\{\mathbb{X}^{\pi_{N,n}}\}_{n=1}^N$ where

$$\Pi(N) := \pi_{N,1} \cup (T + \pi_{N,2}) \cup \ldots \cup ((N-1)T + \pi_{N,N}),$$

[11] For example, $\lambda = 4$ works for all $H \in (1/4, 1/2)$.

and each $\pi_{N,n}$ is a partition of $[0,T]$, i.e. $\mathbb{X}^{\pi_{N,n}} = \{\mathbf{X}_t^{\pi_{N,n}},\ t \in [0,T]\}$ is the linear interpolation of

$$\mathbb{X}^n := \{\mathbf{X}_t^n,\ t \in [0,T]\} \in C([0,T];\mathbb{R}^d), \quad \text{where} \quad \mathbf{X}_t^n = \mathbf{X}_{(n-1)T,(n-1)T+t},\ t \in [0,T],$$

over the partition $\pi_{N,n}$. In this setting, it is natural to assume the discrete-parameter process $\{\mathbb{X}^n,\ n \geq 1\}$ taking values in $C([0,T];\mathbb{R}^d)$ is stationary[12], i.e.

$$(\mathbb{X}^{n_1},\dots,\mathbb{X}^{n_k}) \overset{\mathcal{L}}{=} (\mathbb{X}^{n_1+n},\dots,\mathbb{X}^{n_k+n}), \quad k \in \mathbb{N},\ n_1,\dots,n_k \in \mathbb{N},\ n \geq 0, \tag{4.4.14}$$

to ensure[13] the signatures $S(\mathbb{X}^n)_{[0,T]}$ form a stationary sequence. This implies the signatures are identically distributed and hence the task of estimating $\phi_{\mathbf{I}}(T)$ is easily understood (this is a stronger but more natural assumption than needed to make the estimation task well-posed, see Remark 4.4.18). Note this condition is weaker than requiring $\{\mathbf{X}_t,\ t \geq 0\}$ to be stationary, cf. Proposition 4.4.23.

In both cases, we will be interested in the double asymptotics where, as the number of signature evaluations N increases, the granularity of the discretized paths from which such signatures are computed also increases, i.e.

$$|\Pi(N)| := \max_{1 \leq n \leq N} |\pi_{N,n}| \to 0, \quad N \to \infty.$$

To ensure we are approximating the signature in the in-fill asymptotics, for each $n \geq 1$ we will require the sequence of partitions $\pi_{\cdot,n} := \{\pi_{N,n},\ N \geq n\}$ to be signature-defining, cf. Definition 4.3.6. Note that, when $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a semimartingale, this condition is satisfied by any sequence of partitions with vanishing mesh, cf. Section 4.3.1.

For each $N \geq 1$, we allow the observation partitions $\pi_{N,1},\dots,\pi_{N,N}$ to differ, and hence we briefly revisit the definition of the expected signature estimator (4.4.1) and decomposition (4.4.3). Letting $\{\mathbb{X}^n\}_{n=1}^N$ denote the N "copies" of $\mathbb{X} = \mathbb{X}^1$ underlying the observations $\{\mathbb{X}^{\pi_{N,n}}\}_{n=1}^N$ we consider

$$\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) := \frac{1}{N}\sum_{n=1}^N S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]}, \tag{4.4.15}$$

[12] Throughout we will use the term *stationary* to denote a *strictly stationary* process.

[13] Taking π to be a signature-defining sequence of partitions for the signature of $\mathbb{X}$, as defined in Definition 4.3.6, one has

$$\begin{aligned}\left(S^{\mathbf{I}}(\mathbb{X}^{n_1})_{[0,T]},\dots,S^{\mathbf{I}}(\mathbb{X}^{n_j})_{[0,T]}\right) &= \mathbb{P}\text{–}\lim_{|\pi|\to 0}\left(S^{\mathbf{I}}(\mathbb{X}^{\pi,n_1})_{[0,T]},\dots,S^{\mathbf{I}}(\mathbb{X}^{\pi,n_j})_{[0,T]}\right)\\ &\overset{\mathcal{L}}{=} \mathbb{P}\text{–}\lim_{|\pi|\to 0}\left(S^{\mathbf{I}}(\mathbb{X}^{\pi,n+n_1})_{[0,T]},\dots,S^{\mathbf{I}}(\mathbb{X}^{\pi,n+n_j})_{[0,T]}\right)\\ &= \left(S^{\mathbf{I}}(\mathbb{X}^{n+n_1})_{[0,T]},\dots,S^{\mathbf{I}}(\mathbb{X}^{n+n_j})_{[0,T]}\right),\end{aligned}$$

where the second equality follows by stationarity of $\{\mathbb{X}^n,\ n \geq 1\}$ and $S(\mathbb{X}^\pi)_{[0,T]} = \bigotimes_{[u,v]\in\pi} \exp_\otimes \mathbf{X}_{u,v}$.

where $\mathbf{I}$ is a finite collection of words. By writing

$$\begin{aligned}\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) - \phi_{\mathbf{I}}(T) = \frac{1}{N}\sum_{n=1}^{N}\left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - S^{\mathbf{I}}(\mathbb{X}^{n})_{[0,T]}\right) \\ + \left(\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^{n})_{[0,T]} - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right]\right),\end{aligned} \tag{4.4.16}$$

we note that if one establishes, for example through Theorem 4.4.4 or Theorem 4.4.7, that for all $n \geq 1$,

$$\left\|S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - S^{\mathbf{I}}(\mathbb{X})\right\|_{L^m} \to 0, \quad N \to \infty, \tag{4.4.17}$$

and $S^{\mathbf{I}}(\mathbb{X}^n) \overset{\mathcal{L}}{=} S^{\mathbf{I}}(\mathbb{X})$, $n \geq 1$, then, as long as

$$|\Pi(N)| = \max_{1\leq n\leq N} |\pi_{N,n}| \to 0, \quad N \to \infty,$$

the first term in Equation (4.4.16) vanishes in L^m (and thus also in $\mathbb{P}$). To ensure such in-fill discretization error vanishes when inflated by $\sqrt{N}$, as required to study the asymptotic distribution of the second term, we will additionally need to impose

$$\sqrt{N}\max_{1\leq n\leq N} \rho(N, \pi_{\cdot,n}) \to 0, \quad N \to \infty,$$

where $\{\rho(N, \pi_{\cdot,n}),\ N \geq n\}$ is the rate of convergence of (4.4.17) for the signature-defining sequence of partitions $\pi_{\cdot,n} = \{\pi_{N,n},\ N \geq n\}$. When each $\pi_{\cdot,n}$ is a sequence of refinements the rate of convergence may be obtained explicitly through Theorem 4.4.4 or the second part of Theorem 4.4.7 in the form

$$\rho(N, \pi_{\cdot,n}) = \sum_{n'\geq N} |\pi_{n',n}|^{\epsilon}, \quad N \geq n,$$

with $\epsilon > 0$. It thus remains to determine the limiting behavior of the second term as $N \to \infty$, i.e. the long-span asymptotics. In the following sections, we state combined consistency and asymptotic normality results referring to the assumptions of Theorem 4.4.7 for the in-fill asymptotics but the same results hold with the assumptions of Theorem 4.4.4.

Remark 4.4.17. Here, we focus on estimators obtained by averaging out signatures computed over non-overlapping time intervals, i.e. in the estimator (4.4.15) the processes $\{\mathbb{X}^{\pi_{N,n}},\ n = 1, \ldots, N\}$ are sampled over the disjoint time intervals $[(n-1)T, nT]$. One could alternatively consider estimators obtained by computing the signatures over partially overlapping time intervals: for example, when the process is observed over an evenly spaced grid $\Pi(N)$ with mesh Δ, the signatures can be computed from observations of the process over $[0, T], [\Delta, T + \Delta], \ldots, [(N-1)T, NT]$. For ease of exposition, we restrict our discussion in this chapter to non-overlapping time intervals. However, the asymptotic results for dependent observations – e.g., Theorem 4.4.29 and Theorem 4.4.34 – carry over directly to the overlapping case. On the other hand, the two approaches may exhibit

quantitatively different behaviour in the finite-sample regime: while using partially overlapping observations increases the sample size, it also introduces greater sample correlation. A detailed analysis of this trade-off is an interesting direction for future research but is beyond the scope of this work.

Remark 4.4.18. In this work, we consider the setting where the sequence of successive observations of the process, namely $\{\mathbb{X}^n,\ n \geq 1\}$, is assumed to be stationary. One could alternatively work in the more general setting where such sequence converges ergodically to a stationary regime. In this case, one would aim to estimate the expected signature of such limiting stationary distribution. As a concrete example, one could consider the time-homogeneous Itô diffusion discussed in Example 4.4.32 starting in an arbitrary initial distribution, i.e. $\mathbf{X}_0 \sim \mu_0$, and converging ergodically to the stationary distribution μ. The estimator (4.4.15) would then estimate the expected signature of $\{\mathbf{X}_t,\ t \in [0,T]\}$ in the stationary regime, i.e. with $\mathbf{X}_0 \sim \mu$.

Independent observations

Assuming the $\mathbb{X}^n$'s, and thus the $S(\mathbb{X}^n)_{[0,T]}$'s, are independent we can apply the classic weak law of large numbers and central limit theorem.

Consistency For convergence in L^2 of the sample average, i.e.

$$\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \xrightarrow{L^2} \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right] = \phi_{\mathbf{I}}(T),$$

it is sufficient to have $S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \in L^2$. This is immediately implied by Theorem 4.4.7 with $m = 2$ and $k = \max_{I\in\mathbf{I}} |I|$, which is required for the weakest form of the in-fill asymptotics.

Central Limit Theorem For asymptotic normality,

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}\right]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\mathbf{I}}),$$

it is again sufficient to have $S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \in L^2$. Here one can explicitly compute for $I, J \in \mathbf{I}$

$$\begin{aligned}\Sigma_{I,J} &= \mathrm{Cov}\left(S^I(\mathbb{X})_{[0,T]}, S^J(\mathbb{X})_{[0,T]}\right)\\ &= \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]} S^J(\mathbb{X})_{[0,T]}\right] - \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\right]\mathbb{E}\left[S^J(\mathbb{X})_{[0,T]}\right]\\ &= \sum_{K\in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T),\end{aligned}$$

which, in practice, will also need to be estimated from data.

It is thus clear that under the assumptions of Theorem 4.4.7 with $m = 2$ and $k = \max_{I\in\mathbf{I}} |I|$ one immediately obtains consistency and asymptotic normality of the estimator $\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)$.

Theorem 4.4.19. *For each $N \geq 1$ let $\mathbb{X}^{\pi_{N,1}}, \dots, \mathbb{X}^{\pi_{N,N}}$ be independent discrete-time observations of the canonical geometric stochastic process $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ over the partitions $\pi_{N,1}, \dots, \pi_{N,N}$ of the interval $[0,T]$. Assume that $\mathbb{X}$ satisfies Assumption (Aα) and Assumption (Aδ) with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 2k$ where $k = \max_{I \in \mathbf{I}} |I|$. Assume that for each $n \geq 1$ the sequence of partitions $\pi_{\cdot,n} = \{\pi_{N,n},\ N \geq n\}$ is signature-defining and*

$$|\Pi(N)| = \max_{1 \leq n \leq N} |\pi_{N,n}| \to 0, \quad N \to \infty.$$

Then the expected signature estimator (4.4.15) is

1. *consistent, i.e. $\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) \xrightarrow{L^2} \phi_{\mathbf{I}}(T)$ as $N \to \infty$.*

If, moreover, for each $n \geq 1$, $\pi_{\cdot,n}$ is a sequence of refinements with $\sum_{N \geq n} |\pi_{N,n}|^\epsilon < \infty$ then $p = 2k$ is sufficient and, under the additional assumption

$$\sqrt{N} \max_{1 \leq n \leq N} \sum_{n' \geq N} |\pi_{n',n}|^\epsilon \to 0, \quad N \to \infty,$$

where $\epsilon = \epsilon(\alpha, \delta)$ is given in Equation (4.4.8), the estimator is also

2. *asymptotically normal, i.e. $\sqrt{N}\left(\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) - \phi_{\mathbf{I}}(T)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\mathbf{I}})$ as $N \to \infty$.*

Remark 4.4.20. We state the result with sufficient conditions to establish the in-fill convergence via Theorem 4.4.7, i.e. $\mathbb{X}$ satisfies Assumption (Aα) and Assumption (Aδ) with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 2k$ where $k = \max_{I \in \mathbf{I}} |I|$. These can be replaced by the sufficient conditions of Theorem 4.4.4.

Remark 4.4.21. If $\{\Pi(N),\ N \geq 1\}$ is a sequence of expanding dyadic refinements, i.e. for each $n \geq 1$, $\pi_{\cdot,n}$ is a sequence of dyadic partitions with $|\pi_{N,n}| = 2^{-N}T$, $N \geq n$, then, proceeding as in Remark 4.4.8, $\sum_{n' \geq N} |\pi_{n',n}|^\epsilon = \mathcal{O}(2^{-\epsilon N})$ and hence

$$\sqrt{N} \max_{1 \leq n \leq N} \sum_{n' \geq N} |\pi_{n',n}|^\epsilon = \mathcal{O}(\sqrt{N} 2^{-\epsilon N}) \to 0, \quad N \to \infty.$$

Remark 4.4.22. To make the CLT feasible one requires a consistent estimator for the covariance $\Sigma_{\mathbf{I}}$. The simplest choice is to consider the sample covariance of the $\{S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]},\ n = 1, \dots, N\}$, i.e. the matrix

$$\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)} = \frac{1}{N} \sum_{n=1}^{N} \left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right) \left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right)^{\mathrm{T}},$$

where for $I, J \in \mathbf{I}$ the I,J-th entry is given by

$$\begin{aligned}
\hat{\Sigma}_{I,J}^{\Pi(N)} &= \frac{1}{N} \sum_{n=1}^{N} S^I(\mathbb{X}^{\pi_{N,n}})_{[0,T]} S^J(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - \hat{\phi}_I^{\Pi(N)}(T) \hat{\phi}_J^{\Pi(N)}(T) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{K \in I \sqcup\!\sqcup J} S^K(\mathbb{X}^{\pi_{N,n}})_{[0,T]} - \hat{\phi}_I^{\Pi(N)}(T) \hat{\phi}_J^{\Pi(N)}(T)
\end{aligned}$$

$$= \sum_{K \in I \sqcup\!\sqcup J} \hat{\phi}_K^{\Pi(N)}(T) - \hat{\phi}_I^{\Pi(N)}(T)\hat{\phi}_J^{\Pi(N)}(T).$$

If Assumption (Aα) and Assumption (Aδ) hold with $\alpha \geq 1/2$, $\delta \geq 1$ and $p = 4k$ then, by the consistency result of Theorem 4.4.19, $\hat{\phi}_K^{\Pi(N)}(T)$ consistently estimates $\phi_K(T)$ for all $K \in I \sqcup\!\sqcup J$, $I, J \in \mathbf{I}$, and hence $\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)}$ is consistent for $\Sigma_{\mathbf{I}}$.

Note that, when observations are independent, it suffices to check that Assumption (Aα) and Assumption (Aδ) hold with $\alpha \geq 1/2$ and $\delta \geq 1$ to obtain consistency of the expected signature estimator. We refer to the examples at the end of Section 4.4.1 for sufficient conditions ensuring these assumptions are fulfilled by different classes of stochastic processes. For example, when $\{\mathbf{X}_t,\ t \geq 0\}$ is a Itô diffusion with deterministic initial condition it suffices to consider Lipschitz continuous drift and diffusion coefficients to guarantee the process satisfies Assumption (Aα) with $\alpha = 1/2$ and Assumption (Aδ) with $\delta = 1$ and any $p \geq 1$, cf. Example 4.4.13. We can hence immediately apply Theorem 4.4.19.*1* to obtain consistency of the expected signature estimator for any sequence of partitions $\{\Pi(N),\ N \geq 1\}$ with $|\Pi(N)| \to 0$ since by Theorem 4.3.8 all sequences of partitions with vanishing mesh size are signature-defining. If, moreover, the sequence of partitions $\{\Pi(N),\ N \geq 1\}$ is an expanding sequence of dyadic refinements as in Remark 4.4.21, then Theorem 4.4.19.*2* ensures the estimator is also asymptotically normal.

Dependent observations

We now move to the dependent observations setting, i.e. where samples of the process are obtained under the (chop) observational scheme. Requiring the path-valued sequence $\{\mathbb{X}^n,\ n \geq 1\}$ to be stationary might look like a bit of a high-level condition. A slightly stronger assumption which is sometimes easier to verify for models of practical interest is to assume that $\{\mathbf{X}_t,\ t \geq 0\}$ has jointly stationary increments[14], i.e. for any $n \in \mathbb{N}$, $0 \leq s_i \leq t_i$, $i = 1, \dots, n$ and $t \geq 0$,

$$(\mathbf{X}_{s_1,t_1}, \dots, \mathbf{X}_{s_n,t_n}) \overset{\mathcal{L}}{=} (\mathbf{X}_{t+s_1,t+t_1}, \dots, \mathbf{X}_{t+s_n,t+t_n}). \tag{4.4.18}$$

The following result uses this condition as a middle step to relate stationarity of the process $\{\mathbf{X}_t,\ t \geq 0\}$ and the corresponding chopped up sequence $\{\mathbb{X}^n,\ n \geq 1\}$.

Proposition 4.4.23. *Fix $T > 0$ and let $\{\mathbf{X}_t,\ t \geq 0\}$ be a stochastic process. Then the following sequence of implications holds*

$$\{\mathbf{X}_t,\ t \geq 0\} \text{ stationary} \implies \{\mathbf{X}_t,\ t \geq 0\} \text{ has jointly stationary increments} \implies \{\mathbb{X}^n,\ n \geq 1\} \text{ stationary.}$$

[14] We call condition (4.4.18) *jointly stationarity increments* since the term *stationary increments* usually refers to the weaker condition $\mathbf{X}_{s,t} \overset{\mathcal{L}}{=} \mathbf{X}_{0,t-s}$ for all $0 \leq s \leq t$.

Proof. See Appendix 4.D.1. □

The following somewhat technical result ensures the signature map (can be extended to) a measurable mapping of the observed path, i.e. a random variable over the canonical space.

Proposition 4.4.24. *Let* $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ *denote a canonical geometric stochastic process defined on the probability space* $(\Omega_{[0,T]} = C([0,T];\mathbb{R}^d), \mathcal{B}_{[0,T]}, \mathbb{P}_{[0,T]})$ *by* $\mathbf{X}_t = \omega_{[0,T]}(t)$ *for* $t \in [0,T]$ *and* $\omega_{[0,T]} \in \Omega_{[0,T]}$. *Then for any collection of words* $\mathbf{I}$ *there exists a measurable map* $\mathcal{S}^{\mathbf{I}} : C([0,T];\mathbb{R}^d) \to \mathbb{R}^{|\mathbf{I}|}$ *such that*

$$\mathcal{S}^{\mathbf{I}}(\omega_{[0,T]}) = S^{\mathbf{I}}(\mathbb{X})_{[0,T]}, \quad \mathbb{P}_{[0,T]} - \text{a.s.}$$

Proof. See Appendix 4.D.2. □

Remark 4.4.25. Let $(\Omega = C([0,\infty);\mathbb{R}^d), \mathcal{F} = \mathcal{B}_{[0,\infty)}, \mathbb{P})$ denote the canonical space on which $\{\mathbf{X}_t,\ t \geq 0\}$ is defined. Consider $\{\mathbb{X}^n,\ n \geq 1\}$ as the sequence of $C([0,T];\mathbb{R}^d)$-valued random variables on the space[15] $(C([0,T];\mathbb{R}^d)^\infty, \mathcal{B}^\infty_{[0,T]}, \mathbb{P}^\infty_{[0,T]})$, where the probability measure $\mathbb{P}^\infty_{[0,T]}$ is obtained by pushing forward $\mathbb{P}$ by the measurable mapping

$$\omega \in \Omega \mapsto \{\omega^n_{[0,T]},\ n \geq 1\} \in C([0,T];\mathbb{R}^d)^\infty,$$

where

$$\omega^n_{[0,T]} := \{\omega((n-1)T + t) - \omega((n-1)T),\ t \in [0,T]\}.$$

If $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary and $\mathbb{X} = \mathbb{X}^1$ is a canonical geometric stochastic process then each $\mathbb{X}^n$ is also a canonical geometric stochastic process on $(C([0,T];\mathbb{R}^d), \mathcal{B}_{[0,T]}, \mathbb{P}_*\mathbb{X}^n)$ with signature given by the same measurable mapping $\mathcal{S}^{\mathbf{I}} : C([0,T];\mathbb{R}^d) \to \mathbb{R}^{|\mathbf{I}|}$, i.e.

$$\mathcal{S}^{\mathbf{I}}(\omega^n_{[0,T]}) = S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]}, \quad \mathbb{P}-\text{a.s.}$$

Given a model (i.e. a probability measure $\mathbb{P}$) for $\{\mathbf{X}_t,\ t \geq 0\}$ such that $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary the signature can thus be interpreted as a measurable map from $C([0,T];\mathbb{R}^d)$ to $\mathbb{R}^{|\mathbf{I}|}$ for every $\mathbb{X}^n$. This ensures the sequence $\{S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]},\ n \geq 1\}$ is defined $\mathbb{P}$-a.s. and is also stationary.

[15] Recall that if $(E, \mathcal{B}(E))$ is a Borel measurable space and we equip E^∞, i.e. the space of sequences with values in E, with the product topology then

$$\mathcal{B}(E^\infty) = \otimes_{n\geq 1}\mathcal{B}(E) := \sigma\left(\cup_{n\geq 1}(\mathcal{B}(E)^n \times E^\infty)\right),$$

as long as E is second-countable.

Consistency If $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary[16] and ergodic[17] we can apply Birkhoff's Ergodic Theorem (Kallenberg, 2021, Theorem 25.6), to deduce

$$\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \overset{\mathbb{P}-\text{a.s.}}{\to} \mathbb{E}[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}], \quad N \to \infty.$$

Remark 4.4.26. Under the weakest condition for the in-fill asymptotics, i.e. Assumption (Aα) and Assumption (Aδ) with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 2k$ with $k = \max_{I\in\mathbf{I}} |I|$, $S^{\mathbf{I}}(\mathbb{X})_{[0,T]} \in L^2$ and hence the above convergence extends to L^2.

As for stationarity, cf. Proposition 4.4.23, one might expect ergodicity of $\{\mathbf{X}_t,\ t \geq 0\}$ to imply ergodicity of $\{\mathbb{X}^n,\ n \geq 1\}$. This turns out not to be the case as illustrated by the following counterexample.

Example 4.4.27. Let $\{X_t,\ t \geq 0\}$ be an $\mathbb{R}$-valued process such that

$$X_t = \sin\left(\frac{\pi t}{T} + \phi\right), \quad t \geq 0,$$

and $\phi \sim U([0, 2\pi])$. The process is stationary and ergodic. Stationarity of $\{X_t,\ t \geq 0\}$ implies stationarity of $\{\mathbb{X}^n,\ n \geq 1\}$ by Proposition 4.4.23. But the shift-invariant set

$$I = \prod_{n\geq 1} \{\omega^n_{[0,T]} \in I_{\geq 0} \cup I_{\leq 0}\} \in \mathcal{B}^{\infty}_{[0,T]},$$

where

$$I_{\geq 0} := \{\omega_{[0,T]} : \omega_{[0,T]}(t) \geq 0,\ \forall t \in [0,T]\},\ I_{\leq 0} := \{\omega_{[0,T]} : \omega_{[0,T]}(t) \leq 0,\ \forall t \in [0,T]\} \in \mathcal{B}_{[0,T]},$$

has measure $\mathbb{P}^{\infty}_{[0,T]}(I) = \mathbb{P}(\phi \in [\pi/2, \pi] \cup [3\pi/2, 2\pi]) = 1/2 \notin \{0, 1\}$ and hence $\{\mathbb{X}^n,\ n \geq 1\}$ is not ergodic. Two realizations of $\{X_t,\ t \geq 0\}$, one in I and the other not, are displayed in Figure 4.4.

On the other hand, we will see in Proposition 4.4.28 that strong mixing of $\{\mathbf{X}_t,\ t \geq 0\}$, a stronger condition than ergodicity, implies strong mixing (and hence ergodicity) of $\{\mathbb{X}^n,\ n \geq 1\}$. As we shall see in the next paragraph, when $\{\mathbb{X}^n,\ n \geq 1\}$ is strongly mixing the estimator can be proven to be not just consistent but also asymptotically normal.

---

[16] Note that, stationarity of $\{\mathbb{X}^n,\ n \geq 1\}$ means that $\theta\{\mathbb{X}^n,\ n \geq 1\} \overset{\mathcal{L}}{=} \{\mathbb{X}^n,\ n \geq 1\}$ where the shift operator $\theta : C([0,T];\mathbb{R}^d)^{\infty} \to C([0,T];\mathbb{R}^d)^{\infty}$ maps

$$\theta\{\xi_n,\ n \geq 1\} = \{\xi_{n+1},\ n \geq 1\},$$

or equivalently, using the definition more familiar to the stochastic processes community,

$$(\mathbb{X}_{n_1}, \ldots, \mathbb{X}_{n_k}) \overset{\mathcal{L}}{=} (\mathbb{X}_{n_1+n}, \ldots, \mathbb{X}_{n_k+n}),$$

for all $n_1, \ldots, n_k, n \geq 0$.

[17] Recall, a stationary sequence is ergodic if all shift-invariant sets, i.e. $I \in \mathcal{B}^{\infty}_{[0,T]}$ such that $\theta^{-1}I = I$, have trivial probability, i.e. $\mathbb{P}^{\infty}_{[0,T]}(I) \in \{0, 1\}$.

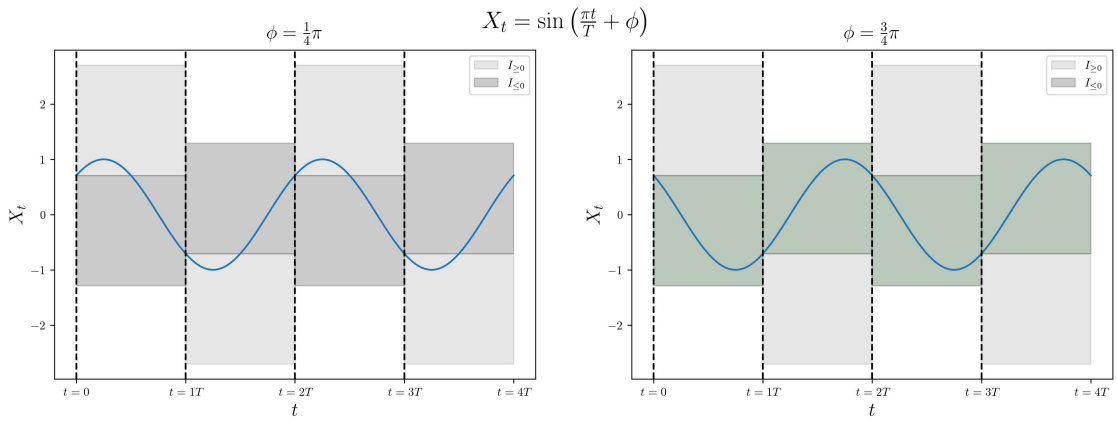


Figure 4.4: Realization of X_t when $\phi = 1/4\pi$ and $\phi = 3/4\pi$ overlayed with the shift-invariant set $I \subseteq C([0,T];\mathbb{R})^\infty$.

Central Limit Theorem We will require the notion of strong-mixing (also known as α-mixing). Let $\{Z_t,\ t \in \mathcal{T}\}$ be a stochastic process, for $s \in \mathcal{T}$ let $\alpha(s)$ be the strong mixing coefficient

$$\alpha(s) := \sup_{t\in\mathcal{T},\, A\in\mathcal{F}^t_{-\infty},\, B\in\mathcal{F}^\infty_{t+s}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \tag{4.4.19}$$

where $\mathcal{F}^b_a := \sigma(Z_u,\ u \in \mathcal{T}, a \leq u \leq b)$ for $a, b \in \overline{\mathcal{T}}$ with $a \leq b$. Suppose that $\alpha(s) \to 0$ as $s \to \infty$, the idea being that for any $t \in \mathcal{T}$, Z_t and Z_{t+s} are approximately independent for large s. In this case the stochastic process $\{Z_t,\ t \in \mathcal{T}\}$ is said to be α-mixing. The definition allows both for discrete-time ($\mathcal{T} \in \{\mathbb{N}, \mathbb{Z}\}$) and continuous-time ($\mathcal{T} \in \{\mathbb{R}_+, \mathbb{R}\}$) stochastic processes. Note this definition does not require $\{Z_t,\ t \in \mathcal{T}\}$ to be strictly stationary.

We start by establishing the following result, which, in a similar spirit to Proposition 4.4.23, describes how strong mixing of the signature sequence is related to strong mixing of the underlying model for $\{\mathbf{X}_t,\ t \geq 0\}$.

Proposition 4.4.28. *Fix $T > 0$ and let $\{\mathbf{X}_t,\ t \geq 0\}$ be a stochastic process such that $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a canonical geometric stochastic process and $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary. Then*

$$\begin{aligned}&\{\mathbf{X}_t,\ t \geq 0\} \textit{ strongly mixing}\\ &\qquad\implies \{\mathbb{X}^n,\ n \geq 1\} \textit{ strongly mixing}\\ &\qquad\qquad\implies \{S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]},\ n \geq 1\} \textit{ strongly mixing},\end{aligned}$$

for any collection of words $\mathbf{I}$.

Proof. See Appendix 4.D.3. □

Under the strong mixing assumption we have access to a wide range of dependent CLT results, see Bradley (2005) for a survey. Usually, these require $S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \in L^{2+\zeta}$ for $\zeta > 0$,

which we immediately obtain when applying Theorem 4.4.7 with $p > 2k$ and $k = \max_{I\in\mathbf{I}} |I|$. To give a concrete example, let us consider the dependent CLT Ibragimov (1962, Theorem 1.7) extended to the multivariate setting via the Cramér-Wold theorem. To apply this result we require $\{S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]},\ n \geq 1\}$ to be stationary with $S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \in L^{2+\zeta}$ and strongly mixing with strong mixing coefficient $\alpha''(n),\ n \in \mathbb{N}$ satisfying

$$\sum_{n\geq 1} \alpha''(n)^{\zeta/(2+\zeta)} < \infty,$$

so that

$$\Sigma_{\mathbf{I}} = \mathrm{Var}\left(S^{\mathbf{I}}(\mathbb{X}^1)_{[0,T]}\right) + 2\sum_{n\geq 2} \mathrm{Cov}\left(S^{\mathbf{I}}(\mathbb{X}^1)_{[0,T]}, S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]}\right) < \infty,$$

and, if $\Sigma_{\mathbf{I}}$ is strictly positive definite,

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} - \mathbb{E}[S^{\mathbf{I}}(\mathbb{X})_{[0,T]}]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\mathbf{I}}), \quad N \to \infty.$$

We can now combine this result with the in-fill asymptotics of Theorem 4.4.7 to obtain the asymptotic properties of the estimator $\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)$ in the dependent observation setting.

Theorem 4.4.29. *Fix $T > 0$ and let $\{\mathbf{X}_t,\ t \geq 0\}$ be a stochastic process such that $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a canonical geometric stochastic process satisfying Assumption (Aα) and Assumption (Aδ) with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 2k$ with $k = \max_{I\in\mathbf{I}} |I|$. Assume $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary and ergodic. For each $N \geq 1$ let $\mathbb{X}^{\pi_{N,1}}, \ldots, \mathbb{X}^{\pi_{N,N}}$ be discrete-time observations of $\mathbb{X}^1, \ldots, \mathbb{X}^N$ obtained by chopping up (and shifting in time and space) a single long discrete-time observation of $\{\mathbf{X}_t,\ t \geq 0\}$ over the partition*

$$\Pi(N) = \pi_{N,1} \cup (T + \pi_{N,2}) \cup \ldots \cup ((N-1)T + \pi_{N,N}),$$

of the interval $[0, NT]$ such that each $\pi_{N,n}$ partitions the interval $[0,T]$, for each $n \geq 1$ the sequence of partitions $\pi_{\cdot,n} = \{\pi_{N,n},\ N \geq n\}$ is signature-defining and

$$|\Pi(N)| = \max_{1\leq n\leq N} |\pi_{N,n}| \to 0, \quad N \to \infty.$$

Then the expected signature estimator (4.4.15) is

1. *consistent, i.e. $\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) \xrightarrow{L^2} \phi_{\mathbf{I}}(T)$ as $N \to \infty$.*

If, moreover, $\{\mathbb{X}^n,\ n \geq 1\}$ is strongly mixing with strong mixing coefficient $\{\alpha(n),\ n \geq 1\}$ such that for $\zeta = p/k - 2 > 0$,

$$\sum_{n\geq 1} \alpha(n)^{\zeta/(2+\zeta)} < \infty, \tag{4.4.20}$$

and for each $n \geq 1$, $\pi_{\cdot,n}$ is a sequence of refinements with $\sum_{N \geq n} |\pi_{N,n}|^\epsilon < \infty$ with

$$\sqrt{N} \max_{1 \leq n \leq N} \sum_{n' \geq N} |\pi_{n',n}|^\epsilon \to 0, \quad N \to \infty,$$

where $\epsilon = \epsilon(\alpha, \delta)$ is given in Equation (4.4.8), then the estimator is also

2. *asymptotically normal, i.e. $\sqrt{N}\left(\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) - \phi_{\mathbf{I}}(T)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\mathbf{I}})$ as $N \to \infty$, as long as*

$$\Sigma_{\mathbf{I}} = \mathrm{Var}\left(S^{\mathbf{I}}(\mathbb{X}^1)_{[0,T]}\right) + 2 \sum_{n \geq 2} \mathrm{Cov}\left(S^{\mathbf{I}}(\mathbb{X}^1)_{[0,T]}, S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]}\right), \tag{4.4.21}$$

is strictly positive definite.

Remark 4.4.30. To make the CLT feasible one requires a consistent estimator for the long-run covariance of the sequence of random variables $\{S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]},\ n \geq 1\}$,

$$\Sigma_{\mathbf{I}} = \Sigma_{\mathbf{I}}^0 + 2 \sum_{n \geq 1} \Sigma_{\mathbf{I}}^n, \quad \text{where} \quad \Sigma_{\mathbf{I}}^n = \mathrm{Cov}\left(S^{\mathbf{I}}(\mathbb{X}^1)_{[0,T]}, S^{\mathbf{I}}(\mathbb{X}^{1+n})_{[0,T]}\right),\ n \geq 0.$$

In a similar spirit to Remark 4.4.22 we can consider the *non-overlapping* sample (cross-)covariances of the sequence $\{S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]},\ n = 1, \ldots, N\}$, i.e. for $|n| \leq N - 1$,

$$\hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)} = \frac{1}{\lfloor N/(n+1) \rfloor} \sum_{m=1}^{\lfloor N/(n+1) \rfloor} \left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,(n+1)m-n}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right)\left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,(n+1)m}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right)^{\mathrm{T}},$$

as estimators for $\Sigma_{\mathbf{I}}^n$. Note that, for a fixed observation partition $\Pi(N)$, we are only able to estimate $\Sigma_{\mathbf{I}}^n$ up to $n = N - 1$ with the quality of the estimator decreasing as n increases[18]. A natural choice would thus be to put less weight on $\hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)}$ with n large than on $\hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)}$ with small n. To do so, one can consider the kernel estimator

$$\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)} = \sum_{n=-(N-1)}^{N-1} k\left(\frac{n}{h_N}\right) \hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)}, \quad \hat{\Sigma}_{\mathbf{I}}^{-n,\Pi(N)} := \left(\hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)}\right)^{\mathrm{T}} \text{ for } n = 1, \ldots, N-1,$$

where $k : \mathbb{R} \to [0, 1]$ is a decreasing kernel function continuous at zero with with $k(0) = 1$ and h_N is an appropriately chosen band-width parameter[19]. In what follows we set k to be the truncation

[18] In this context, one might exploit the available data more efficiently by considering the full sample cross-covariances of the sequence $\{S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,n}})_{[0,T]},\ n = 1, \ldots, N\}$, i.e. for $|n| \leq N - 1$,

$$\frac{1}{N-n} \sum_{m=1}^{N-n} \left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,m}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right)\left(S^{\mathbf{I}}(\mathbb{X}^{\pi_{N,m+n}})_{[0,T]} - \hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T)\right)^{\mathrm{T}},$$

as estimators for $\Sigma_{\mathbf{I}}^n$ with $n \geq 0$. The reason for using the non-overlapping sample covariance will become apparent once we introduce the process $\mathbb{Y}^n$, which will be used to show consistency of the estimator $\hat{\Sigma}_{\mathbf{I}}^{n,\Pi(N)}$ for $\Sigma_{\mathbf{I}}^n$ by re-applying Theorem 4.4.29.*1.* To show consistency of the full sample covariance estimator we would instead require the generalization of such result for time-overlapping expected signature estimators, cf. Remark 4.4.17.

[19] In the presence of conditional heteroskedasticity such estimators for the long-run covariance are known as Heteroskedasticity and Autocorrelation Consistent (HAC) estimators (Newey and West, 1987). Note that decomposing

kernel $k(x) = \mathbb{1}_{[-1,1]}(x)$ for simplicity but other choices, such as the Bartlett kernel, might lead to better finite sample properties. For each $I, J \in \mathbf{I}$, if $\hat{\Sigma}^{n,\Pi(N)}_{I,J}$ is consistent for $\Sigma^n_{I,J}$ in L^2 with monotonically decreasing rate $r(M) \to 0$ as the *effective* sample size $M = \lfloor N/(n+1) \rfloor \to \infty$, then

$$\begin{aligned}\|\hat{\Sigma}^{\Pi(N)}_{I,J} - \Sigma_{I,J}\|_{L^2} &\leq \sum_{|n|\leq h_N} \left\|\hat{\Sigma}^{n,\Pi(N)}_{I,J} - \Sigma^n_{I,J}\right\|_{L^2} + \left|\sum_{|n|>h_N} \Sigma^n_{I,J}\right| \\ &\leq \sum_{|n|\leq h_N} r(\lfloor N/(n+1)\rfloor) + \left|\sum_{|n|>h_N} \Sigma^n_{I,J}\right|.\end{aligned}$$

If the band-width $h_N \to \infty$ as $N \to \infty$ then $\Sigma_{\mathbf{I}} < \infty$ ensures the second term vanishes. If, moreover, we set the rate at which $h_N \to \infty$ to be slow enough to ensure also the first term converges to zero[20], then consistency of the estimators $\hat{\Sigma}^{n,\Pi(N)}_{I,J}$ for $n \geq 0$ is inherited by $\hat{\Sigma}^{\Pi(N)}_{I,J}$. Under such choice of kernel estimator for the long-run variance we can hence focus on determining under which conditions, other than those of Theorem 4.4.29.*2*, the estimator $\hat{\Sigma}^{n,\Pi(N)}_{I,J}$ is consistent for $\Sigma^n_{I,J}$ for any $n \geq 0$ and $I, J \in \mathbf{I}$.

Note that, as in Remark 4.4.22, we can apply the shuffle identity to write for $I, J \in \mathbf{I}$,

$$\hat{\Sigma}^{0,\Pi(N)}_{I,J} = \sum_{K \in I \sqcup\!\sqcup J} \hat{\phi}^{\Pi(N)}_K(T) - \hat{\phi}^{\Pi(N)}_I(T)\hat{\phi}^{\Pi(N)}_J(T),$$

and show consistency of this estimator for $\Sigma^0_{I,J}$ by applying the consistency result for the estimator of the expected signature terms to $K \in I \sqcup\!\sqcup J$ for $I, J \in \mathbf{I}$. We now attempt to apply a similar approach in the dependent observations setting, where the products of signature terms over non-overlapping time intervals appear. To do so, we introduce an $\mathbb{R}^{nd}$-valued auxiliary process[21] defined as $\mathbb{Y}^n = \{\mathbf{Y}^n_t,\ t \in [0, (n+1)T]\}$ where

$$\mathbf{Y}^n_t = \begin{pmatrix} \mathbf{X}^1_{t\wedge T} \\ \mathbf{X}^2_{(t-T)^+\wedge T} \\ \cdots \\ \mathbf{X}^{n+1}_{(t-nT)^+\wedge T} \end{pmatrix},\ t \in [0,(n+1)T] \iff \mathbf{Y}^n_t = \begin{pmatrix} \mathbf{X}^1_T \\ \cdots \\ \mathbf{X}^i_T \\ \mathbf{X}^{i+1}_s \\ 0 \\ \cdots \\ 0 \end{pmatrix},\ t = iT + s, s \in [0,T], 0 \leq i \leq n.$$

the estimation of $2\Sigma^n_{\mathbf{I}}$ into the sum of $\hat{\Sigma}^{n,\Pi(N)}_{\mathbf{I}}$ and its transpose ensures the resulting long-run covariance estimator $\hat{\Sigma}^{\Pi(N)}_{\mathbf{I}}$ is symmetric.

[20] For example, if $r(M) \sim M^{-\upsilon}$ for $\upsilon \in (0,1)$ one can set $h_N = N^{\upsilon/2}$. Unfortunately, Theorem 4.4.29.*1* does not yield the rate of convergence $r(\cdot)$, and hence in practice the band-with h_N needs to be chosen from the data.

[21] We would like to thank Nicola Muca Cirone for suggesting this clever trick.

By construction, we have that for any two words $I, J \in \mathbf{I}$ over the letters $\{1, \dots, d\}$,

$$S^I(\mathbb{X}^1)_{[0,T]} = S^I(\mathbb{Y}^n)_{[0,(n+1)T]} \quad \text{and} \quad S^J(\mathbb{X}^n)_{[0,T]} = S^{nd+J}(\mathbb{Y}^n)_{[0,(n+1)T]},$$

i.e. we have re-written the two signature components of $\mathbb{X}^1$ and $\mathbb{X}^n$ over $[0,T]$ as time-overlapping signature components of $\mathbb{Y}^n$ over $[0,(n+1)T]$. We can thus apply the shuffle product to deduce

$$S^I(\mathbb{X}^1)_{[0,T]} S^J(\mathbb{X}^n)_{[0,T]} = S^I(\mathbb{Y}^n)_{[0,(n+1)T]} S^{nd+J}(\mathbb{Y}^n)_{[0,(n+1)T]} = \sum_{K \in I \sqcup\!\sqcup (nd+J)} S^K(\mathbb{Y}^n)_{[0,(n+1)T]}.$$

We can thus re-write the (I,J)-th entry of $\Sigma^n_{\mathbf{I}}$ as

$$\begin{aligned}
\Sigma^n_{I,J} &= \mathbb{E}\left[S^I(\mathbb{X}^1)_{[0,T]} S^J(\mathbb{X}^{1+n})_{[0,T]}\right] - \mathbb{E}\left[S^I(\mathbb{X}^1)_{[0,T]}\right] \mathbb{E}\left[S^J(\mathbb{X}^{1+n})_{[0,T]}\right] \\
&= \mathbb{E}\left[S^I(\mathbb{Y}^n)_{[0,(n+1)T]} S^{d+J}(\mathbb{Y}^n)_{[0,(n+1)T]}\right] - \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\right] \mathbb{E}\left[S^J(\mathbb{X})_{[0,T]}\right] \\
&= \sum_{K \in I \sqcup\!\sqcup (nd+J)} \mathbb{E}\left[S^K(\mathbb{Y}^n)_{[0,(n+1)T]}\right] - \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\right] \mathbb{E}\left[S^J(\mathbb{X})_{[0,T]}\right] \\
&= \sum_{K \in I \sqcup\!\sqcup (nd+J)} \psi_K((n+1)T) - \phi_I(T)\phi_J(T),
\end{aligned}$$

and estimate it by

$$\hat{\Sigma}^{n,\Pi(N)}_{I,J} = \sum_{K \in I \sqcup\!\sqcup (nd+J)} \hat{\psi}^{\Pi(N;n)}_K((n+1)T) - \hat{\phi}^{\Pi(N)}_I(T)\hat{\phi}^{\Pi(N)}_J(T),$$

where, setting $M = \lfloor N/(n+1) \rfloor$,

$$\hat{\psi}^{\Pi(N;n)}_K((n+1)T) := \frac{1}{M} \sum_{m=1}^{M} S^K(\mathbb{Y}^{\pi_{N,m;n}})_{[0,(n+1)T]},$$

and

$$\Pi(N;n) = \pi_{N,1;n} \cup \ldots \cup ((M-1)(n+1)T + \pi_{N,M;n}),$$

where for each $m = 1, \dots, M$,

$$\pi_{N,m;n} = \pi_{N,(m-1)(n+1)+1} \cup \ldots \cup (nT + \pi_{N,m(n+1)}),$$

partitions $[0,(n+1)T]$ and

$$|\Pi(N;n)| := \max_{1 \le m \le M} |\pi_{N,m;n}| = \max_{1 \le m \le M} \max_{1 \le i \le (n+1)} |\pi_{N,(m-1)(n+1)+i}| \le |\Pi(N)|.$$

In order to apply the consistency result of Theorem 4.4.29 to $\hat{\psi}^{\Pi(N;n)}_K(nT)$ we need to understand under which additional conditions on $\{\mathbf{X}_t,\ t \ge 0\}$, other than those of Theorem 4.4.29, the process

$\{\mathbf{Y}^n_t,\ t \geq 0\}$, defined by

$$\mathbf{Y}^n_t = \sum_{m\geq 1} \begin{pmatrix} \mathbf{X}^m_{(t-(m-1)T)^+\wedge T} \\ \mathbf{X}^{m+1}_{(t-mT)^+\wedge T} \\ \cdots \\ \mathbf{X}^{m+n}_{(t-(m+n-1)T)^+\wedge T} \end{pmatrix},\ t \geq 0,$$

satisfies itself the assumptions of Theorem 4.4.29. Note that canonical geometricity of the stochastic process $\mathbb{X}$ (with lift-defining sequence of partitions ρ, $|\rho| \to 0$) and stationarity of $\{\mathbb{X}^m,\ m \geq 1\}$ imply that each $\mathbb{X}^m$ is a canonical geometric stochastic process and hence so is $\mathbb{Y}^n$ (with lift-defining sequence of partitions $\rho^n = \rho \cup (T+\rho) \cup \cdots \cup (nT+\rho)$, $|\rho| \to 0$). Moreover, stationarity and ergodicity of $\{\mathbb{X}^m,\ m \geq 1\}$ imply stationarity and ergodicity of $\{(\mathbb{X}^m, \ldots, \mathbb{X}^{m+n}),\ m \geq 1\}$ and hence of the measurable transformation $\{(\mathbb{X}^m, \ldots, \mathbb{X}^{m+n}) \mapsto \mathbb{Y}^{n,m},\ m \geq 1\}$.

For fixed $n \geq 0$ and $m \geq 1$, each $\pi_{\cdot,m;n} = \{\pi_{N,m;n},\ \lfloor N/(n+1) \rfloor \geq m\}$ is a signature-defining sequence for $\mathbb{Y}^n = \mathbb{Y}^{n,1}$ over $[0,(n+1)T]$ since each $\pi_{\cdot,n} = \{\pi_{N,n},\ N \geq n\}$ is a signature-defining sequence for $\mathbb{X}$ over $[0,T]$. Moreover, by construction of $\Pi(N;n)$, $|\Pi(N;n)| \leq |\Pi(N)| \to 0$ whenever $|\Pi(N)| \to 0$ as $N \to \infty$.

It thus remains to check that $\mathbb{Y}^n = \mathbb{Y}^{n,1}$ satisfies Assumption (Aα) and Assumption (Aδ) over $[0,(n+1)T]$ with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 4k$ with $k = \max_{I\in\mathbf{I}} |I|$. Clearly, as we are now considering the process over an arbitrarily long time span $[0,(n+1)T]$, we will require $\{\mathbf{X}_t,\ t \geq 0\}$ to satisfy Assumption (Aα) and Assumption (Aδ) with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 4k$ not just over $[0,T]$ but for any $0 \leq s \leq t$ with $|t-s| \leq T$. This condition is automatically fulfilled when $\{\mathbf{X}_t,\ t \geq 0\}$ is stationary (or, more generally, has jointly stationary increments) and satisfies Assumption (Aα) and Assumption (Aδ) over $[0,T]$ with $\alpha \geq 1/2$, $\delta \geq 1$ and $p > 4k$ with $k = \max_{I\in\mathbf{I}} |I|$. It turns out that these conditions, only slightly stronger than those already required in Theorem 4.4.29 for the asymptotic normality of $\hat{\phi}^{\Pi(N)}_{\mathbf{I}}(T)$, are sufficient to show that $\hat{\psi}^{\Pi(N;n)}_K((n+1)T)$ is consistent for $\psi_K((n+1)T)$, and thus $\hat{\Sigma}^{n,\Pi(N)}_{I,J}$ is consistent for $\Sigma^n_{I,J}$ for any $n \geq 0$ and $I, J \in \mathbf{I}$.

To show Assumption (Aα) holds for $\mathbb{Y}^n$ with the same $\alpha \geq 1/2$ as $\{\mathbf{X}_t,\ t \geq 0\}$ note that for $0 \leq s \leq t \leq (n+1)T$,

$$\begin{aligned} \|\mathbf{Y}^n_{s,t}\|_{L^p} &\leq \sum_{i=0}^{n} \|\mathbf{X}^{i+1}_{(s-iT)^+\wedge T,(t-iT)^+\wedge T}\|_{L^p} \\ &\leq \sum_{i=0}^{n} \|\mathbf{X}_{iT+(s-iT)^+\wedge T,iT+(t-iT)^+\wedge T}\|_{L^p} \\ &\lesssim \sum_{i=0}^{n} |(t-iT)^+ \wedge T - (s-iT)^+ \wedge T|^\alpha \lesssim |t-s|^\alpha. \end{aligned}$$

Next, to show that Assumption (Aδ) holds for $\mathbb{Y}^n$ with the same $\delta \geq 1$ as $\{\mathbf{X}_t,\ t \geq 0\}$. Note that

for $0 \leq s \leq t \leq (n+1)T$,

$$\begin{aligned}
\|\mathbb{E}_s[\mathbf{Y}^n_{s,t}]\|_{L^p} &\leq \sum_{i=0}^{n} \|\mathbb{E}_s[\mathbf{X}^{i+1}_{(s-iT)^+\wedge T,(t-iT)^+\wedge T}]\|_{L^p} \\
&\leq \sum_{i=0}^{n} \|\mathbb{E}_s[\mathbf{X}_{iT+(s-iT)^+\wedge T,iT+(t-iT)^+\wedge T}]\|_{L^p} \\
&\leq \sum_{i=0}^{n} \mathbb{1}_{[0,(i+1)T]}(s)\|\mathbb{E}_{iT+(s-iT)^+\wedge T}[\mathbf{X}_{iT+(s-iT)^+\wedge T,iT+(t-iT)^+\wedge T}]\|_{L^p} \\
&\leq \sum_{i=0}^{n} |(t-iT)^+ \wedge T - (s-iT)^+ \wedge T|^\delta \lesssim |t-s|^\delta.
\end{aligned}$$

Example 4.4.31 (Add-time transform cont.). When applying the add-time transform as in Example 4.4.16 we add the same clock to all observations $\mathbb{X}^n$, i.e. the add-time augmented process $\{\hat{\mathbb{X}}^n,\ n \geq 1\}$ is formed by first chopping and shifting the path $\mathbb{X}$ into $\{\mathbb{X}^n,\ n \geq 1\}$ and then adding time $\hat{\mathbb{X}}^n = \{(\psi(t), \mathbf{X}^n_t),\ t \in [0,T]\}$. The process $\{\hat{\mathbb{X}}^n,\ n \geq 1\}$ thus trivially inherits stationarity, ergodicity and strong mixing from $\{\mathbb{X}^n,\ n \geq 1\}$.

We would now wish to revisit the examples presented at the end of Section 4.4.1, providing classic conditions for stationarity and strong mixing of $\{\mathbf{X}_t,\ t \geq 0\}$ and hence, by Proposition 4.4.23 and Proposition 4.4.28, for stationarity and strong mixing of $\{\mathbb{X}^n,\ n \geq 1\}$. Together with the previously discussed class-specific conditions implying Assumption (Aα) and Assumption (Aδ), this would yield a set of sufficient conditions for Theorem 4.4.29. Unfortunately, checking whether a process is strongly mixing is a hard task. A large class of processes for which we do have a thorough understanding of the ergodic and strong mixing conditions is that of Markov processes. Itô diffusions represent a large subclass of Markov processes[22], often used to model real-world phenomena, and hence we will start by expanding on Example 4.4.13.

Example 4.4.32 (Itô diffusions cont.). When developing conditions ensuring stationarity and ergodicity of an Itô diffusion it is natural to restrict $\{\mathbf{X}_t,\ t \geq 0\}$ to the case where it is time-homogeneous, i.e. satisfies

$$d\mathbf{X}_t = f(\mathbf{X}_t)\mathrm{d}t + \sigma(\mathbf{X}_t)\mathrm{d}\mathbf{W}_t, \quad t \geq 0,$$

where $\mathbb{W} = \{\mathbf{W}_t,\ t \geq 0\}$ is a q-dimensional Brownian motion, $f : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times q}$. Assume:

- The diffusion coefficient $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times q}$ is Lipschitz continuous and $\Sigma := \sigma\sigma^{\mathrm{T}} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is bounded and uniformly elliptic, i.e.

$$\inf_{x\in\mathbb{R}^d, \xi\in\mathbb{R}^d\setminus\{0\}} \frac{\langle \xi, \Sigma(x)\xi\rangle}{\|\xi\|^2} > 0,$$

  this is a classic PDE condition which ensures the transition densities are "nice", i.e. continuous

[22] All strong Feller continuous-path Markov processes that are semimartingales are Itô diffusions (Çinlar et al., 1980).

and bounded away from zero (Friedman, 1964).

- The drift $f : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous and has negative radial part at ∞, i.e.

$$\limsup_{\|x\|\to\infty} \left\langle f(x), \frac{x}{\|x\|^{\kappa+1}} \right\rangle =: -C_\kappa \in [-\infty, 0),$$

pushing the process towards the origin with strength[23] controlled by $\kappa \in [-1, \infty)$. When $\kappa = -1$ assume further that $2C_{-1} > \sup_{x\in\mathbb{R}^d} \operatorname{Tr} \Sigma(x)$ and define

$$\eta^*_{f,\Sigma} := \begin{cases} \infty, & \text{if } \kappa > 0, \\ 2C_0/|||\Sigma|||, & \text{if } \kappa = 0, \\ 2C_\kappa/((1+\kappa)|||\Sigma|||), & \text{if } \kappa \in (-1, 0), \\ \left(2C_{-1} - \sup_{x\in\mathbb{R}^d} \operatorname{Tr} \Sigma(x)\right)/|||\Sigma|||, & \text{if } \kappa = -1, \end{cases}$$

where $|||\Sigma||| := \sup_{x\in\mathbb{R}^d} \|\Sigma(x)\|$.

