Statistical Inference for Stochastic Differential Equations using Splitting Schemes

Predrag Pilipović

PhD Thesis

this thesis has been submitted to
the phd school of the faculty of science, university of copenhagen and
faculty of business administration and economics, university of bielefeld

Department of Mathematical Sciences
University of Copenhagen

August 2024

Predrag Pilipović
predrag@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
2100 Copenhagen, Denmark
predrag.pilipovic@uni-bielefeld.de
Bielefeld Graduate School of Economics and Management
University of Bielefeld
33501 Bielefeld, Germany

Thesis title:	Statistical Inference for Stochastic Differential Equations using Splitting Schemes
Supervisors:	Professor Susanne Ditlevsen University of Copenhagen
	Professor Roland Langrock University of Bielefeld
Assessment Committee:	Professor Michael Sørensen (chair) University of Copenhagen
	Professor Alexandros Beskos University College London
	Professor Frank Riedel University of Bielefeld
Date of Submission:	August 31, 2024
Date of Defense:	November 14, 2024
ISBN:	978-87-7125-232-3

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen and Faculty of Business Administration and Economics, University of Bielefeld. It received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)".

Мојој породици
која је дала више него што је имала
да бих стигао довде.

Preface

This PhD thesis was conducted as part of the Marie Skłodowska-Curie Actions Innovative Training Networks, *Economic Policy in Complex Environments* (EPOC), which included 15 doctoral students from 7 European universities.

The thesis represents the culmination of two years of research at the Department of Mathematical Sciences at the University of Copenhagen (KU) and one year at the Faculty of Business Administration and Economics at the University of Bielefeld (UNIBI) as part of a double degree program between the two institutions. During my time in Copenhagen, I spent two months at the Greenland Institute of Natural Resources (GINR) as part of a mandatory non-academic secondment.

Susanne Ditlevsen (KU) was the main supervisor of this PhD project, with Roland Langrock (UNIBI) and Mads-Peter Heide-Jørgensen (GINR) contributing as co-supervisors.

Acknowledgment

Disclaimer: If this acknowledgment is too long, skip to the last two paragraphs.

First, I am incredibly thankful to my main supervisor, Susanne, who is much more than just a supervisor: she is a mentor, a teacher, a supporter, and a friend. Throughout my PhD journey, I realized that having the right supervisor is more important than the specific research topic. Topics may evolve during the PhD and beyond, but the knowledge and skills imparted by the supervisor shape the future researcher. Susanne taught me how to conduct research by patiently and consistently developing and building upon an idea. She also taught me how to navigate through challenging times.

I also deeply appreciate my other supervisors, Roland from Bielefeld and Mads-Peter from the Greenland Institute. Although the trajectory of my PhD project did not result in a joint project with them, they kindly welcomed me into their groups. They were exceedingly patient and supportive as I explored and determined the direction I wanted to take with my work.

Another important mentor throughout this project was Adeline Samson from the University of Grenoble Alpes. Adeline co-authored all three of my PhD projects alongside Susanne. I am very grateful for the opportunity to work with and learn from Adeline, who provided an excellent complement to Susanne's and my approaches. I especially appreciate Adeline's sharp, detail-oriented questions and comments, which saved us significant time by preventing technical errors and typos.

During my PhD, I shared many offices with many people, and I am proud to be part of this diverse community of excellent young researchers. They brought excitement and joy

to my daily life with countless funny moments. Without them, the PhD experience would have been far duller and significantly more stressful. Also, having fruitful discussions with office mates was invaluable in helping me get unstuck. Among the many wonderful colleagues, I want to mention three who became not just colleagues and friends but the closest relationships I built in Copenhagen. Shimeng, for all the coffee dates and restaurants she showed me; Alex, for our long conversations in the office and his help with the technical checks on my first paper; and Matt, for existing.

Speaking of friends in Copenhagen, I want to express my gratitude to Siniša, my old friend who took my advice and came to Copenhagen for his Master's studies. Having a Serbian friend while living in a foreign country is great, but having Siniša is priceless. I am also thankful for the great time we had working together on his thesis, for allowing me to step into the role of supervisor, and for his patience with my eccentric ideas. I am equally grateful for all the wonderful people I met through him.

The best part of EPOC was meeting other doctoral students regularly every couple of months. We formed strong friendships that I hope to cherish in the future.

I have been fortunate enough to have an amazing group of friends from around the world. Here, I want to mention some of them who supported me during the most challenging times of my PhD, with the hope that those not mentioned will forgive me.

Starting with my longest-lasting friendships, I am very grateful for all the coffee dates and board game nights in Belgrade with Ana L., Ana N., Duda, Dušan, Đina, Iva, Maksić and Milica. I must highlight how happy and proud I am of the long and close friendship I have built with Ana L. over all these years.

It is truly remarkable that some friendships formed during my one-year Master's program in Grenoble remain just as strong today, five years later, despite us living in different places. I am glad to have in my life friends like Alvaro, Johanna, Lucas, Lucrezia, Tijana, Xheni, and Zeinab. I am extremely happy to call Zeinab my friend — not only because she kindly proofread my thesis but because she is one of the most incredible human beings I know.

My year in Grenoble was significant not only because it led to friendships that would last a lifetime but also because it was where I met my best friend and soon-to-be wife, Carolina. She has been by my side during every moment of my PhD journey, both the highs and the lows, and has been my rock when I felt like falling apart. I am beyond proud and grateful to have her in my life because, among other things, her presence in my life constantly reminds me that having a fantastic career is a blessing but not the true purpose in life.

Finally, I would like to thank my parents and Vanja, to whom I dedicated this PhD thesis. They were a foundation of everything good in me today. Their unconditional acceptance, support, and love made me who I am today, and I couldn't be more proud of that.

Predrag Pilipović
Copenhagen, August 2024

Abstract

This thesis develops and analyzes advanced parameter estimation techniques for discretely observed nonlinear first- and second-order stochastic differential equations (SDEs), focusing on splitting schemes and their applications.

Initially, new numerical properties of splitting schemes, specifically the Lie-Trotter and Strang schemes, are established, enabling more accurate and robust parameter estimation under less restrictive assumptions on the drift parameter. Theoretical advancements include proving the L^p convergence of the Strang splitting scheme and demonstrating the consistency and asymptotic efficiency of the associated estimator, confirmed in a simulation study of the three-dimensional stochastic Lorenz system.

Expanding this work to second-order SDEs, we introduce and adapt the Strang splitting scheme to address hypoelliptic systems and scenarios involving partial observations caused by the unobserved velocity variable. The proposed estimators are shown to be both theoretically robust and computationally fast, with variations in the asymptotic variance depending on the likelihood approach used. The theory is illustrated by applying the Kramers oscillator model to model paleoclimate data.

The thesis further extends to developing multivariate Pearson diffusion models, which generalize existing univariate Pearson diffusion frameworks by incorporating linear drift and a quadratic function in the diffusion structure. The Strang splitting scheme for nonlinear processes with Pearson-type noise is proposed, and the closed-form solutions for the first two moments are derived. The applicability of these models is demonstrated through their appearance in genetic research and epidemiological modeling, as well as a generalization of the Kramers model with the student-type noise. The simulation studies validate the dominance of the proposed estimator in estimating diffusion parameters with higher accuracy compared to existing methods.

Sammenfatning

Denne afhandling udvikler og analyserer avancerede teknikker til at estimere parametre i diskret observerede ikke-lineære første- og andenordens stokastiske differentialligninger (SDE'er), med fokus på splitting schemes og deres anvendelser.

Indledningsvis etableres nye numeriske egenskaber for splitting schemes, specifikt Lie-Trotter- og Strang-skemaerne. Dette muliggør mere nøjagtig og robust parameterestimation under mindre restriktive antagelser om driftparameteren. Desuden bevises L^p-konvergens af Strang-skemaet og konsistens og asymptotisk normalitet af estimatoren. Resultaterne bekræftes i et simulationsstudie af det tredimensionelle stokastiske Lorenz-system.

Vi generaliserer metoderne til andenordens SDE'er og introducerer og tilpasser Strang-skemaet til hypoelliptiske systemer og scenarier, der involverer uobserverede variable, da hastighedsvariablen ikke er observeret. De foreslåede estimatorer er både teoretisk robuste og beregningsmæssigt hurtige, hvor den asymptotiske varians afhænger af hvilken approximation til likelihood funktionen, man vælger. Teorien illustreres på Kramers oscillatormodel, anvendt på palæoklimadata fra den Grønlandske indlandsis.

Afhandlingen udvider metoderne yderligere til multivariate Pearson diffusionsmodeller, som generaliserer eksisterende univariate Pearson diffusioner ved at kombinere lineær drift med en kvadratisk funktion i diffusionsstrukturen. Vi udvikler Strang-skemaet for ikke-lineære processer med Pearson-type støj, og eksplicitte formler for de første to momenter udledes. Anvendeligheden af disse modeller vises gennem eksempler i genetisk forskning og epidemiologisk modellering, såvel som en generalisering af Kramers-modellen med Pearson-type støj. Simulationsstudierne validerer fordelene ved den foreslåede estimator af diffusionsparametre med større nøjagtighed sammenlignet med eksisterende metoder.

Contributions and Structure

This PhD thesis consists of two parts.

The first part, which includes Chapters 1 to 5, outlines the problem and objectives addressed in the thesis. It provides the necessary information to understand the included papers and the additional work beyond them. Instead of serving as a literature review, this part identifies and connects the common themes across the papers, placing them within a unified framework.

Following the introduction, the second part consists of three chapters, each presenting a paper with appendices. All theorems, sections, etc., are numbered according to the paper in which they appear. The three papers are the following.

Paper I (Parameter Estimation in Nonlinear Multivariate SDEs with Additive Noise) introduces the splitting schemes estimators for elliptic SDEs with additive noise. A new asymptotic theory of the estimators is developed for SDEs with non-Lipschitz drift and super-linear growth. The chapter contains the following paper:

- [Pilipovic et al., 2024a] P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics*, 52(2):842 – 867, 2024a. doi: 10.1214/24-AOS2371. URL https://doi.org/10.1214/24-AOS2371.

Paper II (Parameter Estimation in Nonlinear Multivariate Second-order SDEs with Additive Noise) generalizes the splitting schemes estimators to a specific subclass of hypoelliptic diffusions induced by second-order stochastic differential equations. Moreover, the methodology is adapted to work in case of partial observations. The chapter contains the following paper:

- [Pilipovic et al., 2024b] P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting for parametric inference in second-order stochastic differential equations, 2024b.

  Paper status: Submitted.

Paper III (Parameter Estimation in Nonlinear Multivariate SDEs with Pearson-type Noise) introduces a new model class denoted as SDEs with Pearson-type Noise and proposes a method for estimating parameters using Strang splitting together with Gaussian approximation. The chapter contains the following paper:

- P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting parameter estimator for nonlinear multivariate stochastic differential equations with Pearson-type multiplicative noise, 2024.

  Paper status: working paper.

Contents

1 Introduction

Understanding the behavior of complex systems under the influence of random fluctuations is a fundamental challenge in various scientific fields. Our research contributes to this field, focusing on the parameter estimation problem in multivariate nonlinear stochastic differential equations (SDEs) based on discrete observations.

To demonstrate the practical value of SDE-based models, we study in Paper II the Dansgaard-Oeschger (DO) events — abrupt climatic shifts recorded during the last glacial period. DO events are characterized by sudden warming followed by gradual cooling, which takes decades to centuries to millennia. Relevant research questions about the DO events include the mechanisms driving these rapid changes, the factors influencing their occurrence, and the distribution of waiting times between such events. We investigate whether we can fit models incorporating dynamical oscillations and stochastic resonance, such as the Kramers oscillator (also known as the stochastic Duffing oscillator), to address these questions and enhance our understanding of the underlying climate dynamics.

Despite the evident potential of SDE-based models, parameter estimation of such models presents three main challenges. First, a universal framework for parameter estimation applicable to all SDE models is lacking. Most SDE models do not have a closed-form likelihood function, a fundamental element of traditional statistical inference. Second, hypoelliptic SDEs introduce additional complexities, as the noise does not directly affect all system components but can influence them indirectly through the system's dynamics. Third, only partial observations are often available, meaning not all variables affecting the system's dynamics can be directly measured. In cases of partial observation, the behavior of hidden variables must be inferred from the observed data, adding another layer of complexity to parameter estimation in SDEs.

This thesis addresses these three problems by proposing to approximate the SDE using splitting schemes. Splitting schemes are numerical methods that decompose the SDE into simpler sub-problems, which are easier to solve and can be combined to approximate the solution of the original SDE. Using these schemes, we can derive a pseudo-likelihood, which facilitates the construction of estimators without a closed-form likelihood function.

We develop estimators for different setups and model classes using the pseudo-likelihood obtained from the splitting schemes. We rigorously prove convergence properties for the splitting schemes and the asymptotic properties of the obtained estimators, specifically their consistency and asymptotic normality, allowing for standard inferential procedures.

We also conduct extensive numerical simulation studies to demonstrate the practical performance of the estimators. These studies show that our proposed methods can accurately and quickly estimate parameters even in the presence of nonlinearities, hypoellipticity, and partial observations, thus providing a robust framework for analyzing complex systems modeled by SDEs.

The rest of this part of the thesis is organized as follows. Chapter 2 introduces the model class and discusses the assumptions. In Chapter 3, we recall the maximum likelihood estimation (MLE) for SDEs and compare the most commonly applied approximated MLE methods. In Chapter 4, we introduce the splitting schemes, first in the ordinary differential equation (ODE) settings, followed by the extension to the SDE setup. Chapter 5 discusses the computational tools used to make the implementation faster and more robust.

2 Stochastic Differential Equation Model

This chapter introduces and discusses the fundamental setup of models described by SDEs. We consider the following SDE

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(1)})\,\mathrm{d}t + \boldsymbol{\Sigma}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 = \mathbf{x}_0 \in \mathcal{X}. \tag{1}$$

Here, $\mathbf{X}_t \in \mathcal{X} \subset \mathbb{R}^d$ is a unique, strong solution defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\boldsymbol{\theta}})$ with a complete, right-continuous filtration $(\mathcal{F}_t)_{t\geq 0}$. The m-dimensional Wiener process $\mathbf{W} = (\mathbf{W}_t)_{t\geq 0}$ is adapted to $\mathcal{F}_t$. The probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$. The closure of the parameter space Θ is $\overline{\Theta} = \overline{\Theta}_{\boldsymbol{\theta}^{(1)}} \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}}$, where $\Theta_{\boldsymbol{\theta}^{(1)}}$ and $\Theta_{\boldsymbol{\theta}^{(2)}}$ are two open convex bounded subsets of $\mathbb{R}^r$ and $\mathbb{R}^s$, respectively. The initial value $\mathbf{x}_0$ can be either deterministic or random. The drift function and diffusion matrix are defined as $\mathbf{F} : [0, \infty) \times \mathcal{X} \times \overline{\Theta}_{\boldsymbol{\theta}^{(1)}} \to \mathbb{R}^d$ and $\boldsymbol{\Sigma} : [0, \infty) \times \mathcal{X} \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}} \to \mathbb{R}^{d\times m}$, respectively. The matrix-valued function $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top : \mathbb{R}^d \times \overline{\Theta}_{\boldsymbol{\theta}^{(1)}} \to \mathbb{R}^{d\times d}$ defined as $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)}) = \boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})^\top$ is assumed to be positive semidefinite.

This setup represents a general form of SDE models. However, certain conditions are typically imposed on the drift $\mathbf{F}$ and diffusion function $\boldsymbol{\Sigma}$ to ensure the existence and uniqueness of a strong solution. In the context of parameter estimation problems, these conditions also help guarantee the theoretical and numerical properties of the estimators.

A unique, strong solution is one where the initial conditions and the driving noise uniquely determine the path of the SDE. In contrast, a weak solution only requires that the distribution of the paths matches the SDE. In Paper I, we are also interested in the numerical aspects of discretization schemes, which is why we base our work on the assumption of a strong solution. Moreover, we are interested in how the strong order of convergence influences parameter estimation.

According to [Kloeden and Platen, 1992], approximation $\widetilde{\mathbf{X}}$ converges strongly with order q to the strong solution $\mathbf{X}$ of SDE (1) at time T, if there exists a constant C that does not depend on h, such that

$$\mathbb{E}[\|\mathbf{X}_T - \widetilde{\mathbf{X}}_T\|] \leq Ch^q,$$

where h is the time step size between two consecutive observations. Sometimes, the strong order of convergence is defined by L^2 instead of L^1 norm as in Milstein [1988]. Then, the order of convergence is also called the mean-square convergence. In Paper I, we further explore the L^p convergence of SDE approximations. In this context, L^p convergence implies mean-square convergence when $p = 2$, and mean-square convergence implies strong order convergence.

Additionally, we examine one-step convergence, which measures the accuracy of a single discretization step. This can be expressed as

$$\|\mathbb{E}[\mathbf{X}_{t_k} - \widetilde{\Phi}_h(\mathbf{X}_{t_{k-1}}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]\| \leq Ch^q,$$

where $\widetilde{\Phi}_h$ is the one-step approximation h-flow. While strong order convergence is important to simulate trajectories, the one-step approximation error has important implications for estimating process parameters. MLE relies on the distribution of a one-step ahead state as described in Chapter 3. The exact distribution is usually unavailable, so approximation methods are required. Thus, a better convergence rate for the one-step approximation error leads to a better likelihood approximation.

In Paper I, we present novel results concerning the numerical properties of the Strang splitting scheme, highlighting how these convergence aspects correlate with the non-asymptotic accuracy of parameter estimation based on splitting schemes.

2.1 Common Assumptions

1. **Lipschitz Continuity and Linear Growth**. It is common to assume that both drift function $\mathbf{F} : [0, \infty) \times \mathcal{X} \times \overline{\Theta}_{\boldsymbol{\theta}^{(1)}} \to \mathbb{R}^d$ and diffusion function $\boldsymbol{\Sigma} : [0, \infty) \times \mathcal{X} \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}} \to \mathbb{R}^{d \times m}$ satisfy the Lipschitz continuity (2) and linear growth (3) conditions. Specifically, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, and all $t \geq 0$, there exist constants $L_{\boldsymbol{\theta}}, C_{\boldsymbol{\theta}} \in (0, \infty)$ that do not depend on t, such that

$$\|\mathbf{F}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)}) - \mathbf{F}(t, \mathbf{y}; \boldsymbol{\theta}^{(1)})\| + \|\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)}) - \boldsymbol{\Sigma}(t, \mathbf{y}; \boldsymbol{\theta}^{(2)})\| \leq L_{\boldsymbol{\theta}} \|\mathbf{x} - \mathbf{y}\|, \quad (2)$$

   and

$$\|\mathbf{F}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)})\| + \|\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})\| \leq C_{\boldsymbol{\theta}}(1 + \|\mathbf{x}\|). \quad (3)$$

   These conditions are widely adopted because they simplify the proofs of the existence and uniqueness of strong solutions and the estimators' asymptotic properties.

2. **Constant Diffusion Function**. Another common assumption is that $\boldsymbol{\Sigma}$ does not depend on state $\mathbf{x}$, that is, $\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)}) = \boldsymbol{\Sigma}(t)$. This assumption simplifies the statistical inference significantly and makes different estimators well-defined. Nonetheless, this assumption is often not realistic in practical applications.

3. **Elliptic Diffusion**. Another important assumption often made in the study of SDEs is that the diffusion function $\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})$ is elliptic. This means that the squared diffusion matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})$ is positive definite for all t and $\mathbf{x}$. This assumption ensures non-degeneracy of the diffusion term, which is crucial for some discretization schemes and estimators, as they are well-defined only if the ellipticity condition is satisfied.

These assumptions are simple. However, they limit the model's applicability in real-world scenarios requiring nonlinear growth and nonlinear state-dependent hypoelliptic diffusion functions.

2.2 Our Approach

In the three Papers I, II, and III, we assume an autonomous SDE, meaning that the functions $\mathbf{F}$ and $\mathbf{\Sigma}$ do not depend on time t. This assumption simplifies the derivations, but our approach can easily be extended to non-autonomous SDEs. Here, we provide an overview of how to generalize the autonomous assumption to handle non-autonomous SDEs. Additionally, we assume that the initial value $\mathbf{x}_0$ is deterministic. While this assumption simplifies the analysis, our method can be adapted to accommodate a random initial value. We maintain this assumption in the thesis for clarity and simplicity.

The following subsection discusses how our work relies upon and extends the three Common Assumptions 2.1.

2.2.1 Non-Lipschitz and Polynomial Drift

We do not impose the strong conditions of linear growth and Lipschitz continuity on $\mathbf{F}$. Instead, we adopt weaker conditions:

1. **One-Sided Lipschitz Condition**. Function $\mathbf{F}$ is twice continuously differentiable with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$. Additionally, for all $\boldsymbol{\theta} \in \overline{\Theta}$ $t \geq 0$, for a sufficiently large $p \geq 1$, there is a constant $L_{\boldsymbol{\theta}} > 0$ such that:

$$(\mathbf{x}-\mathbf{y})^\top (\mathbf{F}(t,\mathbf{x};\boldsymbol{\theta}^{(1)}) - \mathbf{F}(t,\mathbf{y};\boldsymbol{\theta}^{(1)})) + \frac{2p-1}{2}\sum_{i=1}^{d} \|\mathbf{\Sigma}_i(t,\mathbf{x};\boldsymbol{\theta}^{(2)}) - \mathbf{\Sigma}_i(t,\mathbf{y};\boldsymbol{\theta}^{(2)})\|^2 \leq L_{\boldsymbol{\theta}}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{X}.$$

2. **Polynomial Growth**. Function $\mathbf{F}$ grows at most polynomially in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}^{(1)}$, i.e., there exist constants $C_{\boldsymbol{\theta}^{(1)}} > 0$ and $p \geq 1$ such that for $t \geq 0$:

$$\|\mathbf{F}(t,\mathbf{x};\boldsymbol{\theta}^{(1)}) - \mathbf{F}(t,\mathbf{y};\boldsymbol{\theta}^{(1)})\|^2 \leq C_{\boldsymbol{\theta}^{(1)}}(1+\|\mathbf{x}\|^{2p-2}+\|\mathbf{y}\|^{2p-2})\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{X}.$$

   Additionally, the derivatives of $\mathbf{F}$ are of polynomial growth in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}^{(1)}$.

These weaker conditions are sufficient for a unique, strong solution to exist [Tretyakov and Zhang, 2013] and better reflect the needs in practical scenarios. We demonstrate through numerical studies on the Lorenz system (an example that does not satisfy the one-sided Lipschitz condition) that these conditions are sufficient but not necessary. The primary requirement is a unique, strong solution to the SDE (1). By adopting these more flexible assumptions, we aim to broaden the applicability of our methods while ensuring mathematical rigor.

2.2.2 Constant and Pearson-type Diffusion Function

In Papers I and II we make the strong assumption of constant diffusion function, $\mathbf{\Sigma}(t,\mathbf{X}_t;\boldsymbol{\theta}^{(2)}) = \mathbf{\Sigma}$. We refer to SDEs with constant diffusion functions as SDEs with additive noise. On the contrary, if the diffusion function depends on the state vector $\mathbf{X}$, it is said that SDE

has multiplicative noise. The additive noise assumption is necessary for the splitting schemes used in all three papers. While we formally introduce these schemes in Chapter 4, we intuitively go into the main idea here to explain why we need a constant diffusion.

Starting with SDE (1) with additive noise, we can split the drift function $\mathbf{F}$ into a sum of linear and nonlinear functions. Then, we split the original SDE (1) into two differential equations: a linear SDE with additive noise and a nonlinear ODE. Although this is not the only way to split SDE (1) and might not be optimal, it is the most natural approach. This is because then we obtain an Ornstein-Uhlenbeck (OU) process, which is the solution of the linear SDE with additive noise. The OU process is well-studied and understood with an explicit closed-form solution and known Gaussian transition probability. This splitting strategy provides a pseudo-likelihood for statistical inference for SDEs using splitting schemes. It can inspire new ideas and methods for further development of this approximation method. For example, in Paper III, we generalize the constant diffusion function assumption to allow quadratic functions of the state vector $\mathbf{X}$ in the squared diffusion matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})$. We refer to SDEs with this type of diffusion matrices as SDEs with Pearson-type noise as a generalization of Pearson diffusions, a standard and powerful class of one-dimensional models. While allowing for non-constant diffusion, one loses the well-defined properties of the OU process. However, we can still find closed-form formulas for the first two moments of a linear SDE with this type of diffusion function. These formulas allow for approximating the transition density of the linear SDE with a Gaussian density with the first two correct moments.