These conditions are enough to ensure there exists a unique invariant probability measure μ on $\mathbb{R}^d$ with

$$\begin{cases} \int_{\mathbb{R}^d} e^{\eta\|x\|}\mu(\mathrm{d}x) < \infty, & \text{if } \kappa \geq 0, \\ \int_{\mathbb{R}^d} e^{\eta\|x\|^{1+\kappa}}\mu(\mathrm{d}x) < \infty, & \text{if } \kappa \in (-1, 0), \\ \int_{\mathbb{R}^d} \|x\|^\eta \mu(\mathrm{d}x) < \infty, & \text{if } \kappa = -1, \end{cases}$$

for all $\eta \in (0, \eta^*_{f,\Sigma})$, such that for any $x \in \mathbb{R}^d$ the transition probabilities[24] $\{P_t(x, \cdot),\ t \geq 0\}$ converge to μ in total variation distance with rates

$$\|P_t(x,\cdot) - \mu\|_{\mathrm{TV}} \leq \begin{cases} c_1 e^{-c_2 t}(e^{\eta\|x\|} + c_3), & \text{if } \kappa \geq 0, \\ c_1 e^{-c_2 t^{(1+\kappa)/(1-\kappa)}}(e^{\eta\|x\|^{1+\kappa}} + c_3), & \text{if } \kappa \in (-1, 0), \\ c_1(1 + c_2 t)^{-\eta/2}(\|x\|^\eta + c_3), & \text{if } \kappa = -1, \end{cases}$$

with $c_1, c_2, c_3 > 0$. (Kulik, 2018, Theorem 3.3.4, 3.3.5 and 3.3.6).

Assuming $\mathbf{X}_0 \sim \mu$ the Itô diffusion $\{\mathbf{X}_t,\ t \geq 0\}$ defines a stationary Markov process with $\mathbf{X}_0 \in L^p$ for all $p \geq 2$ when $\kappa > -1$ and for all $2 \leq p < \eta^*_{f,\Sigma}$ when $\kappa = -1$. Recall that, by the discussion in Example 4.4.13, when f and σ are Lipschitz continuous it is enough to have $\mathbf{X}_0 \in L^p$ to ensure the process $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0, T]\}$ satisfies Assumption (Aα) and Assumption (Aδ) with $\alpha = 1/2$ and $\delta = 1$, implying $\epsilon = 1/2$. By Proposition 4.4.23 the chain $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary

[23] Note that, for large $\|x\|$, one has $\left\langle f(x), \frac{x}{\|x\|} \right\rangle \approx -C_\kappa \|x\|^\kappa$, and hence the strength of the pull grows as $\|x\|$ increases when $\kappa > 0$ and decays as $\|x\|$ increases when $\kappa < 0$.

[24] For the time-homogeneous Itô diffusion $\{\mathbf{X}_t,\ t \geq 0\}$ these are defined by

$$P_t(x, A) := \mathbb{P}(\mathbf{X}_t \in A | \mathbf{X}_0 = x) = \mathbb{P}(\mathbf{X}_{s+t} \in A | \mathbf{X}_s = x),$$

for all $t, s \geq 0,\ x \in \mathbb{R}^d,\ A \in \mathcal{B}(\mathbb{R}^d)$.

and hence it remains to establish strong mixing of $\{\mathbb{X}^n,\ n \geq 1\}$ to apply Theorem 4.4.29. The strong mixing coefficient of the stationary Markov process $\{\mathbf{X}_t,\ t \geq 0\}$ can be easily[25] bounded by

$$\alpha(t) \leq \int_{\mathbb{R}^d} \|P_t(x,\cdot) - \mu\|_{\mathrm{TV}}\, \mu(\mathrm{d}x) \lesssim \begin{cases} e^{-c_2 t}, & \text{if } \kappa \geq 0, \\ e^{-c_2 t^{(1+\kappa)/(1-\kappa)}}, & \text{if } \kappa \in (-1,0), \\ (1+c_2 t)^{-\eta/2}, & \text{if } \kappa = -1, \end{cases}$$

and hence the process is strongly mixing. By Proposition 4.4.28 the chain $\{\mathbb{X}^n,\ n \geq 1\}$ is also strongly mixing with coefficient $\alpha''(n) \leq \alpha((n-3)T)$, $n \geq 3$. It follows immediately that $\{\mathbb{X}^n,\ n \geq 1\}$ is ergodic and hence, we can apply Theorem 4.4.29.*1* to deduce that, letting $|\Pi(N)| \to 0$ as $N \to \infty$, the expected signature estimator (4.4.15) is consistent for any expected signature term when $\kappa > -1$ and for all expected signature terms with $|I| < \frac{1}{2}\eta^*_{f,\Sigma}$ when $\kappa = -1$.

Finally, we note that for any $\zeta > 0$

$$\sum_{n\geq 0} \alpha''(n)^{\zeta/(2+\zeta)} \lesssim \begin{cases} \sum_{n\geq 1} e^{-c_2 nT\zeta/(2+\zeta)} < \infty, & \text{if } \kappa \geq 0, \\ \sum_{n\geq 1} e^{-c_2 (nT)^{(1+\kappa)/(1-\kappa)}\zeta/(2+\zeta)} < \infty, & \text{if } \kappa \in (-1,0), \\ \sum_{n\geq 1} (1+c_2 nT)^{-\eta\zeta/(4+2\zeta)} = \infty, & \text{if } \kappa = -1, \end{cases}$$

and hence, if $\kappa > -1$ and $\Pi(N)$ is a sequence of expanding dyadic refinements, we can apply Theorem 4.4.29.*2* to show that the expected signature estimator (4.4.15) is also asymptotically normal for any expected signature term.

Another class of processes for which we can obtain a characterization of the strong mixing condition in terms of a more fundamental object is that of Gaussian processes. As discussed in the following example, for this class of processes there is a well understood relationship between strong mixing and the properties of the spectral density.

Example 4.4.33 (Gaussian processes cont.). Let $\{\mathbf{X}_t,\ t \in \mathbb{R}\}$ be a stationary Gaussian process, i.e.

[25] If $\{\mathbf{X}_t,\ t \geq 0\}\}$ is a stationary Markov process with stationary distribution μ and $A \in \mathcal{F}^t_{-\infty}, B \in \mathcal{F}^\infty_{t+s}$ for $t, s \geq 0$,

$$\begin{aligned} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| &= |\mathbb{E}[\mathbb{1}_A\mathbb{1}_B] - \mathbb{E}[\mathbb{1}_A]\mathbb{E}[\mathbb{1}_B]| \\ &= \left|\mathbb{E}[\mathbb{1}_A\mathbb{E}[\mathbb{E}[\mathbb{1}_B|\mathcal{F}^{t+s}_{-\infty}]|\mathcal{F}^t_{-\infty}]] - \mathbb{E}[\mathbb{1}_A\mathbb{E}[\mathbb{1}_B]]\right| \\ &= \left|\mathbb{E}\left[\mathbb{1}_A\Big(\mathbb{E}[\mathbb{E}[\mathbb{1}_B|\mathbf{X}_{t+s}]|\mathbf{X}_t] - \mathbb{E}[\mathbb{E}[\mathbb{1}_B|\mathbf{X}_{t+s}]]\Big)\right]\right| \\ &= \left|\mathbb{E}\left[\mathbb{1}_A\left(\int_{\mathbb{R}^d} h_B(x)P_s(\mathbf{X}_t,\mathrm{d}x) - \int_{\mathbb{R}^d} h_B(x)\mu(\mathrm{d}x)\right)\right]\right| \\ &\leq \mathbb{E}\left[\mathbb{1}_A\left|\int_{\mathbb{R}^d} h_B(x)(P_s(\mathbf{X}_t,\mathrm{d}x) - \mu(\mathrm{d}x))\right|\right] \\ &\leq \mathbb{E}\left[\|P_s(\mathbf{X}_t,\cdot) - \mu\|_{\mathrm{TV}}\right] \\ &= \int_{\mathbb{R}^d} \|P_s(x,\cdot) - \mu\|_{\mathrm{TV}}\, \mu(\mathrm{d}x), \end{aligned}$$

where $h_B : \mathbb{R}^d \mapsto [0,1]$ is the measurable function defining the conditional expectation $\mathbb{E}[\mathbb{1}_B|\mathbf{X}_{t+s}] = h(\mathbf{X}_{t+s})$.

its mean function $\mu(t) = \mathbb{E}[\mathbf{X}_t] = \mu$ is constant and its covariance function $C(s,t) = \mathrm{Cov}(\mathbf{X}_s, \mathbf{X}_t) = C(|t-s|)$ is homogeneous in the sense that it depends on $|t-s|$ only. The spectral measure of the stationary Gaussian process $\{\mathbf{X}_t,\ t \geq 0\}$ is defined implicitly as the positive definite matrix-valued countably additive set function $F(\mathrm{d}\lambda)$ such that

$$C(t) = \int_{\mathbb{R}} e^{i\lambda t} F(\mathrm{d}\lambda), \quad t \in \mathbb{R},$$

and, if F is absolutely continuous, the spectral density $f(\lambda)$ is defined by $F(\mathrm{d}\lambda) = f(\lambda)\mathrm{d}\lambda$, cf. Gikhman and Skorokhod (2004, Chapter IV, Theorem 5). Ibragimov and Rozanov (1978, Chapter VI) provides in depth discussions on the connection between the spectral density function and complete regularity of a one-dimensional continuous-time Gaussian process (which, in the case of Gaussian processes, is equivalent to strongly mixing, see Ibragimov and Rozanov (1978, Chapter IV)). In Chapter VI.3, Theorem 2 and Chapter VI.5, Theorems 4 and 5 of Ibragimov and Rozanov (1978) the authors give necessary and sufficient conditions in terms of $f(\lambda)$ for the process to be completely regular and, hence, strongly mixing. In the sufficiency setting the results also yield characterizing upper bounds for the coefficient of complete regularity, and hence for the strong mixing coefficient. Generalizing these result to the multivariate case is an interesting avenue of research for which we do not foresee any major difficulties, but is beyond the scope of this work.

Requiring the process $\{\mathbf{X}_t,\ t \geq 0\}$ to be stationary and strongly mixing is clearly a stronger condition than required to obtain asymptotic theorems. If we drop the strong mixing assumption, checking that $\{\mathbb{X}^n,\ n \geq 1\}$ is ergodic is a challenging task since, as discussed in Example 4.4.27, it is not sufficient to show $\{\mathbf{X}_t,\ t \geq 0\}$ is ergodic. In the next section we thus give alternative conditions in terms of $\{\mathbf{X}_t,\ t \geq 0\}$ to establish consistency of the expected signature estimator for Gaussian processes, including fractional Brownian motion (cf. Example 4.4.15), which do not require neither stationarity of $\{\mathbf{X}_t,\ t \geq 0\}$ nor ergodicity of $\{\mathbb{X}^n,\ n \geq 1\}$.

Dependent observations, Gaussian processes

Given a model for $\{\mathbf{X}_t,\ t \geq 0\}$, ergodicity of $\{\mathbb{X}^n,\ n \geq 1\}$ might be hard to establish. It is thus of practical relevance to ask whether one can obtain sufficient conditions for consistency that do not rely directly on this assumption. For example, when $\{\mathbf{X}_t,\ t \geq 0\}$ is a Gaussian process, it is natural to look for a condition in terms of the covariance function, as this completely specifies the law of the process. As an alternative approach to establish consistency of the estimator, we thus aim to apply the following weak law of large number under for dependent variables[26]. Assume that $\{S_n,\ n \in \mathbb{N}\}$ is a stationary sequence of random variables such that:

1. $S_n \in L^2$;

[26]This result is originally due to Markov but often referred to as Bernstein's weak law of large numbers. It can be proved by a simple application of Chebyshev's inequality. We refer to Seneta (2013) for a nice historical review of the law of large numbers.

2. $\mathrm{Cov}\,(S_n, S_m) \to 0$ as $|m-n| \to \infty$,

then the weak law of large numbers holds, i.e.

$$\frac{1}{N}\sum_{n=1}^{N} S_n \xrightarrow{\mathbb{P}} \mathbb{E}[S_n], \quad N \to \infty.$$

Theorem 4.4.34. *Fix $T > 0$ and let $\{\mathbf{X}_t,\ t \geq 0\}$ be a Gaussian process such that $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a canonical geometric stochastic process satisfying Assumption (Aα) with $\alpha \geq 1/2$ and $p = 2$. When $\alpha = 1/2$, assume furthermore $\mathbb{X}$ satisfies Assumption (Aδ) with $\delta \geq 1$ and $p > 2k$ where $k = \max_{I\in\mathbf{I}} |I|$. Assume the sequence of dyadic partitions of $[0,T]$ is signature-defining for $\mathbb{X}$ and for each $N \geq 1$ let $\mathbb{X}^{\pi_{N,1}}, \ldots, \mathbb{X}^{\pi_{N,N}}$ be discrete-time observations of $\mathbb{X}^1, \ldots, \mathbb{X}^N$ obtained by chopping up (and shifting in time and space) a single long discrete-time observation of $\{\mathbf{X}_t,\ t \geq 0\}$ over the partition*

$$\Pi(N) = \pi_{N,1} \cup (T + \pi_{N,2}) \cup \ldots \cup ((N-1)T + \pi_{N,N}),$$

of the interval $[0, NT]$ such that each $\pi_{N,n}$ is the dyadic partition the interval $[0,T]$ with mesh $|\pi_{N,n}| = 2^{-N}T$.

Suppose $\{\mathbf{X}_t,\ t \geq 0\}$ has constant mean function $\mu : \mathbb{R}_+ \to \mathbb{R}^d$, i.e. $\mu(t) \equiv \mu$ for all $t \geq 0$, and has time-homogeneous covariance of the increments $C : \mathbb{R}_+^4 \to \mathbb{R}^d \times \mathbb{R}^d$, i.e. for all $u, v, s, t, r \geq 0$,

$$\mathrm{Cov}(\mathbf{X}_{u,v}, \mathbf{X}_{s,t}) =: C(u,v,s,t) = C(r+u, r+v, r+s, r+t),$$

satisfying, for some decreasing $\theta : \mathbb{R}_+ \to \mathbb{R}_+$ with $\theta(t) \to 0$, $t \to \infty$ and $\int_0^T \theta(t)\mathrm{d}t < \infty$ and $m \in \mathbb{N}$,

$$(\mathrm{A}\theta)\ \ \|\mathrm{Cov}(\mathbf{X}_{u,v}, \mathbf{X}_{s,t})\| \lesssim \theta(|s-v|)|v-u||t-s| \text{ for all } 0 \leq u \leq v < s \leq t \text{ with } |s-v| \geq \frac{m}{2}(|t-s|+|v-u|).$$

Then the expected signature estimator (4.4.15) is consistent, i.e. $\hat{\phi}_{\mathbf{I}}^{\Pi(N)}(T) \xrightarrow{\mathbb{P}} \phi_{\mathbf{I}}(T)$ as $N \to \infty$.

Remark 4.4.35. We state and prove the theorem for expanding dyadic refinements but more general forms might be obtained via similar arguments.

Note that if Assumption (Aα) holds with $\alpha > 1/2$ then we have no restriction on k and hence the result holds for any collection of words $\mathbf{I}$.

Proof. See Appendix 4.D.4. □

Example 4.4.36 (Ornstein-Uhlenbeck process). If $\{\mathbf{X}_t,\ t \geq 0\}$ is a stationary mean-zero d-dimensional Ornstein-Uhlenbeck process, i.e. a mean-zero Gaussian process with covariance

$$C(s,t) := \mathrm{Cov}(\mathbf{X}_s, \mathbf{X}_t) = e^{-A|t-s|}\Sigma,$$

where $\Sigma = \mathrm{Var}(\mathbf{X}_t)$ and the drift matrix parameter $A \in \mathbb{R}^{d\times d}$ has positive real parts of all eigenvalues, then we can show that:

- Assumption (Aα) holds with $\alpha = 1/2$ and $p = 2$. Note that for all $0 \leq s \leq t \leq T$,

$$\begin{aligned}
\|\mathbf{X}_{s,t}\|_{L^2}^2 &= \mathbb{E}[\mathrm{tr}(\mathbf{X}_{s,t} \otimes \mathbf{X}_{s,t})] \\
&= \mathrm{tr}(\mathrm{Var}(\mathbf{X}_t) + \mathrm{Var}(\mathbf{X}_s) - 2\mathrm{Cov}(\mathbf{X}_s, \mathbf{X}_t)) \\
&= 2\,\mathrm{tr}((I_d - e^{-A|t-s|})\Sigma) \\
&\leq 2\,\|I_d - e^{-A|t-s|}\|\,\|\Sigma\| \\
&\leq 2\,(\|A\||t-s|e^{\|A\||t-s|})\,\|\Sigma\| \\
&\lesssim |t-s|.
\end{aligned}$$

- Assumption (Aδ) holds for any $p \geq 2$ with $\delta = 1$. Using the integral representation of the OU process, one can easily verify that for all $0 \leq s \leq t \leq T$, $\mathbb{E}_s[\mathbf{X}_t] := \mathbb{E}[\mathbf{X}_t|\mathcal{F}_s] = e^{-A|t-s|}\mathbf{X}_s$ where $\{\mathcal{F}_t,\ t \in [0,T]\}$ is the natural filtration of $\mathbb{X}$. Then, for all $0 \leq s \leq t \leq T$,

$$\|\mathbb{E}_s[\mathbf{X}_{s,t}]\|_{L^p} = \|(e^{-A|t-s|} - I_d)\mathbf{X}_s]\|_{L^p} \leq \|I_d - e^{-A|t-u|}\|\,\|\mathbf{X}_s\|_{L^p} \lesssim |t-s|.$$

- The covariance of the increments is homogeneous since for all $u, v, s, t \geq 0$,

$$\begin{aligned}
\mathrm{Cov}\,(\mathbf{X}_{u,v}, \mathbf{X}_{s,t}) &= \mathrm{Cov}\,(\mathbf{X}_v, \mathbf{X}_t) - \mathrm{Cov}\,(\mathbf{X}_v, \mathbf{X}_s) - \mathrm{Cov}\,(\mathbf{X}_u, \mathbf{X}_t) + \mathrm{Cov}\,(\mathbf{X}_u, \mathbf{X}_s) \\
&= (e^{-A|t-v|} - e^{-A|s-v|} - e^{-A|t-u|} + e^{-A|s-u|})\Sigma,
\end{aligned}$$

  depends only on the relative distances $|t-v|, |s-v|, |t-u|$ and $|s-u|$.

- The covariance of the increments satisfies Assumption (Aθ) with $m = 0$ and $\theta(t) = e^{-\lambda_A t}$ where λ_A is a constant depending on the drift matrix $A \in \mathbb{R}^{d\times d}$. For all $0 \leq u \leq v < s \leq t$,

$$\begin{aligned}
\|\mathrm{Cov}\,(\mathbf{X}_{u,v}, \mathbf{X}_{s,t})\,\| &= \|e^{-A|s-v|}(e^{-A|t-s|} - I_d - e^{-A|t-s|-A|v-u|} + e^{-A|v-u|})\Sigma\| \\
&= \|e^{-A|s-v|}(e^{-A|t-s|} - I_d)(I_d - e^{-A|v-u|})\Sigma\| \\
&= \|e^{-A|s-v|}\|\,\|I_d - e^{-A|t-s|}\|\,\|I_d - e^{-A|v-u|}\|\,\|\Sigma\| \\
&\lesssim e^{-\lambda_A|s-v|}|t-s||v-u|,
\end{aligned}$$

  where in the last step we use the fact that $A \in \mathbb{R}^{d\times d}$ has positive real parts of all eigenvalues to find $\lambda_A \in (0, \min_{\lambda\in\sigma(A)} \mathrm{Re}(\lambda))$ such that $\|e^{-At}\| \lesssim e^{-\lambda_A t}$ for all $t \geq 0$.

Note that Assumption (Aα) and Assumption (Aδ) could have been alternatively established by noting that the OU process is an Itô diffusion with Lipschitz continuous coefficients and applying the results of Example 4.4.13.

Example 4.4.37 (Fractional Brownian motion (cont.)). If $\{X_t^H,\ t \geq 0\}$ is a (one-dimensional) fractional Brownian motion with Hurst parameter $H > 1/2$, i.e. a mean-zero Gaussian process with

covariance

$$C^H(s,t) := \mathrm{Cov}(X_s^H, X_t^H) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

then we can show that:

- Assumption (Aα) holds with $\alpha = H > 1/2$ (and $p = 2$) by Example 4.4.15.
- The covariance of the increments is homogeneous since for all $u, v, s, t \geq 0$,

$$\begin{aligned}\mathrm{Cov}\left(X_{u,v}^H, X_{s,t}^H\right) &= \mathrm{Cov}\left(X_v^H, X_t^H\right) - \mathrm{Cov}\left(X_v^H, X_s^H\right) - \mathrm{Cov}\left(X_u^H, X_t^H\right) + \mathrm{Cov}\left(X_u^H, X_s^H\right) \\ &= \frac{1}{2}(|s-v|^{2H} + |t-u|^{2H} - |t-v|^{2H} - |s-u|^{2H}),\end{aligned}$$

  depends only on the relative distances $|t-v|, |s-v|, |t-u|$ and $|s-u|$.
- The covariance of the increments satisfies Assumption (Aθ) with $m = 3$ and $\theta(t) = t^{2H-2}$. For all $0 \leq u \leq v < s \leq t$ with $|s-v| \geq 3/2(|t-s| + |v-u|)$,

$$\begin{aligned}&|\mathrm{Cov}\left(X_{u,v}^H, X_{s,t}^H\right)| \\ &= \frac{1}{2}||s-v|^{2H} + |t-u|^{2H} - |t-v|^{2H} - |s-u|^{2H}| \\ &= \frac{1}{2}\Big||s-v|^{2H} \\ &\qquad + \sum_{n=0}^{\infty} \frac{(2H)\cdots(2H-n+1)}{n!}|s-v|^{2H-n}(|t-u| - |s-v|)^n \\ &\qquad\quad - \sum_{n=0}^{\infty} \frac{(2H)\cdots(2H-n+1)}{n!}|s-v|^{2H-n}(|t-v| - |s-v|)^n \\ &\qquad\quad - \sum_{n=0}^{\infty} \frac{(2H)\cdots(2H-n+1)}{n!}|s-v|^{2H-n}(|s-u| - |s-v|)^n\Big| \\ &= \frac{1}{2}\Big|\sum_{n=2}^{\infty} \frac{(2H)\cdots(2H-n+1)}{n!}|s-v|^{2H-n}\Big[(|t-s| + |v-u|)^n - |t-s|^n - |v-u|^n\Big]\Big| \\ &= \frac{1}{2}|s-v|^{2H-2}\Big|\sum_{n=2}^{\infty} \frac{(2H)\cdots(2H-n+1)}{n!}|s-v|^{-(n-2)}\sum_{j=1}^{n-1}\binom{n}{j}|t-s|^j|v-u|^{n-j}\Big| \\ &\leq \frac{1}{2}|s-v|^{2H-2}|t-s||v-u| \\ &\qquad \times \sum_{n=2}^{\infty} \frac{|(2H)\cdots(2H-n+1)|}{(n-2)!}|s-v|^{-(n-2)}\sum_{j=0}^{n-2}\binom{n-2}{j}|t-s|^j|v-u|^{(n-2)-j} \\ &\leq \frac{1}{2}|s-v|^{2H-2}|t-s||v-u||2H||2H-1|\sum_{n=0}^{\infty}|s-v|^{-n}(|t-s| + |v-u|)^n \\ &\leq \frac{1}{2}|s-v|^{2H-2}|t-s||v-u||2H||2H-1|\left(1 - \frac{|t-s| + |v-u|}{|s-v|}\right)^{-1} \\ &\leq |2H||2H-1||s-v|^{2H-2}|t-s||v-u|,\end{aligned}$$

by using the fact that that $x \mapsto f(x) = x^{2H}$ is analytic around $|s - v| > 0$ and $|s - v| \geq 3/2(|t - s| + |v - u|)$.

4.5 Simulations

In this section we explore the finite sample properties of the expected signature estimator given in Equation (4.4.15) whose asymptotic properties we established in Section 4.4 in Theorem 4.4.19, Theorem 4.4.29 and Theorem 4.4.34. We consider four different processes, displaying a variety of different dynamics:

(BM) A bidimensional standard Brownian motion $\{\mathbf{B}_t, t \geq 0\}$. This is arguably the most fundamental stochastic process, satisfying many desirable properties and serving as a building block for more complex models. Since Brownian motion is a semimartingale, it is a canonical geometric stochastic process and any sequence of partitions with vanishing mesh size is signature defining, cf. Theorem 4.3.8. It can be easily checked it satisfies Assumption (Aα) and Assumption (Aδ) for any $\alpha \geq 1/2, \delta \geq 1$ and $p \geq 2$. Moreover, $\{\mathbf{B}_t, t \geq 0\}$ has stationary and independent increments and hence the i.i.d. and chopped sampling schemes are equivalent: we can thus apply Theorem 4.4.19 to deduce consistency and asymptotic normality of the expected signature estimator under both sampling schemes. As an additional remark note that Brownian motion is also a Gaussian process with constant mean function and time-homogeneous covariance of the increments trivially satisfying Assumption (Aθ) with $\theta \equiv 0$ and $m = 0$, it thus also falls under the scope of Theorem 4.4.34.

(fBm) A bidimensional fractional Brownian motion $\{\mathbf{B}^H_t, t \geq 0\}$ with Hurst parameter $H = 0.75$, i.e. such that its components are independent one-dimensional fractional Brownian motions with Hurst parameter H. This provides an example of a process which is not a semimartingale but that, as shown in Coutin and Qian (2002), is a canonical geometric stochastic process with dyadic signature defining sequence (as long as $H > 1/4$). Note that, by Example 4.4.15, when $H > 1/2$ the assumptions of Theorem 4.4.4 are fulfilled and hence, when paths are independently sampled, we can apply Theorem 4.4.19. When instead paths are obtained by chopping up a single long realization we can apply Theorem 4.4.34, cf. Example 4.4.37, to deduce consistency of the expected signature estimator. Note that the increments of fractional Brownian motion are not strongly mixing (Mandelbrot and Van Ness, 1968) and hence we cannot apply the second part of Theorem 4.4.29 to deduce asymptotic normality.

(MCAR) A bidimensional continuous-time autoregressive process $\{\mathbf{Y}_t, t \geq 0\}$ of order $p = 2$ driven by a standard Brownian motion with drift parameters $\mathbf{A} = (A_1, A_2)$,

$$A_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

These processes represent the continuous-time generalization of classic discrete-time AR(p) models. They are defined as the first d entries of their state space representation $\{\mathbf{X}_t,\ t \geq 0\}$: an Ornstein-Uhlenbeck process with a specific drift matrix $A_{\mathbf{A}}$ and diffusion matrix Σ, cf. Chapter 3. In the case of a bidimensional MCAR process of order $p = 2$ driven by a standard Brownian motion these are given by

$$A_{\mathbf{A}} = -\begin{pmatrix} 0_{2\times 2} & I_{2\times 2} \\ -A_2 & -A_1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & I_{2\times 2} \end{pmatrix}.$$

The MCAR process is therefore a canonical geometric stochastic process (since it is a semimartingale) and any sequence of partitions with vanishing mesh size is signature defining. Moreover, note that since $\{\mathbf{X}_t,\ t \geq 0\}$ is a Gaussian process, so is $\{\mathbf{Y}_t,\ t \geq 0\}$. We can apply Example 4.4.13 to deduce that $\{\mathbf{X}_t,\ t \geq 0\}$ (and hence $\{\mathbf{Y}_t,\ t \geq 0\}$) satisfies Assumption (Aα) and Assumption (Aδ) for $\alpha = 1/2$, $\delta = 1$ and any $p \geq 2$. When paths are independently sampled we can hence apply Theorem 4.4.19 to obtain both consistency and asymptotic normality results when choosing the sequence of partitions appropriately. Moreover, under the classic assumption that $A_{\mathbf{A}}$ has positive real parts of all eigenvalues[27] and the process is started in its stationary distribution, then $\{\mathbf{X}_t,\ t \geq 0\}$ and $\{\mathbf{Y}_t,\ t \geq 0\}$ are stationary, ergodic and strongly mixing with strong mixing coefficient $\alpha(t) = \mathcal{O}(e^{-at})$ for some $a > 0$ (Marquardt and Stelzer, 2007). We can hence apply Proposition 4.4.23 and Proposition 4.4.28 to deduce that $\{\mathbb{Y}^n,\ n \geq 1\}$ is stationary and strongly mixing (and hence ergodic) with strong mixing coefficient $\alpha(n) = \mathcal{O}(e^{-anT})$. The strong mixing coefficients thus satisfy Equation (4.4.20) for (any) $\zeta > 0$ and hence we can apply both the consistency and asymptotic normality results of Theorem 4.4.29 under a suitable choice of partitions $\{\Pi(N),\ N \geq 1\}$. As a final remark, note that, like for Brownian motion, the MCAR process is also a Gaussian process which, by Example 4.4.36, satisfies Assumption (Aθ). Under a sequence of expanding dyadic refinements, we could hence alternatively establish consistency of the expected signature estimator by applying Theorem 4.4.34.

(Heston) The joint price-variance dynamics of a Heston model under the risk-neutral measure[28] $\mathbb{Q}$ with zero interest rate and no dividends, i.e. $\{(S_t, V_t),\ t \geq 0\}$ such that

$$\begin{aligned} \mathrm{d}S_t &= \sqrt{V_t} S_t \mathrm{d}W_t^S \\ \mathrm{d}V_t &= \kappa(\theta - V_t)\mathrm{d}t + \xi\sqrt{V_t}\mathrm{d}W_t^V \end{aligned}$$

[27] In this chapter we consider an OU process to be defined by $\mathrm{d}\mathbf{X}_t = -A\mathbf{X}_t\mathrm{d}t + \mathrm{d}\mathbf{W}_t^{\mathbf{\Sigma}}$. In the referenced work Marquardt and Stelzer (2007) and in Chapter 3 we defined the OU process through $\mathrm{d}\mathbf{X}_t = A\mathbf{X}_t\mathrm{d}t + \mathrm{d}\mathbf{W}_t^{\mathbf{\Sigma}}$, the conditions on the eigenvalues of the matrix A are hence of opposite sign.

[28] To be precise, the Heston model is incomplete (there is only one asset $\{S_t,\ t \geq 0\}$ and two sources of risk) and hence the risk free measure is not uniquely defined. To obtain a proper pricing measure one would need to add a volatility dependent asset to the market and choose the values of κ, θ and ξ under which such (discounted) asset is also a martingale.

where $\{W_t^S,\ t \geq 0\}$ and $\{W_t^V,\ t \geq 0\}$ are standard Brownian motions with correlation $\langle W^S, W^V \rangle_t = \rho t$. Note that we can rewrite the Heston dynamics as

$$\mathrm{d}\begin{pmatrix} S_t \\ V_t \end{pmatrix} = \begin{pmatrix} 0 \\ \kappa(\theta - V_t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sqrt{V_t} S_t & 0 \\ \rho\xi\sqrt{V_t} & \sqrt{1-\rho^2}\xi\sqrt{V_t} \end{pmatrix} \mathrm{d}\begin{pmatrix} W_t^1 \\ W_t^2 \end{pmatrix},$$

where $\{\mathbf{W}_t = (W_t^1, W_t^2),\ t \geq 0\}$ is a bidimensional standard Brownian motion. The Heston process is therefore a canonical geometric stochastic process (since it is a semimartingale) and any sequence of partitions with vanishing mesh size is signature defining. Under the Feller condition $2\kappa\theta > \xi^2$ the variance process is strictly positive (and so is $\{S_t,\ t \geq 0\}$). The Heston model is thus an Itô diffusion with Lipschitz drift $f : \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}^2$ and 1/2-Hölder continuous diffusion $\sigma : \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}^{2\times 2}$. We can thus apply the third case of Example 4.4.13 to prove that $\{(S_t, V_t),\ t \geq 0\}$ satisfies Assumption (Aα) and Assumption (Aδ) with $\alpha = 1/2$, $\delta = 1$ and any $p > 2$ for deterministic initial conditions $S_0 = s_0$ and $V_0 = v_0$. When paths are sampled independently from this model we can hence apply Theorem 4.4.19 to deduce consistency and asymptotic normality of the expected signature estimator. In the experiments below we set $s_0 = 1, v_0 = 0.1, \theta = 0.1, \kappa = 0.6, \xi = 0.2$ and $\rho = -0.15$.

In all the simulations we fix $T = 1$ and consider the two observational schemes described in Section 4.4.2, i.e. for each $N \geq 1$ we consider

(ind) N independent observations over the partitions $\pi_{N,1}, \ldots, \pi_{N,N}$,

(chop) a single long observation over the partition $\Pi(N) = \pi_{N,1} \cup (T + \pi_{N,2}) \cup \cdots \cup ((n-1)T + \pi_{N,N})$,

where in both cases the partitions $\pi_{N,1} = \ldots = \pi_{N,N}$ are uniform with mesh $|\pi_{N,n}| = 2^{-\lfloor N/10 \rfloor + 1}$. This choice ensures the sequences of partitions are signature-defining for all four processes described above and satisfy the conditions necessary for consistency and asymptotic normality, cf. Remark 4.4.21. In Figure 4.5 we plot one realization of each of the four stochastic processes, noting the very different dynamics they model. The code used to produce the simulations is available at `https://github.com/lorenzolucchese/esig` and leverages the `signatory` and `iisignature` packages for efficient signature computations in Python.

4.5.1 Consistency

In Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9 we plot the histograms over 10,000 samples of the expected signature estimates (4.4.15) up to level $k = 3$. We report exact values of the BM, fBm and MCAR expected signatures: for BM these are given by Theorem 4.2.1, for fBm with $H > 1/2$ these can be found in Baudoin and Coutin (2007, Theorem 31) and for MCAR we extract them from the expected signature of the OU state-space process computed using Ni's PDE approach (Ni, 2012). For the Heston model we do not compute the exact expected signature values as this goes beyond the scope of this work.

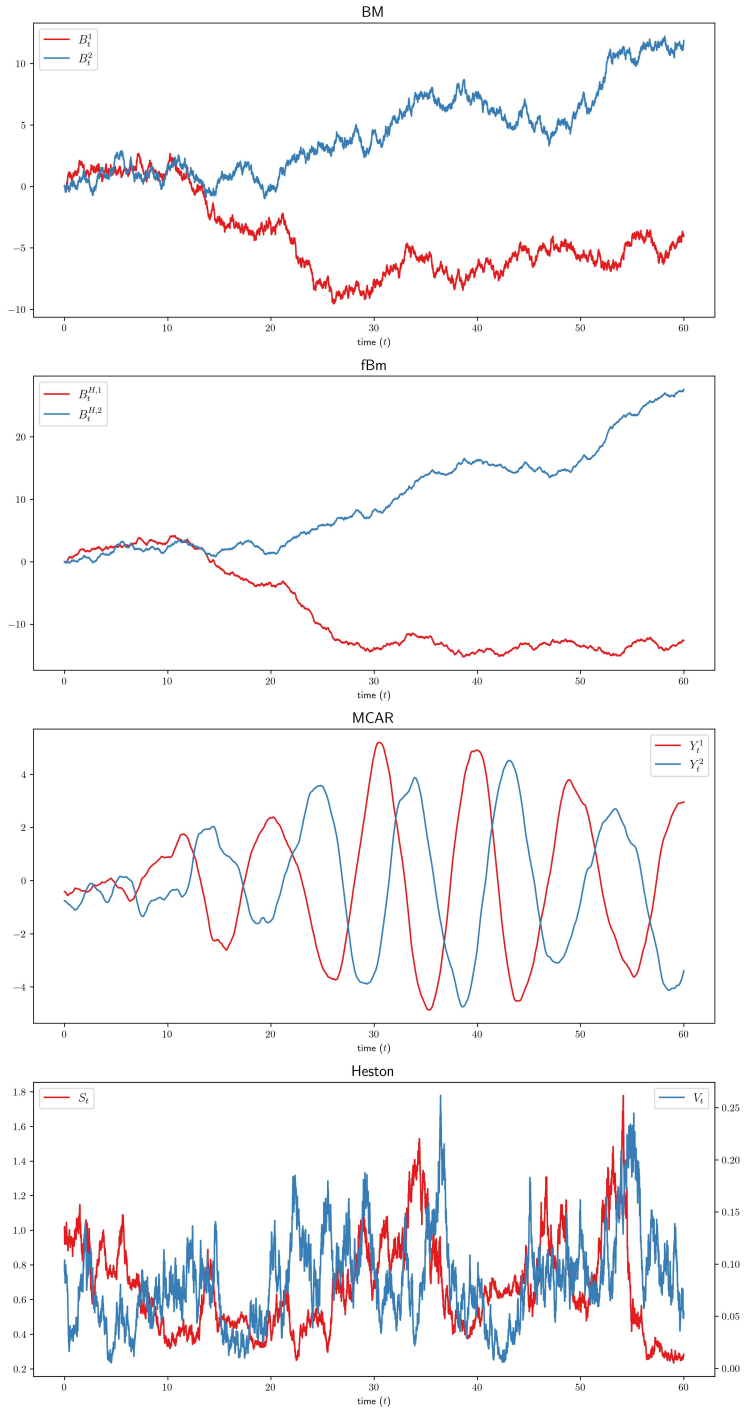


Figure 4.5: Path realizations of the four processes. From top to bottom: a Brownian motion, a fractional Brownian motion, a multivariate CAR process and a price-variance process with Heston dynamics.

In Section 4.4 we proved the asymptotic properties of the expected signature estimator by decomposing it as a sum of the average error due to discretization (in-fill error) and the sampling error around the true signature of the process (long-span error), cf. Equation (4.4.3). An alternative

decomposition, which can be more easily distinguished in the plots, is given by

$$\hat{\phi}_{\mathbf{I}}^{\pi,n}(T) - \phi_{\mathbf{I}}(T) = \left(\frac{1}{N} \sum_{n=1}^{N} S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]} - \mathbb{E}[S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]}] \right) + \left(\mathbb{E}[S^{\mathbf{I}}(\mathbb{X}^{\pi,n})_{[0,T]}] - \mathbb{E}\left[S^{\mathbf{I}}(\mathbb{X})_{[0,T]} \right] \right),$$

where the first term is a mean-zero sampling error and the second term is the bias due to the discretization (independent of the sampling scheme). Note that in all four cases the theoretical consistency results are verified for BM, fBm and MCAR under both observational regimes, and for the Heston model under the (iid) sampling scheme: the L^2 error, measured as $\text{bias}^2 + \text{Var}$, decreases as N increases.

As expected, for BM the (iid) and (chop) schemes coincide[29]. The effect of the in-fill discretization is clear for low values of N (e.g. observe the bias of the distributions for $N = 20$ and $I = (0,0),(1,1)$) while for larger N the error can be mainly attributed to sampling. For fBm and MCAR the expected signature under (chop) observations performs worse than (iid) observations. It is interesting to note that for fBm the difference in performance, due to the long-range persistency in the increments, occurs mainly in the first level. On the other hand, for the MCAR process, the performance gap between the two schemes is marked also in the second signature level. For fBm the discretization error is small (the distributions are close to being unbiased, even for small values of N) while for the MCAR process we observe significant discretization error in the second level of the signature. For the Heston model, without knowing the exact values, we cannot comment on the discretization error. The sampling error though clearly decreases with N.

4.5.2 Asymptotic normality

In Figure 4.10, Figure 4.11 and Figure 4.12 we explore the distributions[30] of the z-statistics

$$\mathbf{Z}^{\Pi(N)} = \sqrt{N} \left(\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)} \right)^{-1/2} \left(\hat{\phi}_{\mathbf{I}}^{\Pi(N)} - \phi_{\mathbf{I}}(T) \right) \in \mathbb{R}^{|\mathbf{I}|},$$

where $\left(\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)} \right)^{-1/2}$ denotes the inverse of the Cholesky decomposition of the estimator for the covariance matrix $\Sigma_{\mathbf{I}}$ required to make the asymptotic normality result feasible. The covariance estimators are discussed in Remark 4.4.22 and Remark 4.4.30 for the independent and the chopped observational schemes respectively. As discussed at the beginning of this section, the theoretical results presented in this chapter ensure asymptotic normality of the expected signature estimators under the independent sampling scheme for all processes under consideration. Moreover, since in all cases the in-fill conditions hold for arbitrary $p \geq 2$, the covariance estimator is also consistent, cf. Remark 4.4.22, and hence the z-statistic will follow a multivariate standard normal distribution asymptotically. Under the chopped observational scheme instead we only have asymptotic normality results for BM and MCAR processes. In both cases the processes are stationary and satisfy the in-fill

[29] Perfect correspondence in the distributions is due to the way the path samples were seeded.

[30] We only consider BM, fBm and MCAR as these are the processes for which we computed the exact value of the expected signature, $\phi_{\mathbf{I}}(T)$.

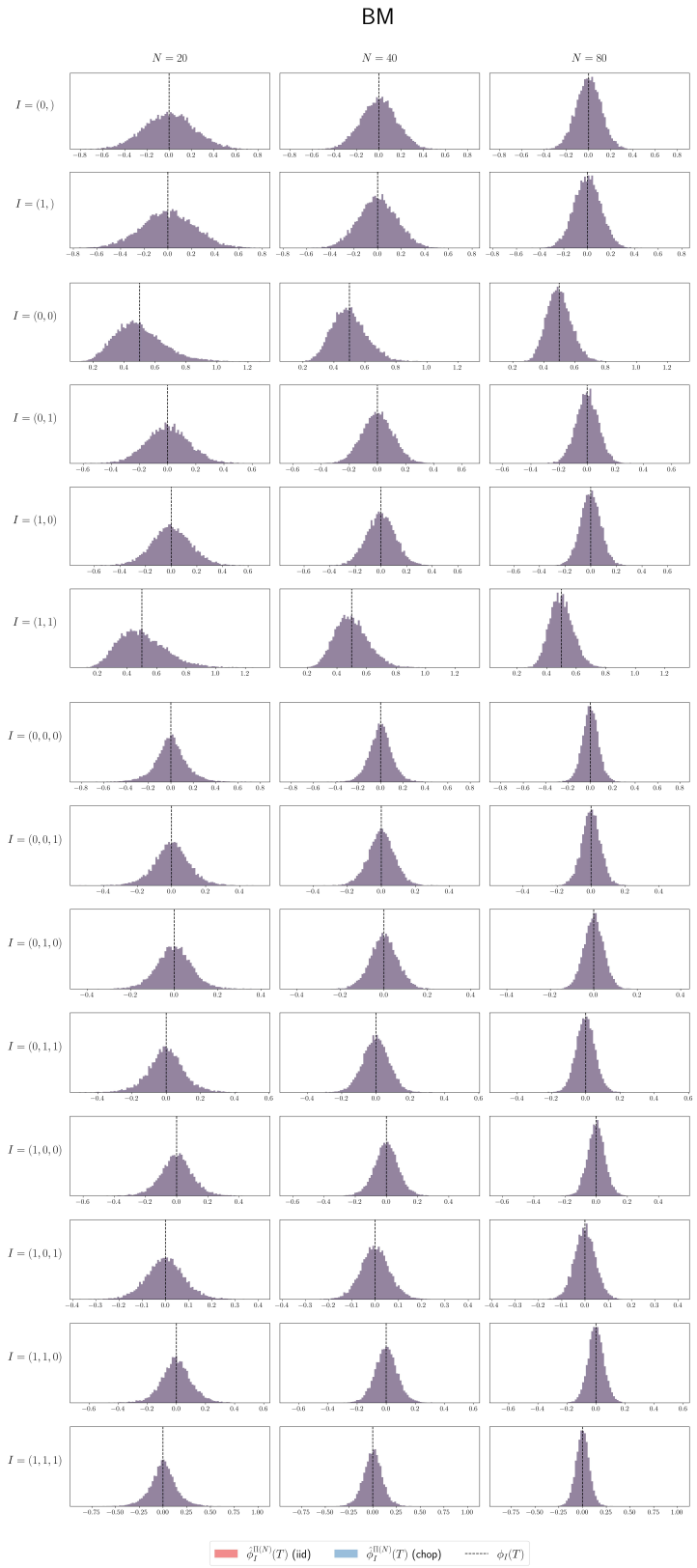


Figure 4.6: Distributions of the expected signature estimators for the BM process: (iid) sampling scheme in red, (chop) sampling scheme in blue and exact values in black. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$.

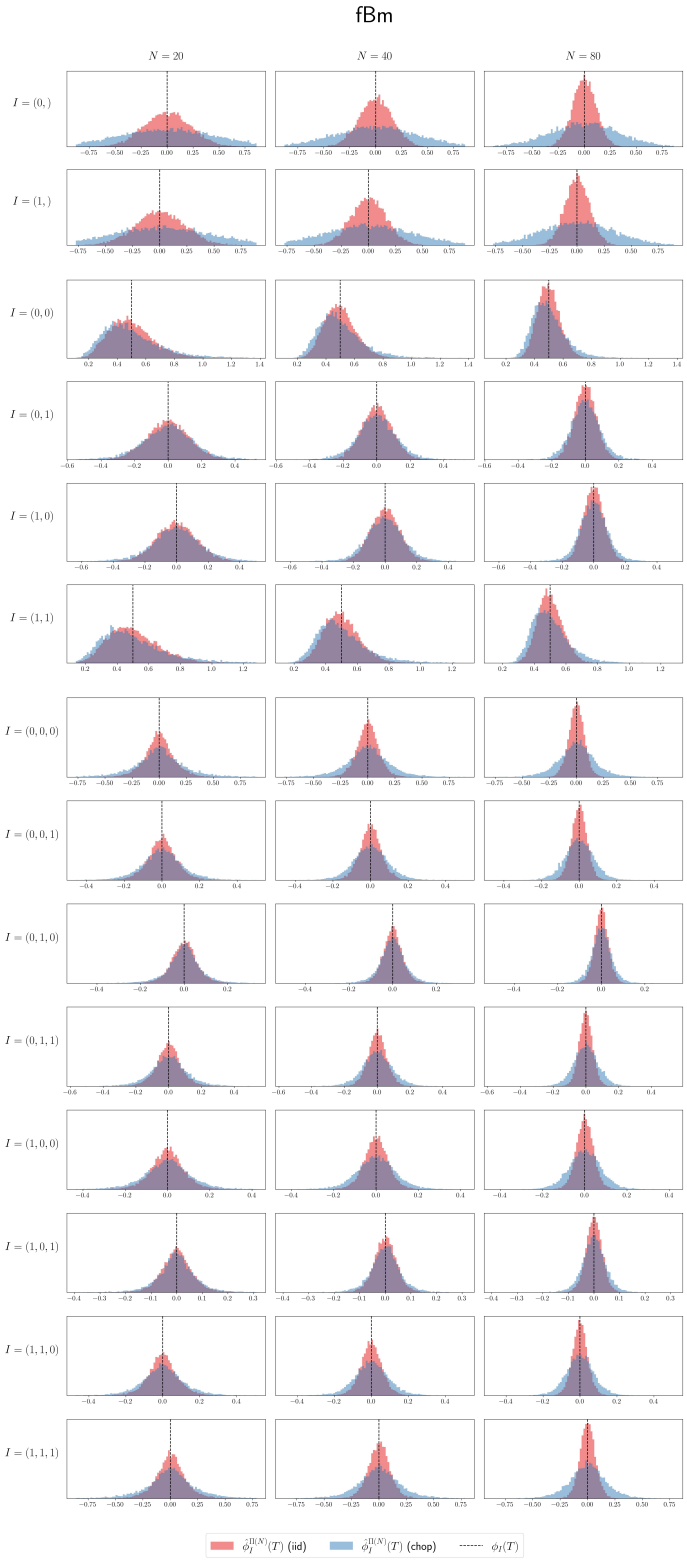


Figure 4.7: Distributions of the expected signature estimator for the fBm process: (iid) sampling scheme in red, (chop) sampling scheme in blue and exact values in black. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$.

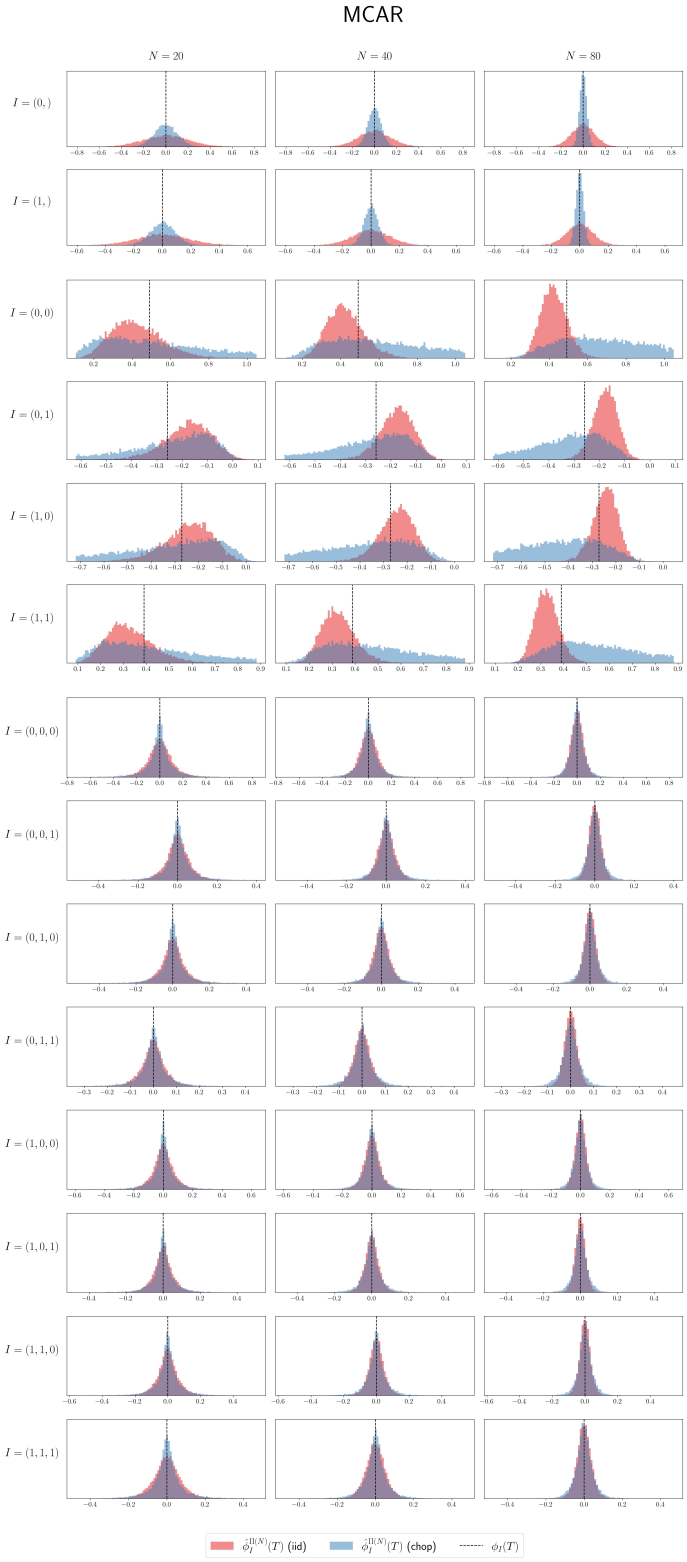


Figure 4.8: Distributions of the expected signature estimator for the MCAR process: (iid) sampling scheme in red, (chop) sampling scheme in blue and exact values in black. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$.

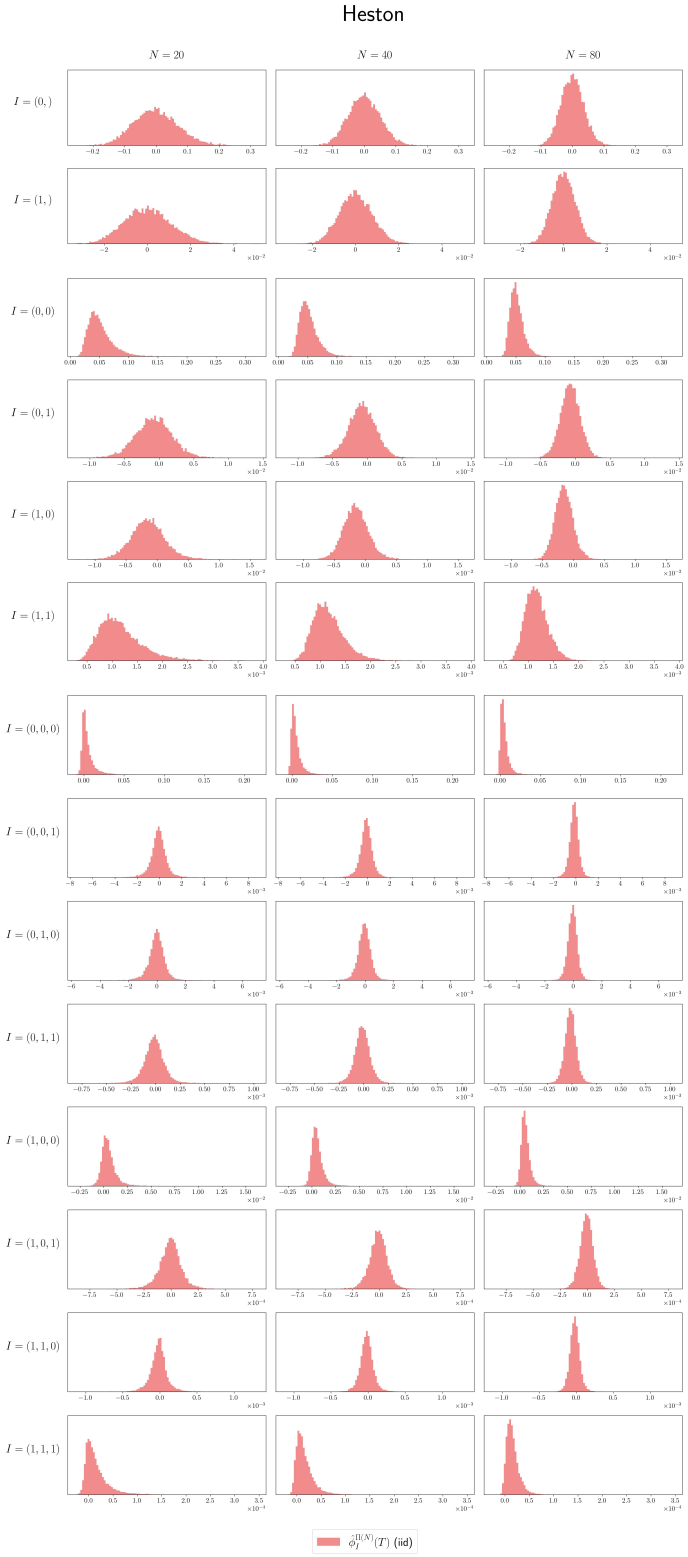


Figure 4.9: Distribution of the expected signature estimator for the Heston process with (iid) sampling scheme. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$.

conditions for arbitrary $p \geq 2$ and hence the covariance estimator is consistent, cf. Remark 4.4.30, ensuring $\mathbf{Z}^{\Pi(N)}$ is indeed an asymptotic z-statistic. Note that, under the chopped observational scheme, the covariance matrix estimator $\hat{\Sigma}_{\mathbf{I}}^{\Pi(N)}$ might not be strictly positive definite. When this is the case, we set the corresponding z-statistic to zero.