2.2.3 Hypoelliptic Diffusion

The SDE (1) is said to be hypoelliptic if $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})$ is not of full rank, while the solution admits a smooth transition density with respect to the Lebesgue measure. According to Hörmander's theorem [Nualart, 2006], this is fulfilled if the SDE in its Stratonovich form satisfies the weak Hörmander condition.

The Stratonovich form of SDE (1) is given as

$$\begin{aligned} \mathrm{d}\mathbf{X}_t = &\left(\mathbf{F}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(1)}) - \frac{1}{2} \sum_{k=1}^{d} \sum_{j=1}^{m} \partial_k \boldsymbol{\Sigma}_{\cdot j}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) \Sigma_{kj}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) \right) \mathrm{d}t \\ &+ \boldsymbol{\Sigma}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) \circ \mathrm{d}\mathbf{W}_t, \end{aligned} \tag{4}$$

where $\boldsymbol{\Sigma}_{\cdot j}$ is the jth column of $\boldsymbol{\Sigma}$. Then, the Stratonovich SDE (4) has drift

$$\mathbf{G}(t, \mathbf{X}_t; \boldsymbol{\theta}) \coloneqq \mathbf{F}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(1)}) - \frac{1}{2} \sum_{k=1}^{d} \sum_{j=1}^{m} \partial_k \boldsymbol{\Sigma}_{\cdot j}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) \Sigma_{kj}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}). \tag{5}$$

To describe the weak Hörmander condition, we start by introducing the Lie bracket. The Lie bracket $[\boldsymbol{f}, \boldsymbol{g}]$ of two smooth vector fields $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$[\boldsymbol{f}, \boldsymbol{g}] \coloneqq (D_{\mathbf{x}} \boldsymbol{g}(\mathbf{x})) \boldsymbol{f}(\mathbf{y}) - (D_{\mathbf{x}} \boldsymbol{f}(\mathbf{x})) \boldsymbol{g}(\mathbf{y}),$$

where $D_{\mathbf{x}}\boldsymbol{f}$ is the Jacobian matrix of function $\boldsymbol{f}$.

We define the set $\mathcal{H}$ of vector fields by initially including $\boldsymbol{\Sigma}_{\cdot j}$, $j = 1, 2, ..., d$, and then recursively adding Lie brackets

$$H \in \mathcal{H} \Rightarrow [\mathbf{G}, H], [\boldsymbol{\Sigma}_{\cdot 1}, H], \ldots, [\boldsymbol{\Sigma}_{\cdot d}, H] \in \mathcal{H}.$$

The weak Hörmander condition is met if the vectors in $\mathcal{H}$ span $\mathbb{R}^d$ at every point $(t, \mathbf{x}; \boldsymbol{\theta}) \in [0, \infty) \times \mathcal{X} \times \overline{\Theta}$.

In Paper II, we work with second-order SDEs. We show that they are hypoelliptic under the assumption of additive noise. In Paper III, we create a framework that includes hypoelliptic SDEs, also with multiplicative noise.

Now that we set up the model class, we focus on parameter estimation.

3 Parameter Estimation from Discrete Observations

Parameter estimation is a common challenge in practical modeling with SDEs. While collaboration with domain experts might provide us with the parametric form of the SDE, the specific parameter values often remain unknown. Typically, we have experimental data that we use to estimate these parameters. This section aims to give an overview of solutions to these problems, focusing specifically on statistical likelihood-based inference methods.

Working with SDEs with additive noise is much more manageable because many methods are designed explicitly for SDEs with constant diffusion coefficients. Additive noise simplifies the SDE, making it more tractable for various methods and analyses. However, many real-world systems exhibit state-dependent diffusion, leading to two possible approaches: 1) transforming the SDE to reduce the multiplicative noise to additive noise or 2) generalizing methods that work with additive noise to handle multiplicative noise. We first discuss the transformation approach, as our initial focus is on models with additive noise. In Section 3.4, we recall different estimation methods and discuss which methods can work only with additive noise and which can be used or generalized to work with multiplicative noise.

3.1 Lamperti Transform

The Lamperti transform converts an SDE with state-dependent diffusion into an SDE with unit diffusion. This transformation is always possible for univariate SDEs. Sometimes, extending the univariate Lamperti transform to the multivariate setting is possible. For example, the following theorem (Theorem 4 in [Møller and Madsen, 2010]) outlines a possible multivariate Lamperti transform.

Theorem 1 (Multivariate Lamperti Transform)**.** *Let $\mathbf{X}_t$ be a solution of SDE (1), where $\boldsymbol{\Sigma}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal elements $\Sigma^{(i,i)}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)})$ that depend only on $X_t^{(i)}$, i.e.,*

$$\Sigma^{(i,i)}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)}) = \Sigma^{(i,i)}(t, X_t^{(i)}; \boldsymbol{\theta}^{(2)}). \tag{6}$$

Then, the ith element of the Lamperti transformation

$$Y_t^{(i)} = \psi^{(i)}(t, X_t^{(i)}; \boldsymbol{\theta}^{(2)}) := \int \frac{1}{\Sigma^{(i,i)}(t, x; \boldsymbol{\theta}^{(2)})}\, \mathrm{d}x \Big|_{x=X_t^{(i)}}, \tag{7}$$

is given by the following SDE

$$\mathrm{d}Y_t^{(i)} = \left(\left. \frac{\partial \psi^{(i)}(t,x;\boldsymbol{\theta}^{(2)})}{\partial t} \right|_{x=\psi^{(i)}(t,Y_t^{(i)};\boldsymbol{\theta}^{(2)})^{-1}} + \frac{F^{(i)}(t,\psi^{(i)}(t,Y_t^{(i)};\boldsymbol{\theta}^{(2)})^{-1};\boldsymbol{\theta}^{(1)})}{\Sigma^{(i,i)}(t,\psi^{(i)}(t,Y_t^{(i)};\boldsymbol{\theta}^{(2)})^{-1};\boldsymbol{\theta}^{(2)})} \right) \mathrm{d}t + \frac{1}{2} \left. \frac{\partial \Sigma^{(i,i)}(t,x;\boldsymbol{\theta}^{(2)})}{\partial t} \right|_{x=\psi^{(i)}(t,Y_t^{(i)};\boldsymbol{\theta}^{(2)})^{-1}} \mathrm{d}t + \mathrm{d}W_t^{(i)}.$$

Applying the Lamperti transform, we obtain a so-called reduced SDE with additive noise with identity diffusion matrix

$$\mathrm{d}\mathbf{Y}_t = \mathbf{F}_{\mathbf{Y}}(t,\mathbf{Y}_t;\boldsymbol{\theta})\,\mathrm{d}t + \mathrm{d}\mathbf{W}_t, \quad \mathbf{Y}_0 = \mathbf{y}_0. \tag{8}$$

Here, the drift term $\mathbf{F}_{\mathbf{Y}}$ is expressed as a function of the original drift $\mathbf{F}$, diffusion $\boldsymbol{\Sigma}$, the transformation $\boldsymbol{\psi}$, and the transformed state $\mathbf{Y}_t$. Sometimes, we apply a similar transformation that leads to an SDE with additive noise but not necessarily with an identity diffusion matrix. This can be useful to avoid a transformation depending on the parameters we wish to estimate. We thus abuse the notation of the Lamperti transform and refer to any transformation that converts an SDE with state-dependent diffusion into an SDE with constant diffusion.

We note that the Lamperti transform is applicable in more general multivariate SDE settings with a more complex structure of diffusion matrix $\boldsymbol{\Sigma}$. See, for example, Aït-Sahalia [2008] and his necessary and sufficient condition for a multivariate Lamperti transform to exist.

In the next section, we assume that the SDE has additive noise, either by model construction or, if possible, obtained from the Lamperti transform.

3.2 Stochastic Differential Equations with Additive Noise

To compare common parameter estimators in Section 3.4 and introduce the splitting schemes in Chapter 4, we assume that SDE (1) has additive noise, that is,

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}(t,\mathbf{X}_t;\boldsymbol{\theta}^{(1)})\,\mathrm{d}t + \boldsymbol{\Sigma}(t;\boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 = \mathbf{x}_0 \in \mathcal{X}. \tag{9}$$

Instead of estimating $\boldsymbol{\theta}^{(2)}$ directly, we only estimate parameters in $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t;\boldsymbol{\theta}^{(2)})$ because the covariance matrix of $\mathbf{X}_t$ depends on $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. For any orthogonal matrix $\mathbf{Q}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}\mathbf{Q}$ induce the same distribution due to the properties of the Wiener process $\mathbf{W}_t$. More formally, any two matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ that satisfy $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top = \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top$ generate the same covariance structure for $\mathbf{X}_t$. Hence, the matrix $\boldsymbol{\Sigma}$ is only identifiable up to orthogonal transformations. That means that we cannot uniquely determine $\boldsymbol{\Sigma}$ itself, but rather we can only estimate the equivalence class of $\boldsymbol{\Sigma}$ defined by $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$.

For that reason, we half-vectorize $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ as

$$\boldsymbol{\varsigma} \coloneqq \mathrm{vech}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) = ([\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{11}, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{12}, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{22}, ..., [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{1d}, ..., [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{dd}).$$

Since $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is a symmetric $d \times d$ matrix, $\boldsymbol{\varsigma}$ is of dimension $s = d(d+1)/2$. For a diagonal matrix, instead of a half-vectorization, we use $\boldsymbol{\varsigma} \coloneqq \mathrm{diag}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$. Then, when we refer to the diffusion parameter, $\boldsymbol{\theta}^{(2)}$, we refer to the parameter of $\boldsymbol{\varsigma}(t;\boldsymbol{\theta}^{(2)})$.

3.3 Maximum Likelihood Estimation

We might have a set of observations of the state $\mathbf{X}_t$ at a finite number of time points. Alternatively, we might only have partial observations of the state possibly corrupted by noise. In this thesis, we focus on the case where we observe the state $\mathbf{X}_t$ directly without additional observational errors. When dealing with partial observations, we assume a specific structure indicating which state components are observed and which are not. This structure is motivated by the second-order SDE framework, which systematically handles partial observations in the modeling process.

Assume we observe $N+1$ values of the SDE, $\mathbf{X}_{t_0}, \mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_N}$. A classical method for SDE parameter estimation is the MLE. We assume the observations are equidistant, i.e., $t_{k+1} - t_k =: h$ for all $k = 0, 1, \ldots, N-1$. Additionally, we consider high-frequency asymptotics where the number of observations N goes to infinity, the step size h goes to zero, and the length of observed time interval $T = Nh$ goes to infinity. This last condition is necessary to prove the consistency of the drift parameter. Moreover, when proving the estimator's asymptotic normality, we require that Nh^2 goes to zero, a condition sometimes referred to as a rapidly increasing experimental design.

Due to the Markov property of SDEs, we can write down the likelihood of the observed values given the parameters as follows

$$p(\mathbf{X}_{t_1}, \ldots \mathbf{X}_{t_N} \mid \boldsymbol{\theta}) = \prod_{k=0}^{N-1} p(\mathbf{X}_{t_{k+1}} \mid \mathbf{X}_{t_k}; \boldsymbol{\theta}),$$

where $p(\mathbf{X}_{t_{k+1}} \mid \mathbf{X}_{t_k}; \boldsymbol{\theta})$ is the transition density of the SDE and can be obtained as a solution to the Kolmogorov forward equation.

In the MLE method, we wish to maximize the preceding likelihood expression or, equivalently, minimize the negative log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\mathbf{X}_{t_1}, \ldots \mathbf{X}_{t_N} \mid \boldsymbol{\theta}) = -\sum_{k=0}^{N-1} \log p(\mathbf{X}_{t_{k+1}} \mid \mathbf{X}_{t_k}; \boldsymbol{\theta}). \tag{10}$$

Thus, the MLE of the parameters is obtained by finding the vector of parameters that minimizes the negative log-likelihood $\mathcal{L}(\boldsymbol{\theta})$, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}).$$

The minimum can be computed analytically by setting derivatives to zero or using numerical optimization methods. However, we need to evaluate the likelihood, which is generally intractable because the transition densities are not available in closed form. For linear SDEs, we know the transition densities, allowing explicit evaluation of the likelihood. We cannot analytically solve the Kolmogorov forward equation in the multivariate nonlinear case, making the transition density intractable. In this situation, a typical approach is to replace the SDE or its transition density with a tractable approximation. We can use various SDE discretization methods, such as Euler-Maruyama

(EM), strong order 1.5 scheme, or local linearization (LL), to form an SDE approximation whose transition density we can evaluate. Alternatively, we can directly approximate the transition density of the SDE, for example, using Gaussian approximations or Hermite expansions.

The following section briefly overviews different MLE approximation methods. It serves as a literature review and a comparative analysis of other approaches. By examining these methods, we can evaluate their effectiveness, understand the assumptions they rely on, and determine how they can be generalized or adapted for specific classes of models. This discussion highlights the similarities and differences between various approximation techniques, their outcomes, and their practical implementations.

3.4 Approximate Maximum Likelihood Estimators

The intractability of the transition density challenges parameter estimation in nonlinear SDEs. We use approximations for the likelihood or transition densities to facilitate evaluating the likelihood function. This section explores various methods for approximating the solution of SDE or its transition density to enable maximum likelihood estimation.

In the rest of this section, we suppress the notation for the parameter, so for example, $\mathbf{F}(t, \mathbf{x})$ stands for $\mathbf{F}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)})$, and so on.

3.4.1 Estimators Based on Approximated Solutions

One approach to parameter estimation in nonlinear SDEs is to approximate the solution of the SDE with a continuous- or discrete-time system for which the transition density is known or can be computed more easily. Below, we discuss three standard methods: the EM, the strong order 1.5, and LL estimators.

3.4.1.1 Euler-Maruyama Estimator

The EM method is a straightforward discretization scheme used to approximate the solution of SDEs. It originates from the Itô-Taylor expansion, which provides a framework for approximating the solution of an SDE by expanding it in terms of the increments of the Wiener process [Kloeden and Platen, 1992]. The EM method uses only the first-order terms of the expansion.

The EM approximation of the SDE (1) is defined as follows

$$\hat{\mathbf{X}}_{t_{k+1}} = \hat{\mathbf{X}}_{t_k} + h\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k}) + \boldsymbol{\Sigma}(t_k, \hat{\mathbf{X}}_{t_k})\Delta\mathbf{W}_k,$$

where $\Delta\mathbf{W}_k \sim \mathcal{N}(\mathbf{0}, h)$. The simplicity of this method makes it a versatile tool, as it does not restrict the diffusion function $\boldsymbol{\Sigma}$ to be constant and can be easily applied to SDEs of any dimension. Furthermore, the method avoids the need to compute the Jacobian, making it computationally fast and easy to implement.

The discrete process $\hat{\mathbf{X}}_{t_k}$ has a Gaussian transition density, leading to the transition density approximation:

$$p(\mathbf{x}_{t_{k+1}} \mid \mathbf{x}_{t_k}; \boldsymbol{\theta}) \approx \mathcal{N}\left(\mathbf{x}_{t_{k+1}}; \boldsymbol{\mu}_h^{[\mathrm{EM}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}), \boldsymbol{\Omega}_h^{[\mathrm{EM}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(2)})\right),$$

where

$$\begin{aligned}\boldsymbol{\mu}_h^{[\mathrm{EM}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) &= \mathbf{x}_{t_k} + h\mathbf{F}(t_k, \mathbf{x}_{t_k}),\\ \boldsymbol{\Omega}_h^{[\mathrm{EM}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(2)}) &= h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k}).\end{aligned}$$

This results in the following approximation of the negative log-likelihood:

$$\begin{aligned}\mathcal{L}^{[\mathrm{EM}]}(\mathbf{X}; \boldsymbol{\theta}) = &\frac{1}{2}\sum_{k=0}^{N-1} \log\det(2\pi\boldsymbol{\Omega}_h^{[\mathrm{EM}]}(t_k, \mathbf{X}_{t_k}))\\ &+ \frac{1}{2}\sum_{k=0}^{N-1}(\mathbf{X}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{EM}]}(t_k, \mathbf{X}_{t_k}))^\top \boldsymbol{\Omega}_h^{[\mathrm{EM}]}(t_k, \mathbf{X}_{t_k})^{-1}(\mathbf{X}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{EM}]}(t_k, \mathbf{X}_{t_k})).\end{aligned}$$

While the method is highly straightforward to implement and seemingly versatile with a non-constant diffusion function, it does have some major limitations. For instance, Hutzenthaler et al. [2010] proved that the EM discretization diverges from the true solution of SDEs with super-linear drift terms. Moreover, it is unsuitable for hypoelliptic SDEs, as it approximates the covariance by $h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$, which is not of full rank.

Even for linear SDEs that have explicit transition densities, the EM approximation does not have the correct covariance matrix since the true covariance is not $h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$.

For all the previous reasons, the EM method can lead to significant bias in the parameter estimators.

3.4.1.2 Strong Order 1.5 Estimator

By taking more terms into the Itô-Taylor expansion, we can form methods of arbitrary order. One such method is the strong order 1.5 scheme (SO1.5) [Kloeden and Platen, 1992], which includes noise terms up to order $h^{3/2}$ and deterministic terms up to order h^2. The SO1.5 scheme can be used to discretize SDEs with non-constant noise because all involved iterated Itô integrals have known distributions. However, higher-order iterated Itô integrals in the SO1.5 scheme become challenging since the final noise distribution consists of a summation of different distributions like Gaussian and chi-square, making it impossible to find the closed-form likelihood.

Thus, we only consider the SO1.5 approximation for SDEs with additive noise:

$$\begin{aligned}\hat{\mathbf{X}}_{t_{k+1}} = &\hat{\mathbf{X}}_{t_k} + h\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k}) + \boldsymbol{\Sigma}(t_k)\Delta\mathbf{W}_k\\ &+ \frac{h^2}{2}\left(\partial_t\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k})\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k}) + \frac{1}{2}\sum_{i,j=1}^{d}\partial_{i,j}^2\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k)]_{ij}\right)\\ &+ D_{\mathbf{x}}\mathbf{F}(t_k, \hat{\mathbf{X}}_{t_k})\boldsymbol{\Sigma}(t_k)\Delta\boldsymbol{\zeta}_k + \boldsymbol{\Sigma}'(t_k)(h\Delta\mathbf{W}_k - \Delta\boldsymbol{\zeta}_k),\end{aligned}$$

where $\Delta\boldsymbol{\zeta}_k$ and $\Delta\mathbf{W}_k$ are jointly normally distributed as

$$\begin{bmatrix} \Delta\boldsymbol{\zeta}_k \\ \Delta\mathbf{W}_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \frac{h^3}{3}\mathbf{I} & \frac{h^2}{2}\mathbf{I}h \\ \frac{h^2}{2}\mathbf{I} & h\mathbf{I} \end{bmatrix} \right). \tag{11}$$

This additional noise structure allows the method to capture more complex details of the SDE's dynamics, leading to a more accurate approximation.

The transition density is approximated as

$$p(\mathbf{x}_{t_{k+1}} \mid \mathbf{x}_{t_k}; \boldsymbol{\theta}) \approx \mathcal{N}\left(\mathbf{x}_{t_{k+1}}; \boldsymbol{\mu}_h^{[\mathrm{SO1.5}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}), \boldsymbol{\Omega}_h^{[\mathrm{SO1.5}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta})\right),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_h^{[\mathrm{SO1.5}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}) &= \mathbf{x}_{t_k} + h\mathbf{F}(t_k, \mathbf{x}_{t_k}) + \frac{h^2}{2}\left(\partial_t \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\mathbf{F}(t_k, \mathbf{x}_{t_k})\right) \\
&\quad + \frac{h^2}{4}\sum_{i,j=1}^{d} \partial_{i,j}^2 \mathbf{F}(t_k, \mathbf{x}_{t_k})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k)]_{ij}, \\
\boldsymbol{\Omega}_h^{[\mathrm{SO1.5}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}) &= h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k) + \frac{h^2}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k)\right) \\
&\quad + \frac{h^2}{2}\partial_t \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k) + \frac{h^3}{3}\left(D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) + \partial_t \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k)\right) \\
&\quad + \frac{h^3}{6}\left(\boldsymbol{\Sigma}'(t_k)\boldsymbol{\Sigma}(t_k)^\top D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\boldsymbol{\Sigma}(t_k)\boldsymbol{\Sigma}'(t_k)^\top\right).
\end{aligned}$$

Under the assumption of a constant diffusion function, the SO1.5 method provides an accurate MLE approximation even in hypoelliptic scenarios [Ditlevsen and Samson, 2019]. It offers significantly higher precision than the EM method. However, despite its increased accuracy, this method has two practical issues. First, the covariance matrix $\boldsymbol{\Omega}_h^{[\mathrm{SO1.5}]}$ might not be positive definite. This issue is more likely to occur for large step size h. A common way to avoid this is to approximate $\log\det \boldsymbol{\Omega}_h^{[\mathrm{SO1.5}]}$ and $(\boldsymbol{\Omega}_h^{[\mathrm{SO1.5}]})^{-1}$ around $h = 0$ using a Taylor expansion. Second, the SO1.5 method requires additional computations, specifically the computation and implementation of the Jacobian and Hessian of the drift function $\mathbf{F}$ and the derivative of $\boldsymbol{\Sigma}$, making it computationally more complex and slower than the EM method.

3.4.1.3 Local Linearization Estimator

The LL method offers an alternative approach to approximating solutions to SDEs with constant diffusion functions [Ozaki, 1985, Shoji and Ozaki, 1998, Shoji, 1998, Ozaki et al., 2000]. This method involves approximating the nonlinear SDE locally (between each two consecutive points) with a linear one, allowing for a more tractable solution.

The basic idea behind the LL is to approximate SDE (8) in the interval $[t, t+h)$ by the following linear SDE

$$\mathrm{d}\hat{\mathbf{X}}_s = \mathbf{A}(t, \hat{\mathbf{X}}_t; \boldsymbol{\theta}^{(1)})\hat{\mathbf{X}}_s\,\mathrm{d}s + \mathbf{a}(s, t, \hat{\mathbf{X}}_t, \boldsymbol{\theta})\,\mathrm{d}s + \boldsymbol{\Sigma}(s)\,\mathrm{d}\mathbf{W}_s, \qquad s \in [t, t+h), \tag{12}$$

where functions $\mathbf{A}$ and $\mathbf{a}$ are derived by linearizing the drift $\mathbf{F}$. The linear SDE (12) has the following solution

$$\hat{\mathbf{X}}_{t+h} = \exp(\mathbf{A}(t,\hat{\mathbf{X}}_t)h)\hat{\mathbf{X}}_t + \int_t^{t+h} \exp(\mathbf{A}(t,\hat{\mathbf{X}}_t)(t+h-u))\mathbf{a}(u,t,\hat{\mathbf{X}}_t;\boldsymbol{\theta})\,\mathrm{d}u + \int_t^{t+h} \exp(\mathbf{A}(t,\hat{\mathbf{X}}_t)(t+h-u))\boldsymbol{\Sigma}(u)\,\mathrm{d}\mathbf{W}_u.$$

To derive $\mathbf{A}$ and $\mathbf{a}$, we start by linearizing $\mathbf{F}$ in each interval $[t, t+h)$ using the Itô formula. Specifically, we get

$$\mathbf{A}(t,\hat{\mathbf{X}}_t) = D_{\mathbf{x}}\mathbf{F}(t,\hat{\mathbf{X}}_t),$$

and

$$\mathbf{a}(s,t,\hat{\mathbf{X}}_t,\boldsymbol{\theta}) = \mathbf{F}(t,\hat{\mathbf{X}}_t) - D_{\mathbf{x}}\mathbf{F}(t,\hat{\mathbf{X}}_t)\hat{\mathbf{X}}_t + \left(\partial_t\mathbf{F}(t,\hat{\mathbf{X}}_t) + \frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{i,j}\mathbf{F}(t,\hat{\mathbf{X}}_t)[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t)]_{ij}\right)(s-t).$$

By construction, the LL yields a Gaussian transition density. After some algebraic manipulations, the approximation of the transition density becomes:

$$p(\mathbf{x}_{t_{k+1}} \mid \mathbf{x}_{t_k};\boldsymbol{\theta}) \approx \mathcal{N}\left(\mathbf{x}_{t_{k+1}};\boldsymbol{\mu}_h^{[\mathrm{LL}]}(t_k,\mathbf{x}_{t_k};\boldsymbol{\theta}^{(1)}),\boldsymbol{\Omega}_h^{[\mathrm{LL}]}(t_k,\mathbf{x}_{t_k};\boldsymbol{\theta})\right),$$

where

$$\begin{aligned}\boldsymbol{\mu}_h^{[\mathrm{LL}]}(t_k,\mathbf{x}_{t_k};\boldsymbol{\theta}) &= \mathbf{x}_{t_k} + \mathbf{r}_h^0(D_{\mathbf{x}}\mathbf{F}(t_k,\mathbf{x}_{t_k}))\mathbf{F}(t_k,\mathbf{x}_{t_k}) \\ &\quad + \left(h\mathbf{r}_h^0(D_{\mathbf{x}}\mathbf{F}(t_k,\mathbf{x}_{t_k})) - \mathbf{r}_h^1(D_{\mathbf{x}}\mathbf{F}(t_k,\mathbf{x}_{t_k}))\right) \\ &\quad \cdot \left(\partial_t\mathbf{F}(t_k,\mathbf{x}_{t_k}) + \frac{1}{2}\sum_{i,j=1}^{d}\partial^2_{i,j}\mathbf{F}(t_k,\mathbf{x}_{t_k})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k)]_{ij}\right), \\ \boldsymbol{\Omega}_h^{[\mathrm{LL}]}(t_k,\mathbf{x}_{t_k};\boldsymbol{\theta}) &= \int_{t_k}^{t_{k+1}} \exp(D_{\mathbf{x}}\mathbf{F}(t_k,\mathbf{x}_{t_k})(t_{k+1}-u))\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(u) \\ &\qquad \exp(D_{\mathbf{x}}\mathbf{F}^\top(t_k,\mathbf{x}_{t_k})(t_{k+1}-u))\,\mathrm{d}u.\end{aligned} \tag{13}$$

In the previous equation, we introduced the following notation

$$\mathbf{r}_h^n(\mathbf{M}) \coloneqq \int_0^h \exp(\mathbf{M}u)u^n\,\mathrm{d}u, \qquad n = 0, 1. \tag{14}$$

Integrals $\mathbf{r}_h^n$ can be computed in various ways, leading to different numerical implementations of the LL scheme. One approach is to solve the integrals analytically. However, this involves finding the inverse of the Jacobian matrix, yielding numerical instability because it depends on the data points and parameters, making it likely that the inverse

does not exist. Alternatively, we can express these integrals using matrix exponentials, resulting in a more stable numerical implementation.