As in the previous section, we consider all words in the first three levels of the expected signature and hence $\mathbf{Z}^{\Pi(N)}$ is a $|\mathbf{I}| = 2 + 4 + 8 = 14$-dimensional vector (the processes are of dimension $d = 2$). Plotting a 14-dimensional distribution is quite a challenging task, we thus make do with the (superposition) of the marginals and the empirical correlation matrix. Again we consider $N = 20, 40, 80$.

Under the (iid) sampling scheme, the z-statistic appears to be converging to a multivariate normal at a similar rate across the three processes, i.e. the rate at which the cumulative marginal standard deviation approaches one and the mean, skewness and kurtosis approach zero. While the correlation matrix clearly whitens as N increases, the correlations which take longer to disappear are the same across the three processes. These correlations appear to define three clusters of the same dimensionality as the three signature levels. The clustering is much more pronounced when considering the correlation matrix of the expected signature estimator, but we do not report this here.

We now consider the distribution of the z-statistics under the (chop) sampling scheme for the BM and the MCAR process. Note that when $N = 20$ most long-run covariance estimates are not positive definite, yielding point masses at zero. When $N = 40, 80$ instead the z-statistics display qualitatively different behaviors for the two processes. The BM z-statistics distribution is very similar to the (iid) case. Since the estimates are the same under the (iid) and (chop) sampling schemes, the difference is entirely driven by the long-run covariance converging to the sample covariance as N increases, i.e. the cross-covariance terms vanishing. On the other hand, the distribution of the z-statistics for the MCAR expected signature estimates converges to a standard multivariate normal slower under the (chop) sampling than under (iid) sampling, displaying a weird hump even for large values of N.

4.6 Estimating the expected signature of a martingale

4.6.1 Martingale correction: variance reduction via control variates

In Section 4.4 we developed the necessary theory to establish the asymptotic properties of the estimator (4.4.15) for the statistic

$$\phi_I(T) = \mathbb{E}[S^I(\mathbb{X})_{[0,T]}],$$

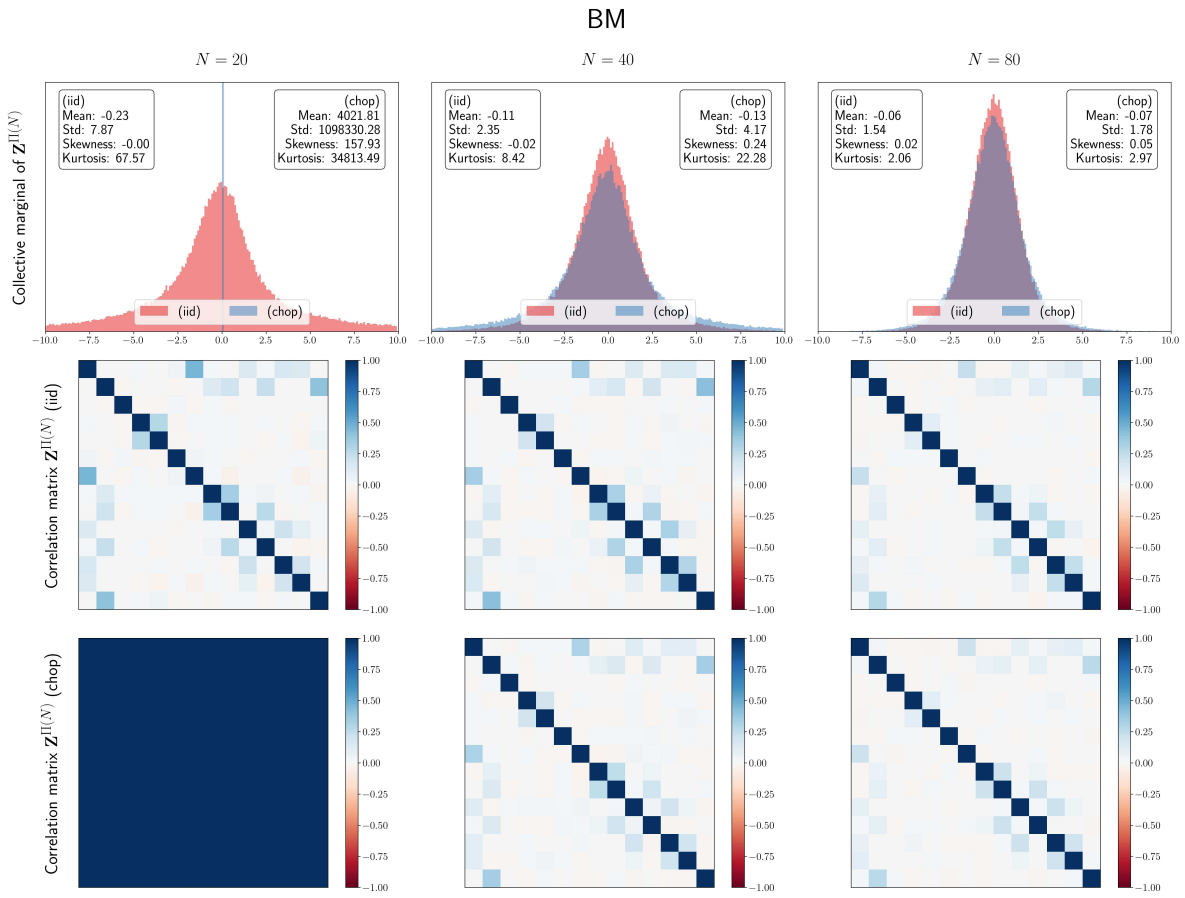


Figure 4.10: Distributions of the z-statistics of the expected signature estimators for the BM process. First row: collective marginal with (iid) sampling in red and with (chop) sampling in blue. Second and third row: correlation matrices with (iid) sampling and (chop) sampling respectively. From left to right: $N = 20, 40, 80$.

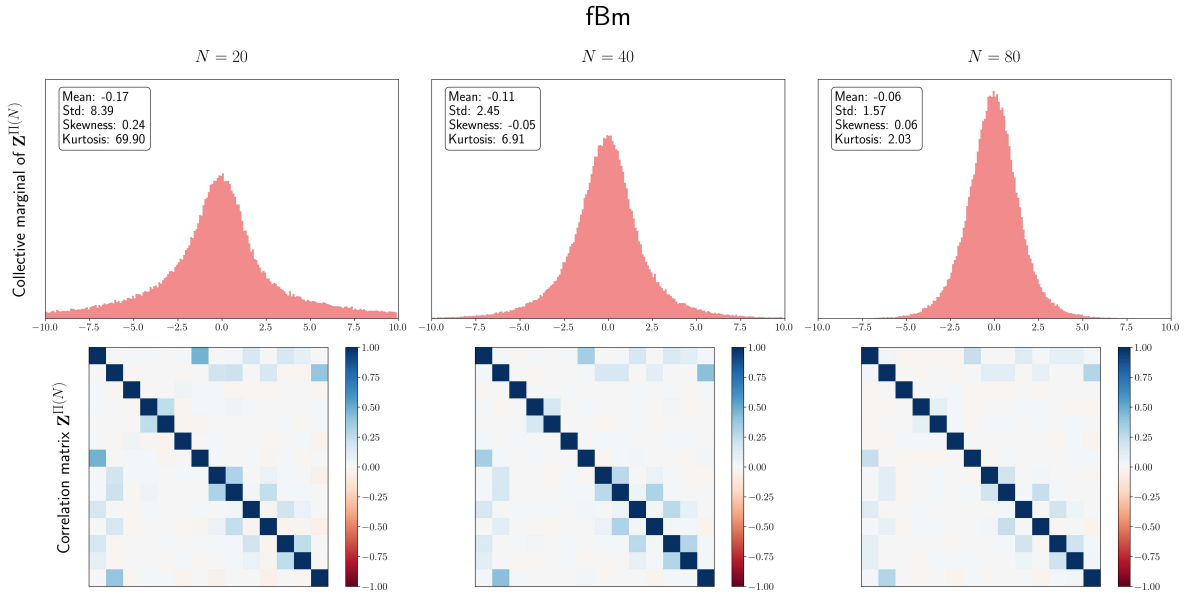


Figure 4.11: Distribution of the z-statistics of the expected signature estimators for the fBm process under the (iid) sampling scheme. Top row: collective marginal. Bottom row: correlation matrix. From left to right: $N = 20, 40, 80$.

for any word $I = (i_1, \dots, i_k)$. This section aims to find an alternative estimator with better finite sample properties when the process $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ is a martingale[31]. We restrict ourselves to the independent observation setting, with the same partition across samples, i.e. $\pi_{N,n} = \pi$ for $n = 1, \dots, N$. We will hence be considering the estimator

$$\hat{\phi}_I^{N,\pi}(T) := \frac{1}{N} \sum_{n=1}^{N} S^I(\mathbb{X}^{n,\pi})_{[0,T]}, \tag{4.6.1}$$

where for each $N \geq 1$ the $\mathbb{X}^{n,\pi}$ are i.i.d. piecewise linear observations of $\mathbb{X}$ over the partition π (note that by the results in Section 4.3.1 any sequence of partitions with vanishing mesh size is signature-defining). Let us start by considering the unfeasible estimator

$$\hat{\phi}_I^{N}(T) := \frac{1}{N} \sum_{n=1}^{N} S^I(\mathbb{X}^{n})_{[0,T]},$$

where the $\mathbb{X}^n$ are i.i.d. copies of $\mathbb{X}$ and, setting $I_{-1} := (i_1, \dots, i_{k-1})$,

$$S^I(\mathbb{X})_{[0,T]} = \int_{0<u_1<\dots<u_k<T} \cdots \int \circ \, \mathrm{d}X_{u_1}^{(i_1)} \otimes \cdots \otimes \circ \, \mathrm{d}X_{u_k}^{(i_k)} = \int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]} \circ \, \mathrm{d}X_s^{(i_k)},$$

[31] To be fully rigorous, in this section we assume $\mathbb{X}$ is a square-integrable martingale "nice enough" to ensure all Itô integrals can be constructed in the L^2 sense and the resulting Itô integral is also a square-integrable martingale (Jacod and Shiryaev, 1987, Theorem I.4.40).

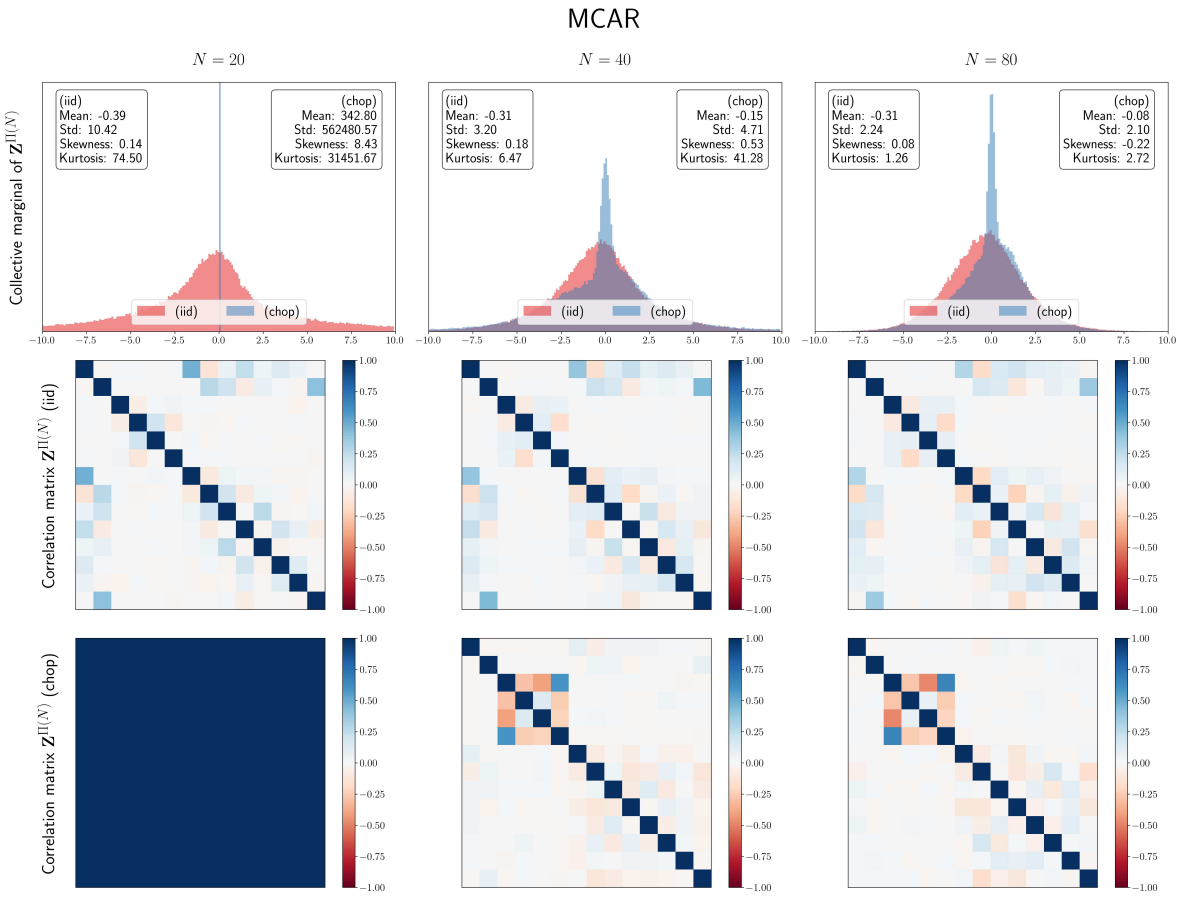


Figure 4.12: Distributions of the z-statistics of the expected signature estimators for the MCAR process. First row: collective marginal with (iid) sampling in red and with (chop) sampling in blue. Second and third row: correlation matrices with (iid) sampling and (chop) sampling respectively. From left to right: $N = 20, 40, 80$.

is defined in the Stratonovich sense. Note that, as long as $S^I(\mathbb{X})_{[0,T]} \in L^2$ or, equivalently, $\phi_K(T)$ is finite for all $K \in I \sqcup\!\sqcup I$, the arguments in Section 4.4.2 imply the estimator is consistent and asymptotically normal for $\phi_I(T)$. To reduce the variance of the estimator while keeping it unbiased we aim to find a mean-zero control $S_c^I(\mathbb{X})_{[0,T]}$ that is highly correlated with $S^I(\mathbb{X})_{[0,T]}$. The form of $S^I(\mathbb{X})_{[0,T]}$ quite naturally suggests the control variate

$$S_c^I(\mathbb{X})_{[0,T]} = \int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]} \, \mathrm{d}X_s^{(i_k)}, \tag{4.6.2}$$

where the outermost integral is now interpreted in the Itô sense. Since $\mathbb{X}$ is a martingale satisfying the conditions for L^2 integration[31], $\{S_c^I(\mathbb{X})_{[0,t]},\ t \in [0,T]\}$ is a square-integrable martingale and hence $\mathbb{E}[S_c^I(\mathbb{X})_{[0,T]}] = 0$. We can thus form the control-variate estimator

$$\hat{\phi}_I^{N,c}(T) := \frac{1}{N} \sum_{n=1}^{N} \left(S^I(\mathbb{X}^n)_{[0,T]} - c S_c^I(\mathbb{X}^n)_{[0,T]} \right), \tag{4.6.3}$$

which, like $\hat{\phi}_I^N(T)$, is unbiased but has reduced variance

$$\mathrm{Var}(\hat{\phi}_I^{N,c^*}(T)) = (1 - \rho_I^2)\mathrm{Var}(\hat{\phi}_I^N(T)), \quad \text{with} \quad \rho_I := \mathrm{Corr}(S^I(\mathbb{X})_{[0,T]}, S_c^I(\mathbb{X})_{[0,T]}),$$

when picking the optimal

$$c^* = c_I^* := \frac{\mathrm{Cov}(S^I(\mathbb{X})_{[0,T]}, S_c^I(\mathbb{X})_{[0,T]})}{\mathrm{Var}(S_c^I(\mathbb{X})_{[0,T]})}.$$

Note that we can compute

$$\mathrm{Var}(S^I(\mathbb{X})_{[0,T]}) = \mathbb{E}[S^I(\mathbb{X})_{[0,T]}^2] - \phi_I(T)^2 = \sum_{J \in I \sqcup\!\sqcup I} \phi_J(T) - \phi_I(T)^2.$$

Introducing the $\mathbb{R}^{d+d^2}$-valued process $(\mathbb{X}, \langle \mathbb{X} \rangle) := \{(\mathbf{X}_t, \mathrm{vec}\langle \mathbf{X} \rangle_t),\ t \in [0,T]\}$ with expected signature $\psi(T)$ over the alphabet $\{1, \ldots, d, (1,1), \ldots, (1,d), \ldots, (d,d)\}$, we can rewrite this as

$$\mathrm{Var}(S^I(\mathbb{X})_{[0,T]}) = \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \psi_I(T)^2,$$

and also compute

$$\begin{aligned}
\mathrm{Var}(S_c^I(\mathbb{X})_{[0,T]}) &= \mathbb{E}\left[\left(\int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]} \, \mathrm{d}X_s^{(i_k)} \right)^2 \right] \\
&= \mathbb{E}\left[\int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]}^2 \, \mathrm{d}\langle X^{(i_k)} \rangle_s \right] \\
&= \sum_{J \in I_{-1} \sqcup\!\sqcup I_{-1}} \mathbb{E}\left[\int_0^T S^J(\mathbb{X})_{[0,s]} \, \mathrm{d}\langle X^{(i_k)} \rangle_s \right]
\end{aligned}$$

$$= \sum_{J \in I_{-1} \sqcup\!\sqcup I_{-1}} \mathbb{E}\left[S^{J*((i_k,i_k))}((\mathbb{X}, \langle \mathbb{X} \rangle))_{[0,T]}\right]$$
$$= \sum_{J \in I_{-1} \sqcup\!\sqcup I_{-1}} \psi_{J*((i_k,i_k))}(T),$$

and

$$\begin{aligned}
\mathrm{Cov}(S^I(\mathbb{X})_{[0,T]}, S^I_c(\mathbb{X})_{[0,T]}) &= \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\left(\int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]}\,\mathrm{d}X_s^{(i_k)}\right)\right] \\
&= \mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\left(S^I(\mathbb{X})_{[0,T]} - \frac{1}{2}\left\langle S^{I_{-1}}(\mathbb{X}),\, X^{(i_k)}\right\rangle_T\right)\right] \\
&= \mathbb{E}\left[S^I(\mathbb{X})^2_{[0,T]}\right] - \frac{1}{2}\mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\left\langle \int_0^{\cdot} S^{I_{-2}}(\mathbb{X})_{[0,s]} \circ \mathrm{d}X_s^{(i_{k-1})},\, X^{(i_k)}\right\rangle_T\right] \\
&= \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \frac{1}{2}\mathbb{E}\left[S^I(\mathbb{X})_{[0,T]}\int_0^T S^{I_{-2}}(\mathbb{X})_{[0,s]}\,\mathrm{d}\langle X^{(i_{k-1})},\, X^{(i_k)}\rangle_s\right] \\
&= \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \frac{1}{2}\mathbb{E}\left[S^I((\mathbb{X}, \langle \mathbb{X} \rangle))_{[0,T]} S^{I_{-2}*((i_{k-1},i_k))}((\mathbb{X}, \langle \mathbb{X} \rangle))_{[0,T]}\right] \\
&= \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \frac{1}{2}\sum_{J \in I \sqcup\!\sqcup I_{-2}*((i_{k-1},i_k))} \mathbb{E}\left[S^J((\mathbb{X}, \langle \mathbb{X} \rangle))_{[0,T]}\right] \\
&= \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \frac{1}{2}\sum_{J \in I \sqcup\!\sqcup I_{-2}*((i_{k-1},i_k))} \psi_J(T),
\end{aligned}$$

where $I * J$ denotes concatenation between word I and J.

In practice, to devise a feasible estimator, i.e. depending on the observed piecewise linear paths $\mathbb{X}^{1,\pi}, \ldots, \mathbb{X}^{N,\pi}$, we need to discretize the estimator (4.6.3). We introduce the control-variate modification of the estimator (4.6.1),

$$\hat{\phi}_I^{N,\pi,c}(T) := \frac{1}{N}\sum_{n=1}^{N}(S^I(\mathbb{X}^{n,\pi})_{[0,T]} - cS^I_c(\mathbb{X}^{n,\pi})_{[0,T]}), \tag{4.6.4}$$

where

$$S^I_c(\mathbb{X}^\pi)_{[0,T]} := \sum_{[u,v] \in \pi} S^{I_{-1}}(\mathbb{X}^\pi)_{[0,u]} X^{(i_k)}_{u,v}.$$

Note that, if $\mathbb{X}$ is a martingale, the correction term is mean-zero and hence the control variate estimator $\hat{\phi}_I^{N,\pi,c}(T)$ has the same bias as $\hat{\phi}_I^{N,\pi}(T)$ but, when picking the optimal[32] c, i.e.

$$c^*_\pi := \frac{\mathrm{Cov}(S^I(\mathbb{X}^\pi)_{[0,T]}, S^I_c(\mathbb{X}^\pi)_{[0,T]})}{\mathrm{Var}(S^I_c(\mathbb{X}^\pi)_{[0,T]})},$$

it has reduced variance

$$\mathrm{Var}(\hat{\phi}_I^{N,\pi,c^*_\pi}(T)) = (1 - \rho^2_{I,\pi})\mathrm{Var}(\hat{\phi}_I^{N,\pi}(T)), \quad \text{where} \quad \rho_{I,\pi} := \mathrm{Corr}(S^I(\mathbb{X}^\pi)_{[0,T]}, S^I_c(\mathbb{X}^\pi)_{[0,T]}).$$

[32] We assume throughout $\mathrm{Var}(S^I_c(\mathbb{X}^\pi)_{[0,T]}) > 0$.

Remark 4.6.1. In this section we focus on considering a single control obtained by substituting the outermost Stratonovich integral with an Itô integral. In principle, for a word of length $|I| = k \geq 2$, one could consider 2^{k-2} distinct controls: for any subset $\mathcal{I} \subseteq \{2, \dots, k-1\}$ one can obtain a control by changing each of the integrals with index in $\mathcal{I} \cup \{k\}$ to Itô integrals. One can then apply the controlled linear regression estimator (with only the intercept term as regressor) described in Appendix 4.F. This family of controls will likely be highly correlated and hence:

- the improvements provided by each additional control might be quite marginal compared to the considerable increase in computational cost;
- the estimator of the (inverse) variance matrix of the controls, needed to make the estimator feasible, might be quite unstable.

Hence, for clarity of exposition and computational ease, throughout the rest of this chapter we only consider the control introduced above.

In practice, to estimate c_π^* we can use one of the two following methods:

- The most straightforward approach would be to use the sample variance and covariance. In this case the estimator for c_π^* is the slope of the simple linear regression of $\{S^I(\mathbb{X}^{n,\pi})_{[0,T]},\ n = 1, \dots, N\}$ against $\{S_c^I(\mathbb{X}^{n,\pi})_{[0,T]},\ n = 1, \dots, N\}$ or, exploiting the mean zero property of the control,
$$\hat{c}_{\pi,1}^* = \frac{\sum_{n=1}^N S^I(\mathbb{X}^{n,\pi})_{[0,T]} S_c^I(\mathbb{X}^{n,\pi})_{[0,T]}}{\sum_{n=1}^N S_c^I(\mathbb{X}^{n,\pi})_{[0,T]}^2}.$$
- Alternatively, we can exploit the explicit form of the covariance and variance of the unfeasible estimator and approximate
$$\begin{aligned}
\mathrm{Cov}(S^I(\mathbb{X}^\pi)_{[0,T]}, S_c^I(\mathbb{X}^\pi)_{[0,T]}) &\approx \mathrm{Cov}(S^I(\mathbb{X})_{[0,T]}, S_c^I(\mathbb{X})_{[0,T]}) \\
&= \sum_{J \in I \sqcup\!\sqcup I} \psi_J(T) - \frac{1}{2} \sum_{J \in I \sqcup\!\sqcup I_{-2*((i_{k-1}, i_k))}} \psi_J(T) \\
&\approx \sum_{J \in I \sqcup\!\sqcup I} \hat{\psi}_J^N(T) - \frac{1}{2} \sum_{J \in I \sqcup\!\sqcup I_{-2*((i_{k-1}, i_k))}} \hat{\psi}_J^N(T) \\
&\approx \sum_{J \in I \sqcup\!\sqcup I} \hat{\psi}_J^{N,\pi}(T) - \frac{1}{2} \sum_{J \in I \sqcup\!\sqcup I_{-2*((i_{k-1}, i_k))}} \hat{\psi}_J^{N,\pi}(T) \\
&\approx \sum_{J \in I \sqcup\!\sqcup I} \hat{\psi}_J^{N,\pi,\prime}(T) - \frac{1}{2} \sum_{J \in I \sqcup\!\sqcup I_{-2*((i_{k-1}, i_k))}} \hat{\psi}_J^{N,\pi,\prime}(T),
\end{aligned}$$

  where
$$\hat{\psi}_J^N(T) := \frac{1}{N} \sum_{n=1}^N S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^n)_{[0,T]}, \quad \hat{\psi}_J^{N,\pi}(T) := \frac{1}{N} \sum_{n=1}^N S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^{n,\pi})_{[0,T]},$$

and

$$\hat{\psi}_J^{N,\pi,\prime}(T) := \frac{1}{N}\sum_{n=1}^{N} S^J((\mathbb{X}^{n,\pi}, \langle\hat{\mathbb{X}}\rangle^{n,\pi}))_{[0,T]},$$

with $\langle\hat{\mathbb{X}}\rangle^{\pi} = \{\langle\hat{\mathbf{X}}\rangle_t^{\pi},\ t \in [0,T]\}$ defined as

$$\langle\hat{\mathbf{X}}\rangle_t^{\pi} = \sum_{[u',v']\in\pi_{[0,u]}} \mathbf{X}_{u',v'}^2 + \frac{t-u}{v-u}\mathbf{X}_{u,v}^2, \quad t \in [u,v].$$

Similarly, we can approximate

$$\mathrm{Var}(S_c^I(\mathbb{X}^{\pi})_{[0,T]}) \approx \mathrm{Var}(S_c^I(\mathbb{X})_{[0,T]}) \approx \sum_{J\in I_{-1}\sqcup I_{-1}} \hat{\psi}_{J*((i_k,i_k))}^{N,\pi}(T) \approx \sum_{J\in I_{-1}\sqcup I_{-1}} \hat{\psi}_{J*((i_k,i_k))}^{N,\pi,\prime}(T),$$

and hence we define the second estimator for c_π^* as

$$\begin{aligned}\hat{c}_{\pi,2}^* &= \frac{\sum_{J\in I\sqcup I} \hat{\psi}_J^{N,\pi,\prime}(T) - \frac{1}{2}\sum_{J\in I\sqcup I_{-2*((i_{k-1},i_k))}} \hat{\psi}_J^{N,\pi,\prime}(T)}{\sum_{J\in I_{-1}\sqcup I_{-1}} \hat{\psi}_{J*((i_k,i_k))}^{N,\pi,\prime}(T)} \\ &= \frac{\sum_{n=1}^{N}\Big(\sum_{J\in I\sqcup I} S^J((\mathbb{X}^{n,\pi}, \langle\hat{\mathbb{X}}\rangle^{n,\pi}))_{[0,T]} - \frac{1}{2}\sum_{J\in I\sqcup I_{-2*((i_{k-1},i_k))}} S^J((\mathbb{X}^{n,\pi}, \langle\hat{\mathbb{X}}\rangle^{n,\pi}))_{[0,T]}\Big)}{\sum_{n=1}^{N}\sum_{J\in I_{-1}\sqcup I_{-1}} S^{J*((i_k,i_k))}((\mathbb{X}^{n,\pi}, \langle\hat{\mathbb{X}}\rangle^{n,\pi}))_{[0,T]}}.\end{aligned}$$

Whether estimator $\hat{c}_{\pi,1}^*$ or estimator $\hat{c}_{\pi,2}^*$ is more precise, in terms of MSE, depends on the properties of the process $\mathbb{X}$ and the expected signature word I being estimated by (4.6.4).

Lemma 4.6.2. *Let $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ be a square-integrable martingale satisfying Assumption (Aα.M) for some $\alpha \geq 1/2$ and $p = 4k$ where $k = |I|$. Assume that π is part of a sequence of refining partitions with mesh vanishing fast enough, i.e.*

$$\sum_{n\geq 1} |\pi_n|^{2\alpha-1/2} < \infty.$$

Then the difference between the mean-square errors of the two estimators $\hat{c}_{\pi,1}^*$ and $\hat{c}_{\pi,2}^*$ for c_π^* is approximately

$$\mathbb{E}[(\hat{c}_{\pi,2}^* - c_\pi^*)^2] - \mathbb{E}[(\hat{c}_{\pi,1}^* - c_\pi^*)^2] \approx \frac{1}{N}\frac{\mu_Y}{\mu_Z^3}\left(\frac{\mu_Y}{\mu_Z}(\mathbb{E}[Z_2^2] - \mathbb{E}[Z_1^2]) - 2(\mathbb{E}[YZ_2] - \mathbb{E}[YZ_1])\right),$$

where

$$Y = S^I(\mathbb{X})_{[0,T]}S_c^I(\mathbb{X})_{[0,T]},\ Z_1 = S_c^I(\mathbb{X})_{[0,T]}^2,\ Z_2 = \sum_{J\in I_{-1}\sqcup I_{-1}} S^{J*((i_k,i_k))}((\mathbb{X}, \langle\mathbb{X}\rangle))_{[0,T]},$$

and $\mu_Y = \mathbb{E}[Y]$, $\mu_Z = \mathbb{E}[Z_1] = \mathbb{E}[Z_2]$.

Proof. See Appendix 4.E.1. □

In practical applications, the above expression cannot be evaluated exactly but we can approximate it by its sample estimate

$$\mathbb{E}[(\hat{c}^*_{\pi,2} - c^*_\pi)^2] - \mathbb{E}[(\hat{c}^*_{\pi,1} - c^*_\pi)^2] \propto \sum_{n=1}^{N} \left(\frac{\bar{\mu}_Y}{\bar{\mu}_Z}((Z^\pi_{2,n})^2 - (Z^\pi_{1,n})^2) - \sum_{j=1}^{2}(Y^\pi_{j,n}Z^\pi_{2,n} - Y^\pi_{j,n}Z^\pi_{1,n}) \right).$$

where

$$\bar{\mu}_Y = \frac{1}{2N}\left(\sum_{n=1}^{N} Y^\pi_{1,n} + \sum_{n=1}^{N} Y^\pi_{2,n}\right), \quad \bar{\mu}_Z = \frac{1}{2N}\left(\sum_{n=1}^{N} Z^\pi_{1,n} + \sum_{n=1}^{N} Z^\pi_{2,n}\right),$$

and $Y^\pi_{1,n}, Z^\pi_{1,n}, Y^\pi_{2,n}, Z^\pi_{2,n}$ are given in Appendix 4.E.1.

Another important discriminant when choosing between estimators $\hat{c}^*_{\pi,1}$ and $\hat{c}^*_{\pi,2}$ is usually computational cost. To compute $\hat{c}^*_{\pi,1}$ it suffices to regress $\{S^I(\mathbb{X}^{n,\pi})_{[0,T]},\ n = 1, \dots, N\}$ against $\{S^I_c(\mathbb{X}^{n,\pi})_{[0,T]},\ n = 1, \dots, N\}$. Both samples need to be computed to evaluate the control-variate estimator (4.6.4) and thus the extra computational cost of $\hat{c}^*_{\pi,1}$ is just the cost of a simple linear regression with sample size N, namely $\mathcal{O}(N)$. On the other hand, to compute $\hat{c}^*_{\pi,2}$, one needs to compute all the higher-order expected signature estimates $\hat{\psi}^{N,\pi,\prime}_J(T)$ with $J \in I \sqcup\!\sqcup I$, $J \in I \sqcup\!\sqcup I_{-2} * ((i_{k-1}, i_k))$ and $J = J' * ((i_k, i_k))$ for $J' \in I_{-1} \sqcup\!\sqcup I_{-1}$ which has (naive) extra computational cost $\mathcal{O}(|\pi|^{-1}Tk(d^{2k} + (d+1)^{2k-1}))$ when parallelizing across the N samples. In the in-fill limit, $|\pi|^{-1}T \gg N$ and hence computing $\hat{c}^*_{\pi,2}$ is significantly more expensive than computing $\hat{c}^*_{\pi,1}$.

Remark 4.6.3. This variance reduction technique is not limited to processes $\mathbb{X}$ that are "full" martingales but can also be applied to "partial" martingales, i.e. $\mathbb{X}$ such that only a subset of the components is a martingale. In this case, we can use the control variate expected signature estimator (4.6.4) for any word $I = (i_1, \dots, i_k)$ where $i_k \in \{1, \dots, d\}$ is such that $\mathbb{X}^{(i_k)}$ is a martingale. For example, if the add-time transformation is applied to the data, i.e. as in Example 4.4.16 and Example 4.4.31 the d-dimensional martingale $\mathbb{X}$ is augmented with time to form the $(d+1)$-dimensional process $\hat{\mathbb{X}} = \{(t, \mathbf{X}_t),\ t \in [0,T]\}$, we can use the control variate expected signature estimator (4.6.4) to estimate $\mathbb{E}[S^I(\hat{\mathbb{X}})_{[0,T]}]$ for any word $I = (i_1, \dots, i_k) \in \{1, \dots, d+1\}^k$ with $i_k \neq 1$.

4.6.2 Machine learning applications

Signatures of paths have found widespread use in the machine learning community, with applications ranging from character recognition (Graham, 2013; Xie et al., 2018) to medical diagnosis (Pérez Arribas et al., 2018). Taking the signature of a stream of data is essentially a feature extraction method mapping raw stream-like data to a lower-dimensional but highly-informative latent space. Originally developed in the context of rough path theory, i.e. to study differential equations driven by irregular paths, the theoretical foundations for their efficacy in machine learning applications (where paths are of bounded variation) are multiple, ranging from a characterization result (Hambly

and Lyons, 2005; Cass and Turner, 2022) to a universal approximation theorem (Levin, Lyons, and Ni, 2016, Theorem 3.1). We refer to Lyons and McLeod (2024) for a recent review of the signature method in machine learning applications.

When dealing with collections of paths it is natural to consider the expected signature as a feature extraction method. A natural question is thus whether the desirable characterization properties of the signature carry over to the distributional setting, i.e. is the expected signature characteristic for the full law of the signature and, in turn, of the path? Positive answers to this question can be found in Fawcett (2003), Chevyrev and Lyons (2016) and Chevyrev and Oberhauser (2018) under progressively more general[33] settings and assumptions. While dealing with a collection of paths is arguably a less common setting than a single stream of data, the literature still provides a wide range of machine learning algorithms leveraging expected signatures[34]. These cover many different tasks (from distributional regression to generative modeling) and applications (from ECG classification to option pricing). In this section, we review five of these machine learning algorithms discussing how the martingale correction introduced in Section 4.6.1 can be applied to the expected signature computation step to improve performance. Before diving into each specific application in detail, we discuss some general considerations on the use of the martingale correction term in practice.

In the previous section, we considered the same framework as in the rest of this work, namely the setting where $\mathbb{X}$ is a continuous-time stochastic process and $\mathbb{X}^\pi$ is a piecewise-linear interpolation of the discrete-time observation of such process along the partition π. It is important to note that, while some applications have a natural underlying latent continuous-time model (Lyons, Nejad, and Pérez Arribas, 2021) others do not (Lemercier et al., 2021). In any case, we do not necessarily require the "background" continuous-time model $\mathbb{X}$ to be defined to apply the control-variate estimator (4.6.4) with $c = \hat{c}^*_{1,\pi}$. Letting $\pi = \{0 = t_1 < \cdots < t_M = T\}$, one can easily see that it is sufficient for the discrete-time process $\{\mathbf{X}_m = \mathbf{X}_{t_m},\ m = 1, \dots, M\}$ to be a discrete-time martingale with respect to $\mathcal{F}_m = \sigma(\mathbf{X}_1, \dots, \mathbf{X}_m)$ for the control variate estimator to have the same bias but lower variance than the naive expected signature estimator. In what follows all path observations are inevitably sampled at discrete points in time and hence, abusing notation slightly, in some places we drop the dependence on the partition π. Whether $\mathbb{X}$ is a continuous-time process or a sequence of observations in discrete time should be clear from the context.

Machine learning methods based on signature methods often apply augmentations to the raw streams of data before computing the signature, cf. Lyons and McLeod (2024, Section 2.5). For example, a path augmentation which is often found to improve model performance is the lead-lag transform. Combining the previous observation on discrete-time martingales with Remark 4.6.3, we can easily see that the control variate expected signature estimator can also be employed when the lead-lag augmentation is applied to the raw data, i.e. when the d-dimensional discrete-time

[33] Even moving beyond bounded variation paths.

[34] The GPES algorithm actually takes as input a single stream of data and applies a data augmentation technique to form a collection of paths.

martingale $\{\mathbf{X}_m,\ m = 1, \dots, M\}$ is embedded into the $2d$-dimensional process

$$\mathbb{X}' = \{(\mathbf{X}_1, \mathbf{X}_1), (\mathbf{X}_2, \mathbf{X}_1), (\mathbf{X}_2, \mathbf{X}_2), \dots, (\mathbf{X}_M, \mathbf{X}_{M-1}), (\mathbf{X}_M, \mathbf{X}_M)\} = \{\mathbf{X}'_{m'},\ m' = 2, \dots, 2M+2\}.$$

Note that for each $m' = 2, \dots, 2M+2$, if $m' = 2m$ then $\mathbf{X}'_{m'} = (\mathbf{X}_m, \mathbf{X}_m),\ \mathbf{X}'_{m'+1} = (\mathbf{X}_{m+1}, \mathbf{X}_m)$ and

$$\mathbb{E}[\mathbf{X}'_{m'+1} | \mathbf{X}'_1, \dots, \mathbf{X}'_{m'}] = (\mathbb{E}[\mathbf{X}_{m+1} | \mathbf{X}_1, \dots, \mathbf{X}_m], \mathbf{X}_m) = (\mathbf{X}_m, \mathbf{X}_m),$$

and if $m' = 2m+1$ then $\mathbf{X}'_{m'} = (\mathbf{X}_{m+1}, \mathbf{X}_m),\ \mathbf{X}'_{m'+1} = (\mathbf{X}_{m+1}, \mathbf{X}_{m+1})$ and

$$\mathbb{E}[\mathbf{X}'_{m'+1} | \mathbf{X}'_1, \dots, \mathbf{X}'_{m'}] = (\mathbf{X}_{m+1}, \mathbf{X}_{m+1}).$$

and hence the leading components, i.e. the first d entries of $\mathbb{X}'$, form a discrete-time martingale with respect to the natural filtration of $\mathbb{X}'$. By Remark 4.6.3, we can hence apply the control variate expected signature estimator (4.6.4) for any word $I = (i_1, \dots, i_k) \in \{1, \dots, 2d\}^k$ with $i_k \in \{1, \dots, d\}$.

Finally, in some applications, we may not have a strong prior on whether the process being modeled is a martingale or not. In this case, we may consider the martingale correction as a model configuration hyperparameter to be tuned, just like the lead-lag path augmentation discussed above. In the model training phase we can then apply a cross-validation procedure to learn whether applying the martingale correction to (some of) the expected signature terms improves the performance of the model.

Remark 4.6.4. Both the signature transform and the expected signature transform are general methods applicable to any machine learning task dealing with (collections of) streams of data. These can thus always be used as out-of-the-box feature extraction methods when little domain knowledge is available. On the other hand, when task-specific information is known, incorporating such knowledge in the machine learning model will most likely improve performance.

Remark 4.6.5. It is important to note there is also a wide range of machine learning methods based on signature kernels (Kiraly and Oberhauser, 2019; Chevyrev and Oberhauser, 2018; Lemercier et al., 2021; Salvi et al., 2021). This bypasses the need to explicitly estimate the expected signature and hence we cannot directly apply the martingale correction developed in Section 4.6.1.

Distributional regression for streams (Lemercier et al., 2021)

The first machine learning model we consider is perhaps the most natural one when working with expected signatures. Introduced in Lemercier et al. (2021), the Signature of the pathwise Expected Signature (SES) model aims to learn a map from a collection of paths $\mathbb{X}^1, \dots, \mathbb{X}^N \in \mathcal{X} \subseteq \mathrm{Lip}([0,T], \mathbb{R}^d)$, understood as an empirical measure on path space $\mu = \frac{1}{N}\sum_{n=1}^N \delta_{\mathbb{X}^n} \in \mathcal{P}(\mathcal{X})$, to a corresponding scalar value $f(\mu) \in \mathbb{R}$. The task of learning such function $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ from a finite number of (possibly noisy) observations $\{(\mu_i, y_i)\}_{i \in \mathcal{I}_{\text{train}}}$ is known as distributional regression.

Under appropriate conditions, by combining the characterizing property of the expected signature and the universality of the signature, the authors show that linear functionals on the signature of the pathwise expected signature are universal for weakly continuous functions $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ (Lemercier et al., 2021, Theorem 3.2). The training and testing of the SES model is summarized in Algorithm 1 with the step at which we can apply the martingale correction highlighted in blue.

Algorithm 1 Signature of pathwise Expected Signature (SES), training and testing

Hyperparameters Signature truncation levels: $k_1, k_2 \in \mathbb{N}$. Linear regression regularizers.

Input $\{(\{\mathbb{X}^n\}_{n\in\mathcal{N}_i}, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$, $\{(\{\mathbb{X}^n\}_{n\in\mathcal{N}_i}, y_i)\}_{i\in\mathcal{I}_{\text{test}}}$ where $\mathbb{X}^n \in \mathbb{R}^{d\times M}$ and $y_i \in \mathbb{R}$.

$d_1 \leftarrow d + \cdots + d^{k_1}$ and $d_2 \leftarrow d_1 + \cdots + d_1^{k_2}$.

for $i \in \mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{test}}$ **do**

  Pathwise ES of $\{\mathbb{X}^n\}_{n\in\mathcal{N}_i}$: $\Phi^i \in \mathbb{R}^{d_1\times M}$, $\Phi^i_{I,m} \leftarrow \hat{\phi}_I^{\mathcal{N}_i,\hat{c}_1^*}(t_m)$, $|I| \le k_1$, $1 \le m \le M$.

  Signature of Φ^i: $\mathbf{S}^i \leftarrow S^{\le k_2}(\Phi^i)_{[0,T]} \in \mathbb{R}^{d_2}$.

end for

Fit linear regression: $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{d_2}) \leftarrow \text{LinearRegressionFit}(\{(\mathbf{S}^i, y_i)\}_{i\in\mathcal{I}_{\text{train}}})$.

Predict using fitted linear regression: $\{\hat{y}_i\}_{i\in\mathcal{I}_{\text{test}}} \leftarrow \text{LinearRegressionPredict}(\hat{\boldsymbol{\beta}}, \{\mathbf{S}^i\}_{i\in\mathcal{I}_{\text{test}}})$.

Output Performance metric: $L(\{\hat{y}_i\}_{i\in\mathcal{I}_{\text{test}}}, \{y_i\}_{i\in\mathcal{I}_{\text{test}}})$.

We repeat two of the synthetic data experiments conducted in Lemercier et al. (2021), analyzing the performance of the SES model without and with martingale correction (MC). In the first experiment (Lemercier et al., 2021, Section 5.2), the task is to infer the temperature of an ideal gas from the paths of $N = 20$ particles moving in a box. The dynamics of the system are inevitably linked to the radius of the particles, with larger particle radii resulting in more frequent collisions. Two settings[35] are therefore considered, one with smaller particle radii $r_1 = 0.35 \times \sqrt[3]{V/N}$ and one with larger particle radii $r_2 = 0.65 \times \sqrt[3]{V/N}$. The second experiment (Lemercier et al., 2021, Section 5.3) concerns the estimation of the mean-reversion parameter in a rough volatility model. More precisely, the task is to infer the value of $a \in [10^{-6}, 1]$ from a sample $\{\sigma_\pi^n\}_{n=1}^N$ of (discretely observed) paths $\sigma^n = \{\sigma_t^n,\ t \in \pi\}$ over the partition $\pi = \{0, 0.01, \ldots, 2\}$ with continuous-time dynamics

$$\mathrm{d}Z_t = -a(Z_t - \mu)\mathrm{d}t + \nu \mathrm{d}B_t^H, \quad \sigma_t = \exp Z_t, \quad t \in [0, 2],$$

where $\{B_t^H,\ t \in [0, 2]\}$ is a fractional Brownian motion with Hurst parameter $H = 0.2$, $\mu = 0.5$, $\nu = 0.3$ and $Z_0 = 0.5$. The performance of the model is evaluated with increasingly large collections of paths as inputs, i.e. $N = 20, 50, 100$. As expected, the model becomes more accurate as the number of paths increases. We refer to Lemercier et al. (2021) for more details on the two experimental setups.

In both experiments, we keep the same training-evaluation pipeline as the one considered in the original paper, namely nested k-fold cross-validation with 5 outer folds for evaluation and 3 inner folds for hyperparameter selection (including the signature truncation k_1 and a

[35] $V = 3\text{cm}^3$ denotes the volume of the box in which the particles are moving.

Lasso regularization parameter). We report the average out-of-sample mean-squared error (MSE) and its standard deviation for the two experiments in Table 4.1 and Table 4.2, as well as the t-statistic and p-value of a pairwise t-test between the MSEs of the two models. While the results do not yield statistical significance there still seems to be a mild benefit in using the martingale correction, especially when considering that the processes of both experiments are not martingales[36]. The code used to produce the results of Table 4.1 and Table 4.2 is available at https://github.com/lorenzolucchese/distribution-regression-streams.

	Predictive MSE $[\times 10^{-2}]$	
	$r_1 = 0.35 \times \sqrt[3]{V/N}$	$r_2 = 0.65 \times \sqrt[3]{V/N}$
SES	1.27 (0.23)	0.09 (0.03)
SES-MC	1.31 (0.45)	0.07 (0.02)
t-stat	-0.29	1.41
p-value	0.79	0.23

Table 4.1: Ideal gas experiment of Lemercier et al. (2021, Section 5.2): SES model without and with martingale correction (MC).

	Predictive MSE $[\times 10^{-3}]$		
	$N = 20$	$N = 50$	$N = 100$
SES	1.49 (0.39)	0.33 (0.13)	0.20 (0.08)
SES-MC	1.26 (0.48)	0.31 (0.09)	0.19 (0.05)
t-stat	0.87	0.63	0.29
p-value	0.43	0.56	0.79

Table 4.2: Rough volatility experiment of Lemercier et al. (2021, Section 5.3): SES model without and with martingale correction (MC).

Time series classification (Triggiano and Romito, 2024)

The next model we consider falls under the general task of time series classification, mapping an input path $\mathbf{x} \in \mathbb{R}^{d \times M_1}$ to a class label $c \in \mathcal{C}$. We will refer to this model as Gaussian Process augmented Expected Signature (GPES) classifier. The input stream $\mathbf{x}$ is interpreted as a discrete-time realization of a Gaussian process $\mathbb{X} \sim \mathrm{GP}(\mu(t), \Sigma(t))$ at points $\pi_1 = \{0 = t_1 < \ldots < t_{M_1} = T\}$, i.e. a realization of $\mathbb{X}^{\pi}$. Values of the process over a fixed set of in-fill points $\pi_2 = \{s_1 < \ldots < s_{M_2}\}$ can thus be sampled[37] from the conditional distribution of $\mathbb{X}^{\pi_2}$ given the input $\mathbb{X}^{\pi_1}$, i.e. the

[36] In the first experiment one could intuitively argue that when the particle radii are large and collisions are more frequent the motion of the gas particles could be somewhat amenable to the motion of pollen grains in water, the original experiment which led to the "discovery" of Brownian motion by Scottish botanist Robert Brown in 1827.

[37] Super-sampling the input data to a collection of realizations from a Gaussian process, can be effectively understood as a regularization by noise technique.

conditional distribution $\mathbb{X}^{\pi_2}|\mathbb{X}^{\pi_1} = \mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x},\pi_1,\pi_2}, \Sigma_{\mathbf{x},\pi_1,\pi_2})$. The expected signature of the process $\mathbb{X}$ is then estimated from a collection of samples $\mathbb{X}^{1,\pi_1\cup\pi_2}, \ldots, \mathbb{X}^{N,\pi_1\cup\pi_2}$ such that $\mathbb{X}^{n,\pi_1} = \mathbf{x}$ and $\mathbb{X}^{n,\pi_2} \sim \mathcal{N}(\mu_{\mathbf{x},\pi_1,\pi_2}, \Sigma_{\mathbf{x},\pi_1,\pi_2})$ for $n = 1, \ldots, N$. In Triggiano and Romito (2024) the authors emphasize the theoretical and empirical importance of the tensor normalization introduced in Chevyrev and Oberhauser (2018), ensuring the resulting expected robust signature is characteristic for a larger class of processes. An important component of the GPES model is thus the (truncated at level k) tensor normalization $\lambda_C : T^K((\mathbb{R}^d)) \to T^K((\mathbb{R}^d))$, controlled by the hyperparameter C. For more details on the effect of the tensor normalization procedure and a sensitivity analysis[38] with respect to the hyper-parameter C we refer to Triggiano and Romito (2024). When applying the martingale correction to the GPES model we subtract the correction term $\hat{c}_1^* S_c^{\mathbf{I}}(\mathbb{X}^{k,\pi_1\cup\pi_2})_{[0,T]}$ to each $S^{\mathbf{I}}(\mathbb{X}^{k,\pi_1\cup\pi_2})_{[0,T]}$, i.e. *before* taking the empirical expectation over the paths. The final layer of the GPES model then maps the expected signature to a class by a combination of a linear transformation and a softmax output activation. The forward pass through the GPES model is summarized in Algorithm 2.

Algorithm 2 Gaussian Process augmented Expected Signature (GPES) classifier, forward pass

Hyperparameters Signature truncation level $k \in \mathbb{N}$, tensor normalization parameter $C \in \mathbb{R}_+$, data augmentation size $N \in \mathbb{N}$, in-fill partition $\pi_2 \in \Delta_{[0,T]}^{M_2}$ s.t. $M_2 \in \mathbb{N}$.

Parameters Biases $\mathbf{b}_\mu \in \mathbb{R}^{d_\mu}$, $\mathbf{b}_\Sigma \in \mathbb{R}^{d_\Sigma}$, $\mathbf{b}_{\text{out}} \in \mathbb{R}^{d_{\text{out}}}$ and weights $W_\mu \in \mathbb{R}^{d_\mu \times d_{\text{in}}}$, $W_\Sigma \in \mathbb{R}^{d_\Sigma \times d_{\text{in}}}$, $W_{\text{out}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{sig}}}$ where $d_{\text{in}} \leftarrow (d\,M_1 + M_1 + M_2)$, $d_\mu \leftarrow d\,M_2$, $d_\Sigma \leftarrow d\,M_2(d\,M_2+1)/2$, $d_{\text{sig}} \leftarrow (d + \ldots + d^k)$ and $d_{\text{out}} \leftarrow |\mathcal{C}|$.

Input $\mathbf{x} \in \mathbb{R}^{d\times M_1}$ and $\pi_1 \in \Delta_{[0,T]}^{M_1}$.
$\mu_{\mathbf{x},\pi_1,\pi_2} \leftarrow \mathbf{b}_\mu + W_\mu(\mathbf{x}, \pi_1, \pi_2)$.
$L_{\mathbf{x},\pi_1,\pi_2} \leftarrow \mathbf{b}_\Sigma + W_\Sigma(\mathbf{x}, \pi_1, \pi_2)$ and $\Sigma_{\mathbf{x},\pi_1,\pi_2} \leftarrow L_{\mathbf{x},\pi_1,\pi_2} L_{\mathbf{x},\pi_1,\pi_2}^{\mathrm{T}}$.
for $n \in \{1, \ldots, N\}$ **do**
    $\mathbb{X}^{n,\pi_1} \leftarrow \mathbf{x}$.
    Sample $\mathbb{X}^{n,\pi_2} \sim \mathcal{N}(\mu_{\mathbf{x},\pi_1,\pi_2}, \Sigma_{\mathbf{x},\pi_1,\pi_2})$.
    Signature of $\mathbb{X}^{n,\pi_1\cup\pi_2}$: $\mathbf{S}^n = S^{\leq k}(\mathbb{X}^{n,\pi_1\cup\pi_2})_{[0,T]} - \hat{c}_1^* S_c^{\leq k}(\mathbb{X}^{n,\pi_1\cup\pi_2})_{[0,T]} \in \mathbb{R}^{d_{\text{sig}}}$.
    Tensor normalization: $\mathbf{S}^n \leftarrow \lambda_C(\mathbf{S}^n)$.
end for
Expected signature $\mathbf{ES} \leftarrow \frac{1}{N}\sum_{n=1}^N \mathbf{S}^n$.
Output $\hat{c} \leftarrow \text{softmax}(\mathbf{b}_{\text{out}} + W_{\text{out}}\mathbf{ES})$.