If $\mathbf{\Sigma}$ does not depend on time t, the integral in the covariance matrix $\mathbf{\Omega}_h^{[\mathrm{LL}]}$ also can be represented using only the matrix exponential, which enhances the stability and efficiency of the computation. Chapter 5 discusses this approach in more detail. For diffusion matrix $\mathbf{\Sigma}$ that depends on t, we can Taylor-expand it and then use the same technique as in the case of constant $\mathbf{\Sigma}$.

It is not hard to see that the LL discretization is of strong order 1.5 as proved in [Jimenez and Biscay, 2002]. Namely, by taking the Taylor expansion of $\boldsymbol{\mu}_h^{[\mathrm{LL}]}$, we get

$$\begin{aligned}\boldsymbol{\mu}_h^{[\mathrm{LL}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}) = \mathbf{x}_{t_k} + h\mathbf{F}(t_k, \mathbf{x}_{t_k}) + \frac{h^2}{2}\left(\partial_t \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\mathbf{F}(t_k, \mathbf{x}_{t_k})\right) \\ + \frac{h^2}{4}\sum_{i,j=1}^{d} \partial_{i,j}^2 \mathbf{F}(t_k, \mathbf{x}_{t_k})[\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k)]_{ij} + \mathbf{R}(h^3, t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}),\end{aligned}$$

where $\mathbf{R}(h^3, t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)})$ is a generic notation for residuals of order at least h^3. We observe that it matches $\boldsymbol{\mu}_h^{[\mathrm{SO1.5}]}$ up to order h^3. Interestingly, the covariance matrices $\mathbf{\Omega}^{[\mathrm{LL}]}$ and $\mathbf{\Omega}_h^{[\mathrm{SO1.5}]}$ coincide only up to order h^2 as it can be seen in the following approximation

$$\begin{aligned}\mathbf{\Omega}_h^{[\mathrm{LL}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}) = h\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) + \frac{h^2}{2}\left(\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k)\right) \\ + \frac{h^2}{2}\partial_t \mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) + \frac{h^3}{6}\left(\partial_t \mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k}\partial_t \mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k))\right) \\ + \frac{h^3}{6}\left(\partial_t^2 \mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) + 2\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k}))\right) \\ + \frac{h^3}{6}\left(\mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k) D_{\mathbf{x}}^\top \mathbf{F}(t_k, \mathbf{x}_{t_k})^2 + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})^2 \mathbf{\Sigma}\mathbf{\Sigma}^\top(t_k)\right) + \mathbf{R}(h^4, t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}).\end{aligned}$$

The LL estimator performs well for SDEs with constant diffusion terms and nonlinear drift, providing robust parameter estimation even in complex scenarios such as for hypoelliptic diffusions [Melnykova, 2020]. One notable advantage of the LL scheme is that it produces the correct solution for linear SDEs, a property not shared by discretizations based on the Itô-Taylor expansion. Moreover, unlike the SO1.5 scheme, the covariance matrix is always positive definite within the parameter space, regardless of the data points or step size h. Thus, the LL estimator is numerically stable and reliable across various scenarios.

The need to compute the integrals $\mathbf{r}_h^n$ and the covariance matrix $\mathbf{\Omega}_h^{[\mathrm{LL}]}$ based on the Jacobian matrix makes the LL scheme slower than more straightforward methods. This increased computational intensity is a trade-off for the method's improved precision and robustness.

3.4.2 Estimators Based on Approximated Transition Densities

Directly approximating the SDE's transition density is an alternative to discretizing it. The most common methods include Gaussian approximations and Hermite expansions.

3.4.2.1 Gaussian Approximated-Likelihood Estimator

Previous sections introduced three methods based on discretization methods that yield Gaussian likelihoods. Alternatively, we can start by assuming that the transition density is Gaussian and derive the likelihood from this premise. The method begins with a Gaussian approximation (GA) of the transition density, expressed as

$$p(\mathbf{x}_{t_{k+1}} \mid \mathbf{x}_{t_k}; \boldsymbol{\theta}) \approx \mathcal{N}(\mathbf{x}_{t_{k+1}}; \boldsymbol{\mu}^{[\mathrm{GA}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}), \boldsymbol{\Omega}_h^{[\mathrm{GA}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta})).$$

Since the first and second moments of SDE (1) are generally unknown, we can approximate them using a Taylor expansion [Kessler, 1997][1]. We approximate the mean and covariance functions using the Taylor series expansion of

$$h \mapsto \mathbb{E}\left[\boldsymbol{\phi}(t+h, \mathbf{X}_{t+h}; \boldsymbol{\theta}) \mid \mathbf{X}_t = \mathbf{x}\right]$$

for some function $\boldsymbol{\phi}$. Consider the Taylor series expansion centered at $h = 0$ for the expectation of the function $\boldsymbol{\phi}$ of the state variable $\mathbf{X}$

$$\mathbb{E}[\boldsymbol{\phi}(t+h, \mathbf{X}_{t+h}; \boldsymbol{\theta})] = \sum_{n=0}^{\infty} \frac{h^n}{n!} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})].$$

Itô lemma yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})] = \mathbb{E}[\mathbb{L}\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})],$$

where $\mathbb{L}$ is the generalized[2] infinitesimal generator of SDE (1)

$$\begin{aligned} \mathbb{L}\boldsymbol{\phi}(t, \mathbf{x}; \boldsymbol{\theta}) = \partial_t \boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta}) + D_{\mathbf{x}} \boldsymbol{\phi}(t, \mathbf{x}; \boldsymbol{\theta}) \mathbf{F}(t, \mathbf{x}; \boldsymbol{\theta}) \\ + \frac{1}{2} \sum_{i,j=1}^{d} \partial_{i,j}^2 \boldsymbol{\phi}(t, \mathbf{x}; \boldsymbol{\theta}) [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})]_{ij}. \end{aligned}$$

Now, repeating the same idea, we derive second-order derivatives

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})] = \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})] \right) = \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[\mathbb{L}\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})] = \mathbb{E}[\mathbb{L}^2 \boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})].$$

Inductively,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})] = \mathbb{E}[\mathbb{L}^n \boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})].$$

Thus, the Taylor series expansion centered at $h = 0$ becomes

$$\mathbb{E}[\boldsymbol{\phi}(t+h, \mathbf{X}_{t+h}; \boldsymbol{\theta})] = \sum_{n=0}^{\infty} \frac{h^n}{n!} \mathbb{E}[\mathbb{L}^n \boldsymbol{\phi}(t, \mathbf{X}_t; \boldsymbol{\theta})].$$

[1] In Papers I-III, we refer to the GA estimator as the Kessler (K) estimator.
[2] Infinitesimal generator plus the time derivative.

Conditioning on $\mathbf{X}_t = \mathbf{x}$, the expectations in the series disappear and we get

$$\mathbb{E}[\boldsymbol{\phi}(t+h, \mathbf{X}_{t+h}; \boldsymbol{\theta}) \mid \mathbf{X}_t = \mathbf{x}] = \sum_{n=0}^{\infty} \frac{h^n}{n!} \mathbb{L}^n \boldsymbol{\phi}(t, \mathbf{x}; \boldsymbol{\theta}).$$

To approximate the conditional mean, choose $\boldsymbol{\phi}(t, \mathbf{x}) = \mathbf{x}$ and get the following approximation up to order h^2

$$\begin{aligned}\boldsymbol{\mu}_h^{[\mathrm{GA2}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) = \mathbf{x}_{t_k} + h\mathbf{F}(t_k, \mathbf{x}_{t_k}) + \frac{h^2}{2}\left(\partial_t \mathbf{F}(t_k, \mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})\mathbf{F}(t_k, \mathbf{x}_{t_k})\right) \\ + \frac{h^2}{4} \sum_{i,j=1}^{d} \partial_{i,j}^2 \mathbf{F}(t_k, \mathbf{x}_{t_k})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k})]_{ij}.\end{aligned}$$

Similarly, to approximate the covariance, choose $\boldsymbol{\phi}(t, \mathbf{x}) = \mathbf{x}\mathbf{x}^\top$, and use the formula

$$\mathrm{Cov}(\mathbf{X}_t) = \mathbb{E}[\mathbf{X}_t \mathbf{X}_t^\top] - \mathbb{E}[\mathbf{X}_t]\mathbb{E}[\mathbf{X}_t]^\top.$$

Then, the approximated conditional covariance matrix up to order h^2 is

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{GA2}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}) &= h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k}) \\ &+ \frac{h^2}{2}\left(D_{\mathbf{x}}\mathbf{F}(t, \mathbf{x}_{t_k})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k}) D_{\mathbf{x}}^\top \mathbf{F}(t, \mathbf{x}_{t_k}) + \partial_t \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k})\right) \\ &+ \frac{h^2}{2}\left(\sum_{i=1}^{d} \partial_i \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k}) F^{(i)}(t_k, \mathbf{x}_{t_k}) + \frac{1}{2}\sum_{i,j=1}^{d} \partial_{i,j}^2 \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t_k, \mathbf{x}_{t_k})]_{ij}\right).\end{aligned}$$

Under the assumption of additive noise, the GA method up to order h^2 provides the same transition density as the SO1.5 scheme. This result illustrates how starting with different ideas can produce the same results. However, unlike SO1.5, the GA method can incorporate multiplicative noise. Additionally, we can add more terms to expand the conditional mean and covariance at the price of increased derivation complexity. Specifically, we advise using a symbolic computation tool like `Mathematica` Wolfram Research, Inc. to derive higher-order GA approximation. The same issues encountered with the SO1.5 method, such as possible non-positive definite covariance matrices, are also the case of the GA method.

This idea of Gaussian approximation and expanding the first two moments of SDE (1) solution inspired a lot of research and was foundational for estimators proposed among others by [Uchida and Yoshida, 2012, Hurn et al., 2013, Gloter and Yoshida, 2020, Iguchi and Beskos, 2023]. Moreover, this idea is vital in Paper III that combines the splitting scheme with a Gaussian approximation.

3.4.2.2 Hermite Expansion Estimator

Another groundbreaking idea is the Hermite expansion (HE) method [Aït-Sahalia, 2002, 2008], which has motivated extensive research and the development of various parameter

estimators for SDEs. This method approximates the transition density of SDE (1) by Hermite series expansion. While the HE method can be applied in non-reducible cases, i.e., for SDEs that cannot be reduced using the Lamperti transform (7), for illustrative purposes, we only focus on scenarios where the SDE (1) can be reduced to a unit diffusion process $\mathbf{Y}_t$ (8). Moreover, this method can be applied to time-inhomogeneous SDEs [Choi, 2013, 2015]. However, in this section, we assume that SDE (1) is autonomous, i.e., the drift and the diffusion functions do not depend on time t.

The core idea is to transform the original SDE into a unit diffusion process, bringing the transition density closer to a Gaussian distribution. This transformation is followed by conditional standardization, which allows an approximation of the transition density as a standard Gaussian. Adding terms in the Hermite expansion further improves the accuracy of this approximation.

The HE method can be categorized into two primary types: the finite expansion, which uses a limited number of Hermite polynomials, and the infinite expansion, where the series of Hermite polynomials extends to infinity. In this discussion, we illustrate the latter, demonstrating how an infinite series of Hermite polynomials can approximate the transition density precisely.

When the SDE (1) is reducible to a unit diffusion process $\mathbf{Y}_t$ (8), the negative log-likelihood of $\mathbf{Y}_t$ can be approximated up to order h^J as

$$\mathcal{L}^{[\mathrm{HE}]}(\mathbf{Y};\boldsymbol{\theta}) = \frac{d}{2}\sum_{k=0}^{N-1}\log(2\pi h) + \sum_{k=0}^{N-1}\frac{C_{\mathbf{Y}}^{(-1)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{h} + \sum_{k=0}^{N-1}\sum_{j=0}^{J}\frac{h^j}{j!}C_{\mathbf{Y}}^{(j)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}),$$

where the coefficients $C_Y^{(j)}$, $j = -1, 0, 1, ..., J$ are obtained from the Kolmogorov backward and forward equations as

$$\begin{aligned}
C_{\mathbf{Y}}^{(-1)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}) &= \frac{1}{2}\|\mathbf{y}_{t_{k+1}} - \mathbf{y}_k\|^2 = \frac{1}{2}\sum_{i=1}^{d}(y_{t_{k+1}}^{(i)} - y_{t_k}^{(i)})^2,\\
C_{\mathbf{Y}}^{(0)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}) &= \sum_{i=1}^{d}(y_{t_{k+1}}^{(i)} - y_{t_k}^{(i)})\int_0^1 F_{\mathbf{Y}}^{(i)}(\mathbf{y}_{t_k} + u(\mathbf{y}_{t_{k+1}} - \mathbf{y}_{t_k}) \mid \mathbf{y}_{t_k})\,\mathrm{d}u,\\
C_{\mathbf{Y}}^{(j)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}) &= j\int_0^1 G_{\mathbf{Y}}^{(j)}(\mathbf{y}_{t_k} + u(\mathbf{y}_{t_{k+1}} - \mathbf{y}_{t_k}) \mid \mathbf{y}_{t_k})u^{j-1}\,\mathrm{d}u, \qquad j = 1, 2, ...J.
\end{aligned}$$

The functions $G_{\mathbf{Y}}^{(j)}$ are calculated as follows

$$
\begin{aligned}
G_{\mathbf{Y}}^{(1)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}) &= \sum_{i=1}^{d} \frac{\partial F_{\mathbf{Y}}^{(i)}(\mathbf{y}_{t_{k+1}})}{\partial y^{(i)}} + \sum_{i=1}^{d} F_{\mathbf{Y}}^{(i)}(\mathbf{y}_{t_{k+1}}) \frac{\partial C_{\mathbf{Y}}^{(0)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial y^{(i)}} \\
&\quad - \frac{1}{2} \sum_{i=1}^{d} \left(\frac{\partial^2 C_{\mathbf{Y}}^{(0)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial (y^{(i)})^2} + \left(\frac{\partial C_{\mathbf{Y}}^{(0)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial y^{(i)}} \right)^2 \right), \\
G_{\mathbf{Y}}^{(j)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k}) &= \sum_{i=1}^{d} F_{\mathbf{Y}}^{(i)}(\mathbf{y}_{t_{k+1}}) \frac{\partial C_{\mathbf{Y}}^{(j-1)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial y^{(i)}} - \frac{1}{2} \sum_{i=1}^{d} \frac{\partial^2 C_{\mathbf{Y}}^{(j-1)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial (y^{(i)})^2} \\
&\quad - \frac{1}{2} \sum_{i=1}^{d} \sum_{m=0}^{j-1} \binom{j-1}{m} \frac{\partial C_{\mathbf{Y}}^{(m)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial y^{(i)}} \frac{\partial C_{\mathbf{Y}}^{(j-1-m)}(\mathbf{y}_{t_{k+1}} \mid \mathbf{y}_{t_k})}{\partial y^{(i)}}, j \geq 2.
\end{aligned}
$$

Finally, use the change of variable formula together with the Lamperti transform $\boldsymbol{\psi}$ (7) to obtain the HE approximated negative log-likelihood based on $\mathbf{X}_t$

$$
\begin{aligned}
\mathcal{L}^{[\mathrm{HE}]}(\mathbf{X}; \boldsymbol{\theta}) &= \frac{1}{2} \sum_{k=0}^{N-1} \log \det \left(2\pi h \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{\top}(\mathbf{x}_{t_k}; \boldsymbol{\theta}) \right) + \frac{1}{2} \sum_{k=0}^{N-1} \frac{\| \boldsymbol{\psi}(\mathbf{x}_{t_{k+1}}; \boldsymbol{\theta}) - \boldsymbol{\psi}(\mathbf{x}_{t_k}; \boldsymbol{\theta}) \|^2}{h} \\
&\quad + \sum_{k=0}^{N-1} \sum_{j=0}^{J} \frac{h^j}{j!} C_{\mathbf{Y}}^{(j)}(\boldsymbol{\psi}(\mathbf{x}_{t_{k+1}}; \boldsymbol{\theta}) \mid \boldsymbol{\psi}(\mathbf{x}_{t_k}; \boldsymbol{\theta}); \boldsymbol{\theta}).
\end{aligned}
$$

A general implementation of this method is highly complex due to iterative calculation of coefficients $C_Y^{(j)}$. In practice, the HE approximation is used up to order h^2, and as in the case of the GA method, it is highly advisable to use symbolic computation tools. In Paper I, we compare our proposed method to the HE estimator and illustrate poor performance for larger h of this method due to small order $J = 2$.

4 Splitting Schemes

This section introduces splitting schemes and their induced estimators. It starts by motivating splitting schemes in deterministic differential equations and then extends the idea to SDEs with additive noise. It also compares the new estimators to those in Section 3.4. Finally, the section ends by extending the idea to SDEs with Pearson-type noise.

First, we assume that the diffusion matrix is zero, $\boldsymbol{\Sigma} \equiv 0$, reducing the SDE (1) to a deterministic ODE

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}(t, \mathbf{x}(t); \boldsymbol{\theta}^{(1)}), \qquad \mathbf{x}(t) = \mathbf{x}_0 \in \mathcal{X}. \tag{15}$$

The basic idea of splitting methods for the time integration of ODEs can be formulated as follows. Given the ODE (15), suppose that function $\mathbf{F}$ can be expressed as a sum of M functions $\mathbf{F}^{[i]} : [0, \infty) \times \mathcal{X} \times \overline{\Theta}_{\boldsymbol{\theta}^{(1)}} \to \mathbb{R}^d$

$$\mathbf{F} = \sum_{m=1}^{M} \mathbf{F}^{[m]}, \tag{16}$$

such that each independent sub-ODE

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}^{[m]}(t, \mathbf{x}(t); \boldsymbol{\theta}^{(1)}), \qquad \mathbf{x}(t) = \mathbf{x}_0, \quad m = 1, \ldots, M \tag{17}$$

can be solved exactly. Their solutions are denoted by h-flows $\Phi_h^{[m]}(\mathbf{x}_0)$ at time step $t = h$. A toy example of splitting a vector field into a sum of two vector fields where each describes an ODE that can be solved explicitly is depicted in Figure 1.

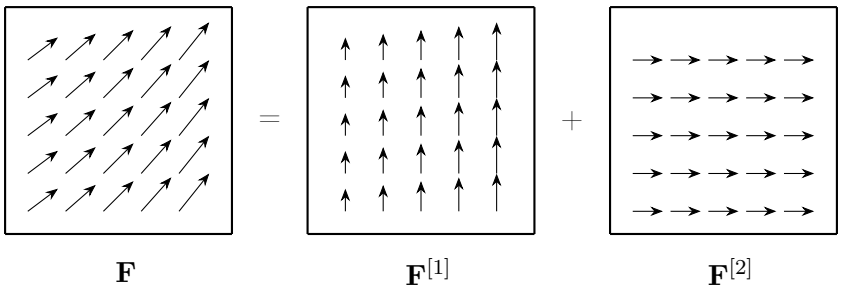

Figure 1: Two-dimensional vector field split into a sum of two uncoupled vector fields.

Having derived the explicit solutions of the sub-equations, we need to compose them properly. Two common ways of composition are the Lie-Trotter (LT) and the Strang

(S) splitting. The Lie-Trotter splitting approximations are obtained by composing the solutions of the sub-equations at time $t = h$ and starting from $\mathbf{x}_0$. Specifically, the Lie-Trotter composition of flows is

$$\Phi_h^{[\mathrm{LT}]} = \Phi_h^{[1]} \circ \cdots \circ \Phi_h^{[M]}. \tag{18}$$

Expanding $\Phi_h^{[\mathrm{LT}]}$ into a Taylor series reveals that it approximates the exact solution $\Phi_h(\mathbf{x}_0)$ to first-order accuracy, i.e., $\Phi_h^{[\mathrm{LT}]}(\mathbf{x}_0) = \Phi_h(\mathbf{x}_0) + \mathcal{O}(h^2)$, where Φ_h is the correct h-flow of ODE (15).

Higher-order approximations can be achieved by introducing additional intermediate steps with appropriately chosen coefficients. An example of higher-order approximation is the Strang splitting, given by

$$\Phi_h^{[\mathrm{S}]} = \Phi_{h/2}^{[M]} \circ \cdots \circ \Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]} \circ \cdots \circ \Phi_{h/2}^{[M]}. \tag{19}$$

It turns out that the S splitting has one order higher accuracy compared to the LT, that is $\Phi_h^{[\mathrm{S}]}(\mathbf{x}_0) = \Phi_h(\mathbf{x}_0) + \mathcal{O}(h^3)$. It is possible to derive a splitting scheme for any arbitrary order of convergence. However, there is a threshold between complexity and precision, where the more precise splitting schemes are naturally more complex to work with. Thus, in this thesis, we focus only on the LT and the S splitting and mainly work with the latter due to its superiority.

Figure 2 illustrates the composition schematic for the LT and S splitting for the example from Figure 1.

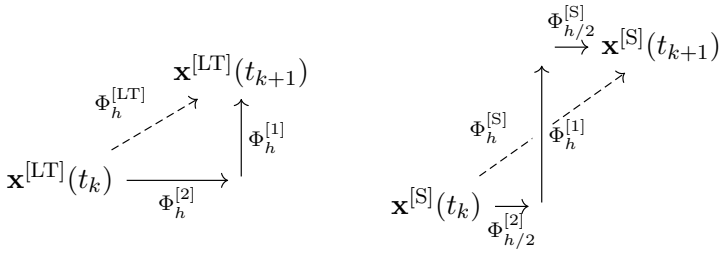

Figure 2: Schematic of the Lie-Trotter (left) and Strang (right) splitting approximations in case of two vector fields $\mathbf{F}^{[1]}$ and $\mathbf{F}^{[2]}$.

Here, we mention just a few reasons why splitting methods are popular in the theory and applications of dynamical systems. According to [Blanes et al., 2009], splitting schemes are simple to implement, explicit, and require modest storage while preserving structural properties of the exact solution, such as symplecticity, volume, and conservation of first integrals. These properties make splitting methods particularly valuable for long-term and geometric numerical integration, where preserving the solution's qualitative features is important.

Before generalizing splitting schemes to the SDE setting, we note three primary ambiguities regarding the splitting schemes.