Note that, unlike classic Gaussian process regression where the prior mean is assumed to be constant $\mu(t) \equiv \mu$ and the prior covariance function

$$\Sigma : [0,T] \to \mathbb{R}^{d\times d},$$

is parameterized by a kernel and the posteriors are computed by combining standard properties of

[38] Choosing a very large value of C is in practice equivalent to not applying a tensor normalization.

the multivariate normal distribution and the kernel trick[39], in the GPES model the conditional mean and covariance functions

$$(\mathbf{x}, \pi_1, \pi_2) \in \mathbb{R}^{d\times M_1} \times \Delta^{M_1}_{[0,T]} \times \Delta^{M_2}_{[0,T]} \cong \mathbb{R}^{d\,M_1+M_1+M_2} \mapsto \begin{cases} \mu_{\mathbf{x},\pi_1,\pi_2} \in \mathbb{R}^{d\times M_2} \cong \mathbb{R}^{d\,M_2}, \\ \Sigma_{\mathbf{x},\pi_1,\pi_2} \in \mathcal{M}(\mathbb{R}^{d\times M_2}) \cong \mathbb{R}^{d\,M_2\times d\,M_2}, \end{cases}$$

are parametrized by linear transformations

$$\begin{aligned} \mu_{\mathbf{x},\pi_1,\pi_2} &= \mathbf{b}_\mu + W_\mu(\mathbf{x}, \pi_1, \pi_2), \\ \Sigma_{\mathbf{x},\pi_1,\pi_2} &= L_{\mathbf{x},\pi_1,\pi_2} L^{\mathrm{T}}_{\mathbf{x},\pi_1,\pi_2}, \quad L_{\mathbf{x},\pi_1,\pi_2} = \mathbf{b}_\Sigma + W_\Sigma(\mathbf{x}, \pi_1, \pi_2), \end{aligned}$$

where $L_{\mathbf{x},\pi_1,\pi_2}$ is a lower triangular matrix. The parameters $\mathbf{b}_\mu, \mathbf{b}_\Sigma, W_\mu, W_\Sigma$ are learned along with the output layer parameters $\mathbf{b}_{\text{out}}, W_{\text{out}}$ in the training phase via numerical optimization, in Triggiano and Romito (2024) the authors use simple stochastic gradient descent (SGD). If the timestamps π_1 of the observations $\mathbf{x}$ are not provided, they need to be fixed and can be regarded as hyperparameters of the model. The (way the) in-fill partition π_2 (is chosen) is instead always set a-priori, with a natural choice for π_2 being the set of mid-points of π_1. Other model hyperparameters include the signature truncation level k, the size of the augmentation N and the constant C controlling the strength of the tensor normalization, as well as the training procedure's hyperparameters (learning rate, batch size etc.).

We replicate the synthetic data experiments of Triggiano and Romito (2024). These consist of three datasets:

(FBM) Two equally balanced classes with samples generated according to a standard Brownian motion and a fractional Brownian motion with Hurst parameter $H = 0.26$ (both in dimension $d = 1$).

(OU) Two equally balanced classes with samples generated according to two different Ornstein-Uhlenbeck (OU) processes (both in dimension $d = 1$).

(Bidim) Six equally balanced classes with samples generated according to six different bidimensional stochastic processes ($d = 2$).

We refer to the original paper for more details on the way the datasets were generated. When fitting the models we take the optimal hyperparameters cross-validated by Triggiano and Romito (2024)[40] and apply cross-validated SGD to the training dataset. That is, we use 80% of the training dataset to iterate through SGD parameter updates, while keeping the remaining 20% of the training

[39]The Gaussian process regression model is then fitted by tuning the kernel hyperparameters (either via maximum likelihood or cross-validation).

[40]The only hyperparameter we modify is the truncation level k which we set to 4 for computational reasons (in the original paper the optimal value was found to be 5 or 6, depending on the dataset). The hyperparameters and the training and testing routines used to produce the results in Table 4.3 can be found at https://github.com/lorenzolucchese/gp-esig-classifier.

dataset (the validation set) to determine when the procedure has converged without overfitting. As described in Triggiano and Romito (2024) the presence of the tensor normalization step often leads to exploding gradients in the training procedure. We thus repeat the SGD routine over 5 different parameter initializations and pick the model with best validation performance. The performance on the out-of-sample testing datasets of the GPES model with and without martingale correction is reported in Table 4.3. The output of the GPES model is by construction stochastic and hence we repeat the evaluation of the model with 10 different seeds. In Table 4.3 we report the mean accuracy and standard error of the model with and without martingale correction (MC), as well as the results of an independent samples t-test between their accuracies. The results suggest the martingale correction indeed improves the performance of the GPES model. Note that, as in the experiments of the previous section, most of the processes in the three datasets are not martingales. Moreover, the GPES algorithm estimates the expected signature conditional on $\mathbb{X}^{n,\pi_1} = \mathbf{x}$, so technically the martingale correction is biasing the estimator. The code used to produce the results of Table 4.3 is available at `https://github.com/lorenzolucchese/gp-esig-classifier`.

	Predictive Accuracy [%]		
	FBM	OU	Bidim
GPES	95.62 (0.18)	62.20 (0.70)	79.33 (0.46)
GPES-MC	95.26 (0.70)	88.26 (0.31)	88.97 (0.44)
t-stat	1.49	−101.92	−45.52
p-value	0.15	0.00	0.00

Table 4.3: Synthetic data experiments of Triggiano and Romito (2024): GPES model without and with martingale correction (MC).

Generative time series modeling (Ni et al., 2022; Liao et al., 2024)

In the two papers Ni et al. (2022) and Liao et al. (2024) the authors introduce two signature-based models for generative time series modeling. The unconditional setting considered in the first paper is "nested" in the conditional framework of the second paper. We thus focus on the generative algorithm introduced in Liao et al. (2024) the Signature-based Conditional Wasserstein Generative Adversarial Network (SIGCWGAN).

We start by recalling the classical Conditional GAN (CGAN) framework. Given a background[41] probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we define two random variables $\mathbf{X} \in (\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $\mathbf{X}^* \in (\mathcal{X}^*, \mathcal{B}(\mathcal{X}^*))$, the aim of the conditional generative model is to simulate samples from the conditional distribution $\mu(\mathbf{x}) = \mathcal{L}(\mathbf{X}^*|\mathbf{X} = \mathbf{x})$. The target distribution is usually approximated by $\nu(\theta^g, \mathbf{x}) = g(\theta^g, \mathbf{x}, \cdot)_*\mu_{\mathbf{Z}}$ where $\mu_{\mathbf{Z}}$ is a standard distribution with values in some space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ which we

[41] All measures are then defined as the push forwards of $\mathbb{P}$ by the respective random variables, e.g. $\mathbf{X} \sim \mu = \mathbf{X}_*\mathbb{P}$ and $\mathbf{Z} \sim \mu_{\mathbf{Z}} = \mathbf{Z}_*\mathbb{P}$.

can simulate from (e.g. a standard multivariate normal distribution) and $g : \Theta^g \times \mathcal{X} \times \mathcal{Z} \to \mathcal{X}^*$ is a "rich-enough" parametric function pushing forward samples of $\mathbf{Z}$ into samples of $\mathbf{X}^*|\mathbf{X} = \mathbf{x}$ (usually parametrized by a neural network architecture). The aim is to find the values of $\theta^g \in \Theta^g$ such that

$$\mu(\mathbf{X}) \approx \nu(\theta^g, \mathbf{X}), \quad \mathbf{X} \sim \mu,$$

where μ denotes the distribution of $\mathbf{X}$. In the (Conditional) Wasserstein GAN setting the quality of the approximation is measured using the Wasserstein metric, i.e.

$$\begin{aligned} \mu(\mathbf{X}) \approx \nu(\theta^g, \mathbf{X}) & \\ \iff & W_1(\mu(\mathbf{X}), \nu(\theta^g, \mathbf{X})) \approx 0 \\ \iff & \sup_{\|f\|_{\text{Lip}} \leq 1} \mathbb{E}_{\mathbf{X}^* \sim \mu(\mathbf{X})}[f(\mathbf{X}^*)] - \mathbb{E}_{\mathbf{X}^* \sim \nu(\theta^g, \mathbf{X})}[f(\mathbf{X}^*)], \end{aligned}$$

where the second equivalence follows by the Kantorovich-Rubenstein duality theorem. Taking the supremum over all Lipschitz-functions with Lipchitz norm bounded by one is not computationally feasible and hence this set is approximated by another "rich-enough" class of parametric functions $d : \Theta^d \times \mathcal{X}^* \to \mathbb{R}$ (this is also usually parametrized by a family of neural networks, with weight clipping or a penalty term added to ensure the derivative remains bounded by one). Training a CWGAN thus reduces to solving the mini-max problem

$$\min_{\theta^g \in \Theta^g} \max_{\theta^d \in \Theta^d} \mathbb{E}_{\mathbf{X} \sim \mu}[\mathbb{E}_{\mathbf{X}^* \sim \mu(\mathbf{X})}[d(\theta^d, \mathbf{X}^*)] - \mathbb{E}_{\mathbf{X}^* \sim \nu(\theta^g, \mathbf{X})}[d(\theta^d, \mathbf{X}^*)]],$$

where the family of functions g is usually referred to as the generator and the family of functions d is the discriminator. As pointed out in Liao et al. (2024) training a classical CWGAN can be hard for two reasons:

- Adversarial training of the min-max problem is in practice difficult to tune and very sensitive to the choice of hyperparameters and discriminator.
- In the conditional setting, for each parameter $\theta^d \in \Theta^d$ encountered in the optimization procedure one needs to compute $\mathbb{E}_{\mathbf{X}^* \sim \mu(\mathbf{X})}[d(\theta^d, \mathbf{X}^*)]$ (and the corresponding derivative in θ^d). By the Doob-Dynkin lemma this is a measurable function of $\mathbf{X}$ satisfying

  $$\mathbb{E}_{\mathbf{X} \sim \mu}\Big[\Big(d(\theta^d, \mathbf{X}) - \mathbb{E}_{\mathbf{X}^* \sim \mu(\mathbf{X})}[d(\theta^d, \mathbf{X}^*)]\Big)^2\Big] = \inf_{h:(\mathcal{X}, \mathcal{B}(\mathcal{X})) \to \mathbb{R}} \mathbb{E}_{\mathbf{X} \sim \mu}\Big[\Big(d(\theta^d, \mathbf{X}) - h(\mathbf{X})\Big)^2\Big],$$

  and hence, in practice, a third function approximation is usually employed approximating $\mathbb{E}_{\mathbf{X}^* \sim \mu(\mathbf{X})}[d(\theta^d, \mathbf{X}^*)] \approx h_{\theta^d}(\theta^h, \mathbf{X})$. This leads to additional complexity in the training procedure.

In the context of conditional time series generation the random variables $\mathbf{X} \in \mathbb{R}^{d \times M}$ and $\mathbf{X}^* \in \mathbb{R}^{d \times M_*}$ are taken to be past and future values of a times series over $M + M_*$ steps. Given the

values $\mathbf{X} = \mathbf{x}$ of the time series over the past M time steps one wishes to sample from the distribution of the time series over the future M_* times steps. The random variables $\mathbf{X}$ and $\mathbf{X}^*$ can be understood as discrete-time observations of the processes[42] $\mathbb{X} \in \mathrm{BV}_0([0,T],\mathbb{R}^d)$ and $\mathbb{X}^* \in \mathrm{BV}_0([T,T+T_*],\mathbb{R}^d)$ (these latent continuous-time processes may just be the linear interpolations of $\mathbf{X}$ and $\mathbf{X}^*$). The authors in Liao et al. (2024) hence propose to adapt the CWGAN model in the context of time series data by exploiting properties of the signature. Motivated by the universality theorem of Levin, Lyons, and Ni (2016) and the characteristic property of the signature, the authors propose to replace the W_1 metric in the CWGAN model by the Sig-W_1 metric, which has the analytic representation

$$\text{Sig-}W_1(\mu,\nu) = \|\mathbb{E}_{\mathbf{S}\sim\mathcal{S}_*\mu}[\mathbf{S}] - \mathbb{E}_{\mathbf{S}\sim\mathcal{S}_*\nu}[\mathbf{S}]\|,$$

where $\mathcal{S}_*\mu$ denotes the push-forward of the measure μ on $\mathrm{BV}_0([0,T],\mathbb{R}^d)$ to $T((\mathbb{R}^{d+1}))$ by the signature transform $\mathcal{S} : \cup_{0\le\tau_1\le\tau_2}\mathrm{BV}_0([\tau_1,\tau_2],\mathbb{R}^d) \to T((\mathbb{R}^{d+1}))$. Since the Sig-$W_1$ metric admits such analytical representation this means there is no need to resort to a duality theorem to compute it and hence there is no need for a discriminator d. The min-max problem is thus reduced to a standard supervised learning framework for the parameters $\theta^g \in \Theta^g$. Moreover, by applying again the bijective property of the signature, note that letting $\mathbf{S} = \mathcal{S}(\mathbb{X})$ we can write

$$\mathbb{E}_{\mathbf{S}^*\sim\mathcal{S}_*\mu(\mathbb{X})}[\mathbf{S}^*] = \mathbb{E}_{\mathbf{S}^*\sim\mathcal{S}_*\mu_{\mathcal{S}}(\mathbf{S})}[\mathbf{S}^*],$$

where $\mu_{\mathcal{S}}(\mathbf{S}) := \mu(\mathcal{S}^{-1}(\mathbf{S})) = \mu(\mathcal{S}^{-1}(\mathcal{S}(\mathbb{X}))) = \mu(\mathbb{X})$. The universality of linear functionals on the signature (Levin, Lyons, and Ni, 2016, Theorem 3.1) then motivates the approximation

$$\begin{aligned}\mathbb{E}_{\mathbf{S}\sim\mathcal{S}_*\mu}[(\mathbf{S}^* - \mathbb{E}_{\mathbf{S}^*\sim\mathcal{S}_*\mu_{\mathcal{S}}(\mathbf{S})}[\mathbf{S}^*])^2] &= \inf_{\substack{h:\, T((\mathbb{R}^{d+1}))\to\mathbb{R}\\ \text{measurable}}} \mathbb{E}_{\mathbf{S}\sim\mathcal{S}_*\mu}[(\mathbf{S}^* - h(\mathbf{S}))^2] \\ &\approx \inf_{\substack{\mathbf{L}:\, T((\mathbb{R}^{d+1}))\to\mathbb{R}\\ \text{linear}}} \mathbb{E}_{\mathbf{S}\sim\mathcal{S}_*\mu}[(\mathbf{S}^* - \mathbf{L}(\mathbf{S}))^2],\end{aligned}$$

and hence $\mathbb{E}_{\mathbf{S}^*\sim\mathcal{S}_*\mu_{\mathcal{S}}(\mathbf{S})}[\mathbf{S}^*] \approx \hat{\mathbf{L}}(\mathbf{S})$. Training a SIGCWGAN thus reduces to linearly regressing the future signatures $\mathbf{S}^* = \mathcal{S}(\mathbb{X}^*)$ against the past signatures $\mathbf{S} = \mathcal{S}(\mathbb{X})$ and then finding

$$\min_{\theta^g\in\Theta^g} \mathbb{E}_{\mathbb{X}\sim\mu}[\|\hat{\mathbf{L}}(\mathcal{S}(\mathbb{X})) - \mathbb{E}_{\mathbb{X}^*\sim\nu(\theta^g,\mathbb{X})}[\mathcal{S}(\mathbb{X}^*)]\|].$$

We leave the form of the generator g unspecified, in Liao et al. (2024) the authors use an Autoregressive Feed-forward Neural Network (AR-FNN) generator. It is thus clear that the SIGCWGAN model does not suffer from the two challenges associated with the training of classic CWGAN discussed above.

The SIGCWGAN model has been applied to the (conditional) generation of price paths, which

[42] Here $\mathrm{BV}_0(I,\mathbb{R}^d)$ denotes the space of continuous d-dimensional paths of finite 1-variation starting from the origin, with the 0th coordinate being the time dimension.

– up to a reasonable approximation[43] – can be regarded as martingales. We hence now discuss how the martingale correction introduced in Section 4.6.1 can be employed in the context of the SIGCWGAN algorithm. Effectively, there are two steps of the SIGCWGAN algorithm which we can modify to use the martingale correction:

- When estimating the linear regression coefficients $\hat{\mathbf{L}}$ we can use the values of the future signature corrections as controls. More precisely, when regressing $S^I(\mathbb{X}^*)_{[T,T+T_*]}$ onto the past signatures $\{S^J(\mathbb{X})_{[0,T]},\ |J| \leq k_2\}$, we can add $S^I_c(\mathbb{X}^*)_{[T,T+T_*]}$ as a control since, if $\{\mathbf{X}_t,\ t \in [0, T+T_*]\}$ is a martingale, we have

$$\mathbb{E}[S^I_c(\mathbb{X}^*)_{[T,T+T_*]}|S(\mathbb{X})_{[0,T]}] = \mathbb{E}[\mathbb{E}[S^I_c(\mathbb{X}^*)_{[T,T+T_*]}|\mathcal{F}_T]|S(\mathbb{X})_{[0,T]}] = 0.$$

  This is an extension of the martingale correction approach from the simple control variate estimation problem introduced in Section 4.6.1 to a more general controlled linear regression setting. We refer to Appendix 4.F for more details on controlled linear regression.

- When computing the expected signature under the synthetic measure $\mathbb{X}^* \sim \nu(\theta^g, \mathbb{X})$. Here we apply the standard control variate martingale correction described in Section 4.6.1. One might decide to introduce the correction at a later stage of training when samples from $\nu(\theta^g, \mathbb{X})$ are closer to the reference measure $\mu(\mathbb{X}) = \mathcal{L}(\mathbb{X}^*|\mathbb{X})$ under which $\mathbb{X}^*$ is a martingale.

We report the original SIGCWGAN training algorithm in Algorithm 3, highlighting the two points at which we can use the martingale correction. Note that when applying the martingale correction term we are biasing the SIGCWGAN model towards producing paths with martingale-like properties. This may or may not be a desirable feature depending on the task at hand. Code implementing the SIGCWGAN model with martingale correction is available at `https://github.com/lorenzolucchese/conditional-sig-wasserstein-gans`.

Pricing path-dependent derivatives (Lyons, Nejad, and Pérez Arribas, 2021)

The next application we consider is a purely financial one. The objective is to price and hedge path-dependent derivatives by decomposing them into a set of atomic Arrow-Debreu-like securities. While in Lyons, Nejad, and Pérez Arribas (2021) the authors consider a more general setting, for brevity we focus on the case where the (discounted) price process $\mathbb{X}$ is assumed to be a semimartingale. Let $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\}$ be a semimartingale on the probability space $(\Omega = C([0,T],\mathbb{R}), \mathcal{F}, \mathbb{P})$ and denote by $\hat{\Omega}^{\mathrm{LL}}_T$ the set of realized (time and lead-lag augmented) price signatures, i.e.

$$\hat{\Omega}^{\mathrm{LL}}_T = \{S(\hat{\mathbb{X}}^{\mathrm{LL}})_{[0,T]} \in T((\mathbb{R}^4)) : \hat{\mathbb{X}} = \{(t, \mathbf{X}_t),\ t \geq 0\}\},$$

where we refer to Lyons, Nejad, and Pérez Arribas (2021, Definition 2.14 and Example 2.15) for the definition of the lead-lag augmentation but, for the purposes of our discussion, it is sufficient

[43] I.e. modulo risk free rate and market price of risk.

Algorithm 3 Signature Conditional Wasserstein Generative Adversarial Network (SIGCWGAN), training

Hyperparameters Signature truncation levels $k, k^* \in \mathbb{N}$, # of Monte Carlo samples $N \in \mathbb{N}$. SGD hyperparameters (learning rate $\eta \in \mathbb{R}_+$, batch size $B \in \mathbb{N}$, # of epochs $N_{\text{epochs}} \in \mathbb{N}$).
Parameters $\theta^g \in \Theta^g$.

Input $\{(\mathbb{X}^i, \mathbb{X}^{*,i})\}_{i \in \mathcal{I}_{\text{train}}}$ such that $\mathbb{X}^i \in \Omega_0([0,T], \mathbb{R}^d)$ and $\mathbb{X}^{*,i} \in \Omega_0([T, T+T_*], \mathbb{R}^d)$ are linear interpolations of the streams $\mathbf{X}^i \in \mathbb{R}^{d \times M}$ and $\mathbf{X}^{*,i} \in \mathbb{R}^{d \times M_*}$ for all $i \in \mathcal{I}_{\text{train}}$.
for $i \in \mathcal{I}^b_{\text{train}}$ **do**
  Compute $\mathbf{S}_i \leftarrow S^{\leq k}(\mathbb{X}^i)_{[0,T]}$.
  Compute $\mathbf{S}_{i,*} \leftarrow S^{\leq k}(\mathbb{X}^{i,*})_{[T,T+T_*]}$, $\mathbf{S}_{i,c,*} = S^{\leq k}_c(\mathbb{X}^{i,*})_{[T,T+T_*]}$.
end for
Fit linear regression: $\hat{\mathbf{L}}_I \leftarrow$ ControlledLinearRegressionFit$(\{(\mathbf{S}_i, S^I_{i,c,*}, S^I_{i,*})\}_{i \in \mathcal{I}_{\text{train}}})$, $|I| \leq k^*$.
Initialize $\hat{\theta}^g \in \Theta^g$.
for $n \in \{1, \dots, N_{\text{epochs}}\}$ **do**
  Randomly partition $\mathcal{I}_{\text{train}} = \cup_{b=1}^{N_{\text{batch}}} \mathcal{I}^b_{\text{train}}$ such that $|\mathcal{I}^b_{\text{train}}| = B$ for $b = 1, \dots, N_{\text{batch}}$.
  for $b \in \{1, \dots, N_{\text{batch}}\}$ **do**
    Initialize $\ell^b(\hat{\theta}^g) \leftarrow 0$.
    for $i \in \mathcal{I}^b_{\text{train}}$ **do**
      Simulate N samples $\mathbf{Z}^1, \dots, \mathbf{Z}^N \sim \mu_{\mathbf{Z}}$.
      Generate N samples $\hat{\mathbb{X}}^{*,i,1}, \dots, \hat{\mathbb{X}}^{*,i,N}$: $\hat{\mathbb{X}}^{*,i,n} \leftarrow g(\hat{\theta}^g, \mathbb{X}^i, \mathbf{Z}^{i,n})$, $n \in \{1, \dots, N\}$.
      ES $\hat{\Phi}^i$, $\hat{\Phi}^i_I \leftarrow \hat{\phi}^{N,i,\hat{c}_1}_I(\hat{\mathbb{X}}^*) = \frac{1}{N} \sum_{n=1}^N [S^I(\hat{\mathbb{X}}^{*,i,n})_{[T,T+T_*]} - \hat{c}_1 S^I_c(\hat{\mathbb{X}}^{*,i,n})_{[T,T+T_*]}]$, $|I| \leq k^*$.
      Update $\ell^b(\hat{\theta}^g) \leftarrow \ell^b(\hat{\theta}^g) + \|\hat{\mathbf{L}}\mathbf{S}^i - \hat{\Phi}^i\|$.
    end for
    Compute $\nabla_{\theta^g} \ell^b(\theta^g)|_{\theta^g = \hat{\theta}^g}$ via backpropagation.
    Update $\hat{\theta}^g \leftarrow \hat{\theta}^g - \eta \nabla_{\theta^g} \ell^b(\theta^g)|_{\theta^g = \hat{\theta}^g}$.
  end for
end for
Output Trained parameters $\hat{\theta}^g \in \Theta^g$.

to note that it is uniquely defined through Stratonovich integration. The authors then consider the market $(\hat{\Omega}_T^{\text{LL}}, \mathcal{B}(\hat{\Omega}_T^{\text{LL}}), \{\mathcal{F}_t,\, t \in [0,T]\}, \hat{\mathbb{P}}^{\text{LL}})$ where $\hat{\mathbb{P}}^{\text{LL}}$ is the push-forward of $\mathbb{P}$ onto $\hat{\Omega}_T^{\text{LL}}$. By defining the set of derivative payoffs as all measurable $F : \hat{\Omega}_T^{\text{LL}} \to \mathbb{R}$, i.e. for a given price realization $\mathbb{X}$ the holder of the derivative receives $F(S(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]})$, in Lyons, Nejad, and Pérez Arribas (2021, Proposition 4.5) the authors use the universality of the signature to show that any *continuous* payoff F can be arbitrarily well approximated by a linear payoff[44]. In particular, this implies the price of any such F can be decomposed as

$$\mathbb{E}^{\mathbb{Q}}[Z_T F(S(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]})] \approx \langle f, Z_T \mathbb{E}^{\mathbb{Q}}[S(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]}]\rangle,$$

for a set of linear coefficients $f \in T((\mathbb{R}^4)^*)$ where $\mathbb{Q}$ is a pricing measure for $\mathbb{X}$ and Z_T is a deterministic discount factor over $[0,T]$. The set of signature payoffs $\{S^I(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]},\, I \in \mathcal{W}(\{1,2,3,4\})\}$ can thus be understood as a set of Arrow-Debreu securities spanning the set of continuous path-dependent derivatives F. Similarly, in Lyons, Nejad, and Pérez Arribas (2021, Proposition 4.6) the authors show that linear trading strategies are dense in the space of admissible trading strategies $\mathcal{A}$, here defined as the set of all continuous functions $\theta : S(\hat{\mathbb{X}})_{[0,t]} \mapsto \theta(S(\hat{\mathbb{X}})_{[0,t]})$ over the stopped at $t \in [0,T]$ time-augmented price path signatures, i.e.

$$\theta(S(\hat{\mathbb{X}})_{[0,t]}) \approx \langle \ell, S(\hat{\mathbb{X}})_{[0,t]}\rangle, \quad \forall t \in [0,T],$$

for a set of linear coefficients $\ell \in T((\mathbb{R}^2)^*)$. These two results are then combined in Lyons, Nejad, and Pérez Arribas (2021, Theorem 4.7) to show that the solution of the quadratic $\mathbb{P}$-hedging problem[45]

$$\theta^* = \underset{\theta \in \mathcal{A}}{\text{argmin}}\ \mathbb{E}^{\mathbb{P}}\left[\left(F(S(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]}) - p_0 - \int_0^T \theta(S(\hat{\mathbb{X}})_{[0,t]})\mathrm{d}\mathbf{X}_t\right)^2\right], \tag{4.6.5}$$

can be arbitrarily well approximated by the solution of a linear signature quadratic hedging problem, i.e.

$$\theta^*(S(\hat{\mathbb{X}})_{[0,t]}) \approx \langle \ell^*, S(\hat{\mathbb{X}})_{[0,t]}\rangle, \quad t \in [0,T],$$

where $\ell^* \in T((\mathbb{R}^2)^*)$ can be computed by

$$\ell^* = \underset{\ell \in T((\mathbb{R}^2)^*)}{\text{argmin}}\ \langle (f - p_0 \varnothing + \ell \mathbf{4})^{\sqcup\!\sqcup 2}, \mathbb{E}^{\mathbb{P}}[S(\hat{\mathbb{X}}^{\text{LL}})_{[0,T]}]\rangle.$$

These theoretical results suggest both a pricing and a hedging strategy for path-dependent derivatives based on expected signatures[46], summarized in Algorithm 4 and Algorithm 5. Both

[44] Note that the approximation of F by f does not depend on the choice of probability measure $\mathbb{P}$, i.e. it is a pathwise density result. In Algorithm 4 and Algorithm 5 we thus assume that f has been estimated offline (i.e. for any model) by linearly regressing $F(\omega)$ against ω for a large set of $\omega \in \hat{\Omega}_T^{\text{LL}}$.

[45] For conciseness here we only discuss the quadratic hedging problem, in Lyons, Nejad, and Pérez Arribas (2021) the authors obtain results for general polynomials which allows them to also approximate the optimal hedge under exponential utility.

[46] As discussed in Pérez Arribas et al. (2018, Section 6.1) the expected signature under the measure $\mathbb{Q}$ can

Algorithm 4 Pricing Path-Dependent Derivatives with Expected Signatures

Hyperparameters Signature truncation levels k, number of Monte Carlo samples $N \in \mathbb{N}$.

Parameters Risk-neutral measure $\mathbb{Q}$, deterministic discount factor $Z_T \in \mathbb{R}^+$, linear approximator $f \in T^k((\mathbb{R}^4)^*)$.

Input Derivative payoff $F : \hat{\Omega}_T^{\mathrm{LL}} \to \mathbb{R}$.
Sample N trajectories $\mathbb{X}^1, \dots, \mathbb{X}^N \sim \mathbb{Q}$.
Compute time-augmented lead-lag transforms $\hat{\mathbb{X}}^{n,\mathrm{LL}}$ for $n \in \{1, \dots, N\}$.
Compute $\mathbf{S}^n \in T^k((\mathbb{R}^4))$ s.t. $\mathbf{S}_I^n \leftarrow S^I(\hat{\mathbb{X}}^{n,\mathrm{LL}})_{[0,T]}$, $|I| \leq k$ for $n \in \{1, \dots, N\}$
Estimate $\Phi \in T^k((\mathbb{R}^4))$ s.t. $\Phi_I \leftarrow \hat{\phi}_I^{N,\hat{c}_1}(T)$, $|I| \leq k$ from $\{\hat{\mathbb{X}}^{n,\mathrm{LL}}\}_{n=1}^N$.
Output Price $p = \langle f, Z_T \Phi \rangle$.

Algorithm 5 Hedging Path-Dependent Derivatives with Expected Signatures

Hyperparameters Signature truncation levels k, number of Monte Carlo samples $N \in \mathbb{N}$.

Parameters Real-world measure $\mathbb{P}$, initial capital $p_0 \in \mathbb{R}$, linear approximator $f \in T^k((\mathbb{R}^4)^*)$.

Input Derivative payoff $F : \hat{\Omega}_T^{\mathrm{LL}} \to \mathbb{R}$.
Sample N trajectories $\mathbb{X}^1, \dots, \mathbb{X}^N \sim \mathbb{P}$.
Compute time-augmented lead-lag transforms $\hat{\mathbb{X}}^{n,\mathrm{LL}}$ for $n \in \{1, \dots, N\}$.
Compute $\mathbf{S}^n \in T^k((\mathbb{R}^4))$ s.t. $\mathbf{S}_I^n \leftarrow S^I(\hat{\mathbb{X}}^{n,\mathrm{LL}})_{[0,T]}$, $|I| \leq k$ for $n \in \{1, \dots, N\}$
Estimate $\Phi \in T^k((\mathbb{R}^4))$ s.t. $\Phi_I \leftarrow \hat{\phi}_I^{N,\hat{c}_1}(T)$, $|I| \leq k$ from $\{\hat{\mathbb{X}}^{n,\mathrm{LL}}\}_{n=1}^N$.
$\hat{\ell}^* \leftarrow \inf_{\ell \in T^{\lfloor k/2 \rfloor}((\mathbb{R}^2)^*)} \langle (f - p_0 \varnothing + \ell \mathbf{4})^{\sqcup\!\sqcup 2}, \Phi \rangle$.
Output Hedging strategy $t \mapsto \langle \hat{\ell}^*, S(\hat{\mathbb{X}})_{[0,t]} \rangle, \ t \in [0, T]$.

algorithms make use of expected signature estimation via Monte Carlo simulations, an approach that provides a classic setting for applying the martingale correction described in Section 4.6.1 (recall that by Remark 4.6.3 and the Lead-Lag discussion at the start of the section we apply the correction only to signature terms with the process $\mathbb{X}$ appearing in the outer integral). Note that price processes under $\mathbb{P}$ (as considered in the hedging setting) are usually not martingales and hence it is not clear whether the variance reduction for the Monte Carlo estimator of the expected signature obtained via the martingale correction offsets the introduced bias. On the other hand, when pricing under $\mathbb{Q}$, the (discounted) price process $\mathbb{X}$ is a martingale and hence the correction term is exact. In Figure 4.13 we compare the finite sample properties of the expected signature estimator with and without martingale correction when the price process is assumed to follow a Brownian motion, in the context of option pricing this is known as the Bachelier model. Similarly, in Figure 4.14, we plot the densities of the two estimators under the Heston dynamics specified in Section 4.5 (we only correct the estimator for signature terms corresponding to words ending in the index 0, i.e. whose outermost integral is with respect to the price process S_t). In other words, in Figure 4.13 and Figure 4.14 we overlay the density of the expected signature estimator $\hat{\phi}_I^{N,\pi}(T)$ under the i.i.d. sampling scheme reported in Figure 4.6 and Figure 4.9 with the density of the martingale corrected expected signature estimator $\hat{\phi}_I^{N,\pi,c}(T)$. Note that the y-axis is now in log scale.

Systematic trading (Futter, Horvath, and Wiese, 2023)

The last application we consider is also motivated by a financial application and can be understood as a natural extension of the quadratic hedging problem given in Equation (4.6.5). In Futter, Horvath, and Wiese (2023) the authors consider the same setup as in Lyons, Nejad, and Pérez Arribas (2021) but allow the trading strategy $\theta \in \mathcal{A}$ to depend on the signature of the augmented process

$$\hat{\mathbb{Z}} = \{(t, \mathbf{X}_t, \mathbf{f}_t),\ t \in [0,T]\} \in C([0,T], \mathbb{R}^{1+d+q}),$$

where $\mathbb{X} = \{\mathbf{X}_t,\ t \in [0,T]\} \in C([0,T], \mathbb{R}^d)$ is a d-dimensional price process and $\mathbb{F} = \{\mathbf{f}_t,\ t \in [0,T]\} \in C([0,T], \mathbb{R}^q)$ is a set of q observable but not tradable factors (or signals) influencing $\mathbb{X}$. Again motivated by a universality result of the signature the authors approximate any such trading strategy by a linear one, i.e.

$$\theta_i(S(\hat{\mathbb{Z}})_{[0,t]}) \approx \langle \ell_i, S(\hat{\mathbb{Z}})_{[0,t]} \rangle, \quad i \in \{1, \dots, d\},\ t \in [0,T],$$

alternatively be estimated in a model-free way from the market prices of a large enough set of exotic derivatives, yielding the implied expected signature. Here we assume the measures $\mathbb{P}$ and $\mathbb{Q}$ have been appropriately calibrated to market data.

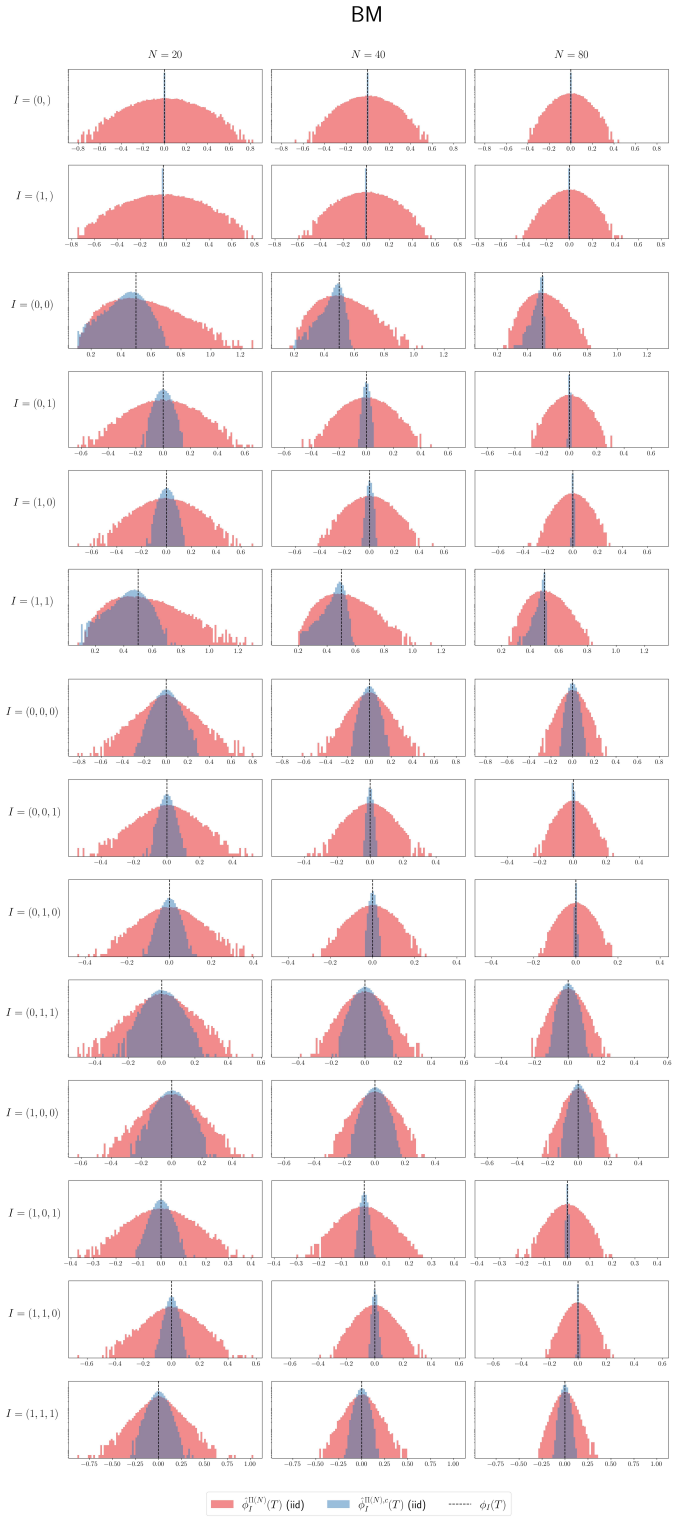


Figure 4.13: Distributions of the expected signature estimators for the BM process: classic estimator in red, martingale corrected estimator in blue and exact values in black. Both estimators are computed under (iid) observations. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$. The y-axis is in log-scale.

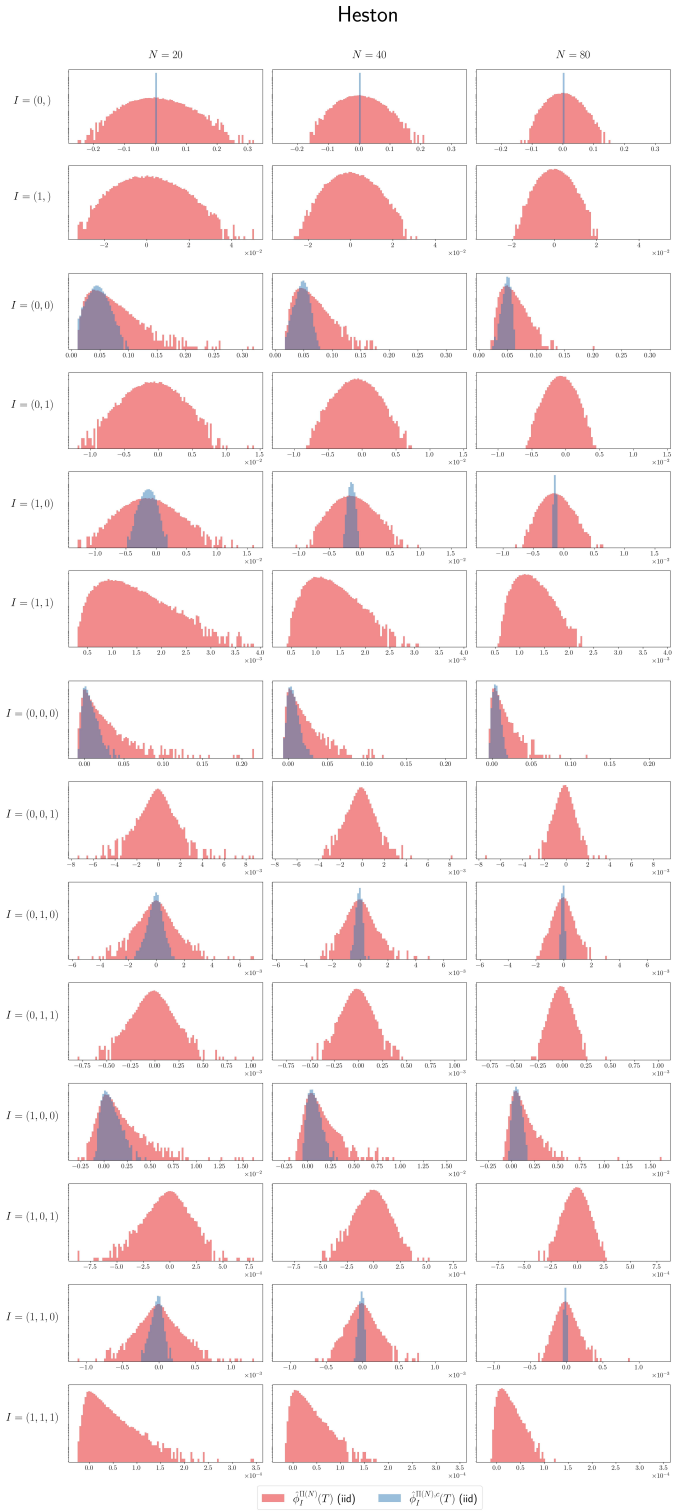


Figure 4.14: Distributions of the expected signature estimators for the Heston process: classic estimator in red and martingale corrected estimator in blue. Both estimators are computed under (iid) observations. From top to bottom: all words up to level $k = 3$. From left to right: $N = 20, 40, 80$. The y-axis is in log-scale.

for some set of linear functionals $\ell_1, \dots, \ell_d \in T((\mathbb{R}^{1+d+q})^*)$. The authors then show that an explicit solution to the path-dependent mean-variance problem

$$\ell_1^*, \dots, \ell_d^* = \underset{\substack{\ell_1,\dots,\ell_d \in T^k((\mathbb{R}^{1+d+q})^*) \\ \mathrm{Var}(\mathrm{PnL}_T) \leq \Delta}}{\operatorname{argmin}} \mathbb{E}^{\mathbb{P}}[\mathrm{PnL}_T], \quad \text{where} \quad \mathrm{PnL}_T = \sum_{i=1}^{d} \int_0^T \langle \ell_i, S(\hat{\mathbb{Z}})_{[0,t]} \rangle \mathrm{d}X_t^i,$$

for arbitrary truncation level $k \in \mathbb{N}$ can be read from the entries of

$$\frac{1}{2\lambda_\Delta} \Sigma_{\mathrm{sig}}^{-1} \mu_{\mathrm{sig}} \in \mathbb{R}^{d_{\mathrm{sig}}}, \quad d_{\mathrm{sig}} = d + \ldots + d(1+d+q)^k,$$

where $\mu_{\mathrm{sig}} \in \mathbb{R}^{d_{\mathrm{sig}}}$ and $\Sigma_{\mathrm{sig}} \in \mathbb{R}^{d_{\mathrm{sig}} \times d_{\mathrm{sig}}}$ and $\lambda_\Delta \in \mathbb{R}_+$ only depend[47] on the expected signature $\mathbb{E}^{\mathbb{P}}[S(\hat{\mathbb{Z}}^{\mathrm{LL}})_{[0,T]}]$.

A standard way of applying the sig-trading strategy in practice is given in Algorithm 6. Note that in real financial markets price and signal paths cannot be resampled and hence the collection $\{(\mathbb{X}^n, \mathbb{F}^n)\}_{n=1}^N$ can only be obtained by chopping-and-shifting a single long observation $\{(\mathbf{X}_t, \mathbf{f}_t),\ t \in [0, NT]\}$. In this respect, the sig-trading algorithm provides a striking example of a setting where the sampling scheme cannot be considered i.i.d. and hence one needs to resort to the results of Theorem 4.4.29 to obtain theoretical guarantees for the expected signature estimator.

A silent but fundamental assumption[48] is that the market dynamics of the collection of *past* price-factor paths $\{(\mathbb{X}^n, \mathbb{F}^n)\}_{n=1}^N$ used to estimate the expected signature will be the same as those of the *future* price-factor process $(\mathbb{X}^*, \mathbb{F}^*)$ to which the trading strategy will be applied. Using the martingale correction for estimating (some of the entries of) the expected signature, induces a bias to "ignore" the drift component of the signature term. For example, the first level of the martingale-corrected expected signature is always zero. This might be a desirable feature to avoid over-fitting the trading strategy to spurious drifts in the data (for example in the price processes $\mathbb{X}$ (Buehler et al., 2022)) while capturing higher order effects. As with other applications discussed in this section, the usefulness of the martingale correction in Algorithm 6 can be empirically assessed via cross-validation.

4.7 Conclusions and outlook

In this chapter, we developed a unified framework for the statistical estimation of expected signatures. Theorem 4.4.19, Theorem 4.4.29 and Theorem 4.4.34 provide quite general conditions on the process $\mathbb{X}$ ensuring the expected signature estimator is consistent and asymptotically normal. In particular, the latter two results relax the independent observations assumption and consider the case where the samples are obtained by chopping a single long realization of the process, as is often the case in

[47] As in the mean-variance optimal portfolio, the variance scaling parameter $\lambda_\Delta \in \mathbb{R}_+$ also depends on the target variance Δ.

[48] Clearly, this assumption is not specific to the sig-trading strategy but applies to any trading strategy that tries to learn patterns from the past to profit in the future.

Algorithm 6 Signature Trading

Hyperparameters Signature truncation level k, maximum variance $\Delta \in \mathbb{R}_+$.

Input Collection of price-factor paths $\{(\mathbb{X}^n, \mathbb{F}^n)\}_{n=1}^N$ where $\mathbb{X}^n = \{\mathbf{X}_t^n,\ t \in [0,T]\} \in C([0,T], \mathbb{R}^d)$ and $\mathbb{F}^n = \{\mathbf{f}_t^n,\ t \in [0,T]\} \in C([0,T], \mathbb{R}^q)$ for $n \in \{1, \dots, N\}$.

Set $\hat{\mathbb{Z}} \leftarrow \{(t, \mathbf{X}_t, \mathbf{f}_t),\ t \in [0,T]\}$.

Compute time-augmented lead-lag transforms $\hat{\mathbb{Z}}^{n,\mathrm{LL}}$ for $n \in \{1, \dots, N\}$.

Estimate $\Phi \in T^k((\mathbb{R}^{1+d+q}))$ s.t. $\Phi_I \leftarrow \hat{\phi}_I^{N,\hat{c}_1}(T)$, $|I| \leq k$ from $\{\hat{\mathbb{Z}}^{n,\mathrm{LL}}\}_{n=1}^N$.

Compute $\lambda_\Delta, \mu_{\mathrm{sig}}, \Sigma_{\mathrm{sig}}$ from corresponding entries of Φ.

Extract $\hat{\ell}_i^*$ from corresponding entries of $(2\lambda_\Delta \Sigma_{\mathrm{sig}})^{-1}\mu_{\mathrm{sig}}$ for $i \in \{1, \dots, d\}$.

Output Trading strategy $t \mapsto \langle \hat{\ell}_i^*, S(\hat{\mathbb{Z}}^*)_{[0,t]} \rangle,\ i \in \{1, \dots, d\},\ t \in [0,T]$ for new $(\mathbb{X}^*, \mathbb{F}^*)$.

applications. When $\mathbb{X}$ is a martingale we can improve the finite sample properties of the estimator, i.e. its mean-squared error, by adding a correction term. The practical advantages of the modified estimator are not limited to the case where the process is a martingale, as illustrated by the empirical results of Section 4.6.2.

We hope this work will serve as a starting point for developing more estimation results for expected signatures. There are many generalizations of our assumptions that one may explore. These include but are not limited to: different sampling schemes, e.g. chopping a single long realization into overlapping samples as discussed in Remark 4.4.17, or different ergodic assumptions, e.g. non-stationary samples converging to a stationary distribution as in Remark 4.4.18. Moreover, as discussed at length in Chevyrev and Oberhauser (2018), by considering a tensor-normalized version of the signature one can obtain more general characterization results than in the un-normalized setting. A natural extension is thus to develop asymptotic results for the expected robust signature estimator, which we briefly touched upon when discussing the GPES model in Section 4.6.2. Finally, there are many machine learning applications that one may consider other than those discussed in Section 4.6.2. These may be tackled by developing new expected signature-based algorithms or by improving already existing ones, like we demonstrated with the martingale corrected estimator.