First, it is not clear how to split a vector field $\mathbf{F}$ and into how many sub-vector fields. To remove this ambiguity and to set up the foundation of splitting estimators in SDEs, we start by splitting $\mathbf{F}$ into a sum of two sub-vector fields. This idea also relies on the recent study of Buckwar et al. [2022] that motivated our first paper. Moreover, following the same research, we decided that it is the most natural to split the drift into a sum of a linear part and a nonlinear part. We mentioned this in the previous section, and the following section explains the idea in detail.

Second, even after we decided to split the drift into linear and nonlinear parts, there are infinitely many ways to do so. The question is, which one is the optimal strategy? The immediate follow-up question is: what is an optimal strategy? We mentioned that in the ODE setting, the generally defined LT and S splitting schemes have the order of convergence h^2 and h^3, respectively, for any number of split vector fields and any choice of splitting. Similar statements are confirmed in the SDE setting. Thus, the order of convergence is the same independently of the splitting strategies, so in that sense, every strategy is optimal. Moreover, in Paper I, we prove that the asymptotic results for parameter estimators based on the LT and S splitting hold for any choice of splitting into linear and nonlinear parts. Therefore, every splitting strategy leads to the same asymptotic results. On the contrary, for a finite step size h and finite sample size N, numerical properties of the splitting schemes and, based on them, induced estimators differ based on the splitting strategy. We discuss this in the first paper, where we suggest, without formal proof, to choose the linear part as the linearization of the drift $\mathbf{F}$ around the system's equilibrium, if it exists. Intuitively, this means that around the equilibrium, the system is mainly governed by the linear part, while far from the equilibrium, the system is governed by the nonlinear dynamics. If this splitting is not possible, we assume the most straightforward possible strategy that can be implemented.

The third and last ambiguity is the order of composition. Given that we choose to split the drift into linear and nonlinear parts and that we know how to choose the linear part, there is still the question of whether we define the LT splitting as $\Phi_h^{[2]} \circ \Phi_h^{[1]}$ or as $\Phi_h^{[1]} \circ \Phi_h^{[2]}$, and similarly for the S splitting. We show in Paper I that the order of splitting is not important for the order of convergence of these two splitting approximations of an SDE. Moreover, we discuss in the same paper why we choose a specific order of composition to define the parameter estimators.

Now, we move on to the splitting schemes for SDEs with additive noise and the corresponding parameter estimators.

4.1 Stochastic Differential Equations with Additive Noise

Following Buckwar et al. [2022], we begin by splitting the drift term $\mathbf{F}$ of SDE (1) into linear and nonlinear components

$$\mathbf{F}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)}) = \mathbf{A}(\boldsymbol{\theta}^{(1)})(\mathbf{x} - \mathbf{b}(t; \boldsymbol{\theta}^{(1)})) + \mathbf{N}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)}). \tag{20}$$

Unlike in Buckwar et al. [2022], in our splitting strategy, we allow the intercept vector $-\mathbf{Ab}$. The intercept is a crucial part of the splitting strategy since it will enable linear

sub-SDE to have an equilibrium different from zero. Another important point is the time-independent matrix $\mathbf{A}$. This assumption is necessary to derive a solution for the linear SDE. While time-dependent $\mathbf{A}$ is possible, we do not discuss this scenario here.

Next, we solve the resulting sub-equations independently. The first sub-equation incorporates the linear part and the noise term

$$\mathrm{d}\mathbf{X}_t^{[1]} = \mathbf{A}(\boldsymbol{\theta}^{(1)})(\mathbf{x} - \mathbf{b}(t; \boldsymbol{\theta}^{(1)}))\,\mathrm{d}t + \boldsymbol{\Sigma}(t)\,\mathrm{d}\mathbf{W}_t, \tag{21}$$

while the second sub-equation deals with the nonlinear part

$$\mathrm{d}\mathbf{X}_t^{[2]} = \mathbf{N}(t, \mathbf{x}; \boldsymbol{\theta}^{(1)})\,\mathrm{d}t. \tag{22}$$

SDE (21) is linear, so we can solve it as

$$\mathbf{X}_{t_{k+1}}^{[1]} = \Phi_h^{[1]}(\mathbf{X}_{t_k}^{[1]}) = \boldsymbol{\mu}_h(t_k, \mathbf{X}_{t_k}^{[1]}; \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k},$$

where

$$\boldsymbol{\xi}_{h,t_k} = \int_{t_k}^{t_{k+1}} \exp(\mathbf{A}(t_{k+1} - u))\boldsymbol{\Sigma}(u)\,\mathrm{d}\mathbf{W}_u \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_h(t_k; \boldsymbol{\theta})), \tag{23}$$

and

$$\boldsymbol{\mu}_h(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) = \exp(\mathbf{A}h)\mathbf{x}_{t_k} - \int_{t_k}^{t_{k+1}} \exp(\mathbf{A}(t_{k+1} - u))\mathbf{A}\mathbf{b}(u)\,\mathrm{d}u, \tag{24}$$

$$\boldsymbol{\Omega}_h(t_k; \boldsymbol{\theta}) = \int_{t_k}^{t_{k+1}} \exp(\mathbf{A}(t_{k+1} - u))\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(u) \exp(\mathbf{A}^\top(t_{k+1} - u))\,\mathrm{d}u. \tag{25}$$

We avoided writing parameter $\boldsymbol{\theta}$ for clarity in the previous two equations. We see that both $\boldsymbol{\mu}_h$ and $\boldsymbol{\Omega}_h$ are not in closed form. To compute $\boldsymbol{\mu}_h$, we either need to split the drift such that intercept $\mathbf{b}$ is time-independent like we do in Paper I, or we somehow numerically compute the integral in $\boldsymbol{\mu}_h$.

On the other hand, calculating $\boldsymbol{\Omega}_h$ is discussed in the next section. Before that, notice how $\boldsymbol{\Omega}_h$ is similar to $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$ (13), with the difference of $\mathbf{A}$ in $\boldsymbol{\Omega}_h$ instead of $D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}_{t_k})$ in $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$, making $\boldsymbol{\Omega}_h$ much simpler to implement and faster. This similarity also suggests that it is a good idea to choose $\mathbf{A}$ as a linearization of $\mathbf{F}$ around the equilibrium, that is, $\mathbf{A} = D_{\mathbf{x}}\mathbf{F}(t_k, \mathbf{x}^\star)$, where $\mathbf{x}^\star$ is the equilibrium.

For the second sub-equation, an ODE (22), we would ideally like to solve it analytically

$$\mathbf{X}_{t_{k+1}}^{[2]} = \Phi_h^{[2]}(\mathbf{X}_{t_k}^{[2]}) = \boldsymbol{f}_h(t_k, \mathbf{X}_{t_k}^{[2]}; \boldsymbol{\theta}^{(1)}).$$

Then, for all $t_k \geq 0$ and $\boldsymbol{\theta}^{(1)} \in \Theta_{\boldsymbol{\theta}^{(1)}}$, the time h-flow $\boldsymbol{f}_h$ fulfills the following semi-group properties

$$\begin{aligned} \boldsymbol{f}_0(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) &= \mathbf{x}_{t_k}, \\ \boldsymbol{f}_{t+s}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) &= \boldsymbol{f}_t(t_k + s, \boldsymbol{f}_s(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}), \quad t, s \geq 0. \end{aligned} \tag{26}$$

Alternatively, we can use numerical methods such as the fourth-order Runge-Kutta method to approximate $\boldsymbol{f}_h$. Then, the semi-group properties (26) will not hold, but we can still use it as an approximation up to a certain order of convergence that will depend on the approximation method.

The splitting approximations are obtained by composing the solutions $\Phi_h^{[1]}$ and $\Phi_h^{[2]}$. The LT splitting is given by

$$\mathbf{X}_{t_{k+1}}^{[\mathrm{LT}]} = \left(\Phi_h^{[1]} \circ \Phi_h^{[2]}\right)(\mathbf{X}_{t_k}^{[\mathrm{LT}]}) = \boldsymbol{\mu}_h(t_k, \boldsymbol{f}_h(t_k, \mathbf{X}_{t_k}^{[\mathrm{LT}]}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k}. \tag{27}$$

If we define

$$\boldsymbol{\mu}_h^{[\mathrm{LT}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) = \boldsymbol{\mu}_h(t_k, \boldsymbol{f}_h(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}),$$

then, the LT splitting becomes

$$\mathbf{X}_{t_{k+1}}^{[\mathrm{LT}]} \coloneqq \boldsymbol{\mu}_h^{[\mathrm{LT}]}(t_k, \mathbf{X}_{t_k}^{[\mathrm{LT}]}; \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k}.$$

In Appendix A, we Taylor-expanded $\boldsymbol{\mu}_h^{[\mathrm{LT}]}$ and $\boldsymbol{\Omega}_h^{[\mathrm{LT}]} \coloneqq \boldsymbol{\Omega}_h$ in case of additive noise as

$$\boldsymbol{\mu}_h^{[\mathrm{LT}]}(\mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) = \mathbf{x}_{t_k} + h\mathbf{F}(\mathbf{x}_{t_k}) + \frac{h^2}{2}\left(\mathbf{A}\mathbf{F}(\mathbf{x}_{t_k}) + D_{\mathbf{x}}\mathbf{F}(\mathbf{x}_{t_k})\mathbf{N}(\mathbf{x}_{t_k})\right) + \mathbf{R}(h^3, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}),$$
$$\boldsymbol{\Omega}_h^{[\mathrm{LT}]}(\mathbf{x}_{t_k}; \boldsymbol{\theta}) = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}^\top + \mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) + \mathbf{R}(h^3, \mathbf{x}_{t_k}; \boldsymbol{\theta}).$$

Compared to methods in Section 3.4, such as SO1.5, it is clear that the LT method is of strong order 1 due to the misalignment between coefficients in front of h^2.

As in the majority of the methods discussed in Section 3.4, LT yields a Gaussian transition density that approximates the negative log-likelihood as

$$\begin{aligned}\mathcal{L}^{[\mathrm{LT}]}(\mathbf{X}; \boldsymbol{\theta}) = &\frac{1}{2}\sum_{k=0}^{N-1} \log\det \boldsymbol{\Omega}_h(t_k) \\ &+ \frac{1}{2}\sum_{k=0}^{N-1}(\mathbf{X}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{LT}]}(t_k, \mathbf{X}_{t_k}))^\top \boldsymbol{\Omega}_h(t_k)^{-1}(\mathbf{X}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{LT}]}(t_k, \mathbf{X}_{t_k})).\end{aligned}$$

On the other hand, the Strang splitting is given by

$$\begin{aligned}\mathbf{X}_{t_{k+1}}^{[\mathrm{S}]} &= \left(\Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]}\right)(\mathbf{X}_{t_k}^{[\mathrm{S}]}) \\ &= \boldsymbol{f}_{h/2}\left(t_k, \boldsymbol{\mu}_h(t_k, \boldsymbol{f}_{h/2}(t_k, \mathbf{X}_{t_k}^{[\mathrm{S}]}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k}; \boldsymbol{\theta}^{(1)}\right).\end{aligned} \tag{28}$$

The Taylor expansion of $\Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]}$ obtained in Appendix A in case of additive noise is

$$\begin{aligned}\mathbf{X}_{t_{k+1}}^{[\mathrm{S}]} = &\mathbf{X}_{t_k}^{[\mathrm{S}]} + h\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) + \boldsymbol{\xi}_{h,t_k} + \frac{h}{2}D_{\mathbf{x}}\mathbf{N}(\mathbf{X}_{t_k}^{[\mathrm{S}]})\boldsymbol{\xi}_{h,t_k} + \frac{h^2}{2}D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]})\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) \\ &+ \frac{h^2}{4}\sum_{i,j=1}^{d}\partial_{i,j}^2\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]})[\boldsymbol{\xi}_{h,t_k}\boldsymbol{\xi}_{h,t_k}^\top]_{ij} + \mathbf{R}(h^{5/2}, \mathbf{X}_{t_k}^{[\mathrm{S}]}, \boldsymbol{\xi}_{h,t_k}; \boldsymbol{\theta}).\end{aligned} \tag{29}$$

Compared to all previous methods, the expansion is fundamentally different due to the appearance of the random variables $\boldsymbol{\xi}_{h,t_k}$. If we also expend $\boldsymbol{\xi}_{h,t_k}$ (23), then approximation (29) becomes

$$\begin{aligned}\mathbf{X}_{t_{k+1}}^{[\mathrm{S}]} &= \mathbf{X}_{t_k}^{[\mathrm{S}]} + h\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) + \boldsymbol{\Sigma}\Delta\mathbf{W}_k + h(\mathbf{A} + \frac{1}{2}D_{\mathbf{x}}\mathbf{N}(\mathbf{X}_{t_k}^{[\mathrm{S}]}))\boldsymbol{\Sigma}\Delta\mathbf{W}_k - \mathbf{A}\boldsymbol{\Sigma}\Delta\boldsymbol{\zeta}_k \\ &+ \frac{h^2}{2}D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]})\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) + \frac{h^2}{4}\sum_{i,j=1}^{d}\partial_{i,j}^2\mathbf{F}(\mathbf{X}_{t_k}^{[\mathrm{S}]})[\boldsymbol{\Sigma}\Delta\mathbf{W}_k\Delta\mathbf{W}_k^\top\boldsymbol{\Sigma}^\top]_{ij} \\ &+ \mathbf{R}(h^{5/2}, \mathbf{X}_{t_k}^{[\mathrm{S}]}; \boldsymbol{\theta}),\end{aligned} \tag{30}$$

where $\Delta\boldsymbol{\zeta}_k$ and $\Delta\mathbf{W}_k$ are jointly normally distributed as in (11). From equation (30), we see that the S splitting does not yield a Gaussian transition density. Moreover, we can see that the conditional mean is correct up to order h^3, matching the other higher-order methods. This property of the S splitting implies that it has order 3 one-step error, unlike the LT method, which has only order 2. In contrast, the conditional covariance is correct only up to order h, making the S splitting have strong order 1 and not 1.5 as other methods. These properties are formally proved in Paper I.

While the LT splitting (27) yields Gaussian transition density, the S splitting (28) is a nonlinear transformation of a Gaussian. Therefore, we shall assume that the nonlinear solution $\boldsymbol{f}_h$ has an inverse $\boldsymbol{f}_h^{-1}$, or equivalently, the backward flow $\boldsymbol{f}_{-h}$. Then, we can rewrite the Strang splitting approximation as

$$\boldsymbol{f}_{-h/2}(t_k, \mathbf{X}_{t_{k+1}}^{[\mathrm{S}]}; \boldsymbol{\theta}^{(1)}) = \boldsymbol{\mu}_h(t_k, \boldsymbol{f}_{h/2}(t_k, \mathbf{X}_{t_k}^{[\mathrm{S}]}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k}. \tag{31}$$

Under the assumption of well definiteness of the backward flow $\boldsymbol{f}_{-h}$, the semigroup property (26) yields

$$\boldsymbol{f}_{h/2}(t_k, \mathbf{x}_{t_k}) = \boldsymbol{f}_h(t_k - h/2, \boldsymbol{f}_{-h/2}(t_k, \mathbf{x}_{t_k})).$$

If we define

$$\begin{aligned}\boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) &\coloneqq \boldsymbol{\mu}_h(t_k, \boldsymbol{f}_h(t_k - h/2, \mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}), \\ \mathbf{Y}_{t_k}^{[\mathrm{S}]} &\coloneqq \boldsymbol{f}_{-h/2}(t_k, \mathbf{X}_{t_k}^{[\mathrm{S}]}; \boldsymbol{\theta}^{(1)}), \qquad k = 0, 1, ..., N,\end{aligned}$$

then the S splitting becomes

$$\mathbf{Y}_{t_{k+1}}^{[\mathrm{S}]} = \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \mathbf{Y}_{t_k}^{[\mathrm{S}]}; \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,t_k}. \tag{32}$$

Notice that if ODE (22) is autonomous, then $\boldsymbol{\mu}_h^{[\mathrm{S}]}(\mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)}) = \boldsymbol{\mu}_h^{[\mathrm{LT}]}(\mathbf{x}_{t_k}; \boldsymbol{\theta}^{(1)})$.

Finally, the Strang approximated negative log-likelihood is

$$\begin{aligned}\mathcal{L}^{[\mathrm{S}]}(\mathbf{Y}; \boldsymbol{\theta}) &= \frac{1}{2}\sum_{k=0}^{N-1}\log\det\boldsymbol{\Omega}_h(t_k) \\ &+ \frac{1}{2}\sum_{k=0}^{N-1}(\mathbf{Y}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \mathbf{Y}_{t_k}))^\top\boldsymbol{\Omega}_h(t_k)^{-1}(\mathbf{Y}_{t_{k+1}} - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \mathbf{Y}_{t_k})).\end{aligned}$$

Since we do not observe $\mathbf{Y}$, after the change of variable, the Strang approximated negative log-likelihood becomes

$$\begin{aligned}\mathcal{L}^{[\mathrm{S}]}(\mathbf{X};\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=0}^{N-1}\log\det\boldsymbol{\Omega}_h(t_k) \\ + \frac{1}{2}\sum_{k=0}^{N-1}(\boldsymbol{f}_{-h/2}(t_{k+1},\mathbf{X}_{t_{k+1}}) - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k,\boldsymbol{f}_{-h/2}(t_k,\mathbf{X}_{t_k})))^\top\boldsymbol{\Omega}_h(t_k)^{-1} \\ (\boldsymbol{f}_{-h/2}(t_{k+1},\mathbf{X}_{t_{k+1}}) - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k,\boldsymbol{f}_{-h/2}(t_k,\mathbf{X}_{t_k}))) \\ - \sum_{k=0}^{N-1}\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_{k+1},\mathbf{X}_{t_{k+1}})|\end{aligned} \tag{33}$$

We can slightly rewrite the previous log-likelihood in a more compact form. We start by converting the last term in (33) from time t_{k+1} to t_k by

$$\begin{aligned}\sum_{k=0}^{N-1}\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_{k+1},\mathbf{X}_{t_{k+1}})| &= -\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_0,\mathbf{X}_{t_0})| \\ &\quad + \sum_{k=0}^{N-1}\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_k,\mathbf{X}_{t_k})| \\ &\quad + \log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_N,\mathbf{X}_{t_N})|.\end{aligned}$$

Notice then that properties of functions log, det, and $\boldsymbol{f}_h$ yield

$$\begin{aligned}-\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_k,\mathbf{X}_{t_k})| &= \log|\det D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{X}_{t_k})| \\ &= \frac{1}{2}\log(\det D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{X}_{t_k}))^2 \\ &= \frac{1}{2}\log\det(D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{X}_{t_k}))^2.\end{aligned}$$

Then, combining the log det terms in (33) yields

$$\begin{aligned}&\frac{1}{2}\sum_{k=0}^{N-1}\left(\log\det\boldsymbol{\Omega}_h(t_k) - 2\log|\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_{k+1},\mathbf{X}_{t_{k+1}})|\right) \\ &= \frac{1}{2}\sum_{k=0}^{N-1}\log\det(\boldsymbol{\Omega}_h(t_k)(D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{X}_{t_k}))^2) \\ &+ \log|\det D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_N,\mathbf{X}_{t_N})\det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_0,\mathbf{X}_{t_0})|.\end{aligned} \tag{34}$$

The last term in equation (34) is irrelevant in the asymptotic case when $N\to\infty$, so we ignore it for now. Then, the previous derivations suggest to define

$$\boldsymbol{\Omega}_h^{[\mathrm{S}]}(t_k,\mathbf{x}_{t_k};\boldsymbol{\theta}) \coloneqq D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{x}_{t_k})\boldsymbol{\Omega}_h(t_k)D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k,\mathbf{x}_{t_k})^\top. \tag{35}$$

Matrix $\boldsymbol{\Omega}_h^{[\mathrm{S}]}$ is richer in information than $\boldsymbol{\Omega}_h^{[\mathrm{LT}]}$ since it includes data, but more importantly, it incorporates both the linear and nonlinear dynamics of the original SDE. The previous argument implies that the Strang splitting yields a much better estimator than the Lie-Trotter. Our Paper I confirms this statement.

Continuing with the previous derivation that is not investigated in other papers, equation (35) yields

$$\boldsymbol{\Omega}_h(t_k)^{-1} = D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k, \mathbf{x}_{t_k})^\top \boldsymbol{\Omega}_h^{[\mathrm{S}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta})^{-1} D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k, \mathbf{x}_{t_k}). \tag{36}$$

Combining all the previous, we get the following representation of the Strang approximated negative log-likelihood

$$\begin{aligned}\mathcal{L}^{[\mathrm{S}]}(\mathbf{X}; \boldsymbol{\theta}) = \log|\det D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_N, \mathbf{X}_{t_N}) \det D_{\mathbf{x}}\boldsymbol{f}_{-h/2}(t_0, \mathbf{X}_{t_0})| + \frac{1}{2}\sum_{k=0}^{N-1} \log\det \boldsymbol{\Omega}_h^{[\mathrm{S}]}(t_k, \mathbf{X}_{t_k}) \\ + \frac{1}{2}\sum_{k=0}^{N-1} \left(D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k, \mathbf{X}_{t_k}) \left(\boldsymbol{f}_{-h/2}(t_{k+1}, \mathbf{X}_{t_{k+1}}) - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \boldsymbol{f}_{-h/2}(t_k, \mathbf{X}_{t_k}))\right)\right)^\top \\ \boldsymbol{\Omega}_h^{[\mathrm{S}]}(t_k, \mathbf{X}_{t_k})^{-1} D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k, \mathbf{X}_{t_k}) \left(\boldsymbol{f}_{-h/2}(t_{k+1}, \mathbf{X}_{t_{k+1}}) - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \boldsymbol{f}_{-h/2}(t_k, \mathbf{X}_{t_k}))\right).\end{aligned}$$

Thus, the Strang splitting also yields a Gaussian likelihood up to an asymptotically negligible term, where the random variable

$$D_{\mathbf{x}}\boldsymbol{f}_{h/2}(t_k, \mathbf{X}_{t_k}) \left(\boldsymbol{f}_{-h/2}(t_{k+1}, \mathbf{X}_{t_{k+1}}) - \boldsymbol{\mu}_h^{[\mathrm{S}]}(t_k, \boldsymbol{f}_{-h/2}(t_k, \mathbf{X}_{t_k}))\right) \tag{37}$$

is conditional zero-mean Gaussian with covariance matrix $\boldsymbol{\Omega}_h^{[\mathrm{S}]}(t_k, \mathbf{x}_{t_k}; \boldsymbol{\theta})$. The previous derivations and conclusions give more intuition about the Strang splitting estimator and are not mentioned or further analyzed in the three papers.

To conclude this section, we note that the LT and S splitting estimators are defined fundamentally differently from the other estimators in Section 3.4, while the LT shares more similarities with them than the S estimator. The S estimator is more complicated to understand with more sophisticated assumptions, such as the existence of the backward flow. This existence is needed only locally for $h \leq h_0$, where h_0 is a known threshold. The S estimator will perform poorly in the case of non-existing backward flow for $h > h_0$. We must approximate the backward flow to be well-defined to solve this issue. However, if this assumption is fulfilled, the complexity of implementation and the computational speed of the Strang splitting estimator will not increase compared to the Lie-Trotter. Still, the accuracy increases drastically, as shown in Paper I.

4.2 Stochastic Differential Equations with Pearson-type Noise

In Paper III, we assume that the SDE (1) has the following coordinate-wise form of the squared diffusion matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)})$

$$[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})]_{ij} = \mathbf{x}^\top \boldsymbol{\alpha}^{ij}(t; \boldsymbol{\theta}^{(2)})\mathbf{x} + \mathbf{x}^\top \boldsymbol{\beta}^{ij}(t; \boldsymbol{\theta}^{(2)}) + \gamma^{ij}(t; \boldsymbol{\theta}^{(2)}), \quad i, j = 1, 2, ..., d, \tag{38}$$

where $\boldsymbol{\alpha}^{ij} : [0, \infty) \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}} \to \mathbb{R}^{d \times d}$, $\boldsymbol{\beta}^{ij} : [0, \infty) \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}} \to \mathbb{R}^d$, and $\gamma^{ij} : [0, \infty) \times \overline{\Theta}_{\boldsymbol{\theta}^{(2)}} \to \mathbb{R}$ are such that, for all $t \geq 0$, $\boldsymbol{\alpha}^{ij}(t)$ are symmetric, and $\boldsymbol{\alpha}^{ij}(t) = \boldsymbol{\alpha}^{ji}(t)$, $\boldsymbol{\beta}^{ij}(t) = \boldsymbol{\beta}^{ji}(t)$, $\gamma^{ij}(t) = \gamma^{ji}(t)$, for all $i, j = 1, \ldots, d$.

This way of defining the SDE (1) is implicit since we do not explicitly define the diffusion matrix $\boldsymbol{\Sigma}(t, \mathbf{x}; \boldsymbol{\theta}^{(2)})$. This leads to the possibility of different SDEs having the same squared diffusion matrix, a problem of identifiability mentioned in Section 3.2, which we do not address here.

In Paper III, we refer to SDEs with a diffusion matrix defined by (38) as SDEs with Pearson-type noise. This term originates from the fact that, under the linear drift $\mathbf{F}$, the SDE (1) with a diffusion matrix defined by (38) can be seen as a multivariate generalization of Pearson diffusions, a univariate class of models with linear drift and squared diffusion that is a quadratic function of the state variable.

In Paper III, we derive explicit formulas for calculating the covariance matrix for multivariate Pearson diffusion. This allows us to combine the strategy of Gaussian approximation with splitting schemes. Specifically, consider the following splitting of (1)

$$\mathrm{d}\mathbf{X}_t^{[1]} = \mathbf{A}(\boldsymbol{\theta}^{(1)})(\mathbf{X}_t^{[1]} - \mathbf{b}(t; \boldsymbol{\theta}^{(1)}))\,\mathrm{d}t + \boldsymbol{\Sigma}(t, \mathbf{X}_t; \boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0^{[1]} = \mathbf{x}_0, \tag{39}$$

$$\mathrm{d}\mathbf{X}_t^{[2]} = \mathbf{N}(t, \mathbf{X}_t^{[2]}; \boldsymbol{\theta}^{(1)})\,\mathrm{d}t, \qquad \mathbf{X}_0^{[2]} = \mathbf{x}_0. \tag{40}$$

Equation (39) describes a multivariate Pearson diffusion whose solution cannot be explicitly obtained in general. However, we can approximate it by assuming the transition density is Gaussian, as seen in Section 3.4. For multivariate Pearson diffusions, the mean and covariance can be calculated explicitly, as presented in Paper III, enabling the approximation of (39) as

$$\mathbf{X}_{t_{k+1}}^{[1]} = \Psi_h^{[1]}(\mathbf{X}_{t_k}^{[1]}) = \boldsymbol{\mu}_h(t_k, \mathbf{X}_{t_k}^{[1]}; \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_h(t_k, \mathbf{X}_{t_k}^{[1]}; \boldsymbol{\theta}), \tag{41}$$

where $\boldsymbol{\xi}_h(t_k, \mathbf{X}_{t_k}; \boldsymbol{\theta}) \overset{i.i.d}{\sim} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h(t_k, \mathbf{X}_{t_k}; \boldsymbol{\theta}))$ for $k = 0, \ldots, N-1$. Note that $\Psi_h^{[1]}$ is not the exact h-flow $\Phi_h^{[1]}$ of SDE (39), but rather an approximation based on the Gaussian transition density assumption.

The function $\boldsymbol{\mu}_h$ in (41) is the same as in (24). However, $\boldsymbol{\Omega}_h(t_k, \mathbf{X}_{t_k}; \boldsymbol{\theta})$ is not equal to that in (25), but is more complicated due to the quadratic term of the state variable in (38). To calculate $\boldsymbol{\Omega}_h(t_k, \mathbf{X}_{t_k}; \boldsymbol{\theta})$, we need the covariance of the multivariate Pearson diffusion, calculated in Paper III for the autonomous case. In the non-autonomous case, we need additional approximations, as discussed in the next section.

5 Computational Tools

This chapter focuses on the computational tools needed for parameter estimation. It comprises two main parts.

The first part addresses the computations of integrals involving matrix exponentials. They are required for the covariance matrices of the splitting and LL schemes and the integrals necessary for implementing the mean vector of the LL scheme. It is also needed to evaluate the covariance matrix of the multivariate Pearson-type diffusion introduced in Paper III.

The second part discusses the implementation and optimization of objective functions in the programming language `R`. Given that base `R` lacks automatic differentiation, we use the `torch` package, which originates from `PyTorch` for `Python` and is implemented in `C++` for `R`. This part also briefly explains automatic differentiation, its importance, and details on coding our estimator in `R`.

5.1 Integrals Involving Matrix Exponentials

Evaluating integrals involving the matrix exponential is fundamental in many areas of control theory and systems analysis. Our workflow requires evaluating covariance matrices such as (13) and (25), and integrals needed for the mean vectors such as (14) and (24).

For now, we assume that we work with autonomous SDE; that is, the previously mentioned integrals can be written as

$$\int_0^t \exp(\mathbf{B}(t-s))\mathbf{A}\,\mathrm{d}s, \qquad \int_0^t \exp(\mathbf{B}(t-s))\mathbf{A}s\,\mathrm{d}s \tag{42}$$

and

$$\int_0^t \exp(\mathbf{B}(t-s))\mathbf{C}\exp(\mathbf{B}^\top(t-s))\,\mathrm{d}s, \tag{43}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are constant matrices of appropriate dimensions. When the matrix $\mathbf{B}$ is invertible, the first integral in equation (42) can be simplified to

$$\int_0^t \exp(\mathbf{B}(t-s))\mathbf{A}\,ds = (\exp(\mathbf{B}t) - \mathbf{I})\mathbf{B}^{-1}\mathbf{A}. \tag{44}$$

However, $\mathbf{B}$ usually depends on a current data point and parameters, so in the optimization step, so it is likely that $\mathbf{B}$ will not be invertible. Moreover, other integrals do not necessarily have a nice closed-form formula. Thus, we use Theorem 1 in Van Loan [1978] to avoid these issues.

Theorem 2 (Theorem 1 in Van Loan [1978]). *Let n_1, n_2, n_3, n_4 be non-negative integers, $m = n_1 + n_2 + n_3 + n_4$, and the $m \times m$ block triangular matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{C}_1 & \mathbf{D}_1 \\ \mathbf{0}_{n_2 \times n_1} & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{C}_2 \\ \mathbf{0}_{n_3 \times n_1} & \mathbf{0}_{n_3 \times n_2} & \mathbf{A}_3 & \mathbf{B}_3 \\ \mathbf{0}_{n_4 \times n_1} & \mathbf{0}_{n_4 \times n_2} & \mathbf{0}_{n_4 \times n_3} & \mathbf{A}_4 \end{bmatrix}.$$

Then, for $t \geq 0$

$$\exp(\mathbf{M}t) = \begin{bmatrix} \mathbf{F}_1(t) & \mathbf{G}_1(t) & \mathbf{H}_1(t) & \mathbf{K}_1(t) \\ \mathbf{0} & \mathbf{F}_2(t) & \mathbf{G}_2(t) & \mathbf{H}_2(t) \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_3(t) & \mathbf{G}_3(t) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{F}_4(t) \end{bmatrix},$$

where

$$\begin{aligned}
\mathbf{F}_j(t) &= \exp(\mathbf{A}_j t), \quad j = 1, 2, 3, 4, \\
\mathbf{G}_j(t) &= \int_0^t \exp(\mathbf{A}_j(t-s))\mathbf{B}_j \exp(\mathbf{A}_{j+1}s)\,\mathrm{d}s, \quad j = 1, 2, 3, \\
\mathbf{H}_j(t) &= \int_0^t \exp(\mathbf{A}_j(t-s))\mathbf{C}_j \exp(\mathbf{A}_{j+2}s)\,\mathrm{d}s \\
&\quad + \int_0^t \int_0^s \exp(\mathbf{A}_j(t-s))\mathbf{B}_j \exp(\mathbf{A}_{j+1}(s-r))\mathbf{B}_{j+1} \exp(\mathbf{A}_{j+2}r)\,\mathrm{d}r\,\mathrm{d}s, \quad j = 1, 2, \\
\mathbf{K}_1(t) &= \int_0^t \exp(\mathbf{A}_1(t-s))\mathbf{D}_1 \exp(\mathbf{A}_4 s)\,\mathrm{d}s \\
&\quad + \int_0^t \int_0^s \exp(\mathbf{A}_1(t-s))\mathbf{C}_1 \exp(\mathbf{A}_3(s-r))\mathbf{B}_3 \exp(\mathbf{A}_4 r)\,\mathrm{d}r\,\mathrm{d}s \\
&\quad + \int_0^t \int_0^s \exp(\mathbf{A}_1(t-s))\mathbf{B}_1 \exp(\mathbf{A}_2(s-r))\mathbf{C}_2 \exp(\mathbf{A}_4 r)\,\mathrm{d}r\,\mathrm{d}s \\
&\quad + \int_0^t \int_0^s \int_0^r \exp(\mathbf{A}_1(t-s))\mathbf{B}_1 \exp(\mathbf{A}_2(s-r))\mathbf{B}_2 \exp(\mathbf{A}_3(r-w))\mathbf{B}_3 e^{\mathbf{A}_4 r}\,\mathrm{d}w\,\mathrm{d}r\,\mathrm{d}s.
\end{aligned}$$

To evaluate the integrals in (42), we start by constructing the following block matrices

$$\mathbf{M}_1 = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}, \qquad \mathbf{M}_2 = \begin{bmatrix} \mathbf{B} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Then, the corresponding values $\mathbf{G}_1(t, \mathbf{M}_1)$, and $\mathbf{H}_1(t, \mathbf{M}_2)$ are given by the integrals in (42) respectively. Here, notation $\mathbf{G}_1(t, \mathbf{M}_1)$ denotes that $\mathbf{G}_1$ from Theorem 2 was obtained by exponentiating block matrix $\mathbf{M}_1$.

Similarly, the integral in (43) equals $\mathbf{G}_1(t, \mathbf{M}_3)\mathbf{F}_1^\top(t, \mathbf{M}_3)$, where $\mathbf{M}_3$ is

$$\mathbf{M}_3 = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & -\mathbf{B}^\top \end{bmatrix}.$$

For a non-autonomous SDE, the target integrals become

$$\int_0^t \exp(\mathbf{B}(t-s))\mathbf{A}(s)\,\mathrm{d}s, \tag{45}$$

and

$$\int_0^t \exp(\mathbf{B}(t-s))\mathbf{C}(s)\exp(\mathbf{B}^\top(t-s))\,\mathrm{d}s, \tag{46}$$

where $\mathbf{B}$ is a constant matrix, while $\mathbf{A}$ and $\mathbf{C}$ are sufficiently smooth, matrix-valued functions with domain $[0,\infty)$ and appropriate co-domain dimensions. Following the ideas from [Carbonell et al., 2008], they first generalize Theorem 2 to allow for an arbitrary number of block matrices within matrix $\mathbf{M}$. Subsequently, they perform a Taylor expansion of matrices $\mathbf{A}$ and $\mathbf{C}$. For example,

$$\mathbf{C}(s) \approx \sum_{i=0}^{p} \frac{s^i}{i!}\mathbf{C}_i = \mathbf{C}_0 + \int_0^s \mathbf{C}_1\,\mathrm{d}r + \int_0^s \int_0^r \mathbf{C}_2\,\mathrm{d}w\,\mathrm{d}r + \ldots,$$

where $\mathbf{C}_i$ is the ith derivative of $\mathbf{C}$ at $s=0$. Then, the integral (46) is approximated as

$$\begin{aligned}
&\int_0^t \exp(\mathbf{B}(t-s))\mathbf{C}(s)\exp(\mathbf{B}^\top(t-s))\,\mathrm{d}s \\
&\approx \int_0^t \exp(\mathbf{B}(t-s))\mathbf{C}_0\exp(\mathbf{B}^\top(t-s)) + \int_0^t\int_0^s \exp(\mathbf{B}(t-s))\mathbf{C}_1\exp(\mathbf{B}^\top(t-s))\,\mathrm{d}r\,\mathrm{d}s \\
&+ \int_0^t\int_0^s\int_0^r \exp(\mathbf{B}(t-s))\mathbf{C}_2\exp(\mathbf{B}^\top(t-s))\,\mathrm{d}w\,\mathrm{d}r\,\mathrm{d}s\ldots
\end{aligned} \tag{47}$$

Finally, using the generalized version of Theorem 2, the approximation (47) can be obtained by taking the exponential of

$$\mathbf{M} = \begin{bmatrix}
\mathbf{B} & \mathbf{C}_p & \mathbf{C}_{p-1} & \ldots & \mathbf{C}_2 & \mathbf{C}_1 & \mathbf{C}_0 \\
\mathbf{0} & -\mathbf{B}^\top & \mathbf{I} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & -\mathbf{B}^\top & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & -\mathbf{B}^\top & \mathbf{I} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & -\mathbf{B}^\top & \mathbf{I} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & -\mathbf{B}^\top
\end{bmatrix}.$$

Namely, if $\mathbf{N}(t) = \exp(\mathbf{M}t)$, then the solution is obtained by multiplying the upper right corner block matrix of $\mathbf{N}$ by the transpose of the upper left corner block matrix. A similar reasoning holds for calculating the integral (45).

The results above enable stable numerical implementations of the LL and splitting schemes estimators. Furthermore, Theorem 2 facilitates the explicit derivation and implementation of formulas for the covariance of multivariate Pearson diffusions, as demonstrated in Paper III.

In the subsequent section, we explore the numerical challenges of estimating parameters by optimizing nonlinear multidimensional objective functions in `R`. Additionally, we provide a brief overview of how the matrix exponential is implemented in the `torch` package.

5.2 Parameter Estimation Using the `torch` Package

In this section, we review different types of gradient-based optimization and discuss the importance of using automatic differentiation over numerical differentiation. We then introduce the `torch` package [Falbel and Luraschi, 2024] as a bridge between automatic differentiation and the `R` programming language. Following this, we provide a subsection on resilient propagation, a gradient descent algorithm used in our optimization process. Finally, we discuss how to implement this estimator in `R` and briefly connect the implementation of the matrix exponential of the previous section, one of the key elements of the log-likelihood objective functions.

5.2.1 Optimization

When doing parameter estimation in SDE models, the goal is to optimize an objective function typically derived from the log-likelihood equation, as defined in (10). Whether exact or approximate, this function requires optimization to estimate parameter values that best fit the data.

Deterministic optimization techniques rely on differentiation to determine the direction and rate of change in function values, mostly known as gradient descent. There are three primary types of differentiation used in gradient descent optimization [Griewank and Walther, 2008]:

1. **Symbolic Differentiation** involves solving derivatives analytically and providing exact expressions. This method is exact and does not introduce numerical error. However, symbolic differentiation can lead to expression swell, where derivatives become increasingly complex, requiring more computational resources in terms of time and memory. The complexity can scale factorially with the number of operations, making it impractical for functions with many variables or complex relationships.

2. **Numerical Differentiation** employs finite difference methods such as forward and central differences to approximate derivatives. While straightforward, this approach introduces two primary types of errors: truncation and round-off errors. Truncation error occurs when the finite difference approximation does not precisely replicate the true derivative, which becomes more pronounced with a larger step size. The truncation error is of order $\mathcal{O}(h)$ for forward differences and of order $\mathcal{O}(h^2)$ for central differences. Conversely, round-off error arises from the limitations of floating-point arithmetic, becoming more significant with a smaller step size. Therefore, as the step size h decreases, the round-off error increases due to floating-point precision limits. Thus, a balance between minimizing truncation and round-off errors is needed. The computational cost for evaluating each derivative is low, at $\mathcal{O}(1)$ per evaluation.

3. **Automatic Differentiation (AD)**, unlike symbolic and numerical differentiation, computes derivatives accurately to machine precision, avoiding truncation

and round-off errors using the splitting of functions into elementary operations and applying the chain rule through techniques known as forward and reverse accumulations. Forward accumulation is efficient for functions with more outputs than inputs. In contrast, reverse accumulation, also known as backpropagation, is suitable for functions with more inputs than outputs, which is common in optimization problems with large parameter spaces and real-valued objective functions. The costs for AD depend on the mode; reverse mode often requires more memory to store intermediate results.

The base `R` programming environment does not support AD. However, `R` can be integrated with other programming languages that support AD through the use of packages such as `TMB` (Template Model Builder) [Kristensen et al., 2016] or `RStan` [Stan Development Team, 2024], which handle AD efficiently for complex, high-dimensional models.

In our research, we employ the `torch` package for `R`, which is based on `PyTorch`, originally a `Python`-based library that has been extended to `R` through `C++`. `PyTorch` is a library mainly intended for deep learning, but its employment of AD is useful for optimizing any objective functions efficiently, making it suitable for MLE.

The `torch` package in `R` offers a syntax that is similar to native `R` syntax, with slight modifications to accommodate object-oriented programming principles. In `torch`, one can apply methods directly to objects, which differs from the usual functional approach in `R`. This design allows for a more intuitive workflow for complex data structures like tensors, a primary data structure in `torch` and equivalent to arrays in `R`. This means that one must convert numerical arrays into tensors to work effectively with `torch`. This conversion is straightforward; for example, a numeric array `par` can be transformed into a tensor using `torch_tensor(par)`. This approach leverages the flexibility of object-oriented programming while maintaining compatibility with `R`'s syntax, making `torch` an accessible and powerful tool for statistical inference in `R`.

Using `torch` is not only useful for AD but also for many different optimizers implemented in `torch`. In our study, we compared all available optimizers in `torch` to find the most effective one for our needs. We chose `optim_rprop` because it consistently demonstrated robustness across different models of interest. It also reliably converged to the true values for different approximations of log-likelihood functions from Chapter 3.4 and different initial parameter values. In the following subsection, we briefly describe this optimization method.

5.2.2 Resilient Propagation

The Resilient Propagation (`RProp`) algorithm is a gradient-based optimization method designed to improve the efficiency of backpropagation used in training neural networks. Developed by Riedmiller and Braun [1993], `RProp` is particularly effective in addressing the limitations of traditional gradient descent methods, which can suffer from slow convergence and instability due to issues like vanishing or exploding gradients.

The main innovation is `RProp`'s ability to adaptively adjust the size of the parameter updates independently, allowing each parameter to be estimated based solely on the

sign of the gradient rather than its magnitude. This feature distinguishes `RProp` from other methods, as it mitigates the adverse effects caused by small or large gradients, which can otherwise lead to inefficient optimization or numerical instability. Its ability to dynamically adapt update sizes makes it particularly useful in parameter estimation with highly nonlinear log-likelihoods, where traditional methods may struggle to achieve robust convergence.

In the `RProp` algorithm, each parameter is associated with an individual update value. This value is dynamically adjusted during optimization based on the gradient's sign. When the sign of the gradient for a particular weight remains consistent between successive iterations, `RProp` increases the update value, allowing the algorithm to take larger steps in the direction of the minimum. Conversely, if the gradient's sign changes, indicating that the algorithm may have overshot the minimum, the update value is decreased to prevent further divergence.

Furthermore, `RProp` is not only robust but also computationally efficient. It requires less tuning of hyperparameters, such as learning rates, which can be a challenge in other gradient-based methods. Despite these advantages, `RProp` is primarily suited for batch updates rather than stochastic or mini-batch approaches. This limitation arises because `RProp` relies on consistent gradient information from the entire dataset to adjust update values effectively. However, this constraint does not affect our work since we focus on parameter optimization in the context of statistical optimization, where batch processing is more suitable and aligns with our goal of optimizing parameters with precision rather than dealing with real-time data updates.

5.2.3 Implementing the Estimator in `torch`

In this subsection, we show the implementation of a parameter estimation step using the `torch` package in `R` to minimize a negative log-likelihood objective function. The `estimator` function, defined in Code 1, optimizes model parameters by minimizing the `objective` function, which depends on the parameters, data, and step size h. The optimization process involves initializing the parameters, setting up an optimizer, and iteratively updating the parameters based on the gradient of the objective function. This iterative process continues until the parameters converge or the maximum number of iterations is reached. Through the `torch` package, we use AD, which significantly improves the computation of gradients required for the optimization algorithm.

5.2.4 Implementing the Matrix Exponential in `torch`

Section 5.1 dealt with avoiding the direct computation of integrals involving matrix exponentials by simply computing and manipulating matrix exponentials. Therefore, having an efficient way to calculate the matrix exponential is essential. Here, we briefly present the main idea behind implementing the matrix exponential in `torch`.

The function `torch_matrix_exp` is based on the method proposed by Bader et al. [2019], which introduces an optimized approach to computing the Taylor polynomial for the matrix exponential. Traditional methods like the Paterson-Stockmeyer technique re-

```r
estimator <- function(objective, data, h, par_star, num_iterations) {

    # convert parameters to torch tensors with gradient tracking
    par_star <- torch_tensor(par_star, requires_grad = TRUE)
    # create a RProp optimizer
    optimizer <- optim_rprop(par_star)

    calc_loss <- function() {
        # reset gradients
        optimizer$zero_grad()
        # calculate the objective function value
        value <- objective(par_star, data, h)
        # perform reverse accumulation to compute gradients
        value$backward()
        value
    }

    for (i in 1:num_iterations) {
        # save current parameters for convergence check
        par_old = as.matrix(par_star)
        # perform an optimization step and update par_star
        optimizer$step(calc_loss)
        value_new <- calc_loss() # recompute the loss function
        par_new = as.matrix(par_star)  # update new parameters

        if(norm(par_new - par_old) < 10^-5) break  # convergence check
    }
    convergence <- 0
    # check if the loop is completed without convergence
    if(i == num_iterations) convergence <- 1

    # return the optimized parameters, final value, and iterations
    list(as.numeric(par_star), as.numeric(value_new), i)
}
```

Code 1: Estimator based on `RProp` optimization algorithm using `torch` in `R`

quire many matrix multiplications, which can be computationally expensive. The method of Bader et al. [2019] reduces the number of these multiplications, making the process more efficient. The reduction is achieved by combining the optimized Taylor polynomial approximation with the scaling and squaring technique.

The scaling and squaring technique works by first scaling down the matrix by a power of two, computing the exponential of the smaller matrix, and then squaring the result repeatedly to get the exponential of the original matrix. By optimizing the Taylor polynomial, the proposed method further enhances the efficiency and accuracy of this process.

This approach improves performance and maintains high accuracy for a wide range of matrix sizes and norms. Numerical experiments have shown that this method outperforms traditional Padé approximants, particularly in terms of computational efficiency.

Papers

I Parameter Estimation in Nonlinear Multivariate SDEs with Additive Noise

This chapter contains the following paper:

- [Pilipovic et al., 2024a] P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics*, 52(2):842 – 867, 2024a. doi: 10.1214/24-AOS2371. URL https://doi.org/10.1214/24-AOS2371.

This paper focuses on developing and analyzing efficient parameter estimation techniques for discretely observed nonlinear SDEs. Specifically, we start with splitting schemes for SDE discretization and prove a new numerical property regarding the order of convergence under less restrictive assumptions on the drift parameter. Moreover, we introduce estimators based on these discretization schemes and establish their asymptotic properties.

To develop these new estimation techniques, we first outline the theoretical framework in Section 2 of the paper, motivating and leveraging the use of the Lie–Trotter and Strang splitting schemes and introducing the estimators. This section discusses how these methods allow us to split the solution of the original SDE into an Ornstein-Uhlenbeck process and a nonlinear part, facilitating more accurate and stable parameter estimation. We then move on to prove the first key result of the paper. Section 3 establishes that the Strang splitting scheme achieves an L^p convergence rate of order 1, a property already known for the Lie–Trotter scheme. These proofs are non-trivial due to our assumption of the one-sided Lipschitz condition on the drift parameter, which is less restrictive than the global Lipschitz condition commonly used in the literature. This relaxation broadens the applicability of our methods but also involves more complex proofs.

Following the numerical developments, Section 5 states the second main result of the paper regarding the estimators' consistency and asymptotic efficiency. This section demonstrates that, under the one-sided Lipschitz assumption, the estimators perform reliably and provide robust parameter estimates. The theoretical results are complemented by a numerical study on the three-dimensional stochastic Lorenz system presented in Section 6. The numerical study shows that the Strang estimator performs well in terms of both precision and computational speed, surpassing several state-of-the-art techniques.