Appendix

4.A Brownian motion: proofs

4.A.1 Proof of Theorem 4.2.1

Proof of Theorem 4.2.1. The result can be equivalently restated as $\forall k \geq 0$, $t \in [0,T]$

$$\phi_k(t) = \mathbb{E}\left[S^k(\mathbb{B})_{[0,t]}\right] = \begin{cases} \dfrac{1}{(k/2)!}\left(\dfrac{t}{2}\right)^{k/2}\left(\displaystyle\sum_{i=1}^{d} e_i \otimes e_i\right)^{k/2}, & \text{if } k \text{ is even,} \\ 0, & \text{if } k \text{ is odd.} \end{cases} \tag{4.A.1}$$

We proceed by induction on k. Note that for $k = 0$ the result is trivial while for $k = 1$ we can easily compute

$$\mathbb{E}\left[S^1(\mathbb{B})_{[0,t]}\right] = \mathbb{E}\left[\int_0^t \mathrm{d}\mathbf{B}_u\right] = \mathbb{E}\left[\mathbf{B}_t\right] = 0, \quad t \in [0,T].$$

Now for general $k \geq 2$ we have that

$$\begin{aligned}
\mathbb{E}\left[S^k(\mathbb{B})_{[0,t]}\right] &= \mathbb{E}\left[\int\cdots\int_{0<u_1<\ldots<u_k<T} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_k}\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\frac{1}{2}\left\langle \int\cdots\int_{0<u_1<\ldots<u_{k-2}<\cdot} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-1}}, \mathbf{B}_\cdot \right\rangle_t\right] \\
&\qquad + \mathbb{E}\left[\int_0^T \left(\int\cdots\int_{0<u_1<\ldots<u_{k-1}} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-1}}\right) \otimes \mathrm{d}\mathbf{B}_{u_k}\right] \\
&\overset{(ii)}{=} \frac{1}{2}\mathbb{E}\left[\left\langle \int_0^\cdot \left(\int\cdots\int_{0<u_1<\ldots<u_{k-1}} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-2}}\right) \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-1}}, \mathbf{B}_\cdot \right\rangle_t\right] \\
&\overset{(iii)}{=} \frac{1}{2}\mathbb{E}\left[\left\langle \int_0^\cdot \left(\int\cdots\int_{0<u_1<\ldots<u_{k-1}} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-2}}\right) \otimes \mathrm{d}\mathbf{B}_{u_{k-1}}, \mathbf{B}_\cdot \right\rangle_t\right] \\
&\overset{(iv)}{=} \frac{1}{2}\mathbb{E}\left[\int_0^t \left(\int\cdots\int_{0<u_1<\ldots<u_{k-1}} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-2}}\right) \otimes \mathrm{d}\langle \mathbf{B}_\cdot, \mathbf{B}_\cdot\rangle_{u_{k-1}}\right] \\
&\overset{(v)}{=} \frac{1}{2}\int_0^t \mathbb{E}\left[\int\cdots\int_{0<u_1<\ldots<u_{k-1}} \circ\,\mathrm{d}\mathbf{B}_{u_1} \otimes \cdots \otimes \circ\,\mathrm{d}\mathbf{B}_{u_{k-2}}\right] \otimes \left(\sum_{i=1}^{d} e_i \otimes e_i\right) \mathrm{d}u_{k-1} \\
&\overset{(vi)}{=} \frac{1}{2}\int_0^t \mathbb{E}\left[S^{k-2}(\mathbb{B})_{[0,u_{k-1}]}\right] \otimes \left(\sum_{i=1}^{d} e_i \otimes e_i\right) \mathrm{d}u_{k-1},
\end{aligned}$$

where in (i) we apply the Stratonovich-to-Itô transformation, in (ii) the mean-zero property of the Itô integral, in (iii) the Stratonovich-to-Itô transformation again and the fact that the Stratonovich-to-Itô correction is a finite variation process and thus has zero quadratic variation, in (iv) the formula to compute the quadratic covariation of two semimartingales (Protter, 1990, Theorem II.29), in (v) the formula for quadratic variation of a d-dimensional Brownian motion and in (vi) the inductive assumption. □

4.A.2 Proof of Theorem 4.2.3

Proof of Theorem 4.2.3. Since $\hat{\mathbb{B}}^M$ is piecewise linear we have that

$$S(\hat{\mathbb{B}}^M)_{[0,m\Delta]} = \bigotimes_{j=0}^{m-1} \exp_{\otimes}\{\mathbf{B}_{t_{j+1}} - \mathbf{B}_{t_j}\},$$

and thus

$$\phi^M(m\Delta) = \mathbb{E}\left[\bigotimes_{j=0}^{m-1} \exp_{\otimes}\{\mathbf{B}_{t_{j+1}} - \mathbf{B}_{t_j}\}\right] = \mathbb{E}\left[\exp_{\otimes}\{\mathbf{B}_{\Delta}\}\right]^{\otimes m},$$

follows from $\mathbb{B}$ having independent and stationary increments and $|t_{j+1} - t_j| = \Delta$ for all $j = 0, \dots, m-1$. Moreover we can write

$$\mathbb{E}\left[\exp_{\otimes}\{\mathbf{B}_{\Delta}\}\right] = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbb{E}\left[(\mathbf{B}_{\Delta})^{\otimes k}\right],$$

where

$$\begin{aligned}\mathbb{E}[(\mathbf{B}_{\Delta})^{\otimes k}] &= \sum_{I \in \mathcal{W}_k} \mathbb{E}[B_{\Delta}^{(i_1)} \cdots B_{\Delta}^{(i_k)}] \otimes_{i \in I} e_i \\ &= \sum_{I \in \mathcal{W}_k} \mathbb{E}[B_{\Delta}^{n_1(I)}] \cdots \mathbb{E}[B_{\Delta}^{n_d(I)}] \otimes_{i \in I} e_i.\end{aligned}$$

Finally, the fact that $\mathbb{E}[B_t^n] = t^{n/2}(n-1)!!$ if n is even and zero if n is odd implies

$$\mathbb{E}[B^{n_1(I)}] \cdots \mathbb{E}[B^{n_d(I)}] = \begin{cases} \Delta^{k/2}(n_1(I)-1)!! \cdots (n_d(I)-1)!! & \text{if } I \text{ is even,} \\ 0 & \text{otherwise.} \end{cases}$$

□

4.A.3 Proof of Corollary 4.2.5

Proof of Corollary 4.2.5. Note that from Theorem 4.2.3 we can write $\phi^M(m\Delta) = \phi^M(\Delta)^{\otimes m}$ and thus

$$\phi_k^M(m\Delta) = \phi_k^M(\Delta)^{\otimes m} = \sum_{\substack{k_1,\dots,k_m \geq 0 \\ k_1+\cdots+k_m=k}} \bigotimes_{l=1}^{m} \phi_{kl}^M(\Delta)$$
$$= \begin{cases} \displaystyle\sum_{\substack{j_1,\dots,j_m \geq 0 \\ 2j_1+\cdots+2j_m=k}} \bigotimes_{l=1}^{m} \phi_{2j_l}^M(\Delta), & \text{if } k \text{ is even,} \\ 0, & \text{if } k \text{ is odd.} \end{cases}$$

since $\phi_k^M(\Delta) = 0$ for all k odd. The result follows by noting that Theorem 4.2.3 gives $\forall j \geq 0$,

$$\phi_{2j}^M(\Delta) = \sum_{I \in \mathcal{K}_{2j}} \frac{C_I}{j!} \left(\frac{\Delta}{2}\right)^j \otimes_{i \in I} e_i.$$

□

4.A.4 Proof of Theorem 4.2.6

Proof of Theorem 4.2.6. We start by considering $\phi^M(\Delta)$ and $\phi(\Delta)$. By setting $m = 1$ in Corollary 4.2.5 and $t = \Delta$ in Equation (4.A.1) we can write for $k \geq 0$:

$$\phi_k^M(\Delta)) - \phi_k(\Delta) = \begin{cases} \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!} \left\{ \sum_{I \in \mathcal{K}_k} C_I \otimes_{i \in I} e_i - \sum_{I \in \mathcal{S}_k} \otimes_{i \in I} e_i \right\}, & \text{if } k \text{ is even,} \\ 0, & \text{if } k \text{ is odd,} \end{cases}$$

where $\mathcal{S}_k \subseteq \mathcal{K}_k$ is the set of all square words of (even) length k, i.e. $I = (i_1, i_1, \dots, i_{k/2}, i_{k/2})$ for $i_1, \dots, i_{k/2} \in \{1, \dots, d\}$. When $k \in \{0, 2\}$ one can easily check that $\mathcal{S}_k = \mathcal{K}_k$ and $C_I = 1$ for all $I \in \mathcal{K}_k$, thus

$$\phi_k^M(\Delta) = \phi_k(\Delta) \quad \text{if } k \text{ is odd or } k \in \{0, 2\}. \tag{4.A.2}$$

For $I \in \mathcal{W}_k$ with $k \geq 4$ and even we have:

$$\phi_I^M(\Delta) - \phi_I(\Delta) = \begin{cases} \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!}(C_I - 1), & \text{if } I \in \mathcal{S}_k, \\ \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!} C_I, & \text{if } I \in \mathcal{K}_k \setminus \mathcal{S}_k \\ 0, & \text{if } I \notin \mathcal{K}_k. \end{cases}$$

Note that C_I denotes the ratio between the total number of ways of arranging k objects in groups of size $n_1(I), \dots, n_d(I)$ with and without fixing pairs. Thus $C_I \in [0,1]$. For $k \geq 4$ even we can thus bound

$$\begin{aligned}\|\phi_k^M(\Delta) - \phi_k(\Delta)\|_1 &= \sum_{I \in \mathcal{S}_k} \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!}(1 - C_I) + \sum_{I \in \mathcal{K}_k \setminus \mathcal{S}_k} \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!} C_I \\ &\leq \left(\frac{\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!} \left[|\mathcal{S}_k| + \sum_{I \in \mathcal{K}_k} C_I\right] \\ &\leq 2\left(\frac{d\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!},\end{aligned}$$

since $|\mathcal{S}_k| = d^{k/2}$ and

$$\begin{aligned}\sum_{I \in \mathcal{K}_k} C_I &= \sum_{\substack{n_1,\dots,n_d \geq 0 \\ 2n_1 + \dots + 2n_d = k}} \frac{|\{I \in \mathcal{K}_k : n_1(I) = 2n_1, \dots, n_d(I) = 2n_d\}| \binom{k/2}{n_1, \dots, n_d}}{\binom{k}{2n_1, \dots, 2n_d}} \\ &= \sum_{\substack{n_1,\dots,n_d \geq 0 \\ 2n_1 + \dots + 2n_d = k}} \binom{k/2}{n_1, \dots, n_d} = d^{k/2}. \end{aligned} \tag{4.A.3}$$

Now, to propagate the estimate to time horizon T we note that

$$\begin{aligned}&\phi^M(T) - \phi(T) \\ &\quad= \sum_{m=0}^{M-1} \left\{\phi(m\Delta) \otimes \phi^M((M-m)\Delta) - \phi((m+1)\Delta) \otimes \phi^M((M-m-1)\Delta)\right\} \\ &\quad= \sum_{m=0}^{M-1} \left\{\phi(m\Delta) \otimes \phi^M(\Delta) \otimes \phi^M((M-m-1)\Delta) - \phi(m\Delta) \otimes \phi(\Delta) \otimes \phi^M((M-m-1)\Delta)\right\} \\ &\quad= \sum_{m=0}^{M-1} \left\{\phi(m\Delta) \otimes \left[\phi^M(\Delta) - \phi(\Delta)\right] \otimes \phi^M((M-m-1)\Delta)\right\},\end{aligned}$$

where we use the fact that $\phi((m+1)\Delta) = \phi(m\Delta) \otimes \phi(\Delta)$ and $\phi^M((m+1)\Delta) = \phi(\Delta) \otimes \phi^M(m\Delta)$ for $m = 0, \dots, M-1$ which follows directly from the explicit forms of ϕ and ϕ^M given in Theorem 4.2.1 and Theorem 4.2.3 respectively. We can hence bound

$$\begin{aligned}&\left\|\phi_k^M(T) - \phi_k(T)\right\|_1 \\ &\leq \sum_{m=0}^{M-1} \left\|\left(\phi(m\Delta) \otimes \left[\phi^M(\Delta) - \phi(\Delta)\right] \otimes \phi^M((M-m-1)\Delta)\right)_k\right\|_1 \\ &\overset{(i)}{\leq} \sum_{m=0}^{M-1} \sum_{\substack{n_1,n_2,n_3 \geq 0 \\ n_1+n_2+n_3 = k}} \|\phi_{n_1}(m\Delta)\|_1 \, \|\phi_{n_2}^M(\Delta) - \phi_{n_2}(\Delta)\|_1 \|\phi_{n_3}^M((M-m-1)\Delta)\|_1\end{aligned}$$

$$\overset{(ii)}{\leq} \begin{cases} \sum_{m=0}^{M-1} \sum_{\substack{k_1,k_2,k_3\geq 0 \\ 2k_1+2k_2+2k_3=k \\ k_2>1}} \left(\frac{dm\Delta}{2}\right)^{k_1} \frac{1}{k_1!} 2 \left(\frac{d\Delta}{2}\right)^{k_2} \frac{1}{k_2!} \left(\frac{d(M-m-1)\Delta}{2}\right)^{k_3} \frac{1}{k_3!}, & \text{if } k \geq 4 \text{ even}, \\ 0, & \text{otherwise.} \end{cases}$$

where in (i) we use the fact that $\|v \otimes w\|_1 \leq \|v\|_1 \|w\|_1$ for $v \in (\mathbb{R}^d)^{\otimes n}, w \in (\mathbb{R}^d)^{\otimes m}$, in (ii) for the middle term we apply Equation (4.A.2) and Equation (4.A.3) and for the first and third terms

$$\|\phi_k(m\Delta)\|_1 = \sum_{I \in \mathcal{W}_k} |\phi_I(m\Delta)| = \begin{cases} \frac{1}{(k/2)!} \left(\frac{m\Delta}{2}\right)^{k/2} |\mathcal{S}_k| = \frac{1}{(k/2)!} \left(\frac{dm\Delta}{2}\right)^{k/2}, & \text{if } k \text{ even}, \\ 0, & \text{if } k \text{ odd}, \end{cases}$$

and

$$\begin{aligned} \left\|\phi_k^M(m\Delta)\right\|_1 &= \sum_{\substack{k_1,\dots,k_m\geq 0 \\ k_1+\cdots+k_m=k}} \prod_{l=1}^{m} \|\phi_{k_l}^M(\Delta)\|_1 \\ &= \begin{cases} \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \prod_{l=1}^{m} \|\phi_{2j_l}^M(\Delta)\|_1 = \frac{1}{(k/2)!} \left(\frac{dm\Delta}{2}\right)^{k/2}, & \text{if } k \text{ even}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

since for k even one can compute

$$\begin{aligned} \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \prod_{l=1}^{m} \|\phi_{2j_l}^M(\Delta)\|_1 &= \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \prod_{l=1}^{m} \sum_{I \in \mathcal{K}_{2j_l}} \frac{C_I}{j_l!} \left(\frac{\Delta}{2}\right)^{j_l} \\ &= \left(\frac{\Delta}{2}\right)^{k/2} \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \prod_{l=1}^{m} \frac{1}{j_l!} \sum_{I \in \mathcal{K}_{2j_l}} C_I \\ &= \left(\frac{\Delta}{2}\right)^{k/2} \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \prod_{l=1}^{m} \frac{1}{j_l!} d^{j_l} \\ &= \left(\frac{d\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!} \sum_{\substack{j_1,\dots,j_m\geq 0 \\ 2j_1+\cdots+2j_m=k}} \binom{k/2}{j_1,\dots,j_m} \\ &= \left(\frac{dm\Delta}{2}\right)^{k/2} \frac{1}{(k/2)!}. \end{aligned}$$

To finish the proof we can manipulate the previous bound as follows when $k \geq 4$ and even

$$\left\|\phi_k^M(T) - \phi_k(T)\right\|_1$$

$$
\begin{aligned}
&\leq \sum_{m=0}^{M-1} \sum_{\substack{k_1,k_2,k_3\geq 0 \\ 2k_1+2k_2+2k_3=k \\ k_2>1}} \left(\frac{dm\Delta}{2}\right)^{k_1} \frac{1}{k_1!} 2 \left(\frac{d\Delta}{2}\right)^{k_2} \frac{1}{k_2!} \left(\frac{d(M-m-1)\Delta}{2}\right)^{k_3} \frac{1}{k_3!} \\
&= \left(\frac{d\Delta}{2}\right)^{k/2} \sum_{k_2=2}^{k/2} \frac{2}{k_2!} \sum_{m=0}^{M-1} \sum_{\substack{k_1,k_3\geq 0 \\ 2k_1+2k_3=k-2k_2}} \frac{1}{k_1!k_3!} m^{k_1}(M-m-1)^{k_3} \\
&= \left(\frac{d\Delta}{2}\right)^{k/2} \sum_{k_2=2}^{k/2} \frac{2}{k_2!(k/2-k_2)!} M(M-1)^{k/2-k_2} \\
&\leq 2M \left(\frac{d\Delta}{2}\right)^{k/2} \sum_{k_2=2}^{k/2} \frac{1}{(k_2-2)!(k/2-k_2)!} (M-1)^{k/2-k_2} \\
&= 2M \left(\frac{d\Delta}{2}\right)^{k/2} \frac{1}{(k/2-2)!} \sum_{k=0}^{n/2-2} \binom{k/2-2}{k} (M-1)^{k/2-k-2} \\
&= 2M \left(\frac{d\Delta}{2}\right)^{k/2} \frac{1}{(k/2-2)!} M^{k/2-2} \\
&= \frac{d^{k/2-1}\Delta}{(k/2-2)!} \left(\frac{T}{2}\right)^{k/2-1}.
\end{aligned}
$$

□

4.A.5 Proof of Theorem 4.2.8

Proof of Theorem 4.2.8. We start by showing that

$$\sigma_{M,I} \to \sigma_I > 0, \quad M \to \infty.$$

Note that

$$\sigma_{M,I}^2 = \sum_{K \in I \sqcup I} \phi_K^M(T) - \phi_I^M(T)^2 \to \sum_{K \in I \sqcup I} \phi_K(T) - \phi_I(T)^2 = \sigma_I^2, \quad M \to \infty,$$

by applying Theorem 4.2.6 to the word I and all words in $I \sqcup I$. Moreover, from Theorem 4.2.1, recall that

$$\phi_I(T) = \begin{cases} \left(\frac{T}{2}\right)^2 \frac{1}{(|I|/2)!}, & I \in \mathcal{S}_{2k} \text{ for some } k \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus if $I = (i_1, \ldots, i_{|I|}) \notin \mathcal{S}_{2k}$ for all $k \geq 0$ we have that $I^2 := (i_1, i_1, \ldots, i_{|I|}, i_{|I|}) \in I \sqcup I$ and thus

$$\sigma_I^2 = \sum_{K \in I \sqcup I} \phi_K(T) - \phi_I(T) \geq \phi_{I^2}(T) > 0.$$

If instead $I \in \mathcal{S}_{2k}$ for some $k \geq 0$ we note that there are $\binom{2k}{k}$ ways of shuffling

$$\{(i_1, i_1), \dots, (i_{|I|}, i_{|I|}), (i_1^*, i_1^*), \dots, (i_{|I|}^*, i_{|I|}^*)\},$$

i.e. the total number of shuffles of a word of length $4k$ with fixed pairs. All of the resulting "fixed-pairs" shuffles are in $\mathcal{S}_{4k}$. The word

$$(i_1, i_1^*, \dots, i_{|I|}, i_{|I|}^*, i_1, i_1^*, \dots, i_{|I|}, i_{|I|}^*),$$

is also in $I \sqcup\!\sqcup I$ but is not a "fixed-pairs" shuffle. We thus have that if $I \in \mathcal{S}_{2k}$

$$\begin{aligned}
\sigma_I^2 &= \sum_{K \in I \sqcup\!\sqcup I} \phi_K(T) - \phi_I(T)^2 \\
&\geq \left[\binom{2k}{k} + 1\right] \left(\frac{T}{2}\right)^{2k} \frac{1}{(2k)!} - \left(\frac{T}{2}\right)^{2k} \frac{1}{(k!)^2} \\
&= \left(\frac{T}{2}\right)^{2k} \frac{1}{(2k)!} > 0.
\end{aligned}$$

We can hence show consistency since

$$\begin{aligned}
&\left\| \hat{\phi}_I^{M,N}(T) - \phi_I(T) \right\|_{L^2} \\
&\qquad \overset{(i)}{\leq} \left\| \hat{\phi}_I^{M,N}(T) - \phi_I^M(T) \right\|_{L^2} + \left| \phi_I^M(T) - \phi_I(T) \right| \\
&\qquad \overset{(ii)}{\leq} \frac{\sigma_{M,I}}{\sqrt{N}} + \left| \phi_I^M(T) - \phi_I(T) \right| \\
&\qquad \overset{(iii)}{\to} 0, \qquad\qquad N, M \to \infty,
\end{aligned}$$

where in (i) we use the triangle inequality, in (ii) the formula for the standard deviation of N i.i.d. random variables and in (iii) $\sigma_{M,I}^2 \to \sigma_I^2$ as $M \to \infty$ and Theorem 4.2.6. Moreover

$$\begin{aligned}
&\frac{\sqrt{N}}{\sigma_I} \left(\hat{\phi}_I^{M,N}(T) - \phi_I(T) \right) \\
&\qquad = \frac{\sigma_{M,I}}{\sigma_I} \frac{\sqrt{N}}{\sigma_{M,I}} \left(\hat{\phi}_I^{M,N}(T) - \phi_I^M(T) \right) + \frac{\sqrt{N}}{\sigma_I} \left(\phi_I^M(T) - \phi_I(T) \right) \\
&\qquad = \begin{cases} \dfrac{\sigma_{M,I}}{\sigma_I} Z_{N,M,I} + \mathcal{O}\left(\dfrac{\sqrt{N}}{M}\right), & \text{if } |I| \geq 4 \text{ is even}, \\ \dfrac{\sigma_{M,I}}{\sigma_I} Z_{N,M,I}, & \text{otherwise}, \end{cases}
\end{aligned}$$

where

$$Z_{N,M,I} := \frac{\sqrt{N}}{\sigma_{M,I}} \left(\hat{\phi}_I^{M,N}(T) - \phi_I^M(T) \right) \to \mathcal{N}(0,1), \quad N, M \to \infty,$$

since, denoting by $F_{N,M,I}$ the c.d.f. of $Z_{N,M,I}$ and by Φ the c.d.f. of a $\mathcal{N}(0,1)$ random variable, we can apply the Berry-Esseen theorem (quantitative CLT) to the i.i.d. variables $\{S^I(\hat{\mathbb{B}}^{M,n})_{[0,T]}\}_{n=1}^N$ to deduce $\forall x \in \mathbb{R}$

$$|F_{N,M,I}(x) - \Phi(x)| \leq \frac{C\rho_{M,I}}{\sigma_{M,I}^3\sqrt{N}} \to 0, \quad M, N \to \infty,$$

since $\sigma_{M,I}^3 \to \sigma_I^3 > 0$ as $M \to \infty$ and

$$\begin{aligned}
\rho_{M,I} &:= \mathbb{E}\left[\left(S^I(\hat{\mathbb{B}}^M)_{[0,T]} - \mathbb{E}[S^I(\hat{\mathbb{B}}^{M,n})_{[0,T]}]\right)^3\right] \\
&= \sum_{K \in I \sqcup I \sqcup I} \phi_K^M(T) - 3\phi_I^M(T) \sum_{K \in I \sqcup I} \phi_K^M(T) + 2\phi_I^M(T)^3 \\
&\to \sum_{K \in I \sqcup I \sqcup I} \phi_K(T) - 3\phi_I(T) \sum_{K \in I \sqcup I} \phi_K(T) + 2\phi_I(T)^3 \\
&=: \rho_I < \infty.
\end{aligned}$$

The CLT result then follows by combining the distributional convergence of $Z_{N,M,I}$ with $\sigma_I/\sigma_{M,I} \to 1$ as $M \to \infty$, under the additional conditional that $\sqrt{N}/M \to 0$ when $|I| \geq 4$ and even. □

4.A.6 Proof of Theorem 4.2.10

Proof of Theorem 4.2.10. Joint consistency follows directly from the element-wise consistency result of Theorem 4.2.8, since element-wise convergence in L^2 implies joint convergence in L^2.

For the joint asymptotic normality result, we apply the Cramér-Wold theorem (Billingsley, 1986, Theorem 29.4). Let $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathbf{I}|} \setminus \{\mathbf{0}\}$ where $|\mathbf{I}|$ is the number of words in $\mathbf{I}$ and note

$$\sigma_{\mathbf{I},\boldsymbol{\lambda}}^2 := \mathrm{Var}\left(\langle \boldsymbol{\lambda}, S^{\mathbf{I}}(\mathbb{B})_{[0,T]} \rangle\right) = \langle \boldsymbol{\lambda}, \Sigma_{\mathbf{I}} \boldsymbol{\lambda} \rangle > 0,$$

since

$$\begin{aligned}
\sigma_{\mathbf{I},\boldsymbol{\lambda}}^2 = \sum_{I,J \in \mathbf{I}} \lambda_I \lambda_J \Sigma_{I,J}^{\mathbf{I}} &= \sum_{I,J \in \mathbf{I}} \lambda_I \lambda_J \mathrm{Cov}\left(S^I(\mathbb{B})_{[0,T]}, S^J(\mathbb{B})_{[0,T]}\right) \\
&= \sum_{I,J \in \mathbf{I}} \lambda_I \lambda_J \left\{ \sum_{K \in I \sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T) \right\}, \\
&\overset{(i)}{\geq} \sum_{I \in \mathbf{I}} \lambda_I^2 \left\{ \sum_{K \in I \sqcup I} \phi_K(T) - \phi_I(T)^2 \right\}, \\
&\overset{(ii)}{\geq} \|\boldsymbol{\lambda}\|^2 \min_{I \in \mathbf{I}} \sigma_I^2 > 0,
\end{aligned}$$

using in (i) the fact that by Theorem 4.2.1 on can deduce[49]

$$\sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T) \geq 0, \quad \forall I, J \neq (),$$

and in (ii) the fact that $\sigma_I^2 > 0$ for any word $I \neq ()$, cf. (proof of) Theorem 4.2.8. Note that since this holds for any $\boldsymbol{\lambda} \in \mathbb{R}^K \setminus \{\mathbf{0}\}$ this implies $\Sigma_{\mathbf{I}}$ is strictly positive definite.

For $M \geq 1$ we an now deduce

$$\sigma^2_{M,\mathbf{I},\boldsymbol{\lambda}} := \mathrm{Var}\left(\langle \boldsymbol{\lambda}, S^{\mathbf{I}}(\hat{\mathbb{B}}^M)_{[0,T]} \rangle\right) = \langle \boldsymbol{\lambda}, \Sigma_{M,\mathbf{I}} \boldsymbol{\lambda} \rangle \to \langle \boldsymbol{\lambda}, \Sigma_{\mathbf{I}} \boldsymbol{\lambda} \rangle = \sigma^2_{\mathbf{I},\boldsymbol{\lambda}}, \quad M \to \infty,$$

since for all $j, k \in \{1, \dots, K\}$

$$\begin{aligned}
\Sigma^{M,\mathbf{I}}_{I,J} &:= \mathrm{Cov}\left(S^I(\hat{\mathbb{B}}^M)_{[0,T]}, S^J(\hat{\mathbb{B}}^M)_{[0,T]}\right) \\
&= \sum_{K \in I \sqcup\!\sqcup J} \phi^M_K(T) - \phi^M_I(T)\phi^M_J(T) \\
&\to \sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T) = \Sigma^{\mathbf{I}}_{I,J}, \quad M \to \infty,
\end{aligned}$$

by applying Theorem 4.2.6 with words I, J and all words in $K \in I \sqcup\!\sqcup J$.

Then, proceeding as in Theorem 4.2.8, we can show asymptotic normality of $\langle \boldsymbol{\lambda}, \hat{\phi}^{M,N}_{\mathbf{I}}(T) \rangle$ by writing

$$\begin{aligned}
&\frac{\sqrt{N}}{\sigma_{\mathbf{I},\boldsymbol{\lambda}}} \left(\langle \boldsymbol{\lambda}, \hat{\phi}^{M,N}_{\mathbf{I}}(T)) \rangle - \langle \boldsymbol{\lambda}, \phi_{\mathbf{I}}(T) \rangle \right) \\
&\quad = \frac{\sigma_{M,\mathbf{I},\boldsymbol{\lambda}}}{\sigma_{\mathbf{I},\boldsymbol{\lambda}}} \frac{\sqrt{N}}{\sigma_{M,\mathbf{I},\boldsymbol{\lambda}}} \left(\langle \boldsymbol{\lambda}, \hat{\phi}^{M,N}_{\mathbf{I}}(T) \rangle - \langle \boldsymbol{\lambda}, \hat{\phi}^{M}_{\mathbf{I}}(T) \rangle \right) + \frac{\sqrt{N}}{\sigma_{\mathbf{I},\boldsymbol{\lambda}}} \left(\langle \boldsymbol{\lambda}, \hat{\phi}^{M}_{\mathbf{I}}(T) - \phi_{\mathbf{I}}(T) \rangle \right) \\
&\quad = \begin{cases} \dfrac{\sigma_{M,\mathbf{I},\boldsymbol{\lambda}}}{\sigma_{\mathbf{I},\boldsymbol{\lambda}}} Z_{N,M,\mathbf{I},\boldsymbol{\lambda}} + \mathcal{O}\left(\dfrac{\sqrt{N}}{M}\right), & \text{if } \exists I \in \mathbf{I} \text{ s.t. } |I| \geq 4 \text{ and even}, \\ \dfrac{\sigma_{M,\mathbf{I},\boldsymbol{\lambda}}}{\sigma_{\mathbf{I},\boldsymbol{\lambda}}} Z_{N,M,\mathbf{I},\boldsymbol{\lambda}}, & \text{otherwise}, \end{cases}
\end{aligned}$$

[49] We proceed as in Theorem 4.2.8 by applying Theorem 4.2.1. Note that if $I \notin \mathcal{S}_{2k}$ for all $k \geq 0$ or $J \notin \mathcal{S}_{2k}$ for all $k \geq 0$ then

$$\sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T) = \sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) \geq 0.$$

If instead $I \in \mathcal{S}_{2k}, J \in \mathcal{S}_{2j}$ for some $k, j \geq 0$ we note that there are $\binom{k+j}{k}$ ways of shuffling

$$\{(i_1, i_1), \dots, (i_k, i_k), (i^*_1, i^*_1), \dots, (i^*_j, i^*_j)\},$$

i.e. the total number of shuffles of a word of length $2(k+j)$ with fixed pairs. All of the resulting "fixed-pairs" shuffles are in $\mathcal{S}_{2(k+j)}$ and hence

$$\sum_{K \in I \sqcup\!\sqcup J} \phi_K(T) - \phi_I(T)\phi_J(T) \geq \binom{k+j}{k} \left(\frac{T}{2}\right)^{k+j} \frac{1}{(k+j)!} - \left(\frac{T}{2}\right)^k \frac{1}{k!} \left(\frac{T}{2}\right)^j \frac{1}{j!} = 0.$$

where

$$Z_{N,M,\mathbf{I},\boldsymbol{\lambda}} := \frac{\sqrt{N}}{\sigma_{M,\mathbf{I},\boldsymbol{\lambda}}}\left(\langle\boldsymbol{\lambda}, \hat{\phi}_{\mathbf{I}}^{M,N}(T)\rangle - \langle\boldsymbol{\lambda}, \hat{\phi}_{\mathbf{I}}^{M}(T)\rangle\right) \to \mathcal{N}(0,1), \quad N, M \to \infty,$$

by applying the Berry-Esseen theorem to the i.i.d. variables $\{\langle\boldsymbol{\lambda}, S^{\mathbf{I}}(\hat{\mathbb{B}}^{M,n})_{[0,T]}\rangle\}_{n=1}^{N}$ as in Theorem 4.2.8. We have thus shown that for any $\boldsymbol{\lambda} \in \mathbb{R}^K \setminus \{\mathbf{0}\}$,

$$\sqrt{N}\left(\langle\boldsymbol{\lambda}, \hat{\phi}_{\mathbf{I}}^{M,N}(T)\rangle - \langle\boldsymbol{\lambda}, \phi_{\mathbf{I}}(T)\rangle\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{\mathbf{I},\boldsymbol{\lambda}}^2 = \langle\boldsymbol{\lambda}, \Sigma_{\mathbf{I}}\boldsymbol{\lambda}\rangle),$$

for any sequence (M, N) such that $M, N \to \infty$ and, when $|I| \geq 4$ and even for some $I \in \mathbf{I}$, $\sqrt{N}/M \to 0$. The Cramér-Wold theorem then implies the desired asymptotic normality result. □

4.A.7 Proof of Proposition 4.2.12

Proof of Proposition 4.2.12. Note

$$S^k(\mathbb{B})_{[0,T]} = \int\cdots\int_{0<u_1<\cdots<u_k<T} \circ \mathrm{d}B_{u_1} \cdots \circ \mathrm{d}B_{u_{k-1}} \circ \mathrm{d}B_{u_k} = \frac{B_T^k}{k!},$$
$$S_c^k(\mathbb{B})_{[0,T]} = \int\cdots\int_{0<u_1<\cdots<u_k<T} \circ \mathrm{d}B_{u_1} \cdots \circ \mathrm{d}B_{u_{k-1}} \circ \mathrm{d}B_{u_{k-1}}\mathrm{d}B_{u_k} = \int_0^T \frac{B_u^{k-1}}{(k-1)!}\mathrm{d}B_u,$$

and, by considering the pathwise evaluation of the stochastic integral (Karandikar, 1995), both terms can be computed from a continuous-time observation $\mathbb{B}_{[0,T]}$.

To compute the variance reduction factor $(1-\rho_k^2)$ note

$$\mathrm{Var}\left(S^k(\mathbb{B})_{[0,T]}\right) = \mathrm{Var}\left(\frac{B_T^k}{k!}\right)$$
$$= \mathbb{E}\left[\frac{B_T^{2k}}{k!^2}\right] - \mathbb{E}\left[\frac{B_T^k}{k!}\right]^2 = \begin{cases} \dfrac{(2k-1)!!}{k!^2}T^k, & \text{if } k \text{ odd}, \\ \left(\dfrac{(2k-1)!!}{k!^2} - \dfrac{(k-1)!!^2}{k!^2}\right)T^k, & \text{if } k \text{ even}, \end{cases}$$

$$\mathrm{Var}\left(S_c^k(\mathbb{B})_{[0,T]}\right) = \mathrm{Var}\left(\int_0^T \frac{B_u^{k-1}}{(k-1)!}\mathrm{d}B_u\right) = \mathbb{E}\left[\left(\int_0^T \frac{B_u^{k-1}}{(k-1)!}\mathrm{d}B_u\right)^2\right]$$
$$= \frac{1}{(k-1)!^2}\int_0^T \mathbb{E}\left[B_u^{2k-2}\right]\mathrm{d}u = \frac{(2k-3)!!}{(k-1)!^2}\int_0^T u^{k-1}\mathrm{d}u = \frac{(2k-3)!!}{(k-1)!k!}T^k,$$

and

$$\mathrm{Cov}\left(S^k(\mathbb{B})_{[0,T]}, S_c^k(\mathbb{B})_{[0,T]}\right)$$

$$
\begin{aligned}
&= \mathrm{Cov}\left(\frac{B_T^k}{k!}, \int_0^T \frac{B_u^{k-1}}{(k-1)!}\mathrm{d}B_u\right) \\
&= \mathbb{E}\left[\left(\frac{B_T^k}{k!}\right)\left(\int_0^T \frac{B_u^{k-1}}{(k-1)!}\mathrm{d}B_u\right)\right] \\
&= \mathbb{E}\left[\left(\frac{B_T^k}{k!}\right)\left(\frac{B_T^k}{k!} - \frac{1}{2}\int_0^T \frac{B_u^{k-2}}{(k-2)!}\mathrm{d}u\right)\right] \\
&= \mathbb{E}\left[\frac{B_T^{2k}}{k!^2}\right] - \frac{1}{2(k-2)!k!}\int_0^T \mathbb{E}\left[B_T^k B_u^{k-2}\right]\mathrm{d}u \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2(k-2)!k!}\int_0^T \left(\sum_{i=0}^k \binom{k}{i}\mathbb{E}\left[(B_T - B_u)^{k-i}B_u^{i+k-2}\right]\right)\mathrm{d}u \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2(k-2)!k!}\sum_{i=0}^k \binom{k}{i}\int_0^T \mathbb{E}\left[(B_T - B_u)^{k-i}\right]\mathbb{E}\left[B_u^{i+k-2}\right]\mathrm{d}u \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2(k-2)!k!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \binom{k}{i}\int_0^T (k-i-1)!!(k+i-3)!!(T-u)^{(k-i)/2}u^{(i+k)/2-1}\mathrm{d}u \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2(k-2)!k!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \frac{k!(k-i-1)!!(k+i-3)!!}{i!(k-i)!}T^k\int_0^1 (1-s)^{(k-i)/2}s^{(i+k)/2-1}\mathrm{d}s \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2(k-2)!k!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \frac{k!(k+i-2)!}{i!(k-i)!!(k+i-2)!!}\frac{((k-i)/2)!((i+k)/2-1)!}{k!}T^k \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2^k(k-2)!k!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \frac{(k+i-2)!}{i!}T^k \\
&= \frac{(2k-1)!!}{k!^2}T^k - \frac{1}{2^k k!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \binom{k+i-2}{i}T^k \\
&= \frac{1}{k!}\left(\frac{T}{2}\right)^k\left\{\frac{(2k)!}{k!^2} - \sum_{\substack{i=0 \\ i+k\text{even}}}^k \binom{k+i-2}{i}\right\}.
\end{aligned}
$$

We can hence explicitly compute

$$
\rho_k = \mathrm{Corr}\Big(S^k(\mathbb{B})_{[0,T]}, S_c^k(\mathbb{B})_{[0,T]}\Big) = \delta_k\sqrt{2-\frac{1}{k}}\left\{1 - \frac{k!^2}{(2k)!}\sum_{\substack{i=0 \\ i+k\text{even}}}^k \binom{k+i-2}{i}\right\} \geq \sqrt{\frac{1}{2}-\frac{1}{4k}},
$$

where

$$
\delta_k = \begin{cases} \left(1 - \dfrac{k!^3}{(2k)!(k/2)!^2}\right)^{-1/2} & \text{if } k \text{ even}, \\ 1 & \text{if } k \text{ odd}. \end{cases}
$$

□

4.A.8 Proof of Proposition 4.2.13

Proof of Proposition 4.2.13. Note

$$S^k(\hat{\mathbb{B}}^M)_{[0,T]} = \int\cdots\int_{0<u_1<\cdots<u_k<T} \mathrm{d}\hat{B}^M_{u_1}\cdots\mathrm{d}\hat{B}^M_{u_k} = \frac{B_T^k}{k!}$$
$$S_c^k(\hat{\mathbb{B}}^M)_{[0,T]} = \sum_{m=0}^{M-1} S^{k-1}(\hat{\mathbb{B}}^M)_{[0,t_m]}(B_{t_{m+1}} - B_{t_m}) = \sum_{m=0}^{M-1} \frac{B_{t_m}^{k-1}}{(k-1)!}(B_{t_{m+1}} - B_{t_m}),$$

by using

$$S^k(\hat{\mathbb{B}}^M)_{[0,t]} = \frac{1}{k!}\sum_{m=0}^{M-1} \mathbb{1}_{[t_m,t_{m+1}]}(t)\left(B_{t_m} + \frac{t-t_m}{t_{m+1}-t_m}(B_{t_{m+1}} - B_{t_m})\right)^k, \quad k \geq 1.$$

This can be proved inductively starting from the base case $k = 1$ which follows trivially from the definition of $\hat{\mathbb{B}}^M$ and then writing

$$\begin{aligned}
&S^{k+1}(\hat{\mathbb{B}}^M)_{[0,t]} \\
&= \int_0^t S^k(\hat{\mathbb{B}}^M)_{[0,u]}\mathrm{d}\hat{B}^M_u \\
&= \int_0^t \frac{1}{k!}\sum_{m=0}^{M-1} \mathbb{1}_{[t_m,t_{m+1}]}(u)\left(B_{t_m} + \frac{u-t_m}{t_{m+1}-t_m}(B_{t_{m+1}} - B_{t_m})\right)^k \\
&\qquad\qquad \times \mathrm{d}\left(\sum_{n=0}^{M-1} \mathbb{1}_{[t_n,t_{n+1}]}(u)\left(B_{t_n} + \frac{u-t_n}{t_{n+1}-t_n}(B_{t_{n+1}} - B_{t_n})\right)\right) \\
&= \int_0^t \frac{1}{k!}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} \mathbb{1}_{[t_m,t_{m+1}]}(u)\mathbb{1}_{[t_n,t_{n+1}]}(u)\left(B_{t_m} + \frac{u-t_m}{t_{m+1}-t_m}(B_{t_{m+1}} - B_{t_m})\right)^k \frac{B_{t_{n+1}} - B_{t_n}}{t_{n+1}-t_n}\mathrm{d}u \\
&= \sum_{m=0}^{M-1}\int_{t\wedge t_m}^{t\wedge t_{m+1}} \frac{1}{k!}\left(B_{t_m} + \frac{B_{t_{m+1}} - B_{t_m}}{t_{m+1}-t_m}(u-t_m)\right)^k \frac{B_{t_{m+1}} - B_{t_m}}{t_{m+1}-t_m}\mathrm{d}u \\
&= \sum_{m=0}^{M-1}\left[\frac{1}{(k+1)!}\left(B_{t_m} + \frac{B_{t_{m+1}} - B_{t_m}}{t_{m+1}-t_m}(u-t_m)\right)^{k+1}\right]_{t\wedge t_m}^{t\wedge t_{m+1}} \\
&= \frac{1}{(k+1)!}\sum_{m=0}^{M-1} \mathbb{1}_{[t_m,t_{m+1}]}(t)\left(B_{t_m} + \frac{B_{t_{m+1}} - B_{t_m}}{t_{m+1}-t_m}(u-t_m)\right)^{k+1}.
\end{aligned}$$

To compute the correlation $\rho_{k,M}$ note

$$\mathrm{Var}\left(S^k(\hat{\mathbb{B}}^M)_{[0,T]}\right) = \mathrm{Var}\left(\frac{B_T^k}{k!}\right) = \begin{cases} \dfrac{(2k-1)!!}{k!^2}T^k & \text{if } k \text{ odd}, \\ \left(\dfrac{(2k-1)!!}{k!^2} - \dfrac{(k-1)!!^2}{k!^2}\right)T^k & \text{if } k \text{ even}, \end{cases}$$

$$\begin{aligned}
\mathrm{Var}\left(S_c^k(\hat{\mathbb{B}}^M)_{[0,T]}\right) &= \mathrm{Var}\left(\sum_{m=0}^{M-1}\frac{B_{t_m}^{k-1}}{(k-1)!}(B_{t_{m+1}}-B_{t_m})\right)\\
&= \mathbb{E}\left[\left(\sum_{m=0}^{M-1}\frac{B_{t_m}^{k-1}}{(k-1)!}(B_{t_{m+1}}-B_{t_m})\right)^2\right]\\
&= \sum_{m,n=0}^{M-1}\mathbb{E}\left[\frac{B_{t_m}^{k-1}}{(k-1)!}\frac{B_{t_n}^{k-1}}{(k-1)!}(B_{t_{m+1}}-B_{t_m})(B_{t_{n+1}}-B_{t_n})\right]\\
&= \sum_{m=0}^{M-1}\frac{1}{(k-1)!^2}\mathbb{E}\left[B_{t_m}^{2k-2}\right]\mathbb{E}\left[(B_{t_{m+1}}-B_{t_m})^2\right]\\
&= \sum_{m=0}^{M-1}\frac{1}{(k-1)!^2}(2k-3)!!t_m^{k-1}(t_{m+1}-t_m)\\
&= \frac{(2k-3)!!}{(k-1)!^2}\left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}m^k,
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Cov}\Big(S^k(\hat{\mathbb{B}}^M)_{[0,T]}, S_c^k(\hat{\mathbb{B}}^M)_{[0,T]}\Big)\\
&= \mathrm{Cov}\left(\frac{B_T^k}{k!},\sum_{m=0}^{M-1}\frac{B_{t_m}^{k-1}}{(k-1)!}(B_{t_{m+1}}-B_{t_m})\right)\\
&= \mathbb{E}\left[\left(\frac{B_T^k}{k!}\right)\left(\sum_{m=0}^{M-1}\frac{B_{t_m}^{k-1}}{(k-1)!}(B_{t_{m+1}}-B_{t_m})\right)\right]\\
&= \sum_{m=0}^{M-1}\frac{1}{k!(k-1)!}\mathbb{E}\Big[B_T^kB_{t_m}^{k-1}(B_{t_{m+1}}-B_{t_m})\Big]\\
&= \sum_{m=0}^{M-1}\frac{1}{k!(k-1)!}\sum_{j=0}^{k}\sum_{i=0}^{j}\binom{k}{i,j-i,k-j}\mathbb{E}\Big[B_{t_m}^{2k-j-1}\Big]\mathbb{E}\Big[(B_{t_{m+1}}-B_{t_m})^{i+1}\Big]\mathbb{E}\Big[(B_T-B_{t_{m+1}})^{j-i}\Big]\\
&= \sum_{m=0}^{M-1}\frac{1}{(k-1)!}\sum_{\substack{j=1\\ j\text{ odd}}}^{k}\sum_{\substack{i=1\\ i\text{ odd}}}^{j}\frac{(2k-j-2)!!i!!(j-i-1)!!}{i!(j-i)!(k-j)!}t_m^{k-(j+1)/2}(t_{m+1}-t_m)^{(i+1)/2}(T-t_{m+1})^{(j-i)/2}\\
&= \left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}\frac{1}{(k-1)!}\sum_{\substack{j=1\\ j\text{ odd}}}^{k}\sum_{\substack{i=1\\ i\text{ odd}}}^{j}\frac{(2k-j-2)!!}{(i-1)!!(j-i)!!(k-j)!}m^{k-(j+1)/2}(M-m-1)^{(j-i)/2}\\
&= \left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}\frac{1}{(k-1)!}\sum_{j=0}^{\lfloor(k-1)/2\rfloor}\sum_{i=0}^{j}\frac{(2k-2j-3)!!}{(2i)!!(2j-2i)!!(k-2j-1)!}m^{k-j-1}(M-m-1)^{j-i}\\
&= \left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}\frac{1}{(k-1)!}\sum_{j=0}^{\lfloor(k-1)/2\rfloor}\frac{(2k-2j-3)!!}{2^j(k-2j-1)!j!}m^{k-j-1}\sum_{i=0}^{j}\frac{j!}{i!(j-i)!}(M-m-1)^{j-i}\\
&= \left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}\frac{1}{(k-1)!}\sum_{j=0}^{\lfloor(k-1)/2\rfloor}\frac{(2k-2j-2)!}{2^{k-1}(k-j-1)!(k-2j-1)!j!}m^{k-j-1}(M-m)^j\\
&= \frac{1}{2^{k-1}(k-1)!}\left(\frac{T}{M}\right)^k\sum_{m=0}^{M-1}\sum_{j=0}^{\lfloor(k-1)/2\rfloor}\binom{2k-2j-2}{j,k-j-1,k-2j-1}m^{k-j-1}(M-m)^j.
\end{aligned}$$

□

4.B Semimartingales as canonical geometric stochastic processes: proofs

4.B.1 Proof of Theorem 4.3.8

Proof of Theorem 4.3.8. This result also follows from the stronger convergence in p-variation given in Friz and Victoir (2010, Theorem 14.16). Here we provide an alternative proof based on a simple application of the Wong-Zakai theorem.

It follows directly from the definition that $\{S^{\leq K}(\mathbb{X})_{[0,t]},\ t \in [0,T]\}$ satisfies the system[50] of Stratonovich SDEs

$$\mathrm{d}S_t^k = S_t^{k-1} \otimes \circ\, \mathrm{d}\mathbf{X}_t, \quad k = 0, \dots, K,\ t \in [0,T],$$

with initial condition $S_0 = 1 \in T^K((\mathbb{R}^d))$. We can rewrite this component-wise as

$$\mathrm{d}S_t^I = S_t^{(i_1,\dots,i_{k-1})} \circ \mathrm{d}X_t^{(i_k)} = \sum_{i=1}^{d} f_i^I(S_t) \circ \mathrm{d}X_t^{(i)}, \quad t \in [0,T], \tag{4.B.1}$$

where for each word $I = (i_1, \dots, i_k)$ with $|I| \leq K$ we have $f_i^I : T^K((\mathbb{R}^d)) \to \mathbb{R}$ maps

$$s \mapsto f_i^I(s) = \begin{cases} s^{(i_1,\dots,i_{k-1})}, & \text{if } i_k = i, \\ 0, & \text{otherwise.} \end{cases}$$

We can immediately observe that for any $i = 1, \dots, d$ and word I with $|I| \leq K$ the function f_i^I is "Fisk–Stratonovich acceptable" as defined[51] in Protter (1990, Chapter V.5): it is Lipschitz continuous with continuous partial derivatives given by

$$s \mapsto \frac{\partial f_i^I}{\partial s^J}(s) = \begin{cases} 1, & \text{if } J = (i_1, \dots, i_{k-1}) \text{ and } i = i_k, \\ 0, & \text{otherwise,} \end{cases}$$

[50] Here and in what follows we adopt the convention that $s^k = 0$ for any $k < 0$, $s \in T^K((\mathbb{R}^d))$.

[51] Conditions (i) and (ii) are satisfied with $g_i^I = f_i^I$ because f_i^I is time-independent, non-random and bounded on compacts where condition (i) should be

$$f(t,\omega,\mathbf{x}) - f(0,\omega,\mathbf{x}) = \int_0^t g(s,\omega,\mathbf{x})\mathrm{d}A_s.$$

for any word J with $|J| \leq K$, and the function

$$s \mapsto \frac{\partial f_i^I}{\partial s^J}(s) f_l^L(s) = \begin{cases} s^{(l_1,\dots,l_{m-1})}, & \text{if } J = (i_1,\dots,i_{k-1}) \text{ and } i = i_k \text{ and } l_m = l, \\ 0, & \text{otherwise,} \end{cases}$$

is also Lipschitz continuous for words I, J as above and $L = (l_1, \dots, l_m)$ with $|L| \leq K$. We can hence apply Protter (1990, Chapter V.5 Theorem 22) to the Stratonovich SDE (4.B.1) to conclude that the solution $S_t = S^K(\mathbb{X})_{[0,t]}$ for $t \in [0,T]$ is unique and satisfies the Itô SDE

$$\begin{aligned} \mathrm{d}S_t^I &= \sum_{i=1}^d f_i^I(S_t)\,\mathrm{d}X_t^{(i)} + \frac{1}{2}\sum_{i,j=1}^d \sum_{|J|\leq K} \frac{\partial f_i^I}{\partial s^J}(S_t) f_j^J(S_t)\,\mathrm{d}[X^{(j)}, X^{(i)}]_t, \\ &= S_t^{(i_1,\dots,i_{k-1})}\,\mathrm{d}X_t^{(i_k)} + \frac{1}{2} S_t^{(i_1,\dots,i_{k-2})}\,\mathrm{d}[X^{(i_{k-1})}, X^{(i_k)}]_t, \quad t \in [0,T], \end{aligned}$$

for each word $I = (i_1, \dots, i_k)$ with $|I| \leq K$. This can be rewritten in tensor notation as

$$\mathrm{d}S_t^k = S_t^{k-1} \otimes \mathrm{d}\mathbf{X}_t + \frac{1}{2} S_t^{k-2} \otimes \mathrm{d}[\mathbf{X}]_t, \quad k = 0, \dots, K,\ t \in [0,T]. \tag{4.B.2}$$

The signature of $\mathbb{X}^\pi$ is

$$S(\mathbb{X}^\pi)_{[0,T]} = \left(1,\, S^1(\mathbb{X}^\pi)_{[0,T]},\, S^2(\mathbb{X}^\pi)_{[0,T]},\, \dots\right) \in T((\mathbb{R}^d)),$$

with

$$S^k(\mathbb{X}^\pi)_{[0,T]} = \int\limits_{0<u_1<\dots<u_k<T} \cdots \int \mathrm{d}\mathbf{X}_{u_1}^\pi \otimes \cdots \otimes \mathrm{d}\mathbf{X}_{u_k}^\pi, \quad k \geq 1,$$

where the Riemann and Itô integrals coincide since $\mathbb{X}^\pi$ has almost surely bounded variation. The truncated signature at level K of $\mathbb{X}^\pi$ thus satisfies the (Itô) SDE

$$\mathrm{d}S_t^k = S_t^{k-1} \otimes \mathrm{d}\mathbf{X}_t^\pi, \quad t \in [0,T],\ k = 0, \dots, K.$$

We can then apply Konecny (1983, Theorem 1) to conclude that the truncated signature up to level K computed from $\mathbb{X}^\pi$, i.e. $\{S^{\leq K}(\mathbb{X}^\pi)_{[0,t]},\ t \in [0,T]\}$, converges uniformly on compacts in probability to the solution $\{S_t,\ t \in [0,T]\}$ of the SDE (4.B.2), i.e. to $\{S^{\leq K}(\mathbb{X})_{[0,t]},\ t \in [0,T]\}$ the truncated signature of $\mathbb{X}$. □

4.C In-fill asymptotics: proofs

We first establish a couple of useful lemmas which will be used repeatedly in the proofs of the in-fill asymptotic results. The first is a basic result which is also applied in the proof of the stochastic sewing lemma (Lê, 2020).

Lemma 4.C.1. *Let $\{Z_n,\ n=1,\dots,N\}$ be a finite sequence of E-valued random variables in L^m with $m \in [2,\infty)$ and let $\{\mathcal{G}_n,\ n=1,\dots,N\}$ be a filtration such that for each $n \in \{1,\dots,N\}$ the variables $Z_1,\dots,Z_{n-1}$ are $\mathcal{G}_n$-measurable. Then*

$$\left\|\sum_{n=1}^{N} Z_n\right\|_{L^m} \leq \sum_{n=1}^{N} \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m} + 2C_m \left(\sum_{n=1}^{N} \|Z_n\|_{L^m}^2\right)^{1/2}.$$

Proof of Lemma 4.C.1.

$$\begin{aligned}
\left\|\sum_{n=1}^{N} Z_n\right\|_{L^m} &\overset{(i)}{\leq} \left\|\sum_{n=1}^{N} \mathbb{E}_{\mathcal{G}_n}[Z_n]\right\|_{L^m} + \left\|\sum_{n=1}^{N} (Z_n - \mathbb{E}_{\mathcal{G}_n}[Z_n])\right\|_{L^m} \\
&\overset{(ii)}{\leq} \sum_{n=1}^{N} \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m} + C_m \left\|\sum_{n=1}^{N} \|Z_n - \mathbb{E}_{\mathcal{G}_n}[Z_n]\|^2\right\|_{L^{m/2}}^{1/2} \\
&\overset{(iii)}{\leq} \sum_{n=1}^{N} \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m} + C_m \left(\sum_{n=1}^{N} \|Z_n - \mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m}^2\right)^{1/2} \\
&\overset{(iv)}{\leq} \sum_{n=1}^{N} \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m} + C_m \left(\sum_{n=1}^{N} (\|Z_n\|_{L^m} + \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m})^2\right)^{1/2} \\
&\overset{(v)}{\leq} \sum_{n=1}^{N} \|\mathbb{E}_{\mathcal{G}_n}[Z_n]\|_{L^m} + 2C_m \left(\sum_{n=1}^{N} \|Z_n\|_{L^m}^2\right)^{1/2},
\end{aligned}$$

by using in (i) the triangle inequality, in (ii) triangle inequality and the Burkholder-Davis-Gundy (BDG) inequality applied to the martingale $\{M_n,\ n=1,\dots,N\}$ with $M_n = \sum_{i=1}^{n}(Z_i - \mathbb{E}_{\mathcal{G}_i}[Z_i])$, in (iii) and (iv) the triangle inequality and in (v) the contraction property of conditional expectation. □

Lemma 4.C.2. *Let $p, p_1, \dots, p_l \in (0,\infty) \cup \{+\infty\}$ be such that $p_1^{-1} + \ldots + p_l^{-1} = p^{-1}$ then for any set of tensors $\mathbf{A_1} \in L^{p_1}((\mathbb{R}^d)^{\otimes k_1}), \dots, \mathbf{A}_l \in L^{p_l}((\mathbb{R}^d)^{\otimes k_l})$*

$$\|\mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_l\|_{L^p} \lesssim d^l \|\mathbf{A}_1\|_{L^{p_1}} \cdots \|\mathbf{A}_l\|_{L^{p_l}}.$$

Proof of Lemma 4.C.2.

$$\begin{aligned}
\|\mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_l\|_{L^p} &\leq \sum_{(w_1,\dots,w_l) \in \mathcal{W}_{k_1+\ldots+k_l}} \left\|A_1^{w_1} \cdots A_l^{w_l}\right\|_{L^p} \\
&\overset{(*)}{\leq} \sum_{(w_1,\dots,w_l) \in \mathcal{W}_{k_1+\ldots+k_l}} \|A_1^{w_1}\|_{L^{p_1}} \cdots \|A_l^{w_l}\|_{L^{p_l}} \\
&\leq d^{k_1+\ldots+k_l} \|\mathbf{A}_1\|_{L^{p_1}} \cdots \|\mathbf{A}_l\|_{L^{p_l}},
\end{aligned}$$

where $\mathcal{W}_k = \{1,\dots,d\}^k$ denotes the set of words of length k and in $(*)$ we applied the classical Hölder inequality. □

4.C.1 Proof of Theorem 4.4.4

Proof of Theorem 4.4.4. Denote by $\{\pi_n,\ n \geq 1\}$ the signature-defining sequence of refining partitions of the interval $[0,T]$. Without loss of generality, we can consider $\{\pi_n,\ n \geq 1\}$ to be such that π_{n+1} is obtained from π_n by adding at most one refinement in each sub-interval, i.e. for each $[s,t] \in \pi_n$ either $[s,t] \in \pi_{n+1}$ or $[s,u],[u,t] \in \pi_{n+1}$ for $u \in (s,t)$. If not, one can consider a super-sequence satisfying this property and then pass to the original subsequence.

In the following for any $n \geq 1$ and $[s,t] \in \pi_n$ denote by $\pi_{n,[s,t]}$ the restriction of π_n to $[s,t]$ and, abusing notation slightly, $S(\mathbb{X}^{\pi_n})_{[s,t]} = S(\mathbb{X}^{\pi_{n,[s,t]}})_{[s,t]}$.