The Annals of Statistics
2024, Vol. 52, No. 2, 842–867
https://doi.org/10.1214/24-AOS2371

PARAMETER ESTIMATION IN NONLINEAR MULTIVARIATE STOCHASTIC DIFFERENTIAL EQUATIONS BASED ON SPLITTING SCHEMES

BY PREDRAG PILIPOVIC[1,a], ADELINE SAMSON[2,b] AND SUSANNE DITLEVSEN[1,c]

[1]*Department of Mathematical Sciences, University of Copenhagen,* [a]*predrag@math.ku.dk*
[2]*Université Grenoble Alpes, CNRS, Grenoble INP, LJK,* [b]*adeline.leclercq-samson@univ-grenoble-alpes.fr,* [c]*susanne@math.ku.dk*

The likelihood functions for discretely observed nonlinear continuous time models based on stochastic differential equations are not available except for a few cases. Various parameter estimation techniques have been proposed, each with advantages, disadvantages and limitations depending on the application. Most applications still use the Euler–Maruyama discretization, despite many proofs of its bias. More sophisticated methods, such as Kessler's Gaussian approximation, Ozaki's local linearization, Aït–Sahalia's Hermite expansions or MCMC methods, might be complex to implement, do not scale well with increasing model dimension or can be numerically unstable. We propose two efficient and easy-to-implement likelihood-based estimators based on the Lie–Trotter (LT) and the Strang (S) splitting schemes. We prove that S has L^p convergence rate of order 1, a property already known for LT. We show that the estimators are consistent and asymptotically efficient under the less restrictive one-sided Lipschitz assumption. A numerical study on the 3-dimensional stochastic Lorenz system complements our theoretical findings. The simulation shows that the S estimator performs the best when measured on precision and computational speed compared to the state-of-the-art.

1. Introduction. Stochastic differential equations (SDEs) are popular models for physical, biological and socioeconomic processes. Some recent applications include tipping points in the climate (Ditlevsen and Ditlevsen (2023)), the spread of COVID-19 (Arnst et al. (2022), Kareem and Al-Azzawi (2021)), animal movements (Michelot et al. (2019, 2021)) and cryptocurrency rates (Dipple et al. (2020)). The advantage of SDEs is their ability to capture and quantify the randomness of the underlying dynamics. They are especially applicable when the dynamics are not entirely understood, and the unknown parts act as random. The following parametric form is common for an SDE model with additive noise:

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}(\mathbf{X}_t; \boldsymbol{\beta})\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_0 = \mathbf{x}_0. \tag{1}$$

We want to estimate the underlying drift parameter $\boldsymbol{\beta}$ and diffusion parameter $\boldsymbol{\Sigma}$ based on discrete observations of $\mathbf{X}_t$. The transition density is necessary for likelihood-based estimators, and thus a closed-form solution to (1). However, the transition density is only available for a few SDEs, including the Ornstein–Uhlenbeck (OU) process, which has a linear drift function $\mathbf{F}$. Extensive literature exists on MCMC methods for the nonlinear case (Chopin and Papaspiliopoulos (2020), Fuchs (2013)) however, these are often computationally intensive and do not always converge to the correct values for complex models. Thus, we need a valid approximation of the transition density to perform likelihood-based statistical inference.

Received January 2023; revised February 2024.
MSC2020 subject classifications. Primary 62F12, 62H12, 62M99; secondary 37M15, 60G65.
Key words and phrases. Asymptotic normality, consistency, L^p convergence, splitting schemes, stochastic differential equations, stochastic Lorenz system.

The most straightforward discretization scheme is the Euler–Maruyama (EM) (Kloeden and Platen (1992)). Its main advantage is the easy-to-implement and intuitive Gaussian transition density. Both frequentist and Bayesian approaches extensively employ EM across theoretical and applied studies. However, the EM-based estimator has many disadvantages. First, it exhibits pronounced bias as the discretization step increases (see Florens-Zmirou (1989) for a theoretical study, or Gloaguen, Etienne and Le Corff (2018), Gu, Wu and Xue (2020) for applied studies). Second, Hutzenthaler, Jentzen and Kloeden (2011) showed that it is not mean-square convergent when the drift function $\mathbf{F}$ of (1) grows super-linearly. Consequently, we should avoid EM for models with polynomial drift. Third, it often fails to preserve important structural properties, such as hypoellipticity, geometric ergodicity, and amplitudes, frequencies and phases of oscillatory processes (Buckwar et al. (2022)).

Some pioneering papers on likelihood-based SDE estimators are Dacunha-Castelle and Florens-Zmirou (1986), Dohnal (1987), Florens-Zmirou (1989), Genon-Catalot and Jacod (1993), Kessler (1997). The first two only estimate the diffusion parameter. Florens-Zmirou (1989) used EM to estimate both parameters and derived asymptotic properties. Genon-Catalot and Jacod (1993) generalized to higher dimensions, nonequidistant discretization step, and a generic form of the objective function, however, only estimating the diffusion parameter. Kessler (1997) proposed an estimator (denoted K) approximating the unknown transition density with a Gaussian density using the true conditional mean and covariance, or approximations thereof using the infinitesimal generator. He proved consistency and asymptotic normality under the commonly used, but too restrictive, global Lipschitz assumption on the drift function $\mathbf{F}$.

A competitive likelihood-based approach relies on local linearization (LL), initially proposed by Ozaki (1985) and later extended by Ozaki (1992), Shoji and Ozaki (1998). They approximated the drift between two consecutive observations using a linear function. In the case of additive noise, this corresponds to an OU process with a known Gaussian transition density. Thus, the likelihood approximation is a product of Gaussian densities. Shoji (1998) proved that LL discretization is one-step consistent and L^p convergent with order 1.5. Shoji (2011), Jimenez, Mora and Selva (2017) extended the theory of LL for SDEs with multiplicative noise. Simulation studies show the superiority of the LL estimator compared to other estimators (Gloaguen, Etienne and Le Corff (2018), Gu, Wu and Xue (2020), Hurn, Jeisman and Lindsay (2007), Shoji and Ozaki (1998)). Until recently, the implementation of the LL estimator was numerically ill-conditioned due to the possible singularity of the Jacobian matrix of the drift function $\mathbf{F}$. However, Gu, Wu and Xue (2020) proposed an efficient implementation that overcomes this. The main disadvantage of the LL method is its slow computational speed.

Aït-Sahalia (2002) proposed Hermite expansions (HE) to approximate the transition density, focusing on univariate time-homogeneous diffusions. This method, widely utilized in finance, was later extended to both reducible and irreducible multivariate diffusions (Aït-Sahalia (2008)). Chang and Chen (2011) found conditions under which the HE estimator has the same asymptotic distribution as the exact maximum likelihood estimator (MLE). Choi (2013, 2015) further broadened the technique to time-inhomogeneous settings. Picchini and Ditlevsen (2011) used the method for multidimensional diffusions with random effects. When an SDE is irreducible, Aït-Sahalia (2008) applied Kolmogorov's backward and forward equations to develop a small-time expansion of the diffusion probability densities. Yang, Chen and Wan (2019) introduced a delta expansion method, using Itô–Taylor expansions to derive analytical approximations of the transition densities of multivariate diffusions inspired by Aït-Sahalia (2002). While Aït-Sahalia's approach allows for a broad class of drift and diffusion functions, the implementation can be complex. To our knowledge, there have not been any applications to models with more than four dimensions. Furthermore, computing coefficients

even up to order two can be challenging, while higher-order approximations are often necessary for nonlinear models. Hurn, Jeisman and Lindsay (2007) implemented HE up to third order in univariate cases, emphasizing the importance of symbolic computation tools like `Mathematica` or `Maple`. Their survey concluded that while LL is the best among discrete maximum likelihood estimators, HE is the preferred overall choice. They highlighted that the HE proposed by Aït-Sahalia (2002) has the best trade-off between speed and accuracy, proving more feasible than LL in most financial applications. Similar results are found in Jensen and Poulsen (2002), López-Pérez, Febrero-Bande and González-Manteigav (2021). However, LL's broad applicability contrasts with the limitations of Hermite expansions, particularly for high-dimensional multivariate models exceeding three dimensions.

Apart from the above-mentioned general methods, there are some specific setups. Sørensen and Uchida (2003) investigated a small-diffusion estimator, Ditlevsen and Sørensen (2004), Gloter (2006) worked with integrated diffusion, and Uchida and Yoshida (2012) used adaptive maximum likelihood estimation. Bibby and Sørensen (1995) and Forman and Sørensen (2008) explored martingale estimation functions (EF) in one-dimensional diffusions, but they are difficult to extend to multidimensional SDEs. Ditlevsen and Samson (2019) used the 1.5 scheme to solve the problem of hypoellipticity when the diffusion matrix is not of full rank.

More recently, contributions from Gloter and Yoshida (2021a, 2021b) have extended the research of Uchida and Yoshida (2012). Gloter and Yoshida (2021a) introduced a nonadaptive approach and offered similar analytic asymptotic results as Ditlevsen and Samson (2019) without imposing strict limitations on the model class. Iguchi, Beskos and Graham (2022) proposed sampling schemes for elliptic and hypoelliptic models that often result in conditionally non-Gaussian integrals, distinguishing their approach from prior works. As the transition density of their new scheme is typically complex, Iguchi, Beskos and Graham (2022) created a closed-form density expansion using Malliavin calculus. They recommended a transition density scheme that retained second-order precision through prudent truncation of the expansion. This closed-form expansion aligns with the works of Aït-Sahalia (2002, 2008) and Li (2013) on elliptic SDEs, although with a different approach. Iguchi, Beskos and Graham (2022) deliver asymptotic results with analytically available rates, beneficial for both elliptic and hypoelliptic models.

Table 1 provides a comprehensive overview of estimator properties, finite sample performance and required model assumptions for the most prominent state-of-the-art methods. While asymptotic properties might be similar in most cases, the finite sample properties are often different. The table also includes the Lie–Trotter (LT) and the Strang (S) splitting estimators, which we propose in this paper. The comparison encompasses four key characteristics: (1) Diffusion coefficient allowed in the model class, distinguishing between additive and general noise; (2) Asymptotic regime, the conditions needed to prove the asymptotic properties; (3) Implementation, assessing the complexity of implementation, dependence on model dimension and parameter optimization time and (4) Finite sample properties, evaluating performance for fixed sample size N and discretization step size h.

An essential aspect of any estimator is the practical execution in real-world applications. Although the previously mentioned research contributes significantly to the theoretical development and broadens our understanding of inference for SDEs, its practical implementations tend not to be user friendly. Except for precomputed models, applications by nonspecialists can be challenging. Our main contribution is proposing estimators that are intuitive, easy to implement, computationally efficient and scalable with increasing dimensions. These characteristics make the estimators accessible to researchers in various applied sciences while maintaining desirable statistical properties. Moreover, these estimators remain competitive with the best state-of-the-art methods, particularly concerning estimation bias and variance.

We propose to use the LT or the S splitting schemes for statistical inference. These numerical approximations were first suggested for ordinary differential equations (ODEs) (see,

TABLE 1

Comparison of the proposed Lie–Trotter (LT) and Strang (S) splittings (in bold) with five state-of-the-art estimators: Euler–Maruyama (EM), Kessler (K), Estimating functions (EF), Local linearization (LL) and Hermite expansion (HE). The comparison focuses on four key characteristics: (1) Noise type—additive or general, (2) Asymptotic regime—investigating conditions where asymptotic properties align with the exact MLE, (3) Computational time and implementation—evaluating implementation and parameter optimization costs and (4) Finite sample properties—assessing performance under fixed N and h. The finite sample properties of the estimators are likely influenced by specific experiment designs

Estimator	Noise type	Asymptotic regime	Computational time and implementation	Finite sample properties
EM	General	$h \to 0$, $Nh \to \infty$, $Nh^2 \to 0$ (Florens-Zmirou (1989))	Fastest optimization and implementation. Straightforward for any dimension.	Earliest bias exhibition with increasing h.
K up to order J	General	J fixed: $h \to 0$, $Nh \to \infty$, $Nh^p \to 0$, for any $p \in \mathbb{N}^a$ (Kessler (1997))	Fast optimization. Straightforward for $J \leq 3$.	Unbiased if the exact mean is known. For larger h, a higher order of J is needed. Performance between EM and LL.
EF	General	h fixed: $N \to \infty$ (Bibby and Sørensen (1995))	Fast optimization. Requires moments of the transition density. Mainly suitable for univariate models.	Unbiased also for large h, but not efficient. Good performance.
LL	Additive (possible generalization) (Jimenez, Mora and Selva (2017))	$h \to 0$, $Nh \to \infty$, $Nh^2 \to 0$ (Ozaki (1992))	Slowest discrete ML approximations. (Hurn, Jeisman and Lindsay (2007)) Straightforward for any dimension.	Best among all discrete ML approximations. (Hurn, Jeisman and Lindsay (2007))
HE up to order J	General	h fixed: $N \to \infty$, $J \to \infty$, $Nh^{2J+2} \to 0$, $J \geq 2$ fixed: $N \to \infty$, $h \to 0$, $Nh^3 \to \infty$, $Nh^{2J+1} \to 0$ (Chang and Chen (2011))	Slower than LL in the univariate case. Implementation becomes significantly more complex in higher dimensions or for $J \geq 2$. (Hurn, Jeisman and Lindsay (2007))	For larger h, a higher order of J is needed. Better than LL in the univariate case. (Hurn, Jeisman and Lindsay (2007))
LT (proposed)	Additive (possible generalization)	$h \to 0$, $Nh \to \infty$, $Nh^2 \to 0$	Slower than K, but notably faster than LL. Straightforward implementation for given nonlinear ODE solution. Scales well with the increasing dimension.	Performance relative to EM varies based on splitting strategy and model.
S (proposed)	Additive (possible generalization)	$h \to 0$, $Nh \to \infty$, $Nh^2 \to 0$	Slower than LT, but notably faster than LL. Straightforward implementation for given nonlinear ODE solution. Scales well with the increasing dimension.	As good as LL.

[a]While Kessler (1997) did not explicitly explore the scenario of a fixed h, it is a reasonable assumption that the asymptotic results will hold as $N \to \infty$ and $J \to \infty$.

e.g., Blanes, Casas and Murua (2008), McLachlan and Quispel (2002)), but their extension to SDEs is straightforward. A few studies have investigated numerical properties (Ableidinger and Buckwar (2016), Ableidinger, Buckwar and Hinterleitner (2017), Bensoussan, Glowinski and Răşcanu (1992), Buckwar et al. (2022)). Barbu (1988) applied LT splitting on nonlinear optimal control problems, while Hopkins and Wong (1986) used it for nonlinear filtering. Abdulle, Vilmart and Zygalakis (2015), Bou-Rabee and Owhadi (2010) used LT splitting to investigate conditions for preserving the measure of the ergodic nonlinear Langevin equations. Recently, Bréhier and Goudenége (2019) showed that LT splitting successfully preserved positivity for a class of nonlinear stochastic heat equations with multiplicative space-time white noise. Additional studies on the application of splitting schemes to SDEs include those by Alamo and Sanz-Serna (2016), Leimkuhler and Matthews (2015), Milstein and Tretyakov (2003), Misawa (2001), Bréhier and Goudenége (2019). Regarding statistical applications, to the best of our knowledge, only Buckwar, Tamborrino and Tubikanec (2020), Ditlevsen, Tamborrino and Tubikanec (2023) used splitting schemes for parametric inference in combination with Approximate Bayesian Computation, and Ditlevsen and Ditlevsen (2023) used it for prediction of a forthcoming collapse in the climate.

This paper presents five main contributions:

1. We introduce two new efficient, easy-to-implement, and computationally fast estimators for multidimensional nonlinear SDEs.
2. We establish L^p convergence of the S splitting scheme.
3. We prove consistency and asymptotic normality of the new estimators under the less restrictive assumption of one-sided Lipschitz. This proof requires innovative approaches.
4. We demonstrate the estimators' performance in a stochastic version of the chaotic Lorenz system, in contrast to prior studies that primarily addressed the deterministic Lorenz system.
5. We compare the new estimators to four discrete maximum likelihood estimators from the literature in a simulation study, comparing the accuracy and computational speed.

The rest of this paper is structured as follows. In Section 2, we introduce the SDE model class and define the splitting schemes and the estimators. In Section 3, we show that the S splitting has better one-step predictions than the LT, and we prove that the S splitting is L^p consistent with order 1.5 and L^p convergent with order 1. To the best of our knowledge, this is a new result. Sections 4 and 5 establish the estimator asymptotics under the less restrictive one-sided global Lipschitz assumption. We illustrate in Section 6 the theoretical results in a simulation study on a model that is not globally Lipschitz, the 3-dimensional stochastic Lorenz systems. Since the objective functions based on pseudo-likelihoods are multivariate in both data and parameters, we use automatic differentiation (AD) to get faster and more reliable estimators. We compare the precision and speed of the EM, K, LL, HE, LT and S estimators. We show that the EM and LT estimators become biased before the others with increasing discretization step h, HE (of order 2) works only for the smallest h in the simulation study, and the LL and S perform the best. However, S is much faster than LL because LL calculates a new covariance matrix for each combination of data points and parameter values.

Notation. We use capital bold letters for random vectors, vector-valued functions and matrices, while lowercase bold letters denote deterministic vectors. $\|\cdot\|$ denotes both the L^2 vector norm in $\mathbb{R}^d$ and the matrix norm induced by the L^2 norm, defined as the square root of the largest eigenvalue. Superscript (i) on a vector denotes the ith component, while on a matrix it denotes the ith row. Double subscript ij on a matrix denotes the component in the i-th row and jth column. If a matrix is a product of more matrices, square brackets with subscripts denote a component inside the matrix. The transpose is denoted by $\top$. Operator $\mathrm{Tr}(\cdot)$ returns the trace of a matrix and $\det(\cdot)$ the determinant. Sometimes, we denote by $[a_i]_{i=1}^d$ a

vector with coordinates a_i, and by $[b_{ij}]_{i,j=1}^d$ a matrix with coordinates b_{ij}, for $i, j = 1, \ldots, d$. We denote with $\partial_i g(\mathbf{x})$ the partial derivative of a generic function $g : \mathbb{R}^d \to \mathbb{R}$ with respect to $x^{(i)}$ and $\partial_{ij}^2 g(\mathbf{x})$ the second partial derivative. The nabla operator ∇ denotes the gradient vector of a function g, $\nabla g(\mathbf{x}) = [\partial_i g(\mathbf{x})]_{i=1}^d$. The differential operator D denotes the Jacobian matrix $D\mathbf{F}(\mathbf{x}) = [\partial_i F^{(j)}(\mathbf{x})]_{i,j=1}^d$, for a vector-valued function $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$. $\mathbf{H}$ denotes the Hessian matrix of a real-valued function g, $\mathbf{H}_g(\mathbf{x}) = [\partial_{ij} g(\mathbf{x})]_{i,j=1}^d$. Let $\mathbf{R}$ represent a vector (or a matrix) valued function defined on $(0, 1) \times \mathbb{R}^d$, such that for some constant C, $\|\mathbf{R}(a, \mathbf{x})\| < aC(1 + \|\mathbf{x}\|)^C$ for all a, $\mathbf{x}$. When denoted R, it is a scalar.

The Kronecker delta function is denoted by δ_i^j. For an open set A, the bar $\overline{A}$ indicates closure. We use $\overset{\theta}{=}$ to indicate equality up to an additive constant that does not depend on $\boldsymbol{\theta}$. We write $\xrightarrow{\mathbb{P}}$, $\xrightarrow{d}$ and $\xrightarrow{\mathbb{P}-a.s.}$ for convergence in probability, distribution, and almost surely, respectively. $\mathbf{I}_d$ denotes the d-dimensional identity matrix, while $\mathbf{0}_{d\times d}$ is a d-dimensional zero square matrix. For an event $E \in \mathcal{F}$, we denote by $\mathbb{1}_E$ the indicator function.

2. Problem setup. Let $\mathbf{X}$ in (1) be defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\boldsymbol{\theta}})$ with a complete right-continuous filtration $(\mathcal{F}_t)_{t\geq 0}$, and let the d-dimensional Wiener process $\mathbf{W} = (\mathbf{W}_t)_{t\geq 0}$ be adapted to $\mathcal{F}_t$. The probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$. Rewrite equation (1) as follows:

$$\mathrm{d}\mathbf{X}_t = \mathbf{A}(\boldsymbol{\beta})\big(\mathbf{X}_t - \mathbf{b}(\boldsymbol{\beta})\big)\,\mathrm{d}t + \mathbf{N}(\mathbf{X}_t; \boldsymbol{\beta})\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_0 = \mathbf{x}_0, \tag{2}$$

such that $\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{A}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{b}(\boldsymbol{\beta})) + \mathbf{N}(\mathbf{x}; \boldsymbol{\beta})$. Let $\overline{\Theta} = \overline{\Theta}_\beta \times \overline{\Theta}_\Sigma$ be the parameter space with Θ_β and Θ_Σ being two open convex bounded subsets of $\mathbb{R}^r$ and $\mathbb{R}^{d\times d}$, respectively.

Functions $\mathbf{F}, \mathbf{N} : \mathbb{R}^d \times \overline{\Theta}_\beta \to \mathbb{R}^d$ are locally Lipschitz, and $\mathbf{A}$, $\mathbf{b}$ are defined on $\overline{\Theta}_\beta$ and take values in $\mathbb{R}^{d\times d}$ and $\mathbb{R}^d$, respectively. Parameter matrix $\boldsymbol{\Sigma}$ takes values in $\mathbb{R}^{d\times d}$. The matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is assumed to be positive definite and determines the variance of the process. Since any square root of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ induces the same distribution, $\boldsymbol{\Sigma}$ is only identifiable up to equivalence classes. Thus, instead of estimating $\boldsymbol{\Sigma}$, we estimate $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. The drift function $\mathbf{F}$ in (1) is split up into a linear part given by matrix $\mathbf{A}$ and vector $\mathbf{b}$ and a nonlinear part given by $\mathbf{N}$. This decomposition is essential for defining the splitting schemes and the objective functions used for estimating $\boldsymbol{\theta}$.

We denote the true parameter value by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and assume that $\boldsymbol{\theta}_0 \in \Theta$. Sometimes we write $\mathbf{A}_0$, $\mathbf{b}_0$, $\mathbf{N}_0(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$ instead of $\mathbf{A}(\boldsymbol{\beta}_0)$, $\mathbf{b}(\boldsymbol{\beta}_0)$, $\mathbf{N}(\mathbf{x}; \boldsymbol{\beta}_0)$ and $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top$, when referring to the true parameters. We write $\mathbf{A}$, $\mathbf{b}$, $\mathbf{N}(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ for any parameter $\boldsymbol{\theta}$. Sometimes we suppress the parameter to simplify notation, for example, $\mathbb{E}$ implicitly refers to $\mathbb{E}_{\boldsymbol{\theta}}$.

REMARK 1. The drift function $\mathbf{F}(\mathbf{x})$ can always be rewritten as $\mathbf{A}(\mathbf{x} - \mathbf{b}) + \mathbf{N}(\mathbf{x})$ for any $\mathbf{A}$, $\mathbf{b}$ by setting $\mathbf{N}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{A}(\mathbf{x} - \mathbf{b})$, including choosing $\mathbf{A}$ and $\mathbf{b}$ to be zero. The splitting proposed below will then result in a Brownian motion (3) and a nonlinear ODE (4).

REMARK 2. We assume additive noise, sometimes referred to as constant volatility, meaning that the diffusion matrix does not depend on the current state. This assumption can be restrictive and even rejected by the data in some applications. The proposed methodology can be extended if the diffusion is reducible (Definition 1 in (Aït-Sahalia (2008))) by applying the Lamperti transform to obtain a unit diffusion coefficient. However, if the transform depends on the parameter, estimation is not straightforward. In this paper, we only consider additive noise.

2.1. *Assumptions*. The main assumption is that (2) has a unique strong solution $\mathbf{X}=(\mathbf{X}_t)_{t\in[0,T]}$, adapted to $(\mathcal{F}_t)_{t\in[0,T]}$, which follows from the following first two assumptions (Theorem 2 in Alyushina (1987), Theorem 1 in Krylov (1990), Theorem 3.5 in Mao (2007)). We need the last three assumptions to prove the properties of the estimators.

(A1) Function $\mathbf{N}$ is twice continuously differentiable with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$, that is, $\mathbf{N}\in C^2$. Additionally, it is one-sided globally Lipschitz continuous with respect to $\mathbf{x}$ on $\mathbb{R}^d\times\overline{\Theta}_\beta$, that is, there exists a constant $C>0$ such that

$$(\mathbf{x}-\mathbf{y})^\top\big(\mathbf{N}(\mathbf{x};\boldsymbol{\beta})-\mathbf{N}(\mathbf{y};\boldsymbol{\beta})\big)\le C\|\mathbf{x}-\mathbf{y}\|^2\quad\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

(A2) Function $\mathbf{N}$ grows at most polynomially in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}$, that is, there exist constants $C>0$ and $\chi\ge 1$ such that

$$\big\|\mathbf{N}(\mathbf{x};\boldsymbol{\beta})-\mathbf{N}(\mathbf{y};\boldsymbol{\beta})\big\|^2\le C\big(1+\|\mathbf{x}\|^{2\chi-2}+\|\mathbf{y}\|^{2\chi-2}\big)\|\mathbf{x}-\mathbf{y}\|^2\quad\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

Additionally, its derivatives are of polynomial growth in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}$.

(A3) The solution $\mathbf{X}$ of SDE (1) has invariant probability $\nu_0(\mathrm{d}\mathbf{x})$.

(A4) $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is invertible on $\overline{\Theta}_\Sigma$.

(A5) Function $\mathbf{F}$ is identifiable in $\boldsymbol{\beta}$, that is, if $\mathbf{F}(\mathbf{x},\boldsymbol{\beta}_1)=\mathbf{F}(\mathbf{x},\boldsymbol{\beta}_2)$ for all $\mathbf{x}\in\mathbb{R}^d$, then $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$.

Assumption (A3) is required for the ergodic theorem to ensure convergence in distribution. Assumption (A4) implies that model (1) is elliptic, which is not needed for the S estimator, whereas the EM estimator breaks down in hypoelliptic models. We will treat the hypoelliptic case in a separate paper where the proofs are more involved. Assumption (A5) ensures the identifiability of the parameter.

Assume a sample $(\mathbf{X}_{t_k})_{k=0}^N\equiv\mathbf{X}_{0:t_N}$ from (2) at time steps $0=t_0<t_1<\cdots<t_N=T$. For notational simplicity, we assume equidistant step size $h=t_k-t_{k-1}$.

2.2. *Moments*. Assumption (A1) ensures finiteness of the moments of the solution $\mathbf{X}$ (Tretyakov and Zhang (2013)), that is,

$$\mathbb{E}\Big[\sup_{t\in[0,T]}\|\mathbf{X}_t\|^{2p}\Big]<C\big(1+\|\mathbf{x}_0\|^{2p}\big)\quad\forall p\ge 1.$$

The infinitesimal generator L of (1) is defined on sufficiently smooth functions $g:\mathbb{R}^d\times\Theta\to\mathbb{R}$ given by

$$L_{\boldsymbol{\theta}_0}g(\mathbf{x};\boldsymbol{\theta})=\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)^\top\nabla g(\mathbf{x};\boldsymbol{\theta})+\frac{1}{2}\operatorname{Tr}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\mathbf{H}_g(\mathbf{x};\boldsymbol{\theta})\big).$$

The moments of (1) are expanded using the following lemma (Lemma 1.10 in Sørensen (2012)).

LEMMA 2.1. *Let Assumptions (A1)–(A2) hold. Let* $\mathbf{X}$ *be a solution of* (1). *Let* $g\in C^{(2l+2)}$ *be of polynomial growth and* $p\ge 2$. *Then*

$$\mathbb{E}_{\boldsymbol{\theta}_0}\big[g(\mathbf{X}_{t_k};\boldsymbol{\theta})|\mathcal{F}_{t_{k-1}}\big]=\sum_{j=0}^{l}\frac{h^j}{j!}L_{\boldsymbol{\theta}_0}^j g(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta})+R\big(h^{l+1},\mathbf{X}_{t_{k-1}}\big).$$

We need terms up to order $R(h^3,\mathbf{X}_{t_{k-1}})$. Applying $L_{\boldsymbol{\theta}}$ on $g(\mathbf{x})=x^{(i)}$, Lemma 2.1 yields

$$\mathbb{E}\big[X_{t_k}^{(i)}|\mathbf{X}_{t_{k-1}}=\mathbf{x}\big]=x^{(i)}+hF^{(i)}(\mathbf{x})+\frac{h^2}{2}\Big(\mathbf{F}(\mathbf{x})^\top\nabla F^{(i)}(\mathbf{x})+\frac{1}{2}\operatorname{Tr}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{H}_{F^{(i)}}(\mathbf{x})\big)\Big)+R\big(h^3,\mathbf{x}\big).$$

2.3. *Splitting schemes.* Consider the following splitting of (2):

$$\mathrm{d}\mathbf{X}_t^{[1]} = \mathbf{A}\big(\mathbf{X}_t^{[1]} - \mathbf{b}\big)\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_0^{[1]} = \mathbf{x}_0, \tag{3}$$

$$\mathrm{d}\mathbf{X}_t^{[2]} = \mathbf{N}\big(\mathbf{X}_t^{[2]}\big)\,\mathrm{d}t, \quad \mathbf{X}_0^{[2]} = \mathbf{x}_0. \tag{4}$$

The solution of equation (3) is an OU process given by the following h-flow:

$$\mathbf{X}_{t_k}^{[1]} = \Phi_h^{[1]}\big(\mathbf{X}_{t_{k-1}}^{[1]}\big) = e^{\mathbf{A}h}\mathbf{X}_{t_{k-1}}^{[1]} + \big(\mathbf{I} - e^{\mathbf{A}h}\big)\mathbf{b} + \boldsymbol{\xi}_{h,k}, \tag{5}$$

where $\boldsymbol{\xi}_{h,k} \overset{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h)$ for $k = 1, \ldots, N$ (Vatiwutipong and Phewchean (2019)). The covariance matrix $\boldsymbol{\Omega}_h$ and the conditional mean of the OU process (5) are provided by

$$\begin{aligned}\boldsymbol{\Omega}_h &= \int_0^h e^{\mathbf{A}(h-u)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{\mathbf{A}^\top(h-u)}\,\mathrm{d}u = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}\big(\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}^\top\big) \\ &\quad + \mathbf{R}(h, \mathbf{x}_0),\end{aligned} \tag{6}$$

$$\boldsymbol{\mu}_h(\mathbf{x}; \boldsymbol{\beta}) := e^{\mathbf{A}(\boldsymbol{\beta})h}\mathbf{x} + \big(\mathbf{I} - e^{\mathbf{A}(\boldsymbol{\beta})h}\big)\mathbf{b}(\boldsymbol{\beta}). \tag{7}$$

Assumptions (A1) and (A2) ensure the existence and uniqueness of the solution of (4) (Theorem 1.2.17 in Humphries and Stuart (2002)). Thus, there exists a unique function $\boldsymbol{f}_h : \mathbb{R}^d \times \Theta_\beta \to \mathbb{R}^d$, for $h \geq 0$, such that

$$\mathbf{X}_{t_k}^{[2]} = \Phi_h^{[2]}\big(\mathbf{X}_{t_{k-1}}^{[2]}\big) = \boldsymbol{f}_h\big(\mathbf{X}_{t_{k-1}}^{[2]}; \boldsymbol{\beta}\big). \tag{8}$$

For all $\boldsymbol{\beta} \in \Theta_\beta$, the time flow $\boldsymbol{f}_h$ fulfills the following semigroup properties:

$$\boldsymbol{f}_0(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}, \qquad \boldsymbol{f}_{t+s}(\mathbf{x}; \boldsymbol{\beta}) = \boldsymbol{f}_t\big(\boldsymbol{f}_s(\mathbf{x}; \boldsymbol{\beta}); \boldsymbol{\beta}\big), \quad t, s \geq 0. \tag{9}$$

REMARK 3. Since only one-sided Lipschitz continuity is assumed, the solution to (4) might not exist for all $h < 0$ and all $\mathbf{x}_0 \in \mathbb{R}^d$, implying that the inverse $\boldsymbol{f}_h^{-1}$ might not exist. If it exists, then $\boldsymbol{f}_h^{-1} = \boldsymbol{f}_{-h}$. For the S estimator, we need a well-defined inverse. This is not an issue when $\mathbf{N}$ is globally Lipschitz.

We, therefore, introduce the following and last assumption.

(A6) Function $\boldsymbol{f}_h^{-1}(\mathbf{x}; \boldsymbol{\beta})$ is defined asymptotically, for all $\mathbf{x} \in \mathbb{R}^d$, $\boldsymbol{\beta} \in \Theta_\beta$, when $h \to 0$.

Before defining the splitting schemes, we present a useful proposition for expanding the nonlinear solution $\boldsymbol{f}_h$ (Section 1.8 in (Hairer, Nørsett and Wanner (1993))).

PROPOSITION 2.2. *Let Assumptions (A1)–(A2) hold. When $h \to 0$, the h-flow of (4) is*

$$\boldsymbol{f}_h(\mathbf{x}) = \mathbf{x} + h\mathbf{N}(\mathbf{x}) + \frac{h^2}{2}\big(D\mathbf{N}(\mathbf{x})\big)\mathbf{N}(\mathbf{x}) + \mathbf{R}\big(h^3, \mathbf{x}\big).$$

Now, we introduce the two most common splitting approximations, which serve as the main building blocks for the proposed estimators.

DEFINITION 2.3. Let Assumptions (A1) and (A2) hold. The Lie–Trotter and Strang splitting approximations of the solution of (2) are given by

$$\mathbf{X}_{t_k}^{[\mathrm{LT}]} := \Phi_h^{[\mathrm{LT}]}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\big) = \big(\Phi_h^{[1]} \circ \Phi_h^{[2]}\big)\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\big) = \boldsymbol{\mu}_h\big(\boldsymbol{f}_h\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\big)\big) + \boldsymbol{\xi}_{h,k}, \tag{10}$$

$$\begin{aligned}\mathbf{X}_{t_k}^{[\mathrm{S}]} &:= \Phi_h^{[\mathrm{S}]}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\big) = \big(\Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]}\big)\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\big) \\ &= \boldsymbol{f}_{h/2}\big(\boldsymbol{\mu}_h\big(\boldsymbol{f}_{h/2}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\big)\big) + \boldsymbol{\xi}_{h,k}\big).\end{aligned} \tag{11}$$

REMARK 4. The order of composition in the splitting schemes is not unique. Changing the order in the S splitting leads to a sum of 2 independent random variables, one Gaussian and one non-Gaussian, whose likelihood is not trivial. Thus, we only use the splitting (11). The reversed order in the LT splitting can be treated the same way as the S splitting.

REMARK 5. Splitting the drift $\mathbf{F}(\mathbf{x})$ into a linear and a nonlinear part is not unique. However, all theorems and properties, particularly consistency and asymptotic normality of the estimators, hold for any splitting choice. Yet, for fixed step size h and sample size N, certain splittings perform better than others. In this paper, we present two general and intuitive strategies. The first applies when the system has a fixed point; here, the linear part of the splitting is the linearization around the fixed point. The linear OU performs accurately near the fixed point, with the nonlinear part correcting for nonlinear deviations. Simulations consistently show this approach to perform best. Another strategy is to linearize around the measured average value for each coordinate. An in-depth analysis of the splitting strategies for a specific example is provided in Section 2.5.

REMARK 6. Trajectories of S and LT splittings coincide up to the first $h/2$ and the last $h/2$ steps of the flow $\Phi^{[2]}_{h/2}$. Indeed, when applied k times, the S splitting can be written as

$$\big(\Phi^{[\mathrm{S}]}_h\big)^k(\mathbf{x}_0) = \big(\Phi^{[2]}_{h/2} \circ \big(\Phi^{[\mathrm{LT}]}_h\big)^k \circ \Phi^{[2]}_{-h/2}\big)(\mathbf{x}_0).$$

Thus, it is natural that LT and S have the same order of L^p convergence. We prove this in Section 3. However, the LT and S trajectories differ in their output points (10) and (11). Strang splitting outputs the middle points of the smooth steps of the deterministic flow (8), while LT splitting outputs the stochastic increments in the rough steps. We conjecture that this is one of the reasons why the S splitting has superior statistical properties.

2.4. *Estimators*. In this section, we first introduce two new estimators, LT and S, given a sample $\mathbf{X}_{0:t_N}$. Subsequently, we provide a brief overview of the estimators EM, K, LL and HE, which will be compared in the simulation study.

2.4.1. *Splitting estimators*. The LT scheme (10) follows a Gaussian distribution. Consequently, the objective function corresponds to (twice) the negative pseudo-log-likelihood:

$$\begin{aligned}\mathcal{L}^{[\mathrm{LT}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) &\overset{\theta}{=} N\log\big(\det \boldsymbol{\Omega}_h(\boldsymbol{\theta})\big)\\ &\quad+\sum_{k=1}^{N}\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big)^\top \boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1}\\ &\quad\times\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big).\end{aligned} \tag{12}$$

The S splitting (11) is a nonlinear transformation of the Gaussian random variable $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}\times(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta})+\boldsymbol{\xi}_{h,k}$. We first define

$$\mathbf{Z}_{t_k}(\boldsymbol{\beta}) := \boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h\big(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big). \tag{13}$$

Afterwards, we apply a change of variables to derive the following objective function:

$$\begin{aligned}\mathcal{L}^{[\mathrm{S}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) &\overset{\theta}{=} N\log\big(\det \boldsymbol{\Omega}_h(\boldsymbol{\theta})\big)+\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top \boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta})\\ &\quad-2\sum_{k=1}^{N}\log\big|\det D\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k};\boldsymbol{\beta})\big|.\end{aligned} \tag{14}$$

The last term is due to the nonlinear transformation and is an extra term that does not appear in commonly used pseudo-likelihoods.

The inverse function $\boldsymbol{f}_h^{-1}$ may not exist for all parameters in the search domain of the optimization algorithm. However, this problem can often be solved numerically. When $\boldsymbol{f}_h^{-1}$ is well defined, we use the identity $-\log|\det D\boldsymbol{f}_h^{-1}(\mathbf{x};\boldsymbol{\beta})| = \log|\det D\boldsymbol{f}_h(\mathbf{x};\boldsymbol{\beta})|$ in (14) to increase the speed and numerical stability.

Finally, we define the estimators as

$$\widehat{\boldsymbol{\theta}}_N^{[k]} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{[k]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}), \quad k \in \{\mathrm{LT}, \mathrm{S}\}. \tag{15}$$

2.4.2. *Euler–Maruyama.* The EM method uses first-order Taylor expansion of (1):

$$\mathbf{X}_{t_k}^{[\mathrm{EM}]} := \mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]} + h\mathbf{F}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]};\boldsymbol{\beta}\big) + \boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]}, \tag{16}$$

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]} \overset{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}, h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ for $k = 1, \dots, N$ (Kloeden and Platen (1992)). The transition density $p^{[\mathrm{EM}]}(\mathbf{X}_{t_k}|\mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ is Gaussian, so the pseudo-likelihood follows trivially.

2.4.3. *Kessler's Gaussian approximation.* The K estimator uses Gaussian transition densities $p^{[\mathrm{K}]}(\mathbf{X}_{t_k}|\mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ with the true mean and covariance of the solution $\mathbf{X}$ (Kessler (1997)). When the moments are unknown, they are approximated using the infinitesimal generator (Lemma 2.1). We implement the estimator K based on the 2nd-order approximation:

$$\begin{aligned}\mathbf{X}_{t_k}^{[\mathrm{K}]} &:= \mathbf{X}_{t_{k-1}}^{[\mathrm{K}]} + h\mathbf{F}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta}\big) + \boldsymbol{\xi}_{h,k}^{[\mathrm{K}]}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}\big)\\ &\quad + \frac{h^2}{2}\Big(D\mathbf{F}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta}\big)\mathbf{F}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta}\big) + \frac{1}{2}\big[\mathrm{Tr}\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{H}_{F^{(i)}}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta}\big)\big)\big]_{i=1}^d\Big),\end{aligned} \tag{17}$$

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{K}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_{h,k}^{[\mathrm{K}]}(\boldsymbol{\theta}))$, and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{K}]}(\boldsymbol{\theta}) = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}(D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top D^\top\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]};\boldsymbol{\beta}))$. The covariance matrix is not constant, which makes the algorithm slower for a larger sample size.

2.4.4. *Ozaki's local linearization.* Ozaki's LL method approximates the drift of (1) between consecutive observations by a linear function (Jimenez, Shoji and Ozaki (1999)). The LL method consists of the following steps:

(1) Perform LL of the drift $\mathbf{F}$ in each time interval $[t, t+h)$ by the Itô–Taylor series;
(2) Compute the analytic solution of the resulting linear SDE.

The approximation becomes

$$\mathbf{X}_{t_k}^{[\mathrm{LL}]} := \mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]} + \Phi_h^{[\mathrm{LL}]}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\theta}\big) + \boldsymbol{\xi}_{h,k}^{[\mathrm{LL}]}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}\big), \tag{18}$$

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}))$, and

$$\begin{aligned}\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}) &:= \int_0^h e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})(h-u)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})^\top(h-u)}\,\mathrm{d}u,\\ \Phi_h^{[\mathrm{LL}]}(\mathbf{x};\boldsymbol{\theta}) &:= \mathbf{R}_{h,0}\big(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})\big)\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) + \big(h\mathbf{R}_{h,0}\big(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})\big)\\ &\qquad - \mathbf{R}_{h,1}\big(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})\big)\big)\mathbf{M}(\mathbf{x};\boldsymbol{\theta}),\\ \mathbf{R}_{h,i}\big(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})\big) &:= \int_0^h \exp\big(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})u\big)u^i\,\mathrm{d}u, \quad i = 0, 1,\end{aligned}$$

$$\mathbf{M}(\mathbf{x};\boldsymbol{\theta}) := \frac{1}{2}\big(\operatorname{Tr}\mathbf{H}_1(\mathbf{x};\boldsymbol{\theta}),\dots,\operatorname{Tr}\mathbf{H}_d(\mathbf{x};\boldsymbol{\theta})\big)^\top,$$

$$\mathbf{H}_k(\mathbf{x};\boldsymbol{\theta}) := \left[\big[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big]_{ij}\frac{\partial^2 F^{(k)}}{\partial x^{(i)}\partial x^{(j)}}(\mathbf{x})\right]_{i,j=1}^{d}.$$

We can efficiently compute $\mathbf{R}_{h,i}$ and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ using formulas from (Van Loan (1978)); see (Gu, Wu and Xue (2020)). For more details, see the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

Thus, $p^{[\mathrm{LL}]}(\mathbf{X}_{t_k}|\mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ is Gaussian and standard likelihood inference applies. Similar to K, $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ depends on the previous state $\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}$, which is a major downside since it is harder to implement and slower to run due to the computation of $N-1$ covariance matrices. Unlike K, LL does not Taylor expand the approximated drift and covariance matrix, so the influence of sample size N on computational times is much larger.

2.4.5. *Aït-Sahalia's infinite Hermite expansion.* The HE method (Aït-Sahalia (2002, 2008)) approximates the likelihood using two transformations to make data resemble a normal distribution, facilitating corrections for finite samples. First, $\mathbf{X}_t$ is transformed to unit diffusion $\mathbf{Y}_t$, using the Lamperti transform. Then $\mathbf{Y}_t$ is transformed into a more normal-like $\mathbf{Z}_t$. Finally, the objective function is a Hermite expansion in terms of convergent power series in h, around this normal density before reverting back to $\mathbf{X}_t$. The Lamperti transform can be omitted for nonreducible diffusions (Aït-Sahalia (2008)). For additive noise, the HE objective function of order J is given as

$$\begin{aligned}\mathcal{L}^{[\mathrm{HE}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) &\overset{\theta}{=} N\log\big(\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big)\\ &\quad - 2\sum_{k=1}^{N}\left(\frac{C_Y^{(-1)}(\boldsymbol{\gamma}(\mathbf{X}_{t_k})|\boldsymbol{\gamma}(\mathbf{X}_{t_{k-1}}))}{h} + \sum_{j=0}^{J}\frac{h^j}{j!}C_Y^{(j)}(\boldsymbol{\gamma}(\mathbf{X}_{t_k})|\boldsymbol{\gamma}(\mathbf{X}_{t_{k-1}}))\right).\end{aligned} \tag{19}$$

Function $\boldsymbol{\gamma}$ is the Lamperti transform, and functions $C_Y^{(j)}$, for $j=-1,0,1,\dots,J$ are calculated recursively according to Theorem 1 in (Aït-Sahalia (2008)).

2.5. *An example*: *The stochastic Lorenz system.* The Lorenz system is a 3D system introduced by Lorenz (1963) to model atmospheric convection. The model is originally deterministic exhibiting deterministic chaos, that is, tiny differences in initial conditions lead to unpredictable and widely diverging trajectories. The Lorenz system evolves around two strange attractors, implying that trajectories remain within some bounded region, while points that start in close proximity may eventually separate by arbitrary distances as time progresses (Hilborn (1994)). We add noise to include unmodeled forces and randomness. The stochastic Lorenz system is given by

$$\begin{aligned}\mathrm{d}X_t &= p(Y_t - X_t)\,\mathrm{d}t + \sigma_1\,\mathrm{d}W_t^{(1)},\\ \mathrm{d}Y_t &= (rX_t - Y_t - X_tZ_t)\,\mathrm{d}t + \sigma_2\,\mathrm{d}W_t^{(2)},\\ \mathrm{d}Z_t &= (X_tY_t - cZ_t)\,\mathrm{d}t + \sigma_3\,\mathrm{d}W_t^{(3)}.\end{aligned} \tag{20}$$

The variables X_t, Y_t and Z_t represent convective intensity, and horizontal and vertical temperature differences, respectively. Parameters p, r and c denote the Prandtl number, the Rayleigh number and a geometric factor, respectively (Tabor (1989)). Lorenz (1963) used the values $p=10$, $r=28$ and $c=8/3$, yielding chaotic behavior.

The system does not fulfill the global or the one-sided Lipschitz condition because it is a second-order polynomial (Humphries and Stuart (1994)). However, it has a unique global

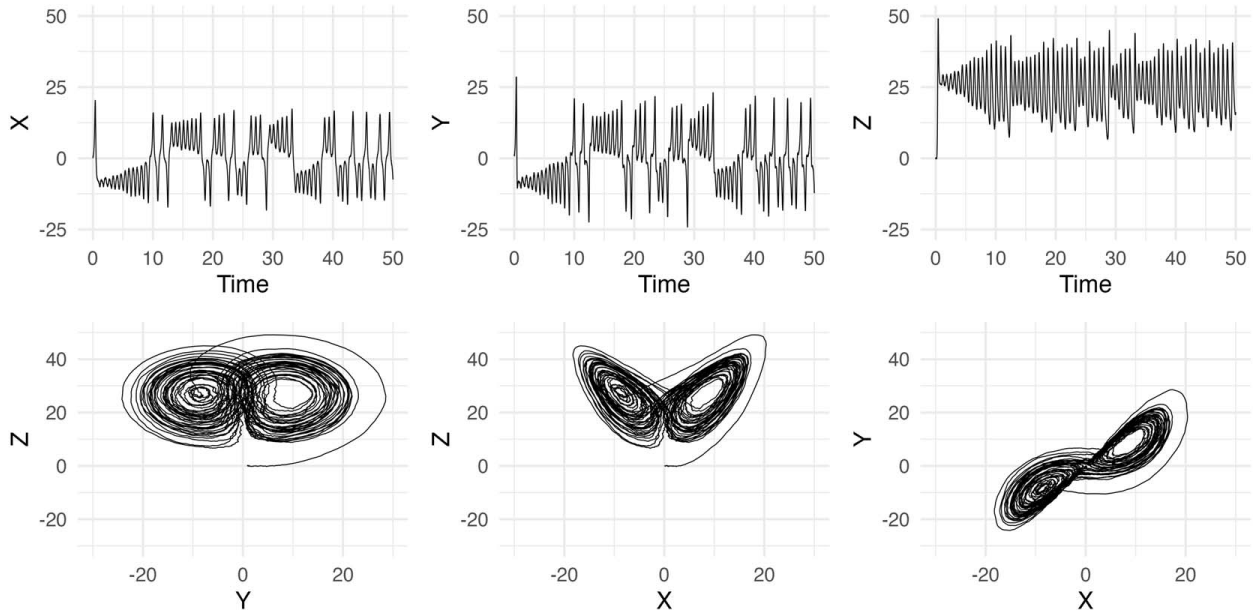

FIG. 1. *An example trajectory of the stochastic Lorenz system* (20) *starting at* $(0, 1, 0)$ *for* $N = 10{,}000$ *and* $h = 0.005$. *The first row shows the evolution of the individual components* X, Y *and* Z. *The second row shows the evolution of component pairs*: (Y, Z), (X, Z) *and* (X, Y). *Parameters are* $p = 10$, $r = 28$, $c = 8/3$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ *and* $\sigma_3^2 = 1.5$.

solution and an invariant probability (Keller (1996)). Thus, all assumptions (A2)–(A5), except (A1) hold. Even so, we show in Section 6 that the estimators work.