Let $[\tau_0,\tau_1] \in \pi_N$ for $N \geq 1$ and note that for any $k \geq 2$ and $n \geq N$ we can write

$$S^k(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,\tau_1]} - S^k(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]} = \sum_{[s,t]\in\pi_{n,[\tau_0,\tau_1]}} \left[S^k(\mathbb{X}^{\pi_{n,t}})_{[\tau_0,\tau_1]} - S^k(\mathbb{X}^{\pi_{n,s}})_{[\tau_0,\tau_1]}\right], \tag{4.C.1}$$

where the partitions $\pi_{n,s}$ are defined as $\pi_{n,s} = \pi_{n+1,[0,s]} \cup \pi_{n,[s,T]}$, i.e. for each $[s,t] \in \pi_n$ the partitions $\pi_{n,s}$ and $\pi_{n,t}$ differ by at most one point $u \in (s,t)$. Using Chen's relation and the definition of the tensor product we can write for each $[s,t] \in \pi_n$ with refinement $u \in (s,t)$,

$$\begin{aligned} &S^k(\mathbb{X}^{\pi_{n,t}})_{[\tau_0,\tau_1]} - S^k(\mathbb{X}^{\pi_{n,s}})_{[\tau_0,\tau_1]} \\ &\quad= \sum_{\substack{i_1,i_2,i_3\geq 0\\ i_1+i_2+i_3=k}} S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]} \otimes \left[S^{i_2}(\mathbb{X}^{\pi_{n+1}})_{[s,t]} - S^{i_2}(\mathbb{X}^{\pi_n})_{[s,t]}\right] \otimes S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]}. \end{aligned} \tag{4.C.2}$$

Note that for $i_2 \in \{0,1\}$,

$$S^{i_2}(\mathbb{X}^{\pi_{n+1}})_{[s,t]} - S^{i_2}(\mathbb{X}^{\pi_n})_{[s,t]} = 0,$$

and applying again Chen's relation when $i_2 \geq 2$ yields

$$S^{i_2}(\mathbb{X}^{\pi_{n+1}})_{[s,t]} = \sum_{j=0}^{i_2} S^j(\mathbb{X}^{\pi_{n+1}})_{[s,u]} \otimes S^{i_2-j}(\mathbb{X}^{\pi_{n+1}})_{[u,t]} = \frac{1}{i_2!}\sum_{j=0}^{i_2}\binom{i_2}{j}\mathbf{X}_{s,u}^{\otimes j} \otimes \mathbf{X}_{u,t}^{\otimes(i_2-j)},$$

where we used the fact that if $\mathbb{Y}$ is linear over $[s,t]$ then $S(\mathbb{Y})_{[s,t]} = \exp_\otimes \mathbf{Y}_{s,t}$ which also implies

$$S^{i_2}(\mathbb{X}^{\pi_n})_{[s,t]} = \frac{\mathbf{X}_{s,t}^{\otimes i_2}}{i_2!} = \frac{(\mathbf{X}_{s,u}+\mathbf{X}_{u,t})^{\otimes i_2}}{i_2!} = \frac{1}{i_2!}\sum_{\mathcal{I}\in\{0,1\}^{i_2}}\bigotimes_{i\in\mathcal{I}}\left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)}\right),$$

denoting by $\mathcal{I} \in \{0,1\}^{i_2}$ a binary number of length i_2 with $|\mathcal{I}| = \sum_{i\in\mathcal{I}} i$ and recalling that $\mathbf{x}^{\otimes 0} = 1, \mathbf{x}^{\otimes 1} = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^d$. We hence have that

$$S^{i_2}(\mathbb{X}^{\pi_{n+1}})_{[s,t]} - S^{i_2}(\mathbb{X}^{\pi_n})_{[s,t]} = \sum_{\mathcal{I}\in\{0,1\}^{i_2}} C_\mathcal{I}\bigotimes_{i\in\mathcal{I}}\left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)}\right),$$

where for $\mathcal{I} \in \{0,1\}^{i_2}$,

$$C_{\mathcal{I}} = \begin{cases} \frac{1}{i_2!}\left[\binom{i_2}{|\mathcal{I}|} - 1\right], & \text{if } \mathcal{I} = (1,\dots,1,0,\dots,0), \\ -\frac{1}{i_2!}, & \text{otherwise.} \end{cases}$$

Plugging this into Equation (4.C.1) via Equation (4.C.2) and noting that $C_{\mathcal{I}} = 0$ for $\mathcal{I} \in \{(0,\dots,0),(1,\dots,1)\}$ we can write for any $N \geq 1, [\tau_0,\tau_1] \in \pi_N, n \geq N$ and $k \geq 2$,

$$\begin{aligned} &S^k(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,\tau_1]} - S^k(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]} \\ &\quad= \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} \sum_{\substack{i_1,i_3\geq 0, i_2\geq 2 \\ i_1+i_2+i_3=k}} \sum_{\substack{\mathcal{I}\in\{0,1\}^{i_2} \\ \mathcal{I}\neq(0,\dots,0),(1,\dots,1)}} C_{\mathcal{I}}\, S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]} \otimes \bigotimes_{i\in\mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)}\right) \otimes S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]}. \end{aligned}$$

We now proceed inductively to show that for any $i \in \{1,\dots,k\}$ and any $[\tau_0,\tau_1] \in \pi_N$ with $N \geq 1$ the sequence $\{S^i(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]},\ n \geq N\}$ converges in $L^{mk/i}$ with rate $\mathcal{O}(\sum_{n'\geq n} |\pi_{n'}|^{\epsilon})$ and

$$\sup_{N\geq 1} \sup_{[\tau_0,\tau_1]\in\pi_N} \|S^i(\mathbb{X}^{\pi_N})_{[\tau_0,\tau_1]}\|_{L^{mk/i}} < \infty. \tag{4.C.3}$$

$k' = 1$. Note that for $[\tau_0,\tau_1] \in \pi_N$ with $N \geq 1$ one has $S^1(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]} = \mathbf{X}_{\tau_0,\tau_1}$ for all $n \geq N$, and

$$\|\mathbf{X}_{\tau_0,\tau_1}\|_{L^{mk}} \lesssim |\tau_1 - \tau_0|^{\alpha} \leq T^{\alpha} < \infty,$$

by Assumption (Aα). Hence $S^1(\mathbb{X})_{[0,T]} = \mathbf{X}_{0,T} \in L^{mk}$ and the statement holds trivially.

Assume the inductive hypothesis holds for all $i \in \{1,\dots,k'\}$ with $k' \in \{1,\dots,k-1\}$. Then for each $[\tau_0,\tau_1] \in \pi_N$ with $N \geq 1$ and $n \geq N$ let

$$\begin{aligned} &\left\|S^{k'+1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,\tau_1]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} \\ &\qquad\leq \sum_{\substack{i_1,i_3\geq 0, i_2\geq 2 \\ i_1+i_2+i_3=k'+1}} \sum_{\substack{\mathcal{I}\in\{0,1\}^{i_2} \\ \mathcal{I}\neq(0,\dots,0),(1,\dots,1)}} |C_{\mathcal{I}}| \left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}}, \end{aligned} \tag{4.C.4}$$

where for each $[\tau_0,\tau_1] \in \pi_N$, π_n with $n \geq N$, $i_1, i_3 \geq 0, i_2 \geq 2$ with $i_1 + i_2 + i_3 = k' + 1$ and $\mathcal{I} \in \{0,1\}^{i_2}$ we define

$$Z^{\mathcal{I}}_{[s,t]} := S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]} \otimes \bigotimes_{i\in\mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)}\right) \otimes S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]}, \quad [s,t] \in \pi_{n,[\tau_0,\tau_1]} \text{ with } u \in (s,t),$$

keeping only the dependence on $\mathcal{I}$ for notational convenience. Note that by applying Lemma 4.C.2, the inductive hypothesis and Assumption (Aα),

$$\|Z^{\mathcal{I}}_{[s,t]}\|_{L^{mk/(k'+1)}} \lesssim \left\|S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]}\right\|_{L^{mk/i_1}} \|\mathbf{X}_{s,u}\|^{|\mathcal{I}|}_{L^{mk}} \|\mathbf{X}_{u,t}\|^{i_2-|\mathcal{I}|}_{L^{mk}} \left\|S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]}\right\|_{L^{mk/i_3}} \lesssim |t-s|^{i_2\alpha},$$

and hence each $Z^{\mathcal{I}}_{[s,t]} \in L^{mk/(k'+1)}$. Moreover, by a simple application of the triangle inequality,

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} |t-s|^{i_2\alpha}. \tag{4.C.5}$$

Hence, if $\alpha > 1/2$, we have for each $[\tau_0,\tau_1] \in \pi_N$, π_n with $n \geq N$, $i_1, i_3 \geq 0, i_2 \geq 2$ with $i_1 + i_2 + i_3 = k'+1$ and $\mathcal{I} \in \{0,1\}^{i_2}$,

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim |\pi_n|^{2\alpha-1}|\tau_1-\tau_0|. \tag{4.C.6}$$

If $\alpha \in (1/3, 1/2]$ note that if $\mathcal{I}$ is such that $i_2 \geq 3$ then

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim |\pi_n|^{3\alpha-1}|\tau_1-\tau_0|, \tag{4.C.7}$$

but if $i_2 = 2$ then the bound Equation (4.C.5) is not strong enough. We can instead apply Lemma 4.C.1 with the filtration $\{\mathcal{G}_{[s,t]},\ [s,t] \in \pi_n\}$ defined by

$$\mathcal{G}_{[s,t]} := \mathcal{F}_s \vee \sigma(\mathbf{X}_{v,w}, [v,w] \in \pi_{n,[t,\tau]}),$$

by noting that each $Z^{\mathcal{I}}_{[v,w]}$ with $w \leq s$ is $\mathcal{G}_{[s,t]}$-measurable and $mk/(k'+1) \geq 2$ for all $k'+1 \leq k$. This implies

$$\begin{aligned}
&\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \\
&\qquad \leq \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} \left\| \mathbb{E}_{\mathcal{G}_{[s,t]}}[Z^{\mathcal{I}}_{[s,t]}] \right\|_{L^{mk/(k'+1)}} + \left(\sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} \|Z^{\mathcal{I}}_{[s,t]}\|^2_{L^{mk/(k'+1)}} \right)^{1/2} \\
&\qquad \leq \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} |t-s|^{\beta} + \left(\sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} |t-s|^{4\alpha} \right)^{1/2} \\
&\qquad \leq |\pi_n|^{(\beta-1)\wedge(2\alpha-1/2)} \left(|\tau_1-\tau_0| + |\tau_1-\tau_0|^{1/2} \right),
\end{aligned} \tag{4.C.8}$$

where we used the fact that for $\mathcal{I} \in \{(0,1),(1,0)\}$,

$$\left\| \mathbb{E}_{\mathcal{G}_{[s,t]}}[Z^{\mathcal{I}}_{[s,t]}] \right\|_{L^{mk/(k'+1)}}$$

$$
\begin{aligned}
&\overset{(i)}{=} \left\| S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]} \otimes \mathbb{E}_{\mathcal{G}_{[s,t]}} \left[\bigotimes_{i\in\mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)} \right) \right] \otimes S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]} \right\|_{L^{mk/(k'+1)}} \\
&\overset{(ii)}{\leq} \left\| S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]} \right\|_{L^{mk/i_1}} \left\| \mathbb{E}_{\mathcal{G}_{[s,t]}} \left[\bigotimes_{i\in\mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes i} \otimes \mathbf{X}_{u,t}^{\otimes(1-i)} \right) \right] \right\|_{L^{mk/2}} \left\| S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]} \right\|_{L^{mk/i_3}} \\
&\overset{(iii)}{\lesssim} \left\| \mathbb{E}_{\mathcal{G}_{[s,t]}} \left[\mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t} \right] \right\|_{L^{mk/2}} \\
&\overset{(iv)}{\lesssim} \left\| \mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}} \left[\mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t} \right] \right\|_{L^{mk/2}} \\
&\overset{(v)}{\lesssim} |t-s|^{\beta},
\end{aligned}
$$

by using in (i) measurability of $S^{i_1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,s]}$ and $S^{i_3}(\mathbb{X}^{\pi_n})_{[t,\tau_1]}$ with respect to $\mathcal{G}_{[s,t]}$, in (ii) Hölder inequality for tensors Lemma 4.C.2, in (iii) the inductive assumption Equation (4.C.3) and the fact that $\|\mathbf{A}\otimes\mathbf{B}\| = \|\mathbf{B}\otimes\mathbf{A}\|$ for any $\mathbf{A},\mathbf{B}\in\mathbb{R}^d$, in (iv) the tower property and the contractive property of conditional expectation applied to $\mathcal{G}_{[s,t]} \subseteq \mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}$ and in (v) Assumption (Aβ). Combining bound Equation (4.C.8) when $i_2 = 2$ and bound Equation (4.C.7) when $i_2 \geq 3$ with $\alpha \in (1/3, 1/2]$, it follows that for each $[\tau_0,\tau_1] \in \pi_N$, π_n with $n \geq N$, $i_1, i_3 \geq 0, i_2 \geq 2$ with $i_1 + i_2 + i_3 = k'+1$ and $\mathcal{I} \in \{0,1\}^{i_2}$,

$$
\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim |\pi_n|^{(\beta-1)\wedge(3\alpha-1)} |\tau_1 - \tau_0|. \tag{4.C.9}
$$

A similar reasoning can be applied when $\alpha \in (1/4, 1/3]$ (and $k \geq 3$) by considering the cases $i_2 \geq 4$, $i_2 = 3$ and $i_2 = 2$ separately. The case $i_2 \geq 4$ follows directly from Equation (4.C.5), the case $i_2 = 2$ follows from Equation (4.C.8) and the case $i_2 = 3$ can be shown in the same way as $i_2 = 2$ with the only difference being that we require Assumption (Aγ) to show that for $\mathcal{I} \in \{(0,0,1),(0,1,0),(1,0,0)\}$,

$$
\left\| \mathbb{E}_{\mathcal{G}_{[s,t]}}[Z^{\mathcal{I}}_{[s,t]}] \right\|_{L^{mk/(k'+1)}} \lesssim \left\| \mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}} \left[\mathbf{X}_{s,u} \otimes \mathbf{X}_{u,t}^{\otimes 2} \right] \right\|_{L^{mk/3}} \lesssim |t-s|^{\gamma},
$$

and for $\mathcal{I} \in \{(0,1,1),(1,0,1),(1,1,0)\}$,

$$
\left\| \mathbb{E}_{\mathcal{G}_{[s,t]}}[Z^{\mathcal{I}}_{[s,t]}] \right\|_{L^{mk/(k'+1)}} \lesssim \left\| \mathbb{E}_{\mathcal{F}_{0,s}\vee\mathcal{F}_{t,T}} \left[\mathbf{X}_{s,u}^{\otimes 2} \otimes \mathbf{X}_{u,t} \right] \right\|_{L^{mk/3}} \lesssim |t-s|^{\gamma},
$$

so that applying again Lemma 4.C.1,

$$
\begin{aligned}
&\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \\
&\qquad \leq \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} \left\| \mathbb{E}_{\mathcal{G}_{[s,t]}}[Z^{\mathcal{I}}_{[s,t]}] \right\|_{L^{mk/(k'+1)}} + \left(\sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]} \\ u\in(s,t)}} \|Z^{\mathcal{I}}_{[s,t]}\|^2_{L^{mk/(k'+1)}} \right)^{1/2}
\end{aligned}
$$

$$\leq \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]}\\ u\in(s,t)}} |t-s|^{\gamma} + \left(\sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]}\\ u\in(s,t)}} |t-s|^{6\alpha}\right)^{1/2}$$
$$\leq |\pi_n|^{(\gamma-1)\wedge(3\alpha-1/2)}\left(|\tau_1-\tau_0| + |\tau_1-\tau_0|^{1/2}\right). \tag{4.C.10}$$

Combining the cases $i_2 = 2$, $i_2 = 3$ and $i_2 \geq 4$ when $\alpha \in (1/4, 1/3]$ yields for each $[\tau_0, \tau_1] \in \pi_N$, π_n with $n \geq N$, $i_1, i_3 \geq 0, i_2 \geq 2$ with $i_1 + i_2 + i_3 = k' + 1$ and $\mathcal{I} \in \{0,1\}^{i_2}$,

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[\tau_0,\tau_1]}\\ u\in(s,t)}} Z^{\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim |\pi_n|^{(\beta-1)\wedge(\gamma-1)\wedge(4\alpha-1)}|\tau_1-\tau_0|. \tag{4.C.11}$$

Defining ϵ as in Equation (4.4.5), we can plug bounds Equation (4.C.6), Equation (4.C.9) and Equation (4.C.11) into Equation (4.C.4) to deduce that

$$\left\|S^{k'+1}(\mathbb{X}^{\pi_{n+1}})_{[\tau_0,\tau_1]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} \lesssim |\tau_1-\tau_0||\pi_n|^{\epsilon},$$

and hence, under the assumption that $\sum_{n\geq 1}|\pi_n|^{\epsilon} < \infty$, for any $[\tau_0, \tau_1] \in \pi_N$ with $N \geq 1$ the sequence $\{S^{k'+1}(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]},\ n \geq N\}$ is Cauchy in $L^{mk/(k'+1)}$. Since $L^{mk/(k'+1)}$ is a Banach space the sequence converges in $L^{mk/(k'+1)}$ to $S^{k'+1}(\mathbb{X})_{[\tau_0,\tau_1]} \in L^{mk/(k'+1)}$ (by uniqueness of limits) with rate

$$\left\|S^{k'+1}(\mathbb{X})_{[\tau_0,\tau_1]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}}$$
$$\leq \sum_{n'\geq n}\left\|S^{k'+1}(\mathbb{X}^{\pi_{n'+1}})_{[\tau_0,\tau_1]} - S^{k'+1}(\mathbb{X}^{\pi_{n'}})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} \lesssim |\tau_1-\tau_0|\sum_{n'\geq n}|\pi_{n'}|^{\epsilon}.$$

And, to complete the inductive step for $k' + 1$, note that for all $N \geq 1$ and $[\tau_0, \tau_1] \in \pi_N$,

$$\left\|S^{k'+1}(\mathbb{X}^{\pi_N})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} \lesssim \left\|S^{k'+1}(\mathbb{X})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} + |\tau_1-\tau_0|\sum_{n'\geq N}|\pi_{n'}|^{\epsilon}$$
$$\lesssim \left\|S^{k'+1}(\mathbb{X})_{[\tau_0,\tau_1]}\right\|_{L^{mk/(k'+1)}} + T\sum_{n'\geq 1}|\pi_{n'}|^{\epsilon}.$$

□

4.C.2 Proof of Lemma 4.4.6

Proof of Lemma 4.4.6. Note that for $k \geq 0$,

$$S^{k+1}(\mathbb{X}^{\pi})_{[0,\tau]} = \sum_{[u,v]\in\pi_{[0,\tau]}}\left[S^{k+1}(\mathbb{X}^{\pi})_{[0,v]} - S^{k+1}(\mathbb{X}^{\pi})_{[0,u]}\right]$$
$$\overset{(*)}{=} \sum_{[u,v]\in\pi_{[0,\tau]}}\left[\sum_{i=0}^{k+1} S^{k+1-i}(\mathbb{X}^{\pi})_{[0,u]} \otimes \frac{\mathbf{X}_{u,v}^{\otimes i}}{i!} - S^{k+1}(\mathbb{X}^{\pi})_{[0,u]}\right]$$

$$= \sum_{[u,v]\in\pi_{[0,\tau]}} \sum_{i=1}^{k+1} S^{k+1-i}(\mathbb{X}^\pi)_{[0,u]} \otimes \frac{\mathbf{X}_{u,v}^{\otimes i}}{i!}$$
$$= \sum_{i=0}^{k} \frac{1}{(1+i)!} \sum_{[u,v]\in\pi_{[0,\tau]}} S^{k-i}(\mathbb{X}^\pi)_{[0,u]} \otimes \mathbf{X}_{u,v}^{\otimes(1+i)},$$

where in $(*)$ we use Chen's relation and $S(\mathbb{X}^\pi)_{[u,v]} = \exp_\otimes \mathbf{X}_{u,v}$ since $\mathbb{X}^\pi$ is linear over $[u,v] \in \pi$. □

4.C.3 Proof of Theorem 4.4.7

Proof of Theorem 4.4.7. By definition of $\{\pi_n,\ n \geq 1\}$, we have

$$S^k(\mathbb{X}^{\pi_n})_{[0,T]} \xrightarrow{\mathbb{P}} S^k(\mathbb{X})_{[0,T]}, \quad n \to \infty,$$

for all $k \geq 0$. To show L^m-convergence it is thus sufficient to show that the sequence $\{S^k(\mathbb{X}^{\pi_n})_{[0,T]},\ n \geq 1\}$ is uniformly bounded in $L^{m'}$ for some $m' > m$, i.e.

$$\sup_{n\geq 1} \|S^k(\mathbb{X}^{\pi_n})_{[0,T]}\|_{L^{m'}} < \infty, \tag{4.C.12}$$

as this implies uniform integrability of $\{\|S^k(\mathbb{X}^{\pi_n})_{[0,T]}\|^m,\ n \geq 1\}$ which, together with convergence in probability, yields L^m-convergence to $S^k(\mathbb{X})_{[0,T]}$ (by uniqueness of limits).

We show Equation (4.C.12) by proving inductively the stronger statement: for each $i \in \{1,\dots,k\}$,

$$\sup_{n\geq 1} \sup_{\tau\in\pi_n} \|S^i(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^{p/i}} < \infty, \tag{4.C.13}$$

Setting $i = k$ and $\tau = T$ yields the required uniform bound in $L^{m'}$ with $m' = p/k > m$.

Note that the case $k' = 1$ is trivial since for any $n \geq 1$ and $\tau \in \pi_n$, by Assumption (Aα)

$$\|S^1(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^p} = \|\mathbf{X}_{0,\tau}\|_{L^p} \lesssim \tau^\alpha \lesssim T^\alpha.$$

Next, for the inductive step, assume that Equation (4.C.13) holds for all $i \in \{1,\dots,k'\}$ with $k' \leq k-1$. Then by using Lemma 4.4.6 we can bound for any $n \geq 1$ and $\tau \in \pi_n$,

$$\begin{aligned}
&\|S^{k'+1}(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^{p/(k'+1)}} \\
&\overset{(i)}{\leq} \sum_{i=0}^{k'} \frac{1}{(1+i)!} \left\| \sum_{[u,v]\in\pi_{n,[0,\tau]}} S^{k'-i}(\mathbb{X}^{\pi_n})_{[0,u]} \otimes \mathbf{X}_{u,v}^{\otimes(i+1)} \right\|_{L^{p/(k'+1)}} \\
&\overset{(ii)}{\leq} \sum_{[u,v]\in\pi_{n,[0,\tau]}} \left\| S^{k'}(\mathbb{X}^{\pi_n})_{[0,u]} \otimes \mathbb{E}_u[\mathbf{X}_{u,v}] \right\|_{L^{p/(k'+1)}} \\
&\qquad + \left(\sum_{[u,v]\in\pi_{n,[0,\tau]}} \left\| S^{k'}(\mathbb{X}^{\pi_n})_{[0,u]} \otimes \mathbf{X}_{u,v} \right\|_{L^{p/(k'+1)}}^2 \right)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{i=1}^{k'} \frac{1}{(1+i)!} \sum_{[u,v]\in\pi_{n,[0,\tau]}} \left\| S^{k'-i}(\mathbb{X}^{\pi_n})_{[0,u]} \otimes \mathbf{X}_{u,v}^{\otimes(i+1)} \right\|_{L^{p/(k'+1)}} \\
&\overset{(iii)}{\lesssim} \sum_{[u,v]\in\pi_{n,[0,\tau]}} \|\mathbb{E}_u[\mathbf{X}_{u,v}]\|_{L^p} + \left(\sum_{[u,v]\in\pi_{n,[0,\tau]}} \|\mathbf{X}_{u,v}\|_{L^p}^2 \right)^{1/2} + \sum_{i=1}^{k'} \sum_{[u,v]\in\pi_{n,[0,\tau]}} \|\mathbf{X}_{u,v}\|_{L^p}^{i+1} \\
&\overset{(iv)}{\lesssim} \sum_{[u,v]\in\pi_{n,[0,\tau]}} |v-u|^{\delta} + \left(\sum_{[u,v]\in\pi_{n,[0,\tau]}} |v-u|^{2\alpha} \right)^{1/2} + \sum_{i=1}^{k'} \sum_{[u,v]\in\pi_{n,[0,\tau]}} |v-u|^{(i+1)\alpha} \\
&\overset{(v)}{\lesssim} \tau + \sqrt{\tau} + \tau \lesssim T + \sqrt{T},
\end{aligned}
$$

where in (i) we applied the triangle inequality, in (ii) we bounded the $i = 0$ term by applying Lemma 4.C.1 to sequence of random variables

$$
Z_{[u,v]} := S^{k'}(\mathbb{X}^{\pi_n})_{[0,u]} \otimes \mathbf{X}_{u,v} \in L^{p/(k'+1)},
$$

with filtration $\{\mathcal{F}_u,\ [u,v] \in \pi_{n,[0,\tau]}\}$ and we bounded the $i = 1, \dots, k'$ terms by applying the triangle inequality, in (iii) we applied the Hölder inequality given in Lemma 4.C.2 and the inductive hypothesis Equation (4.C.13) for all signature levels up to k', in (iv) we used Assumption (Aα) and Assumption (Aδ).

We now consider the case where $\{\pi_n,\ n \geq 1\}$ is a sequence of refinements satisfying $\sum_{n\geq 1} |\pi_n|^{\epsilon} < \infty$. Without loss of generality, we can consider $\{\pi_n,\ n \geq 1\}$ to be such that π_{n+1} is obtained from π_n by adding at most one refinement in each sub-interval, i.e. for each $[s,t] \subseteq \pi_n$ either $[s,t] \subseteq \pi_{n+1}$ or $[s,u],[u,t] \subseteq \pi_{n+1}$ for $u \in (s,t)$. If not, one can consider a super-sequence satisfying this property and then pass to the original subsequence. Note that, assuming Assumption (Aα) and Assumption (Aδ) hold with $p = mk$ we can proceed as in the previous induction to show that

$$
\max_{1\leq i\leq k} \sup_{n\geq 1} \sup_{\tau\in\pi_n} \|S^i(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^{mk/i}} < \infty. \tag{4.C.14}
$$

We shall now proceed again by induction to show that: for each $i \in \{1, \dots, k\}$, for all $N \geq 1$, $\tau \in \pi_N$ and $n \geq N$,

$$
\|S^i(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} - S^i(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^{mk/i}} \lesssim |\pi_n|^{\epsilon}. \tag{4.C.15}
$$

The case $k' = 1$ is again trivial since for all $N \geq 1$, $\tau \in \pi_N$ and $n \geq N$, $S^1(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} = \mathbf{X}_{0,\tau}$, and hence

$$
\|S^1(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} - S^1(\mathbb{X}^{\pi_n})_{[0,\tau]}\|_{L^{mk}} = 0.
$$

For the inductive step assume Equation (4.C.15) holds for all $i \in \{1, \dots, k'\}$ with $k' \leq k$. Fix $N \geq 1$, $\tau \in \pi_N$ and $n \geq N$, then we can write the telescoping sum

$$
S^{k'+1}(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[0,\tau]} = \sum_{[s,t]\in\pi_{n,[0,\tau]}} \left[S^{k'+1}(\mathbb{X}^{\pi_{n,t}})_{[0,\tau]} - S^{k'+1}(\mathbb{X}^{\pi_{n,s}})_{[0,\tau]} \right], \tag{4.C.16}
$$

where the partitions $\pi_{n,s}$ are defined as $\pi_{n,s} = \pi_{n+1,[0,s]} \cup \pi_{n,[s,T]}$, i.e. for each $[s,t] \in \pi_n$ the partitions $\pi_{n,s}$ and $\pi_{n,t}$ differ by at most one point $u \in (s,t)$. Note that for each $[s,t] \in \pi_n$ with refinement $u \in (s,t)$ we can apply Lemma 4.4.6 to write

$$
\begin{aligned}
&S^{k'+1}(\mathbb{X}^{\pi_{n,t}})_{[0,\tau]} - S^{k'+1}(\mathbb{X}^{\pi_{n,s}})_{[0,\tau]} \\
&\quad= \sum_{i=0}^{k'} \frac{1}{(1+i)!} \Bigg\{ S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,s]} \otimes \mathbf{X}_{s,u}^{\otimes(1+i)} + S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,u]} \otimes \mathbf{X}_{u,t}^{\otimes(1+i)} - S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,s]} \otimes \mathbf{X}_{s,t}^{\otimes(1+i)} \\
&\qquad\qquad + \sum_{[v,w] \in \pi_{n,[t,\tau]}} \left[S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,v]} - S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,v]} \right] \otimes \mathbf{X}_{v,w}^{\otimes(1+i)} \Bigg\} \\
&\quad= \sum_{i=0}^{k'} \frac{1}{(1+i)!} \Bigg\{ S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \left[\mathbf{X}_{s,u}^{\otimes(1+i)} - \mathbf{X}_{s,t}^{\otimes(1+i)} \right] + \sum_{j=0}^{k'-i} S^{k'-i-j}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \frac{\mathbf{X}_{s,u}^{\otimes j}}{j!} \otimes \mathbf{X}_{u,t}^{\otimes(1+i)} \\
&\qquad\qquad + \sum_{[v,w] \in \pi_{n,[t,\tau]}} \left[S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,v]} - S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,v]} \right] \otimes \mathbf{X}_{v,w}^{\otimes(1+i)} \Bigg\} \\
&\quad= \sum_{i=0}^{k'} \frac{1}{(1+i)!} \Bigg\{ S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \left[\mathbf{X}_{s,u}^{\otimes(1+i)} + \mathbf{X}_{u,t}^{\otimes(1+i)} - \mathbf{X}_{s,t}^{\otimes(1+i)} \right] \\
&\qquad\qquad + \sum_{j=0}^{k'-i-1} \frac{1}{(1+j)!} S^{k'-i-j-1}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \mathbf{X}_{s,u}^{\otimes(1+j)} \otimes \mathbf{X}_{u,t}^{\otimes(1+i)} \\
&\qquad\qquad + \sum_{[v,w] \in \pi_{n,[t,\tau]}} \left[S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,v]} - S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,v]} \right] \otimes \mathbf{X}_{v,w}^{\otimes(1+i)} \Bigg\} \\
&\quad= \sum_{i=0}^{k'} \frac{1}{(1+i)!} \Bigg\{ - S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \sum_{\substack{\mathcal{I} \in \{0,1\}^{1+i} \\ \mathcal{I} \neq (0,\dots,0),(1,\dots,1)}} \bigotimes_{l \in \mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes l} \otimes \mathbf{X}_{u,t}^{\otimes(1-l)} \right) \\
&\qquad\qquad + \sum_{j=0}^{k'-i-1} \frac{1}{(1+j)!} S^{k'-i-j-1}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \mathbf{X}_{s,u}^{\otimes(1+j)} \otimes \mathbf{X}_{u,t}^{\otimes(1+i)} \\
&\qquad\qquad + \sum_{[v,w] \in \pi_{n,[t,\tau]}} \left[S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,v]} - S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,v]} \right] \otimes \mathbf{X}_{v,w}^{\otimes(1+i)} \Bigg\},
\end{aligned}
$$

by noting that for all $[v,w] \in \pi_n$ with $w \leq s$, $S(\mathbb{X}^{\pi_{n,t}})_{[0,v]} = S(\mathbb{X}^{\pi_{n,s}})_{[0,v]} = S(\mathbb{X}^{\pi_{n+1}})_{[0,v]}$ and when applying Chen's relation to $S(\mathbb{X}^{\pi_{n,t}})_{[0,u]}$ setting $S(\mathbb{X}^{\pi_{n,t}})_{[s,u]} = \exp_{\otimes} \mathbf{X}_{s,u}$ since $\mathbb{X}^{\pi_{n,t}}$ is linear over $[s,u] \in \pi_{n,t}$. Plugging this expression into Equation (4.C.16) and exchanging the orders of the summations we obtain

$$
\begin{aligned}
&S^{k'+1}(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[0,\tau]} \\
&\quad= \sum_{i=0}^{k'} \frac{1}{(1+i)!} \Bigg\{ - \sum_{\substack{\mathcal{I} \in \{0,1\}^{1+i} \\ \mathcal{I} \neq (0,\dots,0),(1,\dots,1)}} \sum_{\substack{[s,t] \in \pi_{n,[0,\tau]} \\ u \in (s,t)}} S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \bigotimes_{l \in \mathcal{I}} \left(\mathbf{X}_{s,u}^{\otimes l} \otimes \mathbf{X}_{u,t}^{\otimes(1-l)} \right) \\
&\qquad\qquad + \sum_{j=0}^{k'-i-1} \frac{1}{(1+j)!} \sum_{\substack{[s,t] \in \pi_{n,[0,\tau]} \\ u \in (s,t)}} S^{k'-i-j-1}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} \otimes \mathbf{X}_{s,u}^{\otimes(1+j)} \otimes \mathbf{X}_{u,t}^{\otimes(1+i)}
\end{aligned}
$$

$$+\sum_{[v,w]\in\pi_{n,[0,\tau]}}\left(\sum_{[s,t]\in\pi_{n,[0,v]}}\left[S^{k'-i}(\mathbb{X}^{\pi_{n,t}})_{[0,v]}-S^{k'-i}(\mathbb{X}^{\pi_{n,s}})_{[0,v]}\right]\right)\otimes\mathbf{X}_{v,w}^{\otimes(1+i)}\Bigg\}$$

$$=\sum_{i=0}^{k'}\frac{1}{(1+i)!}\Bigg\{-\sum_{\substack{\mathcal{I}\in\{0,1\}^{1+i}\\ \mathcal{I}\neq(0,\dots,0),(1,\dots,1)}}\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]}\otimes\bigotimes_{l\in\mathcal{I}}\left(\mathbf{X}_{s,u}^{\otimes l}\otimes\mathbf{X}_{u,t}^{\otimes(1-l)}\right)$$

$$+\sum_{j=0}^{k'-i-1}\frac{1}{(1+j)!}\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}S^{k'-i-j-1}(\mathbb{X}^{\pi_{n+1}})_{[0,s]}\otimes\mathbf{X}_{s,u}^{\otimes(1+j)}\otimes\mathbf{X}_{u,t}^{\otimes(1+i)}$$

$$+\sum_{[v,w]\in\pi_{n,[0,\tau]}}\left[S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,v]}-S^{k'-i}(\mathbb{X}^{\pi_n})_{[0,v]}\right]\otimes\mathbf{X}_{v,w}^{\otimes(1+i)}\Bigg\}$$

$$=\sum_{i=0}^{k'}\frac{1}{(1+i)!}\Bigg\{\sum_{\substack{\mathcal{I}\in\{0,1\}^{1+i}\\ \mathcal{I}\neq(0,\dots,0),(1,\dots,1)}}\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}Z_{[s,t]}^{1,\mathcal{I}}+\sum_{j=0}^{k'-i-1}\frac{1}{(1+j)!}\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}Z_{[s,t]}^{2,i,j}+\sum_{[v,w]\in\pi_{n,[0,\tau]}}Z_{[v,w]}^{3,i}\Bigg\}.$$

We thus proceed to bound each of the summation terms over $\pi_{n,[0,\tau]}$ using Lemma 4.C.1 and Assumption (Aα) and Assumption (Aδ).

Let $\mathcal{I}\in\{0,1\}^{1+i}$ with $\mathcal{I}\neq(0,\dots,0),(1,\dots,1)$ with $i\in\{1,\dots,k'\}$. Note that for all $[s,t]\in\pi_{n,[0,\tau]}$

$$\|Z_{[s,t]}^{1,\mathcal{I}}\|_{L^{mk/(k'+1)}}\leq\left\|S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]}\right\|_{L^{mk/(k'-i)}}\|\mathbf{X}_{s,u}\|_{L^{mk}}^{|\mathcal{I}|}\|\mathbf{X}_{u,t}\|_{L^{mk}}^{1+i-|\mathcal{I}|}\lesssim|t-s|^{(1+i)\alpha},\quad(4.C.17)$$

by applying Lemma 4.C.2, the uniform bound Equation (4.C.14) and Assumption (Aα), hence $Z_{[s,t]}^{1,\mathcal{I}}\in L^{mk/(k'+1)}$. When $i=1$ we can thus apply Lemma 4.C.1 to the sequence $\{Z_{[s,t]}^{1,\mathcal{I}},\ [s,t]\in\pi_{n,[0,\tau]},\ u\in(s,t)\}$ with filtration $\{\mathcal{F}_u,\ [s,t]\in\pi_{n,[0,\tau]},\ u\in(s,t)\}$ to bound

$$\begin{aligned}\left\|\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}Z_{[s,t]}^{1,\mathcal{I}}\right\|_{L^{mk/(k'+1)}}&\leq\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}\|\mathbb{E}_u[Z_{[s,t]}^{1,\mathcal{I}}]\|_{L^{mk/(k'+1)}}+\left(\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}\|Z_{[s,t]}^{1,\mathcal{I}}\|_{L^{mk/(k'+1)}}^2\right)^{1/2}\\&\lesssim\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}|t-s|^{\alpha+\delta}+\left(\sum_{\substack{[s,t]\in\pi_{n,[0,\tau]}\\ u\in(s,t)}}|t-s|^{4\alpha}\right)^{1/2}\\&\lesssim|\pi_n|^{\alpha+\delta-1}\tau+|\pi_n|^{2\alpha-1/2}\sqrt{\tau}\lesssim|\pi_n|^{\epsilon}(\tau+\sqrt{\tau}),\end{aligned}\quad(4.C.18)$$

where we used Equation (4.C.17) with $i=1$ and

$$\begin{aligned}\|\mathbb{E}_u[Z_{[s,t]}^{1,\mathcal{I}}]\|_{L^{mk/(k'+1)}}&=\left\|S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]}\otimes\mathbf{X}_{s,u}\otimes\mathbb{E}_u[\mathbf{X}_{u,t}]\right\|_{L^{mk/(k'+1)}}\\&\leq\left\|S^{k'-1}(\mathbb{X}^{\pi_{n+1}})_{[0,s]}\right\|_{L^{mk/(k'-1)}}\|\mathbf{X}_{s,u}\|_{L^{mk}}\|\mathbb{E}_u[\mathbf{X}_{u,t}]\|_{L^{mk}}\lesssim|t-s|^{\alpha+\delta},\end{aligned}$$

by applying Lemma 4.C.2, the uniform bound Equation (4.C.14) and Assumption (Aα) and Assumption (Aδ). When $i\geq2$ we can directly apply the triangle inequality and Equation (4.C.17)

to bound

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} Z^{1,\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \leq \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} \|Z^{1,\mathcal{I}}_{[s,t]}\|_{L^{mk/(k'+1)}}$$
$$\lesssim \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} |t-s|^{(1+i)\alpha} \lesssim |\pi_n|^{3\alpha-1}\tau \lesssim |\pi_n|^{\epsilon}\tau. \tag{4.C.19}$$

Next, let $i \in \{0,\dots,k'\}$ and $j \in \{0,\dots,k'-i-1\}$. We can proceed exactly as for $Z^{1,\mathcal{I}}_{[s,t]}$ (applying Lemma 4.C.1 when $i=j=0$ or the triangle inequality when $i+j\geq 1$) to show that under Assumption (Aα) and Assumption (Aδ),

$$\left\| \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} Z^{2,i,j}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \lesssim |\pi_n|^{\epsilon}(\tau+\sqrt{\tau}). \tag{4.C.20}$$

Finally, let $i \in \{0,\dots,k'\}$. We proceed in a similar way as for $Z^{1,\mathcal{I}}_{[s,t]}$ and $Z^{2,i,j}_{[s,t]}$ but using the inductive hypothesis Equation (4.C.15) instead of the bound Equation (4.C.14). Note that for all $[s,t]\in\pi_{n,[0,\tau]}$

$$\|Z^{3,i}_{[s,t]}\|_{L^{mk/(k'+1)}} \leq \left\|S^{k'-i}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} - S^{k'-i}(\mathbb{X}^{\pi_n})_{[0,s]}\right\|_{L^{mk/(k'-i)}} \|\mathbf{X}_{s,t}\|^{1+i}_{L^{mk}}$$
$$\lesssim |\pi_n|^{\epsilon}|t-s|^{(1+i)\alpha}, \tag{4.C.21}$$

by applying Lemma 4.C.2, the inductive hypothesis Equation (4.C.15) and Assumption (Aα), hence $Z^{3,i}_{[s,t]} \in L^{mk/(k'+1)}$. When $i=0$ we can hence apply Lemma 4.C.1 to the sequence $\{Z^{3,i}_{[s,t]},\ [s,t]\in\pi_{n,[0,\tau]}\}$ with filtration $\{\mathcal{F}_s,\ [s,t]\in\pi_{n,[0,\tau]}\}$ to bound

$$\left\| \sum_{[s,t]\in\pi_{n,[0,\tau]}} Z^{3,i}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \leq \sum_{[s,t]\in\pi_{n,[0,\tau]}} \|\mathbb{E}_s[Z^{3,i}_{[s,t]}]\|_{L^{mk/(k'+1)}} + \left(\sum_{[s,t]\in\pi_{n,[0,\tau]}} \|Z^{3,i}_{[s,t]}\|^2_{L^{mk/(k'+1)}} \right)^{1/2}$$
$$\lesssim \sum_{[s,t]\in\pi_{n,[0,\tau]}} |\pi_n|^{\epsilon}|t-s|^{\delta} + \left(\sum_{[s,t]\in\pi_{n,[0,\tau]}} |\pi_n|^{2\epsilon}|t-s|^{2\alpha} \right)^{1/2}$$
$$\lesssim |\pi_n|^{\epsilon}(\tau+\sqrt{\tau}), \tag{4.C.22}$$

where we used Equation (4.C.21) with $i=0$ and

$$\|\mathbb{E}_s[Z^{3,i}_{[s,t]}]\|_{L^{mk/(k'+1)}} = \left\| \left(S^{k'}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} - S^{k'}(\mathbb{X}^{\pi_n})_{[0,s]} \right) \otimes \mathbb{E}_s[\mathbf{X}_{s,t}] \right\|_{L^{mk/(k'+1)}}$$
$$\leq \left\| S^{k'}(\mathbb{X}^{\pi_{n+1}})_{[0,s]} - S^{k'}(\mathbb{X}^{\pi_n})_{[0,s]} \right\|_{L^{mk/k'}} \|\mathbb{E}_s[\mathbf{X}_{s,t}]\|_{L^{mk}} \lesssim |\pi_n|^{\epsilon}|t-s|^{\delta},$$

by applying Lemma 4.C.2, the inductive hypothesis Equation (4.C.15) and Assumption (Aδ). When

$i \geq 1$ we can instead directly apply the triangle inequality and Equation (4.C.21) to bound

$$\left\| \sum_{[s,t]\in\pi_{n,[0,\tau]}} Z^{3,i}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \leq \sum_{[s,t]\in\pi_{n,[0,\tau]}} \|Z^{3,i}_{[s,t]}\|_{L^{mk/(k'+1)}} \lesssim \sum_{[s,t]\in\pi_{n,[0,\tau]}} |\pi_n|^{\epsilon}|t-s|^{(1+i)\alpha} \lesssim |\pi_n|^{\epsilon}\tau. \tag{4.C.23}$$

Combining bounds Equation (4.C.18), Equation (4.C.19), Equation (4.C.20), Equation (4.C.22) and Equation (4.C.23) yields

$$\begin{aligned}
&\left\|S^{k'+1}(\mathbb{X}^{\pi_{n+1}})_{[0,\tau]} - S^{k'+1}(\mathbb{X}^{\pi_n})_{[0,\tau]}\right\|_{L^{mk/(k'+1)}} \\
&\quad\leq \sum_{i=0}^{k'} \frac{1}{(1+i)!}\Bigg\{ \sum_{\substack{\mathcal{I}\in\{0,1\}^{1+i} \\ \mathcal{I}\neq(0,\dots,0),(1,\dots,1)}} \left\| \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} Z^{1,\mathcal{I}}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \\
&\qquad\qquad + \sum_{j=0}^{k'-i-1} \frac{1}{(1+j)!} \left\| \sum_{\substack{[s,t]\in\pi_{n,[0,\tau]} \\ u\in(s,t)}} Z^{2,i,j}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \\
&\qquad\qquad + \left\| \sum_{[s,t]\in\pi_{n,[0,\tau]}} Z^{3,i}_{[s,t]} \right\|_{L^{mk/(k'+1)}} \Bigg\} \\
&\quad\lesssim |\pi_n|^{\epsilon}(\tau+\sqrt{\tau}) \lesssim |\pi_n|^{\epsilon}(T+\sqrt{T}),
\end{aligned}$$

which proves Equation (4.C.15) with $i = k'+1$, completing the inductive step.

Setting $i = k$, $N = 1$ and $\tau = T$ in Equation (4.C.15) yields

$$\|S^k(\mathbb{X}^{\pi_{n+1}})_{[0,T]} - S^k(\mathbb{X}^{\pi_n})_{[0,T]}\|_{L^m} \lesssim |\pi_n|^{\epsilon}, \quad n \geq 1,$$

and hence, assuming $\sum_{n\geq 1}|\pi_n|^{\epsilon} < \infty$, the sequence $\{S^k(\mathbb{X}^{\pi_n})_{[0,T]},\ n \geq\}$ is Cauchy in L^m. Since L^m is a Banach space, the sequence converges in L^m to $S^k(\mathbb{X})_{[0,T]} \in L^m$ (by uniqueness of limits) with rate

$$\|S^k(\mathbb{X}^{\pi_n})_{[0,T]} - S^k(\mathbb{X})_{[0,T]}\|_{L^m} \leq \sum_{n'\geq n} \|S^k(\mathbb{X}^{\pi_{n'+1}})_{[0,T]} - S^k(\mathbb{X}^{\pi_{n'}})_{[0,T]}\|_{L^m} \lesssim \sum_{n'\geq n} |\pi_n|^{\epsilon}.$$

□

4.D Long-span asymptotics: proofs

4.D.1 Proof of Proposition 4.4.23

Proof of Proposition 4.4.23. The first implication follows directly from the definitions of stationarity and joint stationarity of the increments. It remains to show the latter implies stationarity of

$\{\mathbb{X}^n,\ n \geq 1\}$, i.e. Equation (4.4.18) implies Equation (4.4.14). Under joint stationarity of the increments, for all $k \in \mathbb{N}$, $n_1, \dots, n_k \in \mathbb{N}$ and $n \geq 0$,

$$\begin{aligned}
&\mathbb{P}(\mathbb{X}^{n_1} \in A_1, \dots, \mathbb{X}^{n_k} \in A_k) \\
&\qquad = \mathbb{P}(\mathbf{X}_{(n_1-1)T,(n_1-1)T+t_1^1} \in B_1^1, \dots, \mathbf{X}_{(n_k-1)T,(n_k-1)T+t_{m_k}^k} \in B_{m_k}^k), \\
&\qquad = \mathbb{P}(\mathbf{X}_{(n_1+n-1)T,(n_1+n-1)T+t_1^1} \in B_1^1, \dots, \mathbf{X}_{(n_k+n-1)T,(n_k+n-1)T+t_{m_k}^k} \in B_{m_k}^k), \\
&\qquad = \mathbb{P}(\mathbb{X}^{n_1+n} \in A_1, \dots, \mathbb{X}^{n_k+n} \in A_k),
\end{aligned}$$

for all $A_1, \dots, A_k$ cylinder sets of the form

$$A_j = \{\omega_{[0,T]} \in C([0,T];\mathbb{R}^d) : \omega(t_1^j) \in B_1^j, \dots, \omega(t_{m_j}^j) \in B_{m_j}^j\},$$

for $B_1^j, \dots, B_{m_j}^j \in \mathcal{B}(\mathbb{R}^d)$, $t_1, \dots, t_{m_j}^j \in [0,T]$, $m_j \geq 1$. Noting that the collection of the sets $A_1 \times \ldots \times A_k$ is a semi-ring that generates[52] the σ-algebra $\mathcal{B}^k_{[0,T]}$, we can apply Caratheodory's extension theorem to conclude that Equation (4.4.14) holds. □

4.D.2 Proof of Proposition 4.4.24

Proof of Proposition 4.4.24. By Remark 4.3.7 there exists a Borel set $\Omega'_{[0,T]} \in \mathcal{B}_{[0,T]}$ and a sequence of partitions ρ^* with vanishing mesh such that for all $\omega_{[0,T]} \in \Omega'_{[0,T]}$,

$$S^{\mathbf{I}}(\omega^{\rho_*}_{[0,T]})_{[0,T]} \to S^{\mathbf{I}}(\omega_{[0,T]})_{[0,T]}, \quad |\rho_*| \to 0,$$

and $\mathbb{P}_{[0,T]}(\Omega'_{[0,T]}) = 1$. For every partition ρ_* the map

$$\omega_{[0,T]} \in \Omega'_{[0,T]} \mapsto S^{\mathbf{I}}(\omega^{\rho_*}_{[0,T]})_{[0,T]} \in \mathbb{R}^{|\mathbf{I}|},$$

is $\Omega'_{[0,T]} \cap \mathcal{B}_{[0,T]}$-measurable (by measurability of the sums and products of coordinate maps appearing in the discretized signature) and hence also

$$\omega_{[0,T]} \in \Omega'_{[0,T]} \mapsto S^{\mathbf{I}}(\omega_{[0,T]})_{[0,T]} \in \mathbb{R}^{|\mathbf{I}|},$$

is $\Omega'_{[0,T]} \cap \mathcal{B}_{[0,T]}$-measurable (by measurability of the pointwise limit of measurable functions). We can hence extend $S^{\mathbf{I}} : \Omega'_{[0,T]} \to \mathbb{R}^{|\mathbf{I}|}$ to a measurable map on the whole of $\Omega_{[0,T]} = C([0,T];\mathbb{R}^d)$. □

[52] Recall that for a set $I \subseteq \mathbb{R}_+$, the Borel σ-algebra $\mathcal{B}_I := \mathcal{B}(C(I;\mathbb{R}^d))$ (w.r.t. the topology induced by $\|\cdot\|_\infty$) can be equivalently defined by

$$\mathcal{B}_I = \sigma\left(\omega \in C(I;\mathbb{R}^d) : \omega(t_1) \in A_1, \dots, \omega(t_n) \in A_n,\ t_1, \dots, t_n \in I,\ A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}^d),\ n \geq 1\right).$$

Moreover, if $(E, \mathcal{B}(E))$ is a Borel measurable space and we equip E^k with the product topology, then

$$\mathcal{B}(E^k) = \mathcal{B}(E)^k := \sigma(A_1 \times \cdots \times A_k,\ A_1, \dots, A_k \in \mathcal{B}(E)) = \sigma(A_1 \times \cdots \times A_k,\ A_1, \dots, A_k \in \mathcal{G}),$$

where $\mathcal{G}$ is a generating collection for $\mathcal{B}(E)$, i.e. $\mathcal{B}(E) = \sigma(\mathcal{G})$.

4.D.3 Proof of Proposition 4.4.28

Proof of Proposition 4.4.28. We start by showing that strong mixing of $\{\mathbf{X}_t,\ t \geq 0\}$ with strong mixing coefficient $\alpha(s),\ s \in \mathbb{R}_+$ implies strong mixing of the progressive increment process $\{\mathbf{X}_t^T,\ t \geq 0\}$ where

$$\mathbf{X}_t^T := \mathbf{X}_{\lfloor t/T \rfloor T, t}, \quad t \geq 0,$$

with strong mixing coefficient $\alpha'(s) \leq \alpha(s - 2T),\ s \geq 2T$. This follows immediately from the definition of strong mixing and the fact that for $t \geq 0$,

$$\sigma(\mathbf{X}_u^T,\ u \leq t) \subseteq \sigma(\mathbf{X}_u,\ u \leq t),$$

and for $s \geq 2T$,

$$\sigma(\mathbf{X}_u^T,\ u \geq t + s) \subseteq \sigma(\mathbf{X}_u,\ u \geq \lfloor (t+s)/T \rfloor T) \subseteq \sigma(\mathbf{X}_u,\ u \geq t + (s - 2T)).$$

Next, we show that strong mixing of $\{\mathbf{X}_t^T,\ t \geq 0\}$ with strong mixing coefficient $\alpha'(s),\ s \in \mathbb{R}_+$ implies strong mixing of $\{\mathbb{X}^n,\ n \geq 1\}$ with strong mixing coefficient $\alpha''(n) \leq \alpha'((n-1)T),\ n \in \mathbb{N}$, and thus $\alpha''(n) \leq \alpha((n-3)T),\ n \geq 3$. Let

$$\mathcal{X}_a^b := \sigma(\mathbf{X}_u^T,\ u \in [a, b]),$$

for $a, b \in \overline{\mathbb{R}}_+$ with $a \leq b$ and let

$$\mathcal{S}_m^n := \sigma(\mathbb{X}^l,\ m \leq l \leq n),$$

for $m, n \in \overline{\mathbb{N}}$ with $m \leq n$. Then, by definition, for each $n \in \mathbb{N}$,

$$\mathcal{X}_{(n-1)T}^{nT} = \sigma(\mathbf{X}_{(n-1)T,(n-1)T+t},\ t \in [0, T]) = \mathcal{S}_n^n.$$

Thus for any $m, n \in \overline{\mathbb{N}}$ with $m \leq n$

$$\mathcal{S}_m^n = \mathcal{X}_{(m-1)T}^{nT}.$$

Letting $k \in \mathbb{N}$ and $A \in \mathcal{S}_{-\infty}^k$, $B \in \mathcal{S}_{k+n}^\infty$ we thus have

$$|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \leq \alpha'((n-1)T) \to 0, \quad k \to \infty.$$

Finally, note that by measurability[53] of $S^{\mathbf{I}} : C([0,T];\mathbb{R}) \to \mathbb{R}$, $\sigma(S^{\mathbf{I}}(\mathbb{X}^n)) \subseteq \sigma(\mathbb{X}^n)$ for all $n \geq 1$, which implies $\{S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]},\ n \geq 1\}$ is also strongly mixing with strong mixing coefficient at most $\alpha''(n),\ n \in \mathbb{N}$. □

[53] Here we abuse notation by writing $S^{\mathbf{I}}$ to denote the extension of $S^{\mathbf{I}} : \Omega'_{[0,T]} \to \mathbb{R}$ to the whole of $C([0,T];\mathbb{R})$.

4.D.4 Proof of Theorem 4.4.34

Proof of Theorem 4.4.34. Note that for any $n \in \mathbb{N}$, $0 \leq s_i \leq t_i$ for $i = 1, \dots, n$ and $t \geq 0$

$$(\mathbf{X}_{t+s_1,t+t_1}, \dots, \mathbf{X}_{t+s_n,t+t_n}) \overset{\mathcal{L}}{=} (\mathbf{X}_{s_1,t_1}, \dots, \mathbf{X}_{s_n,t_n}),$$

since both vectors are normally distributed with means

$$\mathbb{E}[\mathbf{X}_{t+s_i,t+t_i}] = \mathbb{E}[\mathbf{X}_{s_i,t_i}] = 0,$$

for $i = 1, \dots, n$ and covariances

$$\mathrm{Cov}(\mathbf{X}_{t+s_i,t+t_i}, \mathbf{X}_{t+s_j,t+t_j}) = \mathrm{Cov}(\mathbf{X}_{s_i,t_i}, \mathbf{X}_{s_j,t_j}) = C(|t_i - s_i|, |s_j - t_i|, |t_j - s_j|),$$

for $i, j = 1, \dots, n$. This implies $\{\mathbf{X}_t,\ t \geq 0\}$ has jointly stationary increments and hence, by Proposition 4.4.23, $\{\mathbb{X}^n,\ n \geq 1\}$ is stationary. By Remark 4.4.25 the sequence $\{S^{\mathbf{I}}(\mathbb{X}^n),\ n \geq 1\}$ is defined $\mathbb{P}$-a.s. and is stationary.

Note that by Example 4.4.11, $\mathbb{X}$ satisfies Assumption (Aα) for any $p \geq 2$. We can thus apply Theorem 4.4.4 when $\alpha > 1/2$ and Theorem 4.4.7 when $\alpha = 1/2$ to obtain an L^2 in-fill asymptotic result along a sequence of dyadic refinements. By the discussion at the start of Section 4.4.2, we can thus focus on showing the weak law of large numbers holds for $\{S^{\mathbf{I}}(\mathbb{X}^n),\ n \geq 1\}$.

To apply the weak law of large number for dependent random variables, note that:

1. $S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]} \in L^2$, by the in-fill asymptotic result.

2. To show $\mathrm{Cov}\left(S^{\mathbf{I}}(\mathbb{X}^n)_{[0,T]}, S^{\mathbf{I}}(\mathbb{X}^m)_{[0,T]}\right) \to 0$ as $|m - n| \to \infty$, by stationarity of $\{\mathbb{X}_n,\ n \geq 1\}$, it is sufficient to show that

$$\mathrm{Cov}\left(S^{k'}(\mathbb{X}^1)_{[0,T]}, S^{k'}(\mathbb{X}^n)_{[0,T]}\right) \to 0, \quad n \to \infty, \tag{4.D.1}$$

   for any $1 \leq k' \leq k = \max_{I \in \mathbf{I}} |I|$.

To show Equation (4.D.1) we wish to find a bound on $\mathrm{Cov}(S^{k'}(\mathbb{X}^1)_{[0,T]}, S^{k'}(\mathbb{X}^n)_{[0,T]})$ vanishing to zero as $n \to \infty$. Note that, by the in-fill asymptotic result, along the signature-defining sequence of dyadic refinements, for all $n \geq 1$ and $1 \leq k' \leq k$,

$$\mathrm{Cov}\left(S^{k'}(\mathbb{X}^{\rho,1})_{[0,T]}, S^{k'}(\mathbb{X}^{\rho,n})_{[0,T]}\right) \to \mathrm{Cov}\left(S^{k'}(\mathbb{X}^1)_{[0,T]}, S^{k'}(\mathbb{X}^n)_{[0,T]}\right), \quad |\rho| \to 0. \tag{4.D.2}$$

To bound $\mathrm{Cov}(S^{k'}(\mathbb{X}^1)_{[0,T]}, S^{k'}(\mathbb{X}^n)_{[0,T]})$ we can hence start by bounding the covariance between the discretized signatures $\mathrm{Cov}(S^{k'}(\mathbb{X}^{\rho,1})_{[0,T]}, S^{k'}(\mathbb{X}^{\rho,n})_{[0,T]})$ and then let $|\rho| \to 0$.

Let ρ be a dyadic partition of $[0,T]$ and note the form of the discretized signature

$$S(\mathbb{X}^\rho)_{[0,T]} = \bigotimes_{[u,v]\in\rho} \exp_\otimes \mathbf{X}_{u,v},$$

implies

$$S^{k'}(\mathbb{X}^\rho)_{[0,T]} = \sum_{i_\rho \in \mathcal{M}_\rho^{k'}} \bigotimes_{[u,v]\in\rho} \frac{\mathbf{X}_{u,v}^{\otimes i_{[u,v]}}}{i_{[u,v]}!},$$

where

$$\mathcal{M}_\rho^{k'} := \{i_\rho := \{i_{[u,v]}\}_{[u,v]\in\rho} : \ 0 \le i_{[u,v]} \le k', \ [u,v]\in\rho, \ \sum_{[u,v]\in\rho} i_{[u,v]} = k'\},$$

is the set of multindices over ρ with sum of components k'. Note that we can rewrite this as

$$\begin{aligned} S^{k'}(\mathbb{X}^\rho)_{[0,T]} &= \sum_{j=1}^{k'} \sum_{\substack{i_1,\dots,i_j\ge 1 \\ i_1+\cdots+i_j=k'}} \sum_{[u_1,v_1]<\cdots<[u_j,v_j]\in\rho} \cdots \sum \bigotimes_{l=1}^{j} \frac{\mathbf{X}_{u_l,v_l}^{\otimes i_l}}{i_l!} \\ &= \sum_{j=1}^{k'} \sum_{\substack{i_1,\dots,i_j\ge 1 \\ i_1+\cdots+i_j=k'}} \sum_{[u_1,v_1]<\cdots<[u_j,v_j]\in\rho} \cdots \sum \left[\prod_{l=1}^{j} \frac{1}{i_l!}\right] \bigotimes_{x=1}^{k'} \mathbf{X}_{u_{l_x},v_{l_x}}, \end{aligned}$$

where we first split the sum over the number of non-zero $i_{[u,v]}$ for each $i_\rho \in \mathcal{M}_\rho^{k'}$ and then rewrite the tensor product by introducing $(l_1,\cdots,l_{k'}) = (1,\dots,1,2,\dots,2,\dots,j,\dots,j)$ where the index 1 is repeated i_1 times, the index 2 is repeated i_2 times, and so on. We introduce the notation $[u_1,v_1] <_{m'} [u_2,v_2]$ denoting at least m' intervals between $[u_1,v_1]$ and $[u_2,v_2]$ in ρ. When $m'=0$ we equivalently write $<$ or $<_0$ to denote the interval $[u_2,v_2]$ being after $[u_1,v_1]$ in ρ.

We then group the above summations over the intervals $[u_1,v_1],\dots,[u_j,v_j]$ over all possible combinations of time intervals with at least one pair less than m steps away in the partition ρ. To do so, introduce the set[54] variable $\mathcal{I}\oplus_0\mathcal{J} \in \mathcal{S}_{j',j,m}$ where $\mathcal{I}\in\{1\}\times\{0,1\}^{j-1}$ is such that $|\mathcal{I}| = \mathcal{I}_1+\ldots+\mathcal{I}_j = j'$ and $\mathcal{J}\in\{1,\dots,m\}^{j-j'}$ where m is such that Assumption (Aθ) holds. We can then recursively define for $l=1,\dots,j$,

$$[u_l,v_l](\mathcal{I}\oplus_0\mathcal{J}) = \begin{cases} \text{interval } \mathcal{J}_l \text{ steps after } [u_{l-1},v_{l-1}](\mathcal{I}\oplus_0\mathcal{J}) & \text{if } \mathcal{I}_l = 0, \\ [u_{|\mathcal{I}_{1:l}|}, v_{|\mathcal{I}_{1:l}|}] & \text{if } \mathcal{I}_l = 1, \end{cases}$$

where $|\mathcal{I}_{1:l}| = \mathcal{I}_1+\ldots+\mathcal{I}_l$, and write

$$S^{k'}(\mathbb{X}^\rho)_{[0,T]} = \sum_{j=1}^{k'} \sum_{\substack{i_1,\dots,i_j\ge 1 \\ i_1+\cdots+i_j=k'}} \sum_{j'=1}^{j} \sum_{\mathcal{I}\oplus_0\mathcal{J}\in\mathcal{S}_{j',j,m}} \sum_{[u_1,v_1]<_{m_1}\cdots<_{m_{j'-1}}[u_{j'},v_{j'}]\in\rho} \cdots \sum \left[\prod_{l=1}^{j} \frac{1}{i_l!}\right] \bigotimes_{x=1}^{k'} \mathbf{X}_{[u_{l_x},v_{l_x}](\mathcal{I}\oplus_0\mathcal{J})},$$

[54] Here $\mathcal{I}\oplus_0\mathcal{J}$ denotes pairing elements of $\mathcal{J}$ to the elements of $\mathcal{I}$ equal to 0.

where for each $l' = 1, \dots, j'-1$ we have $[u_{l'}, v_{l'}] <_{m_{l'}} [u_{l'+1}, v_{l'+1}]$, i.e. there are at least $m_{l'}$ intervals between $[u_{l'}, v_{l'}]$ and $[u_{l'+1}, v_{l'+1}]$ in ρ, where

$$m_{l'} = m_{l'}(\mathcal{I} \oplus_0 \mathcal{J}) = \sum_{l=1}^{j} \mathcal{J}_l \mathbb{1}_{\{|\mathcal{I}_{1:l}|=l', \mathcal{I}_l=0\}}.$$

The structure of the tensor products in the summation is thus

$$\mathbf{X}_{u_1,v_1}^{\otimes i_1} \otimes \mathbf{X}_{u_1^2,v_1^2}^{\otimes i_2} \otimes \cdots \otimes \mathbf{X}_{u_1^{n_1},v_1^{n_1}}^{\otimes i_{n_1}} \otimes \mathbf{X}_{u_2,v_2}^{\otimes i_{n_1+1}} \otimes \mathbf{X}_{u_2^2,v_2^2}^{\otimes i_{n_1+2}} \otimes \cdots \otimes \mathbf{X}_{u_2^{n_2},v_2^{n_2}}^{\otimes i_{n_1+n_2}} \otimes \cdots,$$

where $[u_1, v_1], [u_1^2, v_1^2], \cdots, [u_1^{n_1}, v_1^{n_1}]$ are all less than m intervals apart and $[u_2, v_2]$ is at least m intervals after $[u_1^{n_1}, v_1^{n_1}]$ (and so on)[55].

Let $n \geq 3$ and define $\rho_n = (n-1)T + \rho$. Then, for each word $I = (w_1, \dots, w_{k'})$ of length k', we can write

$$\begin{aligned}
&\mathrm{Cov}\left(S^I(\mathbb{X}^{\rho,1})_{[0,T]}, S^I(\mathbb{X}^{\rho,n})_{[0,T]}\right) \\
&= \sum_{j_1,j_2=1}^{k'} \sum_{\substack{i_1,\dots,i_{j_1}\geq 1 \\ i_1+\cdots+i_{j_1}=k'}} \sum_{\substack{e_1,\dots,e_{j_2}\geq 1 \\ e_1+\cdots+e_{j_2}=k'}} \sum_{j_1'=1}^{j_1} \sum_{j_2'=1}^{j_2} \sum_{\substack{\mathcal{I}_1\oplus_0\mathcal{J}_1\in S_{j_1',j_1,m} \\ \mathcal{I}_2\oplus_0\mathcal{J}_2\in S_{j_2',j_2,m}}} \\
&\qquad \sum_{[u_1,v_1]<_{m_1^1}\cdots<_{m_{j_1'-1}^1}[u_{j_1'},v_{j_1'}]\in\rho} \cdots \sum_{[s_1,t_1]<_{m_1^2}\cdots<_{m_{j_2'-1}^2}[s_{j_2'},t_{j_2'}]\in\rho_n} \left[\prod_{l=1}^{j_1}\frac{1}{i_l!}\right]\left[\prod_{l=1}^{j_2}\frac{1}{e_l!}\right] \mathrm{Cov}\left(\prod_{x=1}^{k'} X^{(w_x)}_{[u_{l_x^1},v_{l_x^1}](\mathcal{I}_1\oplus_0\mathcal{J}_1)}, \prod_{x=1}^{k'} X^{(w_x)}_{[s_{l_x^2},t_{l_x^2}](\mathcal{I}_2\oplus_0\mathcal{J}_2)}\right) \\
&= \sum_{j_1,j_2=1}^{k'} \sum_{\substack{i_1,\dots,i_{j_1}\geq 1 \\ i_1+\cdots+i_{j_1}=k'}} \sum_{\substack{e_1,\dots,e_{j_2}\geq 1 \\ e_1+\cdots+e_{j_2}=k'}} \sum_{j_1'=1}^{j_1} \sum_{j_2'=1}^{j_2} \sum_{\substack{\mathcal{I}_1\oplus_0\mathcal{J}_1\in S_{j_1',j_1,m} \\ \mathcal{I}_2\oplus_0\mathcal{J}_2\in S_{j_2',j_2,m}}} \sum_{[u_1,v_1]<_{m_1^1}\cdots<_{m_{j_1'-1}^1}[u_{j_1'},v_{j_1'}]\in\rho} \cdots \sum_{[s_1,t_1]<_{m_1^2}\cdots<_{m_{j_2'-1}^2}[s_{j_2'},t_{j_2'}]\in\rho_n} \left[\prod_{l=1}^{j_1}\frac{1}{i_l!}\right]\left[\prod_{l=1}^{j_2}\frac{1}{e_l!}\right] \\
&\qquad \sum_{p\in MP^2_{(2,k')}} \prod_{\substack{\{(\delta_1,x_1),\\(\delta_2,x_2)\}\in p}} \mathrm{Cov}\left(X^{(w_{x_1})\delta_1}_{[u_{l_{x_1}^1},v_{l_{x_1}^1}](\mathcal{I}_1\oplus_0\mathcal{J}_1)} X^{(w_{x_1})(1-\delta_1)}_{[s_{l_{x_1}^2},t_{l_{x_1}^2}](\mathcal{I}_2\oplus_0\mathcal{J}_2)}, X^{(w_{x_2})\delta_2}_{[u_{l_{x_2}^1},v_{l_{x_2}^1}](\mathcal{I}_1\oplus_0\mathcal{J}_1)} X^{(w_{x_2})(1-\delta_2)}_{[s_{l_{x_2}^2},t_{l_{x_2}^2}](\mathcal{I}_2\oplus_0\mathcal{J}_2)}\right),
\end{aligned}$$

by using the fact that, for two collections of mean-zero normal random variables $(Z_{0,1}, \dots, Z_{0,k'})$

[55] To help intuitive understanding consider the case where $k' = 2$ and $m = 1$, then the above expression reduces to

$$S^2(\mathbb{X}^\rho)_{[0,T]} = \sum_{[u_1,v_1]\in\rho} \frac{\mathbf{X}_{u_1,v_1}^{\otimes 2}}{2!} + \sum_{[u_1,v_1]\in\rho} \mathbf{X}_{u_1,v_1} \otimes \mathbf{X}_{v_1,w_1} + \sum_{[u_1,v_1]\in\rho} \sum_{\substack{[u_2,v_2]\in\rho \\ [u_1,v_1]<_1[u_2,v_2]}} \mathbf{X}_{u_1,v_1} \otimes \mathbf{X}_{u_2,v_2},$$

where $[v_1, w_1]$ is the interval right-contiguous to $[u_1, v_1]$ in ρ. If $k' = 2$ and $m = 2$, instead

$$\begin{aligned}
S^2(\mathbb{X}^\rho)_{[0,T]} = &\sum_{[u_1,v_1]\in\rho} \frac{\mathbf{X}_{u_1,v_1}^{\otimes 2}}{2!} + \sum_{[u_1,v_1]\in\rho} \mathbf{X}_{u_1,v_1} \otimes \mathbf{X}_{v_1,w_1} + \sum_{[u_1,v_1]\in\rho} \mathbf{X}_{u_1,v_1} \otimes \mathbf{X}_{w_1,z_1} \\
&+ \sum_{[u_1,v_1]\in\rho} \sum_{\substack{[u_2,v_2]\in\rho \\ [u_1,v_1]<_2[u_2,v_2]}} \mathbf{X}_{u_1,v_1} \otimes \mathbf{X}_{u_2,v_2},
\end{aligned}$$

where $[u_1, v_1], [v_1, w_1], [w_1, z_1]$ are consecutive intervals in ρ.

and $(Z_{1,1}, \dots, Z_{1,k'})$ we can apply Isserlis' Theorem to show

$$\begin{aligned}
&\operatorname{Cov}\left(Z_{0,1}\cdots Z_{0,k'}, Z_{1,1}\cdots Z_{1,k'}\right) \\
&\quad= \mathbb{E}\left[Z_{0,1}\cdots Z_{0,k'}Z_{1,1}\cdots Z_{1,k'}\right] - \mathbb{E}\left[Z_{0,1}\cdots Z_{0,k'}\right]\mathbb{E}\left[Z_{1,1}\cdots Z_{1,k'}\right] \\
&\quad= \sum_{p\in P^2_{(2,k')}} \prod_{\substack{\{(i_1,i_2),\\(j_1,j_2)\}\in p}} \mathbb{E}\left[Z_{i_1,i_2}Z_{j_1,j_2}\right] - \left(\sum_{q\in P^2_{k'}}\prod_{\{i,j\}\in q}\mathbb{E}\left[Z_{0,i}Z_{0,j}\right]\right)\left(\sum_{r\in P^2_{k'}}\prod_{\{i,j\}\in r}\mathbb{E}\left[Z_{1,i}Z_{1,j}\right]\right) \\
&\quad= \sum_{p\in MP^2_{(2,k')}} \prod_{\{(i_1,i_2),(j_1,j_2)\}\in p} \mathbb{E}\left[Z_{i_1,i_2}Z_{j_1,j_2}\right] \\
&\quad= \sum_{p\in MP^2_{(2,k')}} \prod_{\{(i_1,i_2),(j_1,j_2)\}\in p} \operatorname{Cov}\left(Z_{i_1,i_2}, Z_{j_1,j_2}\right),
\end{aligned}$$

where $P^2_{(2,k')}$ denotes the set of all the pairings of $\{0,1\}\times\{1,\dots,k'\}$, $P^2_{k'}$ denotes the set of all the pairings of $\{1,\dots,k'\}$ and $MP^2_{(2,k')}$ denotes the set of all the pairings of $\{0,1\}\times\{1,\dots,k'\}$ that contain at least one "mixed" pair, i.e. for all $p\in MP^2_{(2,k')}$ there exist $\{(i_1,i_2),(j_1,j_2)\}\in p$ such that $i_1\neq j_1$.

Note that each $[u_{l'}, v_{l'}]$ for $l' = 1,\dots,j'_1$ appears in at least one covariance term. We proceed by cases:

- If $[u_{l'}, v_{l'}]$ appears in a pair with an interval $[u_*, v_*]$ such that $[u_{l'-1}, v_{l'-1}] < [u_*, v_*] < [u_{l'+1}, v_{l'+1}]$ then
$$|\operatorname{Cov}(X^{(w)}_{[u_{l'},v_{l'}]}, X^{(q)}_{[u_*,v_*]})| \lesssim |v_{l'} - u_{l'}|,$$
by applying Assumption (Aα) with $\alpha \geq 1/2$ and the fact that ρ is dyadic (and hence uniform).

- If $[u_{l'}, v_{l'}]$ appears in a pair with an interval $[u_*, v_*]$ such that $[u_*, v_*] \leq [u_{l'-1}, v_{l'-1}]$ then we have $|u_{l'} - v_*| \geq m/2(|v_* - u_*| + |v_{l'} - u_{l'}|)$ and
$$|\operatorname{Cov}(X^{(w)}_{[u_{l'},v_{l'}]}, X^{(q)}_{[u_*,v_*]})| \lesssim \theta(|u_{l'} - v_*|)|v_* - u_*||v_{l'} - u_{l'}| \lesssim \theta(|u_{l'} - v_{l'-1}|)|v_* - u_*||v_{l'} - u_{l'}|,$$
by applying Assumption (Aθ).

- Similarly, if $[u_{l'}, v_{l'}]$ appears in a pair with an interval $[u_*, v_*]$ such that $[u_{l'+1}, v_{l'+1}] \leq [u_*, v_*]$ then we have $|u_* - v_{l'}| \geq m/2(|v_{l'} - u_{l'}| + |v_* - u_*|)$ and
$$|\operatorname{Cov}(X^{(w)}_{[u_{l'},v_{l'}]}, X^{(q)}_{[u_*,v_*]})| \lesssim \theta(|u_* - v_{l'}|)|v_{l'} - u_{l'}||v_* - u_*| \lesssim \theta(|u_{l'+1} - v_{l'}|)|v_{l'} - u_{l'}||v_* - u_*|.$$

- If $[u_{l'}, v_{l'}]$ appears in a pair with an interval $[s_*, t_*] \in \rho_n$, then
$$|\operatorname{Cov}(X^{(w)}_{[u_{l'},v_{l'}]}, X^{(q)}_{[s_*,t_*]})| \lesssim \theta((n-2)T)|v_{l'} - u_{l'}||t_* - s_*|,$$
by applying Assumption (Aθ)

As similar reasoning applies to each $[s_{l'}, t_{l'}]$ for $l' = 1, \dots, j_2'$. Noting that at least one pairing is mixed across ρ and ρ_n we can hence bound

$$
\begin{aligned}
&|\mathrm{Cov}\left(S^I(\mathbb{X}^{\rho,1})_{[0,T]}, S^I(\mathbb{X}^{\rho,n})_{[0,T]}\right)| \\
&\lesssim \sum_{j_1,j_2=1}^{k'} \sum_{\substack{i_1,\dots,i_{j_1}\geq 1 \\ i_1+\cdots+i_{j_1}=k'}} \sum_{\substack{e_1,\dots,e_{j_2}\geq 1 \\ e_1+\cdots+e_{j_2}=k'}} \sum_{j_1'=1}^{j_1}\sum_{j_2'=1}^{j_2} \sum_{\substack{\mathcal{I}_1\oplus_0\mathcal{J}_1\in S_{j_1',j_1,m} \\ \mathcal{I}_2\oplus_0\mathcal{J}_2\in S_{j_2',j_2,m}}} \sum_{[u_1,v_1]<_{m_1^1}\cdots<_{m_{j_1'-1}^1}} \cdots \sum_{[s_1,t_1]<_{m_1^2}\cdots<_{m_{j_2'-1}^2}} \left[\prod_{l=1}^{j_1}\frac{1}{i_l!}\right]\left[\prod_{l=1}^{j_2}\frac{1}{e_l!}\right] \\
&\qquad \sum_{p\in MP^2_{(2,k')}} \theta((n-2)T)|v_1-u_1|\theta(|u_2-v_1|)|v_2-u_2|\cdots\theta(|u_{j_1'}-v_{j_1'-1}|)|v_{j_1'}-u_{j_1'}| \\
&\qquad\qquad \times |t_1-s_1|\theta(|s_2-t_1|)|t_2-s_2|\cdots\theta(|s_{j_2'}-t_{j_2'-1}|)|t_{j_2'}-s_{j_2'}| \\
&\lesssim \sum_{j_1,j_2=1}^{k'} \sum_{\substack{i_1,\dots,i_{j_1}\geq 1 \\ i_1+\cdots+i_{j_1}=k'}} \sum_{\substack{e_1,\dots,e_{j_2}\geq 1 \\ e_1+\cdots+e_{j_2}=k'}} \sum_{j_1'=1}^{j_1}\sum_{j_2'=1}^{j_2} \theta((n-2)T) \\
&\qquad \times \left(\sum_{[u_1,v_1]\in\rho}|v_1-u_1|\right)\left(\sum_{[u_2,v_2]\in\rho}\theta(|u_2|)|v_2-u_2|\right)\cdots\left(\sum_{[u_{j_1'},v_{j_1'}]\in\rho}\theta(|u_{j_1'}|)|v_{j_1'}-u_{j_1'}|\right) \\
&\qquad\qquad \times \left(\sum_{[s_1,t_1]\in\rho_n}|t_1-s_1|\right)\left(\sum_{[s_2,t_2]\in\rho_n}\theta(|s_2|)t_2-s_2|\right)\cdots\left(\sum_{[s_{j_2'},t_{j_2'}]\in\rho_n}\theta(|s_{j_2'}|)|t_{j_2'}-s_{j_2'}|\right) \\
&\to \theta((n-2)T)\left(\sum_{j_1,j_2=1}^{k'} \sum_{\substack{i_1,\dots,i_{j_1}\geq 1 \\ i_1+\cdots+i_{j_1}=k'}} \sum_{\substack{e_1,\dots,e_{j_2}\geq 1 \\ e_1+\cdots+e_{j_2}=k'}} \sum_{j_1'=1}^{j_1}\sum_{j_2'=1}^{j_2} T^2\left(\int_0^T\theta(t)\mathrm{d}t\right)^{j_1'+j_2'-2}\right), \quad |\rho|\to 0.
\end{aligned}
$$

Combining this bound with Equation (4.D.2) we can conclude that, for all $n \geq 3$,

$$
|\mathrm{Cov}\left(S^I(\mathbb{X}^1)_{[0,T]}, S^I(\mathbb{X}^n)_{[0,T]}\right)| \lesssim \theta((n-2)T) \to 0, \quad n\to\infty,
$$

i.e. we have shown Equation (4.D.1). Hence, we can apply the weak law of large numbers of dependent random variables and the in-fill asymptotics to obtain the desired consistency result.

□

4.E Martingale correction: proofs

4.E.1 Proof of Lemma 4.6.2

Proof of Lemma 4.6.2. Note that both $\hat{c}^*_{\pi,1}$ and $\hat{c}^*_{\pi,2}$ are ratio estimators of the form

$$
\hat{c}^*_{\pi,j} = \frac{\bar{Y}^\pi_j}{\bar{Z}^\pi_j} = \frac{\sum_{n=1}^N Y^\pi_{j,n}}{\sum_{n=1}^N Z^\pi_{j,n}},
$$

for random variables $Y^\pi_{j,1}, \dots, Y^\pi_{j,N}$ and $Z^\pi_{j,1}, \dots, Z^\pi_{j,N}$ with $j = 1, 2$ given by i.i.d. copies of

$$Y^\pi_1 = S^I(\mathbb{X}^\pi)_{[0,T]} S^I_c(\mathbb{X}^\pi)_{[0,T]}, \quad Z^\pi_1 = S^I_c(\mathbb{X}^\pi)^2_{[0,T]},$$

and

$$Y^\pi_2 = \sum_{J \in I \sqcup I} S^J((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]} - \frac{1}{2} \sum_{J \in I \sqcup I_{-2*((i_{k-1}, i_k))}} S^J((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]},$$
$$Z^\pi_2 = \sum_{J \in I_{-1} \sqcup I_{-1}} S^{J*((i_k, i_k))}((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]}.$$

By the standard theory of ratio estimators, these are biased estimators for

$$c^*_{\pi,j} = \frac{\mathbb{E}[Y^\pi_j]}{\mathbb{E}[Z^\pi_j]},$$

and applying a first-order Taylor expansion[56] we can approximate the mean squared error as

$$\begin{aligned}\mathbb{E}[(\hat{c}^*_{\pi,j} - c^*_{\pi,j})^2] &\approx \frac{1}{\mathbb{E}[Z^\pi_j]^2} \mathrm{Var}(\bar{Y}^\pi_j) + \frac{\mathbb{E}[Y^\pi_j]^2}{\mathbb{E}[Z^\pi_j]^4} \mathrm{Var}(\bar{Z}^\pi_j) - 2 \frac{\mathbb{E}[Y^\pi_j]}{\mathbb{E}[Z^\pi_j]^3} \mathrm{Cov}(\bar{Y}^\pi_j, \bar{Z}^\pi_j) \\ &\approx \frac{1}{N} \left(\frac{1}{\mathbb{E}[Z^\pi_j]^2} \mathrm{Var}(Y^\pi_j) + \frac{\mathbb{E}[Y^\pi_j]^2}{\mathbb{E}[Z^\pi_j]^4} \mathrm{Var}(Z^\pi_j) - 2 \frac{\mathbb{E}[Y^\pi_j]}{\mathbb{E}[Z^\pi_j]^3} \mathrm{Cov}(Y^\pi_j, Z^\pi_j) \right).\end{aligned}$$

Note that

$$c^*_{\pi,1} = \frac{\mathbb{E}[S^I(\mathbb{X}^\pi)_{[0,T]} S^I_c(\mathbb{X}^\pi)_{[0,T]}]}{\mathbb{E}[S^I_c(\mathbb{X}^\pi)^2_{[0,T]}]} = \frac{\mathrm{Cov}(S^I(\mathbb{X}^\pi)_{[0,T]}, S^I_c(\mathbb{X}^\pi)_{[0,T]})}{\mathrm{Var}(S^I_c(\mathbb{X}^\pi)^2_{[0,T]})} = c^*_\pi,$$

and, as $|\pi| \to 0$,

$$\begin{aligned}c^*_{\pi,2} &= \frac{\sum_{J \in I \sqcup I} \mathbb{E}[S^J((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]}] - \frac{1}{2} \sum_{J \in I \sqcup I_{-2*((i_{k-1}, i_k))}} \mathbb{E}[S^J((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]}]}{\sum_{J \in I_{-1} \sqcup I_{-1}} \mathbb{E}[S^{J*((i_k, i_k))}((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,T]}]} \\ &\overset{(4.E.6)}{\approx} \frac{\sum_{J \in I \sqcup I} \mathbb{E}[S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^\pi)_{[0,T]}] - \frac{1}{2} \sum_{J \in I \sqcup I_{-2*((i_{k-1}, i_k))}} \mathbb{E}[S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^\pi)_{[0,T]}]}{\sum_{J \in I_{-1} \sqcup I_{-1}} \mathbb{E}[S^{J*((i_k, i_k))}((\mathbb{X}, \langle \mathbb{X} \rangle)^\pi)_{[0,T]}]}\end{aligned}$$

[56] On the set $|\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]| < |\mathbb{E}[Z^\pi_j]|$,

$$\begin{aligned}\hat{c}_{\pi,j} &= c^*_{\pi,j} \left(1 + \frac{\bar{Y}^\pi_j - \mathbb{E}[Y^\pi_j]}{\mathbb{E}[Y^\pi_j]}\right) \left(1 + \frac{\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]}{\mathbb{E}[Z^\pi_j]}\right)^{-1} \\ &= c^*_{\pi,j} \left(1 + \frac{\bar{Y}^\pi_j - \mathbb{E}[Y^\pi_j]}{\mathbb{E}[Y^\pi_j]}\right) \left(1 - \frac{\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]}{\mathbb{E}[Z^\pi_j]} + \mathcal{O}\left(\left(\frac{\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]}{\mathbb{E}[Z^\pi_j]}\right)^2\right)\right) \\ &= c^*_{\pi,j} + \frac{1}{\mathbb{E}[Z^\pi_j]}(\bar{Y}^\pi_j - \mathbb{E}[Y^\pi_j]) - \frac{\mathbb{E}[Y^\pi_j]}{\mathbb{E}[Z^\pi_j]^2}(\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]) + \mathcal{O}\left(\left(\frac{\bar{Y}^\pi_j - \mathbb{E}[Y^\pi_j]}{\mathbb{E}[Y^\pi_j]}\right)\left(\frac{\bar{Z}^\pi_j - \mathbb{E}[Z^\pi_j]}{\mathbb{E}[Z^\pi_j]}\right)\right).\end{aligned}$$

For the approximation to be rigorous one needs to assume the probability of the set $\{\omega : |\bar{Z}^\pi_j(\omega) - \mathbb{E}[Z^\pi_j]| \geq |\mathbb{E}[Z^\pi_j]|\}$ approaches 0 faster than the speed at which the MSE conditional on this set explodes as $N \to \infty$.

$$\overset{(4.E.3)}{\approx} \frac{\sum_{J\in I\sqcup I}\mathbb{E}[S^J((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]}] - \frac{1}{2}\sum_{J\in I\sqcup I_{-2*((i_{k-1},i_k))}}\mathbb{E}[S^J((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]}]}{\sum_{J\in I_{-1}\sqcup I_{-1}}\mathbb{E}[S^{J*((i_k,i_k))}((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]}]}$$
$$= \frac{\mathrm{Cov}(S^I(\mathbb{X})_{[0,T]}, S^I_c(\mathbb{X})_{[0,T]})}{\mathrm{Var}(S^I_c(\mathbb{X})^2_{[0,T]})}$$
$$\overset{(4.E.1),(4.E.2)}{\approx} \frac{\mathrm{Cov}(S^I(\mathbb{X}^\pi)_{[0,T]}, S^I_c(\mathbb{X}^\pi)_{[0,T]})}{\mathrm{Var}(S^I_c(\mathbb{X}^\pi)^2_{[0,T]})} = c^*_\pi.$$

We refer to Lemma 4.E.2 for the rigorous justification of the approximations Equation (4.E.1), Equation (4.E.2), Equation (4.E.3), Equation (4.E.6). Moreover, as $|\pi| \to 0$, by Equation (4.E.1) and Equation (4.E.2),

$$Y^\pi_1 \overset{L^2}{\to} S^I(\mathbb{X})_{[0,T]}S^I_c(\mathbb{X})_{[0,T]} =: Y_1, \quad Z^\pi_1 \overset{L^2}{\to} S^I_c(\mathbb{X})^2_{[0,T]} =: Z_1,$$

and by combining Equation (4.E.3) and Equation (4.E.6),

$$Y^\pi_2 \overset{L^2}{\to} \sum_{J\in I\sqcup I} S^J((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]} - \frac{1}{2}\sum_{J\in I\sqcup I_{-2*((i_{k-1},i_k))}} S^J((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]} =: Y_2,$$
$$Z^\pi_2 \overset{L^2}{\to} \sum_{J\in I_{-1}\sqcup I_{-1}} S^{J*((i_k,i_k))}((\mathbb{X},\langle\mathbb{X}\rangle))_{[0,T]} =: Z_2.$$

Note that $Y_1 = Y_2 =: Y$ but $Z_1 \neq Z_2$ even though $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] =: \mu_Z$. When $|\pi|$ is small we can thus approximate the MSEs of $\hat{c}^*_{\pi,1}$ and $\hat{c}^*_{\pi,2}$ with respect to c^*_π as

$$\mathbb{E}[(\hat{c}^*_{\pi,1} - c^*_\pi)^2] \approx \frac{1}{N}\left(\frac{1}{\mu^2_Z}\sigma^2_Y + \frac{\mu^2_Y}{\mu^4_Z}\mathrm{Var}(Z_1) - 2\frac{\mu_Y}{\mu^3_Z}\mathrm{Cov}(Y,Z_1)\right),$$

and

$$\mathbb{E}[(\hat{c}^*_{\pi,2} - c^*_\pi)^2] \approx \frac{1}{N}\left(\frac{1}{\mu^2_Z}\sigma^2_Y + \frac{\mu^2_Y}{\mu^4_Z}\mathrm{Var}(Z_2) - 2\frac{\mu_Y}{\mu^3_Z}\mathrm{Cov}(Y,Z_2)\right),$$

which differ by

$$\mathbb{E}[(\hat{c}^*_{\pi,2} - c^*_\pi)^2] - \mathbb{E}[(\hat{c}^*_{\pi,1} - c^*_\pi)^2] \approx \frac{1}{N}\frac{\mu_Y}{\mu^3_Z}\left(\frac{\mu_Y}{\mu_Z}(\mathbb{E}[Z^2_2] - \mathbb{E}[Z^2_1]) - 2(\mathbb{E}[YZ_2] - \mathbb{E}[YZ_1])\right).$$

Remark 4.E.1. Note that we can "mix" the two estimators and form

$$\hat{c}^*_{\pi,2,1} = \frac{\bar{Y}^\pi_2}{\bar{Z}^\pi_1} \quad \text{and} \quad \hat{c}^*_{\pi,1,2} = \frac{\bar{Y}^\pi_1}{\bar{Z}^\pi_2}.$$

The discussion above ensures that, as $|\pi| \to 0$, $\hat{c}^*_{\pi,2,1}$ and $\hat{c}^*_{\pi,1,2}$ have the same MSEs as $\hat{c}^*_{\pi,1,1} = \hat{c}_{\pi,1}$ and $\hat{c}^*_{\pi,2,2} = \hat{c}^*_{\pi,2}$ respectively.

□

Lemma 4.E.2. *Assume that $\mathbb{X}$ satisfies Assumption (Aα.M) for some $\alpha \geq 1/2$ and $p = 4k$ where $k = |I|$. Assume that π is part of a sequence of refining partitions with mesh vanishing fast enough, i.e.*

$$\sum_{n\geq 1} |\pi_n|^{2\alpha - 1/2} < \infty.$$

Then the approximations Equation (4.E.1), Equation (4.E.2), Equation (4.E.3), Equation (4.E.6) hold as $|\pi| \to 0$.

Proof. By Example 4.4.10 $\mathbb{X}$ satsfies Assumption (Aα) with $\alpha \geq 1/2$ and $p = 4k$. Note that, since $\mathbb{X}$ is a martingale, Assumption (Aδ) holds trivially and hence we can set $\epsilon = 2\alpha - 1/2$.

We can apply Theorem 4.4.7 to deduce that

$$S^I(\mathbb{X}^\pi)_{[0,T]} \overset{L^4}{\to} S^I(\mathbb{X})_{[0,T]}, \quad |\pi| \to 0.$$

Moreover,

$$S_c^I(\mathbb{X}^\pi)_{[0,T]} \overset{L^4}{\to} S_c^I(\mathbb{X})_{[0,T]}, \quad |\pi| \to 0.$$

since

$$\begin{aligned}
&\|S_c^I(\mathbb{X})_{[0,T]} - S_c^I(\mathbb{X}^\pi)_{[0,T]}\|_{L^4} \\
&\overset{(i)}{\leq} \left\| S_c^I(\mathbb{X})_{[0,T]} - \sum_{[u,v]\in\pi} S^{I_{-1}}(\mathbb{X})_{[0,u]} X_{u,v}^{(i_k)} \right\|_{L^4} + \left\| \sum_{[u,v]\in\pi} (S^{I_{-1}}(\mathbb{X})_{[0,u]} - S^{I_{-1}}(\mathbb{X}^\pi)_{[0,u]}) X_{u,v}^{(i_k)} \right\|_{L^4} \\
&\overset{(ii)}{\leq} \left\| \int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]} \mathrm{d}X_s^{(i_k)} - \sum_{[u,v]\in\pi} S^{I_{-1}}(\mathbb{X})_{[0,u]} X_{u,v}^{(i_k)} \right\|_{L^4} \\
&\qquad + 2C_2 \left(\sum_{[u,v]\in\pi} \left\| S^{I_{-1}}(\mathbb{X})_{[0,u]} - S^{I_{-1}}(\mathbb{X}^\pi)_{[0,u]} \right\|_{L^{2k/(k-1)}}^2 \|X_{u,v}^{(i_k)}\|_{L^{2k}}^2 \right)^{1/2} \\
&\overset{(iii)}{\lesssim} \left\| \int_0^T S^{I_{-1}}(\mathbb{X})_{[0,s]} \mathrm{d}X_s^{(i_k)} - \sum_{[u,v]\in\pi} S^{I_{-1}}(\mathbb{X})_{[0,u]} X_{u,v}^{(i_k)} \right\|_{L^4} \\
&\qquad + 2C_2 \sup_{u\in\pi} \left\| S^{I_{-1}}(\mathbb{X})_{[0,u]} - S^{I_{-1}}(\mathbb{X}^\pi)_{[0,u]} \right\|_{L^{2k/(k-1)}} \left(\sum_{[u,v]\in\pi} |v-u|^{2\alpha} \right)^{1/2} \\
&\overset{(iv)}{\to} 0, \quad |\pi| \to 0,
\end{aligned}$$

by applying in (i) the triangle inequality, in (ii) Lemma 4.C.1 with the natural filtration of $\mathbb{X}$, in (iii) Theorem 4.4.7 to the word I_{-1} with $m = 2k/(k-1)$ and in (iv) the definition of the Itô integral and Remark 4.4.9.

Under these assumptions, we thus have

$$S^I(\mathbb{X}^\pi)_{[0,T]} S_c^I(\mathbb{X}^\pi)_{[0,T]} \overset{L^2}{\to} S^I(\mathbb{X})_{[0,T]} S_c^I(\mathbb{X})_{[0,T]}, \quad |\pi| \to 0, \tag{4.E.1}$$

and

$$S_c^I(\mathbb{X}^\pi)^2_{[0,T]} \overset{L^2}{\to} S_c^I(\mathbb{X})^2_{[0,T]}, \quad |\pi| \to 0. \tag{4.E.2}$$

Next, we consider the approximation given in Equation (4.E.3). Note that $\mathbb{X}$ satisfies Assumption (Aα) (and trivially Assumption (Aδ)) with $\alpha \geq 1/2$ and $p = 4k$ while $\langle \mathbb{X} \rangle$ satisfies Assumption (Aα) and Assumption (Aδ) both with exponent $2\alpha \geq 1$ and $p = 2k$. By using a slightly more general version[57] of Theorem 4.4.7 applied to the process $(\mathbb{X}, \langle \mathbb{X} \rangle)$, for any word $J \in I \sqcup\!\sqcup I$ or $J \in I \sqcup\!\sqcup I_{-2} * ((i_{k-1}, i_k))$ or $J = J' * ((i_k, i_k))$ with $J' \in I_{-1} \sqcup\!\sqcup I_{-1}$ we have

$$S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^\pi)_{[0,T]} \overset{L^2}{\to} S^J((\mathbb{X}, \langle \mathbb{X} \rangle))_{[0,T]}, \quad |\pi| \to 0, \tag{4.E.3}$$

and hence for fixed $N \geq 1$,

$$\hat{\psi}_J^{N,\pi}(T) \overset{L^2}{\to} \hat{\psi}_J^N(T), \quad |\pi| \to 0.$$

Finally, making Equation (4.E.6) rigorous is a bit more challenging. Let us start by considering the case where $\langle \mathbb{X} \rangle$ only appears in the outermost integral, i.e. $J = (j_1, \dots, j_{k'})$ where $j_1, \dots, j_{k'-1} \in \{1, \dots, d\}$ and $j_{k'} \in \{(1,1), \dots, (1,d), \dots, (d,d)\}$ for some $k' \in \{1, \dots, 2k-1\}$. Then for any $\tau \in \pi$

$$\begin{aligned}
&\left\| S^J((\mathbb{X}, \langle \mathbb{X} \rangle)^\pi)_{[0,\tau]} - S^J((\mathbb{X}^\pi, \langle \hat{\mathbb{X}} \rangle^\pi))_{[0,\tau]} \right\|_{L^{4k/(k'+1)}} \\
&\quad \overset{(i)}{=} \left\| \sum_{i=1}^{k'} \frac{1}{i!} \sum_{[u,v] \in \pi_{[0,\tau]}} S^{J_{-i}}(\mathbb{X}^\pi)_{[0,u]} X_{u,v}^{(j_{k'-i+1})} \cdots X_{u,v}^{(j_{k'-1})} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \langle \hat{X} \rangle_{u,v}^{\pi,(j_{k'})} \right) \right\|_{L^{4k/(k'+1)}} \\
&\quad \overset{(ii)}{\leq} \sum_{i=1}^{k'} \frac{1}{i!} \left\| \sum_{[u,v] \in \pi_{[0,\tau]}} S^{J_{-i}}(\mathbb{X}^\pi)_{[0,u]} X_{u,v}^{(j_{k'-i+1})} \cdots X_{u,v}^{(j_{k'-1})} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})} \right)^2 \right) \right\|_{L^{4k/(k'+1)}} \\
&\quad \overset{(iii)}{\leq} \left(\sum_{[u,v] \in \pi_{[0,\tau]}} \left\| S^{J_{-1}}(\mathbb{X}^\pi)_{[0,u]} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})} \right)^2 \right) \right\|^2_{L^{4k/(k'+1)}} \right)^{1/2} \\
&\qquad\qquad + \sum_{i=2}^{k'} \frac{1}{i!} \sum_{[u,v] \in \pi_{[0,\tau]}} \left\| S^{J_{-i}}(\mathbb{X}^\pi)_{[0,u]} X_{u,v}^{(j_{k'-i+1})} \cdots X_{u,v}^{(j_{k'-1})} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})} \right)^2 \right) \right\|_{L^{4k/(k'+1)}} \\
&\quad \overset{(iii)}{\leq} \left(\sum_{[u,v] \in \pi_{[0,\tau]}} \left\| S^{J_{-1}}(\mathbb{X}^\pi)_{[0,u]} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})} \right)^2 \right) \right\|^2_{L^{4k/(k'+1)}} \right)^{1/2} \\
&\qquad\qquad + \sum_{i=2}^{k'} \frac{1}{i!} \sum_{[u,v] \in \pi_{[0,\tau]}} \left\| S^{J_{-i}}(\mathbb{X}^\pi)_{[0,u]} X_{u,v}^{(j_{k'-i+1})} \cdots X_{u,v}^{(j_{k'-1})} \left(\langle X \rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})} \right)^2 \right) \right\|_{L^{4k/(k'+1)}} \\
&\quad \overset{(iv)}{\leq} \sup_{u \in \pi_{[0,\tau]}} \left\| S^{J_{-1}}(\mathbb{X}^\pi)_{[0,u]} \right\|_{L^{4k/(k'-1)}} \left(\sum_{[u,v] \in \pi_{[0,\tau]}} \left\| \langle \mathbf{X} \rangle_{u,v} - \mathbf{X}_{u,v}^{\otimes 2} \right\|^2_{L^{2k}} \right)^{1/2} \\
&\qquad\qquad + \sum_{i=2}^{k'} \frac{1}{i!} \sup_{u \in \pi_{[0,\tau]}} \left\| S^{J_{-i}}(\mathbb{X}^\pi)_{[0,u]} \right\|_{L^{4k/(k'-i)}} \sum_{[u,v] \in \pi_{[0,\tau]}} \| \mathbf{X}_{u,v} \|^{i-1}_{L^{4k}} \left\| \langle \mathbf{X} \rangle_{u,v} - \mathbf{X}_{u,v}^{\otimes 2} \right\|_{L^{2k}}
\end{aligned}$$

[57] To estimate the expected signature term corresponding to the word $I = (i_1, \dots, i_k)$ it is sufficient to require Assumption (Aα) and Assumption (Aδ) to be satisfied by $X^{(i_1)}, \dots, X^{(i_k)}$ with $p_1, \dots, p_k$ such that $p_1^{-1} + \cdots + p_k^{-1} \leq m^{-1}$. In Theorem 4.4.7 we considered the case where $p_1 = \cdots = p_k = p$ for clarity.

$$\overset{(v)}{\lesssim} |\pi|^{4\alpha-1}\sqrt{\tau} + |\pi|\tau \to 0, \quad |\pi| \to 0,$$

by using in (i) Lemma 4.4.6, in (ii) the triangle inequality and the definition of $\langle\hat{\mathbb{X}}\rangle^\pi$, in (iii) Lemma 4.C.1 with the natural filtration of $\mathbb{X}$ (with respect to which $\mathbb{X} - \langle\mathbb{X}\rangle$ is a martingale) for the $i = 1$ term and the triangle inequality for the $i = 2, \dots, j$ terms, in (iv) Hölder inequality and in (v) Remark 4.4.9 after Theorem 4.4.7 applied to $\mathbb{X}$ and J_{-i} with $p = 4k$, $m = 4k/(k'-i)$ and $|J_{-i}| = k' - i$ for $i = 1, \dots, k'$ and

$$\left\|\langle\mathbf{X}\rangle_{u,v} - \mathbf{X}_{u,v}^{\otimes 2}\right\|_{L^{2k}} \leq \|\langle\mathbf{X}\rangle_{u,v}\|_{L^{2k}} + \|\mathbf{X}_{u,v}\|_{L^{4k}}^2 \lesssim |v-u|^{2\alpha}, \tag{4.E.4}$$

by Assumption (Aα.M) with $\alpha \geq 1/2$ and $p = 4k$. We have thus shown

$$\sup_{\tau\in\pi} \left\|S^J((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,\tau]} - S^J((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,\tau]}\right\|_{L^{4k/(k'+1)}} \lesssim |\pi|. \tag{4.E.5}$$

This extends to any $J = (j_1, \dots, j_{k'}, \dots, j_{k'+m'})$ with $m' \geq 0$ and $1 \leq k', k'+m' \leq 2k-1$ where, as above, $j_{k'} \in \{(1,1), \dots, (1,d), \dots, (d,d)\}$ and $j_i \in \{1, \dots, d\}$ for all other $i \neq k'$. We proceed inductively on $m' = 0, \dots, 2k - k' - 1$ to show Equation (4.E.5) holds for any J of this form in the $L^{4k/(k'+m'+1)}$ norm. The case $m' = 0$ has been covered above. Assume Equation (4.E.5) holds for $J_{k'+m}$ with $m \in \{0, \dots, m'\}$ and $0 \leq m' \leq 2k - k' - 1$, then

$$\begin{aligned}
&\left\|S^{J_{k'+m'+1}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,\tau]} - S^{J_{k'+m'+1}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,\tau]}\right\|_{L^{4k/(k'+m'+2)}} \\
&\overset{(i)}{=} \Bigg\| \sum_{i=1}^{m'+1} \frac{1}{i!} \sum_{[u,v]\in\pi_{[0,\tau]}} \left[S^{J_{k'+m'+1-i}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,u]} - S^{J_{k'+m'+1-i}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,u]}\right] X_{u,v}^{(j_{k'+m'+2-i})} \cdots X_{u,v}^{(j_{k'+m'+1})} \\
&\quad + \sum_{i=m'+2}^{k'+m'+1} \frac{1}{i!} \sum_{[u,v]\in\pi_{[0,\tau]}} S^{J_{k'+m'+1-i}}(\mathbb{X}^\pi)_{[0,u]} \\
&\qquad\qquad \times X_{u,v}^{(j_{k'+m'+2-i})} \cdots \left(\langle X\rangle_{u,v}^{(j_{k'})} - \langle\hat{X}\rangle_{u,v}^{\pi,(j_{k'})}\right) \cdots X_{u,v}^{(j_{k'+m'+1})} \Bigg\|_{L^{4k/(k'+m'+2)}} \\
&\overset{(ii)}{\leq} \left\| \sum_{[u,v]\in\pi_{[0,\tau]}} \left[S^{J_{k'+m'}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,u]} - S^{J_{k'+m'}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,u]}\right] X_{u,v}^{(j_{k'+m'+1})} \right\|_{L^{4k/(k'+m'+2)}} \\
&\quad + \sum_{i=2}^{m'+1} \frac{1}{i!} \Bigg\| \sum_{[u,v]\in\pi_{[0,\tau]}} \left[S^{J_{k'+m'+1-i}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,u]} - S^{J_{k'+m'+1-i}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,u]}\right] \\
&\qquad\qquad \times X_{u,v}^{(j_{k'+m'+2-i})} \cdots X_{u,v}^{(j_{k'+m'+1})} \Bigg\|_{L^{4k/(k'+m'+2)}} \\
&\quad + \sum_{i=m'+2}^{k'+m'+1} \frac{1}{i!} \Bigg\| \sum_{[u,v]\in\pi_{[0,\tau]}} S^{J_{k'+m'+1-i}}(\mathbb{X}^\pi)_{[0,u]} \\
&\qquad\qquad \times X_{u,v}^{(j_{k'+m'+2-i})} \cdots \left(\langle X\rangle_{u,v}^{(j_{k'})} - \left(X_{u,v}^{(j_{k'})}\right)^2\right) \cdots X_{u,v}^{(j_{k'+m'+1})} \Bigg\|_{L^{4k/(k'+m'+2)}} \\
&\overset{(iii)}{\leq} \sup_{u\in\pi_{[0,\tau]}} \left\|S^{J_{k'+m'}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,u]} - S^{J_{k'+m'}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,u]}\right\|_{L^{4k/(k'+m'+1)}} \left(\sum_{[u,v]\in\pi_{[0,\tau]}} \|\mathbf{X}_{u,v}\|_{L^{4k}}^2\right)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=2}^{m'+1} \frac{1}{i!} \sup_{u\in\pi_{[0,\tau]}} \left\| S^{J_{k'+m'+1-i}}((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,u]} - S^{J_{k'+m'+1-i}}((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,u]} \right\|_{L^{4k/(k'+m'-i+2)}} \\
&\qquad\qquad \times \left(\sum_{[u,v]\in\pi_{[0,\tau]}} \|\mathbf{X}_{u,v}\|^i_{L^{4k}} \right) \\
&+ \sum_{i=m'+2}^{k'+m'+1} \frac{1}{i!} \sup_{u\in\pi_{[0,\tau]}} \left\| S^{J_{k'+m'+1-i}}(\mathbb{X}^\pi)_{[0,u]} \right\|_{L^{4k/(k'+m'+1-i)}} \left(\sum_{[u,v]\in\pi_{[0,\tau]}} \|\mathbf{X}_{u,v}\|^{i-1}_{L^{4k}} \|\langle\mathbf{X}\rangle_{u,v} - \mathbf{X}^{\otimes 2}_{u,v}\|_{L^{2k}} \right) \\
&\overset{(v)}{\lesssim} |\pi|\sqrt{\tau} + |\pi|\tau \to 0, |\pi| \to 0,
\end{aligned}
$$

where in (i) we have used Lemma 4.4.6, in (ii) triangle inequality, in (iii) Lemma 4.C.1 with the natural filtration of $\mathbb{X}$ for the $i = 1$ term, traingle inequality for the $i = 2, \dots, k' + m' + 1$ terms and Hölder inequality across all terms, in (iv) the inductive hypothesis for $J_{k'+m}$ with $m = 0, \dots, m'$ for the first two sups, Remark 4.4.9 after Theorem 4.4.7 applied to $\mathbb{X}$ and J_i with $p = 4k$, $m = 4k/i$ and $|J_i| = i$ for $i = 1, \dots, k' - 1$ for the third sup and bounds Equation (4.E.4) and Assumption (Aα) with $p = 4k$ and $\alpha \geq 1/2$ for the summations over $\pi_{[0,\tau]}$.

We have thus shown that for any word $J \in I \sqcup\!\sqcup I$ or $J \in I \sqcup\!\sqcup I_{-2} * ((i_{k-1}, i_k))$ or $J = J' * ((i_k, i_k))$ for $J' \in I_{-1} \sqcup\!\sqcup I_{-1}$ we have

$$
S^J((\mathbb{X}, \langle\mathbb{X}\rangle)^\pi)_{[0,T]} - S^J((\mathbb{X}^\pi, \langle\hat{\mathbb{X}}\rangle^\pi))_{[0,T]} \overset{L^2}{\to} 0, \quad |\pi| \to 0, \tag{4.E.6}
$$

and hence for fixed $N \geq 1$,

$$
\hat{\psi}^{N,\pi}_J(T) - \hat{\psi}^{N,\pi,\prime}_J(T) \overset{L^2}{\to} 0, \quad |\pi| \to 0.
$$

□

4.F Controlled linear regression

In this section, we introduce the notion of controlled linear regression. The main rationale is to exploit as much information as possible in the training phase to make the coefficient estimators as precise as possible. We start by considering the following linear model

$$
y = \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon,
$$

and assume we observe the training data $\{(y_n, x_{n,1}, \dots, x_{n,p}),\ n = 1, \dots, N\}$, i.e.

$$
y_n = \beta_1 x_{n,1} + \ldots + \beta_p x_{n,p} + \epsilon_n = \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_n, \quad n = 1, \dots, N,
$$

where, given the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$, the errors $\{\epsilon_n,\ n = 1, \dots, N\}$ are

- mean zero, i.e. $\mathbb{E}[\epsilon_n|\mathbf{X}] = 0$ for $n = 1, \dots, N$,

- homoscedastic, i.e. $\mathbb{E}[\epsilon_n^2|\mathbf{X}] = \sigma^2 \in [0, \infty)$ for $n = 1, \dots, N$, and
- uncorrelated, i.e. $\mathbb{E}[\epsilon_n \epsilon_m|\mathbf{X}] = 0$ for $n, m = 1, \dots, N$ with $n \neq m$.

For any test observation $\mathbf{x}_* = (x_{*,1}, \dots, x_{*,p}) \in \mathbb{R}^p$ we thus have the best possible prediction (in terms of MSE) for y_* is

$$\mathbb{E}[y_*|\mathbf{x}_*] = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} =: \hat{y}_*(\boldsymbol{\beta}),$$

and plugging in an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ yields the predictor

$$\hat{y}_*(\hat{\boldsymbol{\beta}}) = \mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}.$$

Note that, under the assumptions introduced above, the mean squared error of such a predictor can be decomposed as

$$\begin{aligned}\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}) - y_*)^2|\mathbf{X}, \mathbf{x}_*\right] &= \mathbb{E}\left[\epsilon_*^2|\mathbf{X}, \mathbf{x}_*\right] + \mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right] \\ &= \sigma^2 + \mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right].\end{aligned}$$

Under the assumptions discussed above minimizing the mean squared error of the predictor relative to the target y_* is thus equivalent to minimizing the mean squared error of the predictor relative to the unfeasible best prediction $\hat{y}_*(\boldsymbol{\beta})$.

4.F.1 Controlled Ordinary Least Squares (OLS) estimation

Classic OLS estimation

The usual OLS estimator for $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ is given by[58]

$$\hat{\boldsymbol{\beta}}_{\mathbf{X}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

which, by the Gauss-Markov theorem, is known to be the best linear unbiased estimator (BLUE): for any $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$,

$$\mathbb{E}\left[(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X}} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\beta})^2|\mathbf{X}\right] = \min_{\tilde{\boldsymbol{\beta}}_{\mathbf{X}} \in \mathrm{LUE}(\mathbf{X}, \mathbf{y})} \mathbb{E}\left[(\boldsymbol{\lambda}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\mathbf{X}} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\beta})|\mathbf{X}\right],$$

where $\mathrm{LUE}(\mathbf{X}, \mathbf{y})$ is the set of all linear and unbiased estimator for $\boldsymbol{\beta}$, i.e. $\tilde{\boldsymbol{\beta}}_{\mathbf{X}} = C(\mathbf{X})\mathbf{y}$ for some $\mathbf{X}$-measurable matrix $C(\mathbf{X}) \in \mathbb{R}^{p \times N}$ and $\mathbb{E}[\tilde{\boldsymbol{\beta}}_{\mathbf{X}}|\mathbf{X}] = \boldsymbol{\beta}$. By applying the BLUE property, we can show that $\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}})$ is the best[59] predictor across all predictors formed from linear and unbiased

[58] Here, and in all other estimators discussed in this section, the dependence on $\mathbf{y}$ is dropped from the notation.

[59] In terms of mean squared error (MSE). Recall that for unbiased estimators, the MSE is equal to the estimator's variance.

estimators, i.e.

$$\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2 | \mathbf{X}, \mathbf{x}_*\right] \leq \mathbb{E}\left[(\hat{y}_*(\tilde{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2 | \mathbf{X}, \mathbf{x}_*\right],$$

for all $\tilde{\boldsymbol{\beta}}_{\mathbf{X}} \in \mathrm{LUE}(\mathbf{X}, \mathbf{y})$. Note that here we applied the mean-zero uncorrelated errors assumption.

Controlled OLS estimation

Let us now assume we can additionally observe the "control" random variables $\{\mathbf{z}_n = (z_{n,1}, \ldots, z_{n,k}) \in \mathbb{R}^k,\ n = 1, \ldots, N\}$. We shall assume the controls are available for training, i.e. when estimating $\boldsymbol{\beta}$, but for predicting, i.e. when forecasting y_* we will have access to $\mathbf{x}_*$ but not to $\mathbf{z}_*$. Given the original design matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, we now assume the errors and the controls are jointly

- mean zero, i.e. $\mathbb{E}[(\epsilon_n, \mathbf{z}_n)|\mathbf{X}] = \mathbf{0} \in \mathbb{R}^{k+1}$ for $n = 1, \ldots, N$,
- homoscedastic, i.e. $\mathbb{E}[(\epsilon_n, \mathbf{z}_n)^{\otimes 2}|\mathbf{X}] = \Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ for $n = 1, \ldots, N$ for some Σ symmetric positive definite,
- uncorrelated, i.e. $\mathbb{E}[(\epsilon_n, \mathbf{z}_n) \otimes (\epsilon_m, \mathbf{z}_m)|\mathbf{X}] = \mathbf{0} \in \mathbb{R}^{(k+1)\times(k+1)}$ for $n, m = 1, \ldots, N$ with $n \neq m$.

In what follows we partition

$$\Sigma = \begin{pmatrix} \sigma^2 & \Sigma_{y,\mathbf{z}} \\ \Sigma_{\mathbf{z},y} & \Sigma_{\mathbf{z}} \end{pmatrix},$$

for $\sigma^2 \in (0, \infty)$, $\Sigma_{\mathbf{z},y} = \Sigma_{y,\mathbf{z}}^{\mathrm{T}} \in \mathbb{R}^k$ and $\Sigma_{\mathbf{z}} \in \mathbb{R}^{k\times k}$. Note that we can use the notation $\Sigma_{\mathbf{z},y}$ since for all $n = 1, \ldots, N$ we have $\mathrm{Cov}(\mathbf{z}_n, Y_n|\mathbf{X}) = \mathbb{E}[\mathbf{z}_n \epsilon_n|\mathbf{X}] =: \Sigma_{\mathbf{z},y}$. We use this notation, instead of the possibly more natural $\Sigma_{\mathbf{z},\epsilon}$, as we believe that in applications it is often easier to find highly correlated controls by thinking about the target rather than the residual.

Remark 4.F.1. Throughout the whole section, unless stated otherwise, we consider the original design matrix $\mathbf{X}$ and the new observation $\mathbf{x}_* \in \mathbb{R}^p$ to be fixed with random controls $\mathbf{Z}$ and errors $\boldsymbol{\epsilon}$.

As discussed above, fixing the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ and a test observation $\mathbf{x}_* = (x_{*,1}, \ldots, x_{*,p}) \in \mathbb{R}^p$, the predictor

$$\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) = \mathbf{x}_*^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathbf{X}},$$

is unbiased for the statistic $\hat{y}_*(\boldsymbol{\beta}) = \mathbf{x}_*^{\mathrm{T}} \boldsymbol{\beta} \in \mathbb{R}$. We hence introduce the control variate predictor

$$\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda}) = \hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{Z} \boldsymbol{\lambda}_2,$$

where $\mathbf{Z} \in \mathbb{R}^{N\times k}$ is the control design matrix while $\boldsymbol{\lambda}_1 \in \mathbb{R}^N$ and $\boldsymbol{\lambda}_2 \in \mathbb{R}^k$ are measurable in $\mathbf{X}$ and $\mathbf{x}_*$. Under the assumptions introduced above the controlled predictor can be shown to be unbiased

and attains a minimum[60] in variance when

$$\boldsymbol{\lambda}_1^* = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* \quad \text{and} \quad \boldsymbol{\lambda}_2^* = -\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}.$$

For any $\mathbf{x}_* = (x_{*,1}, \ldots, x_{*,p}) \in \mathbb{R}^p$ we thus have

$$\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda}^*) = \mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y} = \mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma} = \hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}),$$

yielding the unfeasible (in the sense that it depends on the unknown quantity Σ) estimator

$$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}) = \hat{\boldsymbol{\beta}}_{\mathbf{X}} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}.$$

Note that, given $\mathbf{X}$, $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$ is unbiased (by mean zero property of the controls). Moreover, for

---

[60] Note that since $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are assumed to be jointly spherical given $\mathbf{X}$ (and $\mathbf{x}_*$) the variance of the controlled predictor is given by

$$\begin{aligned}
&\mathrm{Var}(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda})|\mathbf{X}, \mathbf{x}_*) \\
&\quad = \mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{Z}\boldsymbol{\lambda}_2|\mathbf{X}, \mathbf{x}_*) \\
&\quad = \mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}|\mathbf{X}, \mathbf{x}_*]\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* + 2\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\lambda}_2^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}|\mathbf{X}, \mathbf{x}_*]\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_1^{\mathrm{T}}\mathbb{E}[\mathbf{Z}\boldsymbol{\lambda}_2\boldsymbol{\lambda}_2^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}|\mathbf{X}, \mathbf{x}_*]\boldsymbol{\lambda}_1 \\
&\quad = \sigma^2\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* + 2\boldsymbol{\lambda}_2^{\mathrm{T}}\Sigma_{\mathbf{z},y}\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2^{\mathrm{T}}\Sigma_{\mathbf{z}}\boldsymbol{\lambda}_2\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\lambda}_1.
\end{aligned}$$

Setting partial derivatives in $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ equal to zero

$$\begin{aligned}
\partial_{\boldsymbol{\lambda}_1}\mathrm{Var}(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda})|\mathbf{X}, \mathbf{x}_*) &= 2\boldsymbol{\lambda}_2^{\mathrm{T}}\Sigma_{\mathbf{z},y}\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} + 2\boldsymbol{\lambda}_2^{\mathrm{T}}\Sigma_{\mathbf{z}}\boldsymbol{\lambda}_2\boldsymbol{\lambda}_1 = 2\boldsymbol{\lambda}_2^{\mathrm{T}}(\Sigma_{\mathbf{z},y}\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} + \Sigma_{\mathbf{z}}\boldsymbol{\lambda}_2\boldsymbol{\lambda}_1) = 0, \\
\partial_{\boldsymbol{\lambda}_2}\mathrm{Var}(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda})|\mathbf{X}, \mathbf{x}_*) &= 2\boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*\Sigma_{\mathbf{z},y} + 2\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\lambda}_1\Sigma_{\mathbf{z}}\boldsymbol{\lambda}_2 = 2\boldsymbol{\lambda}_1^{\mathrm{T}}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*\Sigma_{\mathbf{z},y} + \boldsymbol{\lambda}_1\Sigma_{\mathbf{z}}\boldsymbol{\lambda}_2) = 0,
\end{aligned}$$

yields as non-trivial (i.e. such that $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \neq 0$) stationary point $\boldsymbol{\lambda}_1^* = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*$ and $\boldsymbol{\lambda}_2^* = -\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}$. By taking second derivatives we can compute the Hessian at the stationary point to be

$$\partial_{\boldsymbol{\lambda}}\partial_{\boldsymbol{\lambda}^{\mathrm{T}}}\mathrm{Var}(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{Z}, \boldsymbol{\lambda})|\mathbf{X}, \mathbf{x}_*)\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_*} = 2\begin{pmatrix} \Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}I_{N\times N} & -\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*\Sigma_{y,\mathbf{z}} \\ -\Sigma_{\mathbf{z},y}\mathbf{x}_*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} & \mathbf{x}_*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*\Sigma_{\mathbf{z}} \end{pmatrix},$$

which is positive (semi)definite since Σ is positive definite and hence $\boldsymbol{\lambda}^*$ is a minimum. To show the assumption that Σ is positive definite implies the Hessian is positive (semi)definite we use the following result from linear algebra theory: a symmetric block matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{D} \end{pmatrix},$$

where $\mathbf{A}$ is square is positive (semi)definite if and only $\mathbf{A}$ and $\mathbf{D} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}$ are positive (semi)definite. Applying this result to the Hessian at $\boldsymbol{\lambda}_*$ note that $\mathbf{A} = \Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}I_{N\times N}$ is trivially positive (semi)definite since $\Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y} \geq 0$ by positive definitness of $\Sigma_{\mathbf{z}}$, and

$$\mathbf{D} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B} = \mathbf{x}_*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*\left(\Sigma_{\mathbf{z}} - \frac{\Sigma_{\mathbf{z},y}\Sigma_{y,\mathbf{z}}}{\Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}}\right),$$

is positive (semi)definite since $\mathbf{x}_*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* \geq 0$ by positive definitness of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\Sigma_{\mathbf{z}} - (\Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y})^{-1}\Sigma_{\mathbf{z},y}\Sigma_{y,\mathbf{z}}$ is positive (semi)definite by applying the converse of the previous statement to the positive (semi)definite matrix

$$\mathbf{M}' = \begin{pmatrix} \Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y} & \Sigma_{y,\mathbf{z}} \\ \Sigma_{\mathbf{z},y} & \Sigma_{\mathbf{z}} \end{pmatrix}.$$

any test observation $\mathbf{x}_* = (x_{*,1}, \dots, x_{*,p}) \in \mathbb{R}^p$,

$$\begin{aligned}
&\mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}|\mathbf{X},\mathbf{x}_*) \\
&\quad= \mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X}}|\mathbf{X},\mathbf{x}_*) + \mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}|\mathbf{X},\mathbf{x}_*) \\
&\qquad\qquad - 2\,\mathrm{Cov}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}|\mathbf{X},\mathbf{x}_*) \\
&\quad= \sigma^2\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* + \Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* - 2\,\Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* \\
&\quad= (\sigma^2 - \Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y})\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* \\
&\quad\leq \sigma^2\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_* = \mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X}}|\mathbf{X},\mathbf{x}_*),
\end{aligned}$$

with equality iff $\Sigma_{\mathbf{z},y} = 0$. In other words, as long as the controls are correlated with the target, we obtain a better prediction by using $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$ instead of the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ and the quality of the prediction increases as the correlation between y and $\mathbf{z}$ increases. The variance reduction factor is constant across test observations and is given by

$$\frac{\mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}|\mathbf{X},\mathbf{x}_*)}{\mathrm{Var}(\mathbf{x}_*^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{X}}|\mathbf{X},\mathbf{x}_*)} = \left(1 - \frac{\Sigma_{y,\mathbf{z}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}}{\sigma^2}\right).$$

Remark 4.F.2. Note that when $\mathbf{X} = \mathbb{1} \in \mathbb{R}^N$, $\mathbf{x}_* = 1$ and $\mathbf{Z} \in \mathbb{R}^N$ we estimate $\mu_* = \mathbb{E}[y]$ with the OLS estimator $\hat{\mu}_* = \hat{\beta} = \bar{\mathbf{y}}$ and the simplest control variate estimator

$$\hat{\mu}_*^c = \bar{\mathbf{y}} - \frac{\mathrm{Cov}(y,z)}{\mathrm{Var}(z)}\bar{\mathbf{z}}.$$

since $\boldsymbol{\lambda}_1^* = \frac{1}{N}\mathbb{1} \in \mathbb{R}^N$ and $\lambda_2^* = -\frac{\mathrm{Cov}(y,z)}{\mathrm{Var}(z)}$ which has reduced variance by a factor of $(1-\mathrm{Corr}(y,z))$.

In practice, the correlation matrix Σ is usually unknown; hence, to make the estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$ feasible, we need to estimate it. Under the assumptions discussed above, the most natural candidate is given by the sample estimates

$$\hat{\Sigma}_{\mathbf{z},y} = \frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \quad \text{and} \quad \hat{\Sigma}_{\mathbf{z}} = \frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{Z},$$

yielding the feasible estimator

$$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(I - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{y}.$$

This can be equivalently understood as regressing $\mathbf{y}$ on the control $\mathbf{Z}$ (i.e. projecting $\mathbf{y}$ onto the space spanned by $\mathbf{Z}$) and then regressing the residual onto $\mathbf{X}$. The resulting estimator will likely be biased (given $\mathbf{X}$), with its exact finite sample properties depending on the distribution of $(\mathbf{X},\mathbf{Z})|\boldsymbol{\epsilon}$.

When $\boldsymbol{\epsilon}$ depends linearly on $\mathbf{Z}$, i.e.

$$\boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\eta},$$

we can write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\eta},$$

and hence we know that the joint OLS estimator obtained from the design matrix $(\mathbf{X}\ \mathbf{Z}) \in \mathbb{R}^{N\times(p+k)}$, i.e.

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}\\ \hat{\boldsymbol{\alpha}}_{\mathbf{X},\mathbf{Z}}\end{pmatrix} = \left[\begin{pmatrix}\mathbf{X}^{\mathrm{T}}\\ \mathbf{Z}^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}\right]^{-1}\begin{pmatrix}\mathbf{X}^{\mathrm{T}}\\ \mathbf{Z}^{\mathrm{T}}\end{pmatrix}\mathbf{y} = \begin{pmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{y}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{y}\end{pmatrix}.$$

By using some simple algebraic manipulations for block matrices, we can extract the first p entries of the joint OLS estimator, i.e. the estimator for $\boldsymbol{\beta}$, as

$$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}} = \hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}'} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} - \mathbf{Z}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z})^{-1}(\mathbf{Z}^{\mathrm{T}}\mathbf{y} - \mathbf{Z}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}),$$

i.e. $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$ with Σ estimated by

$$\hat{\Sigma}'_{\mathbf{z},y} = \frac{1}{N}\mathbf{Z}^{\mathrm{T}}(I - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y} \quad \text{and} \quad \hat{\Sigma}'_{\mathbf{z}} = \frac{1}{N}\mathbf{Z}^{\mathrm{T}}(I - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Z}.$$

Note these are the covariance and variance estimators obtained when projecting $\mathbf{Z}$ onto the orthogonal complement of the space spanned by $\mathbf{X}$. Under the assumptions introduced above, these are also unbiased for $\Sigma_{\mathbf{z},y}$ and $\Sigma_{\mathbf{z}}$. This provides a second feasible controlled estimator for $\boldsymbol{\beta}$.

If we fix both $\mathbf{X}$ and $\mathbf{Z}$ and assume $\boldsymbol{\eta}$ satisfies the Gauss-Markov conditions, then the joint OLS estimator is the BLUE for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Extending the classic OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ to $(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \hat{\boldsymbol{\alpha}}_{\mathbf{X},\mathbf{Z}})$ we obtain another linear and unbiased estimator for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. It follows from the Gauss-Markov theorem that for any $\mathbf{x}_* \in \mathbb{R}^p$, setting $\boldsymbol{\lambda} = (\mathbf{x}_*, \mathbb{E}[\mathbf{z}_*|\mathbf{X}, \mathbf{Z}, \mathbf{x}_*] = \mathbf{0})$, one has

$$\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{Z}, \mathbf{x}_*\right] \leq \mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{Z}, \mathbf{x}_*\right],$$

and hence, by the tower property of conditional expectation,

$$\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right] \leq \mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right].$$

Using the forms of the variances of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ we can compute

$$\begin{aligned}\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right] &= \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbb{E}[(\mathbf{X}^{\mathrm{T}}(I - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{X})^{-1}|\mathbf{X}]\mathbf{x}_*,\\ \mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right] &= \sigma^2\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*,\end{aligned}$$

and thus quantify the MSE reduction factor as

$$\frac{\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right]}{\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}}) - \hat{y}_*(\boldsymbol{\beta}))^2|\mathbf{X}, \mathbf{x}_*\right]} = \frac{\mathbf{x}_*^{\mathrm{T}}\mathbb{E}[(\mathbf{X}^{\mathrm{T}}(I - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{X})^{-1}|\mathbf{X}]\mathbf{x}_*}{\mathbf{x}_*^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_*},$$

which we note does not depend on σ^2.

Given $\mathbf{X}$ and $\mathbf{Z}$, augmenting the first feasible controlled estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ to an estimator for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, yields a linear but biased (at least in the $\boldsymbol{\beta}$ components) estimator. Whether $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ or $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ yields a better predictor cannot thus be deduced from the Gauss-Markov theorem. As we will see in the numerical experiments discussed in the next section, which estimator performs better depends on the properties of the data generating process.

4.F.2 Simulation study

In the previous section we introduced two feasible control estimators, $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$, for the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. We showed that when

$$\boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\eta},$$

and $\boldsymbol{\eta}|\mathbf{Z}, \mathbf{X}$ is mean-zero, uncorrelated and homoscedastic then $\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}})$ is always a better predictor than the one formed by the classic OLS estimator $\hat{y}_*(\hat{\boldsymbol{\beta}}_{\mathbf{X}})$. This leaves unanswered the question of how $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ performs relative to $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ (and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$). We address this question empirically in Table 4.F.1 with the following experimental setup.

We consider $N = 1,000$ i.i.d. samples from the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

with $\beta_0 = -1, \beta_1 = 6, \beta_2 = 8$ and

$$\epsilon = \sigma(\rho z + \sqrt{1-\rho^2}\eta),$$

with $\eta \sim \mathcal{N}(0,1) \perp z \sim \mathcal{N}(0,1)$ independently of $\mathbf{X}$. We thus have $\text{Corr}(y, z|\mathbf{x}) = \text{Corr}(\epsilon, z) = \rho$ and $\epsilon \sim \mathcal{N}(0, \sigma^2) \perp \mathbf{X}$. We fix $N = 1,000$ samples $x_1 \sim \mathcal{N}(0,1) \perp x_2 \sim \mathcal{N}(1,1)$ to obtain $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and evaluate the performance of the predictor on the new observation $x_1^* = 0, x_2^* = 1$. For each estimator, we report an estimate of the root mean square error

$$\text{RMSE}(\hat{y}_*(\hat{\boldsymbol{\beta}}), \hat{y}_*(\boldsymbol{\beta})) = \sqrt{\mathbb{E}\left[(\hat{y}_*(\hat{\boldsymbol{\beta}}) - \hat{y}_*(\boldsymbol{\beta}))^2 | \mathbf{X}, \mathbf{x}_*\right]},$$

obtained over $10,000$ Monte Carlo samples (i.e. keeping $\mathbf{X}$ and $\mathbf{x}^*$ fixed and resampling $\mathbf{Z} \in \mathbb{R}^N$ and $\boldsymbol{\eta} \in \mathbb{R}^N$). We highlight with a single asterisk (*) RMSEs that are lower than the uncontrolled OLS estimator's RMSE with high statistical significance (t-test p-value across the Monte Carlo samples less than 0.001). When one of the two proposed estimators outperforms the other with high statistical significance we highlight the corresponding RMSE with double asterisks (**).

As expected, as the correlation between the control and the target ρ increases, the performance gain obtained by using either of the feasible control estimators increases. The joint-OLS estimator outperforms the standard OLS estimator by the same amount across different signal-to-noise ratios

while $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$'s relative performance changes. Comparing $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ we note the results suggest that for high signal-to-noise ratios[61] the joint-OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ slightly outperforms $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ while for lower signal-to-noise ratios the latter estimator performs marginally better. Consider the two edge cases:

- when there is no error, i.e. $\sigma = 0$, $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ performs worse than $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ $(= \hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}})$ as it is adding uninformative variability to the estimator;
- when there is no signal, i.e. $\mathbf{X} = \mathbb{1} \in \mathbb{R}^N$, $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ both reduce to the classic control variates estimators, cf. Remark 4.F.2, but the former is more precise as it estimates $\boldsymbol{\lambda}_2^* = -\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z},y}$ using the knowledge that the control is mean zero.

Next, we investigate the empirical performance of $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$ when the dependency between $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ is non-linear but $(\mathbf{Z}, \epsilon)|\mathbf{X}$ are still jointly mean-zero, uncorrelated (across samples) and homoscedastic. Keeping the same design matrix $\mathbf{X}$ as in Table 4.F.1 and the same parameters $\boldsymbol{\beta} = (-1, 6, 8)$ we now let

$$\epsilon = \sigma\kappa f(z) + \sigma\sqrt{1-\kappa^2}\eta,$$

with $\eta \sim \mathcal{N}(0,1) \perp z \sim \mathcal{N}(0,1)$ independently of $\mathbf{X}$. By choosing $f(z)$ such that $\mathbb{E}[f(z)] = 0$ and $\mathbb{E}[f(z)^2] = 1$ we have $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2$. Moreover, choosing $\kappa = \mathrm{Cov}(z, f(z))^{-1}\rho$ ensures that $\mathrm{Corr}(y, z|\mathbf{x}) = \mathrm{Corr}(\epsilon, z) = \rho$. We investigate the following three dependence functions:

$$\text{(sq)} \quad f(z) = \frac{z^2 + z - 1}{\sqrt{3}};$$

$$\text{(cube)} \quad f(z) = \frac{z^3}{\sqrt{15}};$$

$$\text{(exp)} \quad f(z) = \frac{e^z - \sqrt{e}}{\sqrt{e^2 - e}}.$$

The code used to produce the results of Table 4.F.1 and Table 4.F.2 can be found at `https://github.com/lorenzolucchese/controlled-linear-regression`.

We also experimented with multiplicative noise ($\epsilon \propto f(z)\,\eta$), heavier tailed errors ($\eta \sim t_3$) and different dataset sizes ($N \in \{100, 1000, 10000\}$). The results are similar to the ones reported in Table 4.F.1 and Table 4.F.2: the two control estimators outperform the classic OLS estimator as long as $\rho > 0$ while the differences in performance between the two control estimators are statistically significant but small.

[61] Here we define the signal-to-noise ratio as $\mathrm{std}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})/\sigma$ where $\mathrm{std}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ measures the variablity of the explainable part of the model (signal) and σ measures the variability in the error ϵ (noise). In Table 4.F.1 the signal component of the model is kept fixed and hence higher σ denotes lower signal-to-noise ratio.

		$\hat{\boldsymbol{\beta}}_{\mathbf{X}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$		$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$		$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$	
σ	ρ	**RMSE**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**
5	0.00	0.1572	0.1579	100.46%	0.1573	100.03%	0.1572	100.00%
	0.25	0.1588	0.1550*	97.57%	0.1538**	96.86%	0.1489	93.75%
	0.50	0.1581	0.1368*	86.48%	0.1358**	85.86%	0.1186	75.00%
	0.75	0.1592	0.1052*	66.08%	0.1043**	65.52%	0.0697	43.75%
	1.00	0.1606	0.0164*	10.21%	0.0000**	0.01%	0.0000	0.00%
10	0.00	0.3144	0.3146	100.06%	0.3145	100.03%	0.3144	100.00%
	0.25	0.3177	0.3082*	97.03%	0.3077**	96.86%	0.2978	93.75%
	0.50	0.3163	0.2719*	85.97%	0.2715*	85.86%	0.2372	75.00%
	0.75	0.3185	0.2088*	65.57%	0.2086*	65.52%	0.1393	43.75%
	1.00	0.3211	0.0164*	5.11%	0.0000**	0.00%	0.0000	0.00%
20	0.00	0.6289	0.6286	99.96%	0.6291	100.03%	0.6289	100.00%
	0.25	0.6353	0.6154*	96.86%	0.6154*	96.86%	0.5956	93.75%
	0.50	0.6325	0.5429*	85.83%	0.5431*	85.86%	0.4744	75.00%
	0.75	0.6369	0.4169*	65.46%	0.4173*	65.52%	0.2786	43.75%
	1.00	0.6422	0.0164*	2.55%	0.0000**	0.00%	0.0000	0.00%
40	0.00	1.2578	1.2569	99.93%	1.2582	100.03%	1.2578	100.00%
	0.25	1.2707	1.2300**	96.80%	1.2307*	96.86%	1.1913	93.75%
	0.50	1.2651	1.0853**	85.79%	1.0862*	85.86%	0.9488	75.00%
	0.75	1.2738	0.8336**	65.44%	0.8346*	65.52%	0.5573	43.75%
	1.00	1.2845	0.0164*	1.28%	0.0000**	0.00%	0.0000	0.00%

Table 4.F.1: RMSEs of the classic OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$, the two feasible control estimators $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$, and the unfeasible optimal control estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$. A single asterisk (*) indicates feasible RMSEs that are lower than the uncontrolled classic OLS estimator's RMSE with high statistical significance (t-test p-value across the 10,000 Monte Carlo samples less than 0.001). A double asterisk (**) indicates the feasible control estimator's RMSE is lower than the other feasible control estimator's RMSE with high statistical significance.

		$\hat{\boldsymbol{\beta}}_{\mathbf{X}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$		$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$		$\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$	
f	ρ	**RMSE**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**	**RMSE**	**(% of $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$)**
	0.00	0.3144	0.3146	100.06%	0.3145	100.03%	0.3144	100.00%
	0.25	0.3174	0.3084*	97.14%	0.3075**	96.88%	0.2976	93.75%
(sq)	0.50	0.3202	0.2772*	86.57%	0.2761**	86.25%	0.2401	75.00%
	0.75							
	1.00							
	0.00	0.3144	0.3146	100.06%	0.3145	100.03%	0.3144	100.00%
	0.25	0.3184	0.3088*	97.00%	0.3083*	96.83%	0.2985	93.75%
(cube)	0.50	0.3184	0.2736*	85.95%	0.2734*	85.86%	0.2388	75.00%
	0.75	0.3222	0.2125*	65.96%	0.2123*	65.90%	0.1409	43.75%
	1.00							
	0.00	0.3144	0.3146	100.06%	0.3145	100.03%	0.3144	100.00%
	0.25	0.3178	0.3085*	97.07%	0.3078**	96.85%	0.2979	93.75%
(exp)	0.50	0.3182	0.2741*	86.13%	0.2734**	85.92%	0.2387	75.00%
	0.75	0.3191	0.2091*	65.53%	0.2081**	65.20%	0.1396	43.75%
	1.00							

Table 4.F.2: RMSEs of the classic OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$, the two feasible control estimators $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\hat{\Sigma}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z}}$, and the unfeasible optimal control estimator $\hat{\boldsymbol{\beta}}_{\mathbf{X},\mathbf{Z},\Sigma}$. A single asterisk (*) indicates feasible RMSEs that are lower than the uncontrolled classic OLS estimator's RMSE with high statistical significance (t-test p-value across the 10,000 Monte Carlo samples less than 0.001). A double asterisk (**) indicates the feasible control estimator's RMSE is lower than the other feasible control estimator's RMSE with high statistical significance. Empty values indicate setups that are not achievable for the given correlation ρ and dependence function f.

Bibliography

Abergel, F., Lehalle, C.-A., and Rosenbaum, M. (2014). Understanding the Stakes of High-Frequency Trading. In: *The Journal of Trading* 9.4, pp. 49–73. ISSN: 1559-3967. DOI: 10.3905/jot.2014.9.4.049.

Aït-Sahalia, Y., Fan, J., Xue, L., and Zhou, Y. (2022). *How and When are High-Frequency Stock Returns Predictable?* NBER Working Papers 30366. National Bureau of Economic Research, Inc. URL: https://ideas.repec.org/p/nbr/nberwo/30366.html.

Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton University Press. ISBN: 9780691161433. DOI: 10.23943/princeton/9780691161433.001.0001.

Applebaum, D. (2009). *Lévy Processes and Stochastic Calculus*. 2nd. Cambridge University Press. ISBN: 9780511650581.

Bacry, E. and Muzy, J.-F. (2014). Hawkes model for price and trades high-frequency dynamics. In: *Quantitative Finance* 14.7, pp. 1147–1166. ISSN: 1469-7688, 1469-7696. DOI: 10.1080/14697688.2014.897000.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 167–241. DOI: 10.1111/1467-9868.00282.

Baudoin, F. and Coutin, L. (2007). Operators associated with a stochastic differential equation driven by fractional Brownian motions. In: *Stochastic Processes and their Applications* 117.5, pp. 550–574. ISSN: 0304-4149. DOI: 10.1016/j.spa.2006.09.004.

Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.8, pp. 1798–1828. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50.

Billingsley, P. (1986). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471804789.

Bonnier, P. and Oberhauser, H. (2020). Signature cumulants, ordered partitions, and independence of stochastic processes. In: *Bernoulli* 26.4, pp. 2727–2757.

Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M. (2018). *Trades, Quotes and Prices: Financial Markets Under the Microscope*. Cambridge University Press. ISBN: 9781316659335.

Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. In: *Probability Surveys* 2, pp. 107–144. DOI: 10.1214/154957805100000104.

Briola, A., Bartolucci, S., and Aste, T. (2024). *Deep Limit Order Book Forecasting*. arXiv: 2403.09267 [q-fin.TR].

Brockwell, P. J. (2009). Lévy–Driven Continuous–Time ARMA Processes. In: *Handbook of Financial Time Series*. Ed. by T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen. Springer Berlin Heidelberg, pp. 457–480. ISBN: 978-3-540-71297-8. DOI: 10.1007/978-3-540-71297-8_20.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer Cham, pp. 323–351. ISBN: 978-3-319-29854-2. DOI: 10.1007/978-3-319-29854-2_11.

Brockwell, P. J., Davis, R. A., and Yang, Y. (2007a). Continuous-time Gaussian Autoregression. In: *Statistica Sinica* 17.1, pp. 63–80. ISSN: 10170405, 19968507.

— (2007b). Estimation for Nonnegative Lévy-Driven Ornstein-Uhlenbeck Processes. In: *Journal of Applied Probability* 44.4, pp. 977–989. ISSN: 00219002.

— (2011). Estimation for Non-Negative Lévy-Driven CARMA Processes. In: *Journal of Business & Economic Statistics* 29.2, pp. 250–259. ISSN: 07350015.

Brockwell, P. J. and Lindner, A. (2009). Existence and uniqueness of stationary Lévy-driven CARMA processes. In: *Stochastic Processes and their Applications* 119.8, pp. 2660–2681. ISSN: 0304-4149. DOI: 10.1016/j.spa.2009.01.006.

Buehler, H., Murray, P., Pakkanen, M. S., and Wood, B. (2022). *Deep Hedging: Learning to Remove the Drift under Trading Frictions with Minimal Equivalent Near-Martingale Measures*. arXiv: 2111.07844 [q-fin.CP].

Byrd, D., Hybinette, M., and Balch, T. H. (2020). ABIDES: Towards High-Fidelity Multi-Agent Market Simulation. In: *Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. ACM, pp. 11–22. ISBN: 9781450375924. DOI: 10.1145/3384441.3395986.

Cass, T. and Turner, W. F. (2022). *Topologies on unparameterised path space*. arXiv: 2206.11153 [math.CA].

Cboe (2023). *U.S. Equities Market Volume Summary*. URL: www.cboe.com/us/equities/market_share/ (visited on 08/30/2023).

Chan, K.-S. and Tong, H. (1987). A Note on Embedding a Discrete Parameter ARMA Model in a Continuous Parameter ARMA Model. In: *Journal of Time Series Analysis* 8.3, pp. 277–281. DOI: 10.1111/j.1467-9892.1987.tb00439.x.

Chen, K.-T. (1954). Iterated Integrals and Exponential Homomorphisms. In: *Proceedings of the London Mathematical Society* s3–4.1, pp. 502–512. DOI: 10.1112/plms/s3-4.1.502.

Chevyrev, I. and Lyons, T. (2016). Characteristic functions of measures on geometric rough paths. In: *The Annals of Probability* 44.6, pp. 4049–4082. ISSN: 00911798.

Chevyrev, I. and Oberhauser, H. (2018). *Signature moments to characterize laws of stochastic processes*. arXiv: 1810.10971 [math.ST].

Cho, K., Merriënboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing (EMNLP). Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.

Çinlar, E., Jacod, J., Protter, P., and Sharpe, M. J. (1980). Semimartingales and Markov Processes. In: *Probability Theory and Related Fields* 54, pp. 161–219. DOI: 10.1007/BF00531446.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. In: *Journal of the American Statistical Association* 74.368, pp. 829–836. DOI: 10.1080/01621459.1979.10481038.

Coletta, A., Moulin, A., Vyetrenko, S., and Balch, T. (2022). Learning to simulate realistic limit order book markets from data as a World Agent. In: *Proceedings of the Third ACM International Conference on AI in Finance*. ACM, pp. 428–436. ISBN: 9781450393768. DOI: 10.1145/3533271.3561753.

Cont, R., Kukanov, A., and Stoikov, S. (2013). The Price Impact of Order Book Events. In: *Journal of Financial Econometrics* 12.1, pp. 47–88. ISSN: 1479-8409. DOI: 10.1093/jjfinec/nbt003.

Courgeau, V. and Veraart, A. E. D. (2022). Likelihood theory for the graph Ornstein-Uhlenbeck process. In: *Statistical Inference for Stochastic Processes* 25.2, pp. 227–260. ISSN: 1387-0874, 1572-9311. DOI: 10.1007/s11203-021-09257-1.

Coutin, L. and Qian, Z. (2002). Stochastic analysis, rough path analysis and fractional Brownian motions. In: *Probability Theory and Related Fields* 122, pp. 108–140. DOI: 10.1007/s004400100158.

CSPi (2023). *Tick-to-Trade Latency*. URL: www.cspi.com/wp-content/uploads/2016/06/Tick-to-Trade-Latency_FINAL-2.pdf (visited on 08/30/2023).

Cuchiero, C., Svaluto-Ferro, S., and Teichmann, J. (2023). *Signature SDEs from an affine and polynomial perspective*. arXiv: 2302.01362 [math.PR].

Cybenko, G. V. (1989). Approximation by Superpositions of a Sigmoidal Function. In: *Mathematics of Control, Signals and Systems* 2, pp. 303–314.

Dellacherie, C. and Meyer, P. A. (1978). *Probabilities and Potential*. North-Holland Mathematics Studies. Hermann; North-Holland Pub. Co.; sole distributors for the U.S.A. and Canada, Elsevier North-Holland. ISBN: 9780720407013.

Doob, J. L. (1962). *Stochastic Processes*. Wiley publications in statistics. Wiley.

Dragomir, S. S. (2003). *Some Gronwall Type Inequalities and Applications*. Nova Science, New York.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. In: *Econometrica* 50.4, pp. 987–1007. ISSN: 00129682, 14680262.

Enyx (2020). *NASDAQ Totalview Latency Tests*. URL: www.enyx.com/performance-reports/nxfeed/NASDAQ-Totalview/index.html (visited on 08/30/2023).

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. ISBN: 9780470316658. DOI: 10.1002/9780470316658.

European Parliament and Council (2004). *Directive 2004/39/EC on markets in financial instruments amending Council Directives 85/611/EEC and 93/6/EEC and Directive 2000/12/EC of the European Parliament and of the Council and repealing Council Directive 93/22/EEC*.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. In: *The Journal of Finance* 25.2, pp. 383–417. ISSN: 00221082, 15406261.

Fasen, V. (2013). Statistical estimation of multivariate Ornstein–Uhlenbeck processes and applications to co-integration. In: *Journal of Econometrics* 172.2. Latest Developments on Heavy-Tailed Distributions, pp. 325–337. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2012.08.019.

Fawcett, T. (2003). "Problems in stochastic analysis. Connections between rough paths and non-commutative harmonic analysis". PhD thesis. University of Oxford.

Friedman, A. (1964). *Partial Differential Equations of Parabolic Type*. Prentice-Hall.

Friz, P. K., Hager, P. P., and Tapia, N. (2022). Unified signature cumulants and generalized Magnus expansions. In: *Forum of Mathematics, Sigma* 10. DOI: 10.1017/fms.2022.20.

— (2024). *On expected signatures and signature cumulants in semimartingale models*. arXiv: 2408.05085 [stat.ML].

Friz, P. K. and Victoir, N. B. (2010). *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Futter, O., Horvath, B., and Wiese, M. (2023). *Signature Trading: A Path-Dependent Extension of the Mean-Variance Framework with Exogenous Signals*. arXiv: 2308.15135 [q-fin.PM].

Gegler, A. (2011). Estimation of the Characteristics of a multivariate Lévy Process. In: *Universität Ulm Preprint Series*.

Gikhman, I. I. and Skorokhod, A. V. (2004). *The Theory of Stochastic Processes I*. Classics in Mathematics. Springer Berlin Heidelberg. ISBN: 9783540202844.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Graham, B. (2013). *Sparse arrays of signatures for online character recognition*. arXiv: 1308.0371 [cs.CV].

Gushchin, A., Pavlyukevich, I., and Ritsch, M. (2020). Drift estimation for a Lévy-driven Ornstein–Uhlenbeck process with heavy tails. In: *Statistical Inference for Stochastic Processes* 23.3, pp. 553–570. ISSN: 1387-0874, 1572-9311. DOI: 10.1007/s11203-020-09210-8.

Hambly, B. and Lyons, T. (2005). Uniqueness for the signature of a path of bounded variation and the reduced path group. In: *Annals of Mathematics* 171.1, pp. 109–167.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. In: *Econometrica* 79.2, pp. 453–497. DOI: 10.3982/ECTA5771.

Harvey, A. and Bates, D. (2003). *Multivariate Unit Root Tests and Testing for Convergence*. Cambridge Working Papers in Economics 0301. Faculty of Economics, University of Cambridge. URL: https://ideas.repec.org/p/cam/camdae/0301.html.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In: *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

Hornik, K., Stinchcombe, M., and White, H. (1990). Universal Approximation of an Unknown Mapping and its Derivatives using Multilayer Feedforward Networks. In: *Neural Networks* 3.5, pp. 551–560. ISSN: 0893-6080. DOI: 10.1016/0893-6080(90)90005-6.

Huang, R. and Polak, T. (2011). LOBSTER: Limit Order Book Reconstruction System. In: *Information Systems & Economics eJournal*.

Ibragimov, I. A. (1962). Some Limit Theorems for Stationary Processes. In: *Theory of Probability & Its Applications* 7.4, pp. 349–382. DOI: 10.1137/1107036.

Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian Random Processes*. Applications of mathematics. Springer. ISBN: 9783540903024.

Imperial College Research Computing Service (2022). *High-Performance Computing Cluster*. DOI: 10.14469/hpc/2232.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456.

Itô, K. (1944). Stochastic Integral. In: *Proceedings of the Japan Academy, Series A, Mathematical Sciences* 20.8. ISSN: 0386-2194.

Jacod, J. and Mémin, J. (1976). Caractéristiques locales et conditions de continuité absolue pour les semi-martingales. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 35.1, pp. 1–37. ISSN: 0044-3719, 1432-2064. DOI: 10.1007/BF00532597.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes*. Vol. 288. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg. ISBN: 9783662025161. DOI: 10.1007/978-3-662-02514-7.

Jones, R. H. (1981). Fitting a Continuous-Time Autoregression to Discrete Data. In: *Applied Time Series Analysis II*. Ed. by D. F. Findley. Academic Press, pp. 651–682. ISBN: 978-0-12-256420-8. DOI: 10.1016/B978-0-12-256420-8.50026-5.

Jongbloed, G., Van Der Meulen, F. H., and Van Der Vaart, A. W. (2005). Nonparametric inference for Lévy-driven Ornstein-Uhlenbeck processes. In: *Bernoulli* 11.5. ISSN: 1350-7265. DOI: 10.3150/bj/1130077593.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Probability theory and stochastic modelling. Springer. ISBN: 9783030618728.

Karandikar, R. L. (1995). On pathwise stochastic integration. In: *Stochastic Processes and their Applications* 57.1, pp. 11–18. ISSN: 0304-4149. DOI: 10.1016/0304-4149(95)00002-0.

Karatzas, I. and Shreve, S. E. (1996). *Brownian Motion and Stochastic Calculus*. 2nd. Graduate Texts in Mathematics. Springer New York, NY. ISBN: 9780387976556.

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., and Uttley, P. (2014). Flexible and Scalable Methods for Quantifying Stochastic Variability in the Era of Massive Time-domain Astronomical Data Sets. In: *The Astrophysical Journal* 788.1. DOI: 10.1088/0004-637X/788/1/33.

Kingma, D. P. and Ba, J. (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Kiraly, F. J. and Oberhauser, H. (2019). Kernels for Sequentially Ordered Data. In: *Journal of Machine Learning Research* 20.31, pp. 1–45.

Kirilenko, A. A. and Lamacie, G. (2015). *Latency and Asset Prices*. SSRN Scholarly Paper. DOI: 10.2139/ssrn.2546567.

Klenke, A. (2008). *Probability Theory. A Comprehensive Course*. 1st. Universitext. Springer London. DOI: 10.1007/978-1-84800-048-3.

Kolm, P. N., Turiel, J. D., and Westray, N. (2021). Deep Order Flow Imbalance: Extracting Alpha at Multiple Horizons from the Limit Order Book. In: *Econometric Modeling: Capital Markets – Portfolio Theory eJournal*.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Chelsea Publishing Company. ISBN: 9780821826485.

Konecny, F. (1983). On Wong-Zakai approximation of stochastic differential equations. In: *Journal of Multivariate Analysis* 13.4, pp. 605–611. ISSN: 0047-259X. DOI: 10.1016/0047-259X(83)90043-X.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc.

Küchler, U. and Sørensen, M. (1997). *Exponential Families of Stochastic Processes*. Springer Series in Statistics. Springer New York, NY. ISBN: 9780387949819. DOI: 10.1007/b98954.

Kühn, F. and Schilling, R. L. (2019). Strong convergence of the Euler–Maruyama approximation for a class of Lévy-driven SDEs. In: *Stochastic Processes and their Applications* 129.8, pp. 2654–2680. ISSN: 0304-4149. DOI: 10.1016/j.spa.2018.07.018.

Kulik, A. (2018). *Ergodic Behavior of Markov Processes*. Berlin, Boston: De Gruyter. ISBN: 9783110458930. DOI: 10.1515/9783110458930.

Kunita, H. and Watanabe, S. (1967). On Square Integrable Martingales. In: *Nagoya Mathematical Journal* 30, pp. 209–245. ISSN: 0027-7630, 2152-6842. DOI: 10.1017/S0027763000012484.

Large, J. (2007). Measuring the resiliency of an electronic limit order book. In: *Journal of Financial Markets* 10.1, pp. 1–25. ISSN: 1386-4181. DOI: 10.1016/j.finmar.2006.09.001.

Lê, K. (2020). A stochastic sewing lemma and applications. In: *Electronic Journal of Probability* 25, pp. 1–55. DOI: 10.1214/20-EJP442.

Lemercier, M., Salvi, C., Damoulas, T., Bonilla, E. V., and Lyons, T. (2021). Distribution regression for sequential data. In: *24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*. Proceedings of Machine Learning Research. Journal of Machine Learning Research, pp. 3754–3762.

Levin, D., Lyons, T., and Ni, H. (2016). *Learning from the past, predicting the statistics for the future, learning an evolving system*. arXiv: 1309.0260 [q-fin.ST].

Liao, S., Ni, H., Sabate-Vidales, M., Szpruch, L., Wiese, M., and Xiao, B. (2024). Sig-Wasserstein GANs for conditional time series generation. In: *Mathematical Finance* 34.2, pp. 622–670. DOI: 10.1111/mafi.12423.

Lucchese, L., Pakkanen, M. S., and Veraart, A. E. D. (2023). *Estimation and Inference for Multivariate Continuous-time Autoregressive Processes*. arXiv: 2307.13020 [stat.ME].

— (2024). The Short-Term Predictability of Returns in Order Book Markets: A Deep Learning Perspective. In: *International Journal of Forecasting*. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2024.02.001.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421. DOI: 10.18653/v1/D15-1166.

Lyons, T., Caruana, M., and Lévy, T. (2007). *Differential Equations Driven by Rough Paths: École D'été de Probabilités de Saint-Flour XXXIV-2004*. Springer. ISBN: 9783540712848.

Lyons, T. and McLeod, A. D. (2024). *Signature Methods in Machine Learning*. arXiv: 2206.14674 [stat.ML].

Lyons, T., Nejad, S., and Pérez Arribas, I. (2021). Non-parametric pricing and hedging of exotic derivatives. In: *Applied Mathematical Finance* 27.6, pp. 457–494.

Lyons, T. J. (1998). Differential equations driven by rough signals. In: *Revista Matemática Iberoamericana* 14.2, pp. 215–310. URL: http://eudml.org/doc/39555.

Mai, H. (2014). Efficient maximum likelihood estimation for Lévy-driven Ornstein–Uhlenbeck processes. In: *Bernoulli* 20.2. ISSN: 1350-7265. DOI: 10.3150/13-BEJ510.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional Brownian Motions, Fractional Noises and Applications. In: *SIAM Review* 10.4, pp. 422–437. ISSN: 00361445.

Marquardt, T. and Stelzer, R. (2007). Multivariate CARMA processes. In: *Stochastic Processes and their Applications* 117.1, pp. 96–120. ISSN: 03044149. DOI: 10.1016/j.spa.2006.05.014.

Masuda, H. (2004). On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process. In: *Bernoulli* 10.1. ISSN: 1350-7265. DOI: 10.3150/bj/1077544605.

Masuda, H. (2010). Approximate self-weighted LAD estimation of discretely observed ergodic Ornstein-Uhlenbeck processes. In: *Electronic Journal of Statistics* 4, pp. 525–565. DOI: 10.1214/10-EJS565.

Melis, G., Dyer, C., and Blunsom, P. (2017). *On the State of the Art of Evaluation in Neural Language Models*. arXiv: 1707.05589 [cs.LG].

Menkveld, A. J. and Zoican, M. A. (2017). Need for Speed? Exchange Latency and Liquidity. In: *The Review of Financial Studies* 30.4, pp. 1188–1228. ISSN: 0893-9454, 1465-7368. DOI: 10.1093/rfs/hhx006.

Millar, P. W. (1972). Stochastic integrals and processes with stationary independent increments. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Probability Theory* 6.3, pp. 307–332.

Moreno, J., Vogeley, M. S., Richards, G. T., and Yu, W. (2019). Stochastic Modeling Handbook for Optical AGN Variability. In: *Publications of the Astronomical Society of the Pacific* 131.1000, pp. 1–27. DOI: 10.1088/1538-3873/ab1597.

Nasdaq (2023). *Exchange Trading Technology*. URL: www.nasdaq.com/solutions/nasdaq-trading-technology/exchange-matching (visited on 08/30/2023).

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. In: *Econometrica* 55.3, pp. 703–708. ISSN: 00129682, 14680262.

Ni, H. (2012). "The expected signature of a stochastic process". PhD thesis. University of Oxford.

Ni, H., Szpruch, L., Sabate-Vidales, M., Xiao, B., Wiese, M., and Liao, S. (2022). Sig-wasserstein GANs for time series generation. In: *Proceedings of the Second ACM International Conference on AI in Finance*. ICAIF '21. New York, NY, USA: Association for Computing Machinery. ISBN: 9781450391481. DOI: 10.1145/3490354.3494393.

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

Nualart, D. and Tindel, S. (2011). A construction of the rough path above fractional Brownian motion using Volterra's representation. In: *The Annals of Probability* 39.3, pp. 1061–1096. DOI: 10.1214/10-AOP578.

Nyblom, J. and Harvey, A. (2000). Tests of Common Stochastic Trends. In: *Econometric Theory* 16.2, pp. 176–199. ISSN: 02664666, 14694360.

Passeggeri, R. (2020). On the signature and cubature of the fractional Brownian motion for $H > 1/2$. In: *Stochastic Processes and their Applications* 130.3, pp. 1226–1257. ISSN: 0304-4149. DOI: 10.1016/j.spa.2019.04.013.

Pérez Arribas, I., Goodwin, G. M., Geddes, J. R., Lyons, T., and Saunders, K. E. A. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. In: *Translational Psychiatry* 8.

Priestley, M. B. (1982). *Spectral Analysis and Time Series: Volumes I and II*. Probability and Mathematical Statistics. Elsevier Science. ISBN: 9780125649223.

Protter, P. (1990). *Stochastic Integration and Differential Equations*. Springer Berlin, Heidelberg. ISBN: 9783662026212. DOI: 10.1007/978-3-662-02619-9.

Rajput, B. S. and Rosinski, J. (1989). Spectral Representations of Infinitely Divisible Processes. In: *Probability Theory and Related Fields* 82.3, pp. 451–487. ISSN: 1432-2064. DOI: 10.1007/BF00339998.

Salvi, C., Cass, T., Foster, J., Lyons, T., and Yang, W. (2021). The Signature Kernel Is the Solution of a Goursat PDE. In: *SIAM Journal on Mathematics of Data Science* 3.3, pp. 873–899. DOI: 10.1137/20M1366794.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press. ISBN: 9780521553025.

Sato, K. and Yamazato, M. (1984). Operator-selfdecomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type. In: *Stochastic Processes and their Applications* 17.1, pp. 73–100. ISSN: 03044149. DOI: 10.1016/0304-4149(84)90312-0.

Schell, A. and Oberhauser, H. (2023). Nonlinear independent component analysis for discrete-time and continuous-time signals. In: *Annals of Statistics* 51.2, pp. 487–518.

Schlemm, E. and Stelzer, R. (2012a). Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes. In: *Bernoulli* 18.1, pp. 46–63. DOI: 10.3150/10-BEJ329.

— (2012b). Quasi maximum likelihood estimation for strongly mixing state space models and multivariate Lévy-driven CARMA processes. In: *Electronic Journal of Statistics* 6, pp. 2185–2234. DOI: 10.1214/12-EJS743.

Seneta, E. (2013). A Tricentenary history of the Law of Large Numbers. In: *Bernoulli* 19.4, pp. 1088–1121. ISSN: 13507265.

Sirignano, J. and Cont, R. (2019). Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning. In: *Quantitative Finance* 19.9, pp. 1449–1459. DOI: 10.1080/14697688.2019.1622295.

Slutzky, E. (1937). The Summation of Random Causes as the Source of Cyclic Processes. In: *Econometrica* 5.2, pp. 105–146. ISSN: 00129682, 14680262.

Sørensen, M. (1991). Likelihood Methods for Diffusions with Jumps. In: *Statistical Inference in Stochastic Processes*. Ed. by N. U. Prabhu and I. V. Basawa. CRC Press. Chap. 3, pp. 67–105. ISBN: 9781003066606.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958.

Telgarsky, M. (2015). *Representation Benefits of Deep Feedforward Networks*. arXiv: 1509.08101 [cs.LG].

Tran, D. T., Iosifidis, A., Kanniainen, J., and Gabbouj, M. (2019). Temporal Attention-Augmented Bilinear Network for Financial Time-Series Data Analysis. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.5, pp. 1407–1418. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2018.2869225.

Triggiano, F. and Romito, M. (2024). Gaussian Processes Based Data Augmentation and Expected Signature for Time Series Classification. In: *IEEE Access* 12, pp. 80884–80895. DOI: 10.1109/ACCESS.2024.3408712.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the Theory of the Brownian Motion. In: *Physical Review* 36 (5), pp. 823–841. DOI: 10.1103/PhysRev.36.823.

Urbanik, K. and Woyczynski, W. A. (1967). A random integral and Orlicz spaces. In: *Bulletin de l'Académie Polonaise des Sciences*. Série des sciences mathématiques, astronomiques et physiques 15, pp. 161–169.

US Securities and Exchange Commission (2005). *Regulation NMS - Proposed rules and amendments to joint industry plans*.

VanderPlas, J. T. (2018). Understanding the Lomb–Scargle Periodogram. In: *The Astrophysical Journal Supplement Series* 236.1. DOI: 10.3847/1538-4365/aab766.

Willett, D. W. (1964). Nonlinear vector integral equations as contraction mappings. In: *Archive for Rational Mechanics and Analysis* 15, pp. 79–86. DOI: 10.1007/bf00257405.

Wu, Y., Mahfouz, M., Magazzeni, D., and Veloso, M. (2021). *Towards Robust Representation of Limit Orders Books for Deep Learning Models*. arXiv: 2110.05479 [q-fin.TR].

Xie, Z., Sun, Z., Jin, L., Ni, H., and Lyons, T. (2018). Learning Spatial-Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.8, pp. 1903–1917. DOI: 10.1109/TPAMI.2017.2732978.

Xu, K., Gould, M. D., and Howison, S. D. (2019). Multi-level Order-flow Imbalance in a Limit Order Book. In: *Market Microstructure and Liquidity* 4. ISSN: 2424-8037.

Yule, G. U. (1927). On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226.636-646, pp. 267–298.

Zhang, Z. and Zohren, S. (2021). *Multi-Horizon Forecasting for Limit Order Books: Novel Deep Learning Approaches and Hardware Acceleration using Intelligent Processing Units*. arXiv: 2105.10430 [cs.LG].

Zhang, Z., Zohren, S., and Roberts, S. (2019). DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. In: *IEEE Transactions on Signal Processing* 67.11, pp. 3001–3012. DOI: 10.1109/TSP.2019.2907260.