Different approaches for estimating parameters in the Lorenz system have been proposed, mostly in the deterministic case. Zhuang et al. (2020) and Lazzús, Rivera and López-Caraballo (2016) used sophisticated optimization algorithms to achieve better precision. Dubois et al. (2020) and Ann et al. (2022) used deep neural networks in combination with other machine learning algorithms. Ozaki, Jimenez and Haggan-Ozaki (2000) used Kalman filtering based on LL on the stochastic Lorenz system.

Figure 1 shows an example trajectory of the stochastic Lorenz system. The trajectory was generated by subsampling from an EM simulation, such that $N = 10{,}000$ and $h = 0.05$, with parameter values $p = 10$, $r = 28$, $c = 8/3$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\sigma_3^2 = 1.5$. Even if the trajectory had not been stochastic, the unpredictable jumps in the first row of Figure 1 would still have been there due to the chaotic behavior.

We suggest to split SDE (20) by choosing the OU part (3) as the linearization around one of the two fixed points $(x^\star, y^\star, z^\star) = (\pm\sqrt{c(r-1)}, \pm\sqrt{c(r-1)}, r-1)$. For simplicity, we exclude the fixed point $(0, 0, 0)$ since X and Y spend little time around this point; see Figure 1. Specifically, we apply a mixture of two splittings, linearizing around $(\sqrt{c(r-1)}, \sqrt{c(r-1)}, r-1)$ when $X > 0$ and around $(-\sqrt{c(r-1)}, -\sqrt{c(r-1)}, r-1)$ when $X < 0$. We denote these estimators by $\mathrm{LT}_{\mathrm{mix}}$ and $\mathrm{S}_{\mathrm{mix}}$. The splitting is given by

$$\mathbf{A}_{\mathrm{mix}} = \begin{bmatrix} -p & p & 0 \\ 1 & -1 & -x^\star \\ y^\star & x^\star & -c \end{bmatrix}, \qquad \mathbf{b}_{\mathrm{mix}} = \begin{bmatrix} x^\star \\ y^\star \\ z^\star \end{bmatrix}, \qquad \mathbf{N}_{\mathrm{mix}}(x, y, z) = \begin{bmatrix} 0 \\ -(x - x^\star)(z - z^\star) \\ (x - x^\star)(y - y^\star) \end{bmatrix}.$$

The OU process is mean-reverting toward $\mathbf{b}_{\mathrm{mix}} = (x^\star, y^\star, z^\star)$. The nonlinear solution is

$$\boldsymbol{f}_{\mathrm{mix},h}(x, y, z) = \begin{bmatrix} x \\ (y - y^\star)\cos\big(h(x - x^\star)\big) - (z - z^\star)\sin\big(h(x - x^\star)\big) + y^\star \\ (y - y^\star)\sin\big(h(x - x^\star)\big) + (z - z^\star)\cos\big(h(x - x^\star)\big) + z^\star \end{bmatrix}.$$

The solution is a composition of a 3D rotation and translation of (y, z) around the fixed point. The inverse always exists, and thus, Assumption (A6) holds. Moreover, $\det D\boldsymbol{f}_{\mathrm{mix},h}^{-1}(\cdot) = 1$.

The mixing strategy does not increase the complexity of the implementation significantly, and it is straightforward to incorporate into the existing framework. Thus, this splitting strategy is convenient when the model has several fixed points.

An alternative splitting linearizes around the average of the observations. Let (μ_x, μ_x, μ_z) be the average of the data, where we put $\mu_x = \mu_y$ since the difference of their averages is small, around 10^{-3}. We denote these estimators by $\mathrm{LT}_{\mathrm{avg}}$ and $\mathrm{S}_{\mathrm{avg}}$. The splitting is given by

$$\mathbf{A}_{\mathrm{avg}} = \begin{bmatrix} -p & p & 0 \\ r-\mu_z & -1 & -\mu_x \\ \mu_x & \mu_x & -c \end{bmatrix}, \qquad \mathbf{b}_{\mathrm{avg}} = \begin{bmatrix} \mu_x \\ \mu_x \\ \mu_z \end{bmatrix},$$

$$\mathbf{N}_{\mathrm{avg}}(x,y,z) = \begin{bmatrix} 0 \\ -(x-\mu_x)(z-\mu_z) + (r-1-\mu_z)\mu_x \\ (x-\mu_x)(y-\mu_x) + \mu_x^2 - c\mu_z \end{bmatrix}.$$

The nonlinear solution is

$$\boldsymbol{f}_{\mathrm{avg},h}(x,y,z) = \begin{bmatrix} \mu_x \\ \mu_x + \frac{c\mu_z - \mu_x^2}{x-\mu_x} \\ \mu_z + \frac{\mu_x(r-1-\mu_z)}{x-\mu_x} \end{bmatrix} + \begin{bmatrix} x-\mu_x \\ \left(y-\mu_x - \frac{c\mu_z-\mu_x^2}{x-\mu_x}\right)\cos\big(h(x-\mu_x)\big) - \left(z-\mu_z - \frac{\mu_x(r-1-\mu_z)}{x-\mu_x}\right)\sin\big(h(x-\mu_x)\big) \\ \left(y-\mu_x - \frac{c\mu_z-\mu_x^2}{x-\mu_x}\right)\sin\big(h(x-\mu_x)\big) + \left(z-\mu_z - \frac{\mu_x(r-1-\mu_z)}{x-\mu_x}\right)\cos\big(h(x-\mu_x)\big) \end{bmatrix},$$

where $\boldsymbol{f}_{\mathrm{avg},h}(\mu_x, y, z) := (\mu_x, y + h\mu_x(r-1-\mu_z), z + h\mu_x^2 - c\mu_z)^\top$ and $\det D\boldsymbol{f}^{-1}_{\mathrm{avg},h}(\cdot) = 1$.

3. Order of one-step predictions and L^p convergence. In this section, we investigate L^p convergence of the splitting schemes and the order of the one-step predictions. Theorem 2.1 in Tretyakov and Zhang (2013) extends Milstein's fundamental theorem on L^p convergence for global Lipschitz coefficients (Milstein (1987)) to Assumptions (A1) and (A2). This theorem provides the theoretical underpinning for our approach, drawing on the key concepts of L^p consistency and boundedness of moments.

DEFINITION 3.1 (L^p consistency of a numerical scheme). The one-step approximation $\widetilde{\Phi}_h$ of the solution $\mathbf{X}$ is L^p consistent, $p \geq 1$, of order $q_2 - 1/2 \geq 0$, if for $k = 1, \ldots, N$ and some $q_1 \geq q_2 + 1/2$:

$$\big\|\mathbb{E}\big[\mathbf{X}_{t_k} - \widetilde{\Phi}_h(\mathbf{X}_{t_{k-1}})|\mathbf{X}_{t_{k-1}} = \mathbf{x}\big]\big\| = R\big(h^{q_1}, \mathbf{x}\big),$$

$$\big(\mathbb{E}\big[\big\|\mathbf{X}_{t_k} - \widetilde{\Phi}_h(\mathbf{X}_{t_{k-1}})\big\|^{2p}|\mathbf{X}_{t_{k-1}} = \mathbf{x}\big]\big)^{\frac{1}{2p}} = R\big(h^{q_2}, \mathbf{x}\big).$$

DEFINITION 3.2 (Bounded moments of a numerical scheme). A numerical approximation $\widetilde{\mathbf{X}}$ of the solution $\mathbf{X}$ has bounded moments, if for all $p \geq 1$, there exists constant $C > 0$, such that, for $k = 1, \ldots, N$:

$$\mathbb{E}\big[\|\widetilde{\mathbf{X}}_{t_k}\|^{2p}\big] \leq C\big(1 + \|\mathbf{x}_0\|^{2p}\big).$$

The following theorem (Theorem 2.1 in Tretyakov and Zhang (2013)) gives sufficient conditions for L^p convergence of a numerical scheme in a one-sided Lipschitz framework.

THEOREM 3.3 (L^p convergence of a numerical scheme). *Let Assumptions (A1) and (A2) hold, and let $\widetilde{\mathbf{X}}_{t_k}$ be a numerical approximation of the solution $\mathbf{X}_{t_k}$ of (1) at time t_k. If*:

(1) *The one-step approximation $\widetilde{\mathbf{X}}_{t_k} = \widetilde{\Phi}_h(\widetilde{\mathbf{X}}_{t_{k-1}})$ is L^p consistent of order $q_2 - 1/2$; and*
(2) *$\widetilde{\mathbf{X}}$ has bounded moments,*

then $\widetilde{\mathbf{X}}$ is L^p convergent, $p \geq 1$, of order $q_2 - 1/2$, that is, for $k = 1, \ldots, N$, it holds:

$$\big(\mathbb{E}\big[\|\mathbf{X}_{t_k} - \widetilde{\mathbf{X}}_{t_k}\|^{2p}\big]\big)^{\frac{1}{2p}} = R\big(h^{q_2-1/2}, \mathbf{x}_0\big).$$

3.1. *Lie–Trotter splitting*. We first show that the one-step LT approximation is of order $R(h^2, \mathbf{x}_0)$ in mean. The following proposition is proved in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)) for scheme (10), as well as for the reversed order of composition. We demonstrate that the order of one-step prediction cannot be improved unless the drift $\mathbf{F}$ is linear.

PROPOSITION 3.4 (One-step prediction of LT splitting). *Assume (A1)–(A2), let $\mathbf{X}$ be the solution to (1) and let $\Phi_h^{[\mathrm{LT}]}$ be the LT approximation (10). Then, for $k = 1, \ldots, N$, it holds*:

$$\big\|\mathbb{E}\big[\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{LT}]}(\mathbf{X}_{t_{k-1}})|\mathbf{X}_{t_{k-1}} = \mathbf{x}\big]\big\| = R\big(h^2, \mathbf{X}_{t_{k-1}}\big).$$

L^p convergence of the LT splitting scheme is established in Theorem 2 in Buckwar et al. (2022), which we repeat here for convenience.

THEOREM 3.5 (L^p convergence of the LT splitting). *Assume (A1)–(A2), let $\mathbf{X}^{[\mathrm{LT}]}$ be the LT approximation defined in (10) and let $\mathbf{X}$ be the solution of (1). Then there exists $C \geq 1$ such that for all $p \geq 2$, and $k = 1, \ldots, N$, it holds*:

$$\big(\mathbb{E}\big[\big\|\mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{LT}]}\big\|^p\big]\big)^{\frac{1}{p}} = R(h, \mathbf{x}_0).$$

Now, we investigate the same properties for the S splitting.

3.2. *Strang splitting*. The following proposition states that the S splitting (11) has higher order one-step predictions than the LT splitting (10). The proof can be found in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

PROPOSITION 3.6. *Assume (A1)–(A2), let $\mathbf{X}$ be the solution to (1) and let $\Phi_h^{[\mathrm{S}]}$ be the S splitting approximation (11). Then, for $k = 1, \ldots, N$, it holds*:

$$\big\|\mathbb{E}\big[\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})|\mathbf{X}_{t_{k-1}} = \mathbf{x}\big]\big\| = R\big(h^3, \mathbf{X}_{t_{k-1}}\big). \tag{21}$$

REMARK 7. Even though LT and S have the same order of L^p convergence, the crucial difference is in the one-step prediction. The approximated transition density between two consecutive data points depends on the one-step approximation. Thus, the objective function based on pseudo-likelihood from the S splitting is more precise than the one from the LT.

To prove L^p convergence of the S splitting scheme for (1) with one-sided Lipschitz drift, we follow the same procedure as in Buckwar et al. (2022). The proof of the following theorem is in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

THEOREM 3.7 (L^p convergence of S splitting). *Assume (A1), (A2) and (A6), let* $\mathbf{X}^{[\mathrm{S}]}$ *be the S splitting defined in (11) and let* $\mathbf{X}$ *be the solution of (1). Then there exists* $C \geq 1$ *such that for all* $p \geq 2$ *and* $k = 1, \ldots, N$, *it holds*:

$$\big(\mathbb{E}\big[\big\|\mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{S}]}\big\|^p\big]\big)^{\frac{1}{p}} = R(h, \mathbf{x}_0).$$

Before we move to parameter estimation, we prove a useful corollary.

COROLLARY 3.8. *Let all assumptions from Theorem 3.7 hold. Then* $(\mathbb{E}[\|\mathbf{Z}_{t_k} - \boldsymbol{\xi}_{h,k}\|^p])^{1/p} = R(h, \mathbf{x}_0)$.

PROOF. From the definition of $\mathbf{Z}_{t_k}$ in (13), it is enough to prove that

$$\big(\mathbb{E}\big[\big\|\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h\big(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}})\big) - \boldsymbol{\xi}_{h,k}\big\|^p\big]\big)^{1/p} = R(h, \mathbf{x}_0).$$

From (11), we have that $\boldsymbol{\xi}_{h,k} = \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}))$. Then

$$\begin{aligned}
&\mathbb{E}\big[\big\|\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h\big(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}})\big) - \boldsymbol{\xi}_{h,k}\big\|^p\big]^{1/p} \\
&\quad \leq C\big(\mathbb{E}\big[\big\|\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}^{[\mathrm{S}]})\big\|^p\big] + \mathbb{E}\big[\big\|\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}) - \boldsymbol{f}_{h/2}\big(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\big)\big\|^p\big]\big)^{1/p} \\
&\quad \leq C\big(\mathbb{E}\big[\big\|\mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{S}]}\big\|^p\big] + \mathbb{E}\big[\big\|\mathbf{X}_{t_{k-1}} - \mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\big\|^p\big]\big)^{1/p} + R(h, \mathbf{x}_0).
\end{aligned}$$

We used Proposition 2.2 that $\mathbf{X}, \mathbf{X}^{[\mathrm{S}]}$ have finite moments and $\boldsymbol{f}_{h/2}, \boldsymbol{f}_{h/2}^{-1}$ grow polynomially. The result follows from L^p convergence of the S splitting scheme, Theorem 3.7. □

4. Auxiliary properties. This paper centers around proving the properties of the S estimator. There are two reasons for this. First, most numerical properties in the literature are proved only for LT splitting because proofs for S splitting are more involved. Here, we establish both the numerical properties of the S splitting as well as the properties of the estimator. Second, the S splitting introduces a new pseudo-likelihood that differs from the standard Gaussian pseudo-likelihoods. Consequently, standard tools, like those proposed by Kessler (1997), do not directly apply.

The asymptotic properties of the LT estimator are the same as for the S estimator. However, the following auxiliary properties will be stated and proved only for the S estimator. They can be reformulated for the LT estimator following the same logic.

Before presenting the central results for the estimator, we establish the groundwork with two essential lemmas that rely on the model assumptions. Lemma 4.1 (Lemma 6 in Kessler (1997)) deals with the pth moments of the SDE increments and also provides a moment bound of a polynomial map of the solution. The proof of this lemma in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)) differs from that in Kessler (1997) due to our relaxation of the global Lipschitz assumption of the drift $\mathbf{F}$. Instead, we use a one-sided Lipschitz condition in conjunction with the generalized Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020) to establish the result, see the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024))).

Lemma 4.2 (Lemma 8 in Kessler (1997), Lemma 2 in Sørensen and Uchida (2003)) constitutes a central ergodic property that is essential for establishing the asymptotic behavior of the estimator. The proof when the drift $\mathbf{F}$ is one-sided Lipschitz is identical to the one presented in Kessler (1997), particularly when combined with Lemma 4.1.

LEMMA 4.1. *Assume (A1)–(A2). Let* $\mathbf{X}$ *be the solution of (1). For* $t_k \geq t \geq t_{k-1}$, *where* $h = t_k - t_{k-1} < 1$, *the following two statements hold*:

(1) *For $p \geq 1$, there exists $C_p > 0$ that depends on p, such that*

$$\mathbb{E}\big[\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\|^p | \mathcal{F}_{t_{k-1}}\big] \leq C_p (t - t_{k-1})^{p/2} \big(1 + \|\mathbf{X}_{t_{k-1}}\|\big)^{C_p}.$$

(2) *If $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ is of polynomial growth in $\mathbf{x}$ uniformly in $\boldsymbol{\theta}$, then there exist constants C and $C_{t-t_{k-1}}$ that depends on $t - t_{t_{k-1}}$, such that*

$$\mathbb{E}\big[\big|g(\mathbf{X}_t; \boldsymbol{\theta})\big| | \mathcal{F}_{t_{k-1}}\big] \leq C_{t-t_{k-1}} \big(1 + \|\mathbf{X}_{t_{k-1}}\|\big)^{C}.$$

LEMMA 4.2. *Assume (A1), (A2), (A3) and let $\mathbf{X}$ be the solution to (1). Let $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ be a differentiable function with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$ with derivative of polynomial growth in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}$. If $h \to 0$ and $Nh \to \infty$, then*

$$\frac{1}{N} \sum_{k=1}^{N} g(\mathbf{X}_{t_k}, \boldsymbol{\theta}) \xrightarrow[\substack{Nh \to \infty \\ h \to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \int g(\mathbf{x}, \boldsymbol{\theta}) \, \mathrm{d}\nu_0(\mathbf{x}),$$

uniformly in $\boldsymbol{\theta}$.

Lastly, we state the moment bounds needed for the estimator asymptotics. The proof is in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

PROPOSITION 4.3 (Moment bounds). *Assume (A1), (A2), (A6). Let $\mathbf{X}$ be the solution of (1), and $\mathbf{Z}_{t_k}$ as defined in (13). Let $\mathbf{g}(\mathbf{x}; \boldsymbol{\beta})$ be a generic function with derivatives of polynomial growth, and $\boldsymbol{\beta} \in \Theta_\beta$. Then, for $k = 1, \ldots, N$, the following moment bounds hold*:

(i) $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) | \mathbf{X}_{t_{k-1}} = \mathbf{x}] = \mathbf{R}(h^3, \mathbf{X}_{t_{k-1}})$

(ii) $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta})^\top | \mathbf{X}_{t_k} = \mathbf{x}] = \frac{h}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top \mathbf{g}(\mathbf{x}; \boldsymbol{\beta}) + D\mathbf{g}(\mathbf{x}; \boldsymbol{\beta}) \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}})$;

(iii) $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top | \mathbf{X}_{t_{k-1}} = \mathbf{x}] = h \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}})$.

5. Asymptotics. The estimators $\hat{\boldsymbol{\theta}}_N$ are defined in (15). However, the full objective functions (12) and (14) are not needed to prove consistency and asymptotic normality. It is enough to approximate $\boldsymbol{\Omega}_h$ up to the second order by $h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}(\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}^\top)$ (see equation (6)). Indeed, after applying Taylor series on the inverse of $\boldsymbol{\Omega}_h$, we get

$$\begin{aligned}
\boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1} &= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \Big(\mathbf{I} + \frac{h}{2}\big(\mathbf{A}(\boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\big)^{-1} \Big) + R(h, \mathbf{x}_0) \\
&= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \big(\mathbf{I} - \frac{h}{2}\big(\mathbf{A}(\boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\big) + R(h, \mathbf{x}_0) \\
&= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} - \frac{1}{2}\big((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{A}(\boldsymbol{\beta}) + \mathbf{A}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\big) + R(h, \mathbf{x}_0).
\end{aligned}$$

Similarly, we approximate the log-determinant as

$$\begin{aligned}
\log\det \boldsymbol{\Omega}_h(\boldsymbol{\theta}) &= \log\det\Big(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}\big(\mathbf{A}(\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}(\boldsymbol{\beta})^\top\big)\Big) + R\big(h^2, \mathbf{x}_0\big) \\
&\overset{\theta}{=} \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \log\det\Big(\mathbf{I} + \frac{h}{2}\big(\mathbf{A}(\boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\big)\Big) + R\big(h^2, \mathbf{x}_0\big) \\
&= \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h}{2}\operatorname{Tr}\big(\mathbf{A}(\boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\big) + R\big(h^2, \mathbf{x}_0\big) \\
&= \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + h \operatorname{Tr}\mathbf{A}(\boldsymbol{\beta}) + R\big(h^2, \mathbf{x}_0\big).
\end{aligned}$$

Using the same approximation, we obtain

$$\begin{aligned}2\log\left|\det D\boldsymbol{f}_{h/2}(\mathbf{x};\boldsymbol{\beta})\right| &= 2\log\left|\det\left(\mathbf{I}+\frac{h}{2}D\mathbf{N}(\mathbf{x};\boldsymbol{\beta})\right)\right| \\ &= 2\log\left|1+\frac{h}{2}\operatorname{Tr}D\mathbf{N}(\mathbf{x};\boldsymbol{\beta})\right| + R(h,\mathbf{x}) \\ &= h\operatorname{Tr}D\mathbf{N}(\mathbf{x};\boldsymbol{\beta}) + R\big(h^2,\mathbf{x}_0\big).\end{aligned}$$

Retaining terms up to order $R(Nh^2,\mathbf{x}_0)$ from (12) and (14), we establish the approximate objective functions:

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{LT}]}(\boldsymbol{\theta}) :=\ & N\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + Nh\operatorname{Tr}\mathbf{A}(\boldsymbol{\beta}) \\ & + \frac{1}{h}\sum_{k=1}^{N}\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big)^\top\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big)^{-1} \\ & \times\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big) \\ & - \sum_{k=1}^{N}\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big)^\top\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big)^{-1} \\ & \times\mathbf{A}(\boldsymbol{\beta})\big(\mathbf{X}_{t_k}-\boldsymbol{\mu}_h\big(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}\big)\big),\end{aligned} \tag{22}$$

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\theta}) :=\ & N\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + Nh\operatorname{Tr}\mathbf{A}(\boldsymbol{\beta}) + \frac{1}{h}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}) \\ & - \sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top\big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\big)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta}) + h\sum_{k=1}^{N}\operatorname{Tr}D\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}).\end{aligned} \tag{23}$$

Unlike other likelihood-based methods, such as Kessler (1997), Aït-Sahalia (2002, 2008), Choi (2013, 2015), Yang, Chen and Wan (2019), our estimators do not involve expansions. The objective functions are formulated in simple terms without hyperparameters, such as the order of the expansions. Hence, our approach is robust and user friendly, as we directly employ (12) and (14). The approximations (22) and (23) are only used for the proofs.

5.1. *Consistency.* Now, we state the consistency of $\hat{\boldsymbol{\beta}}_N$ and $\widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top$. The proof of Theorem 5.1 is in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

THEOREM 5.1. *Assume (A1)–(A6). Let* $\mathbf{X}$ *be the solution of* (1) *and* $\widehat{\boldsymbol{\theta}}_N = (\widehat{\boldsymbol{\beta}}_N, \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top)$ *be the estimator that minimizes either* (22) *or* (23). *If* $h\to 0$ *and* $Nh\to\infty$, *then*

$$\hat{\boldsymbol{\beta}}_N \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \boldsymbol{\beta}_0, \qquad \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top.$$

5.2. *Asymptotic normality.* First, we need some preliminaries. Let $\rho>0$ and $\mathcal{B}_\rho(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta}\in\Theta \mid \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\rho\}$ be a ball around $\boldsymbol{\theta}_0$. Since $\boldsymbol{\theta}_0\in\Theta$, for sufficiently small $\rho>0$, $\mathcal{B}_\rho(\boldsymbol{\theta}_0)\in\Theta$. Let $\mathcal{L}_N$ be either (22) or (23). For $\hat{\boldsymbol{\theta}}_N\in\mathcal{B}_\rho(\boldsymbol{\theta}_0)$, the mean value theorem yields

$$\left(\int_0^1 \mathbf{H}_{\mathcal{L}_N}\big(\boldsymbol{\theta}_0 + t(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0)\big)\,\mathrm{d}t\right)(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0) = -\nabla\mathcal{L}_N(\boldsymbol{\theta}_0). \tag{24}$$

With $\boldsymbol{\varsigma} := \operatorname{vech}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) = ([\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{11}, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{12}, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{22}, \ldots, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{1d}, \ldots, [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{dd})$, we half-vectorize $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ to avoid working with tensors when computing derivatives with respect

to $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. Since $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is a symmetric $d \times d$ matrix, $\boldsymbol{\varsigma}$ is of dimension $s = d(d+1)/2$. For a diagonal matrix, instead of a half-vectorization, we use $\boldsymbol{\varsigma} := \operatorname{diag}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$. Define

$$\mathbf{C}_N(\boldsymbol{\theta}) := \begin{bmatrix} \frac{1}{Nh}\partial_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathcal{L}_N(\boldsymbol{\theta}) & \frac{1}{N\sqrt{h}}\partial_{\boldsymbol{\beta}\boldsymbol{\varsigma}}\mathcal{L}_N(\boldsymbol{\theta}) \\ \frac{1}{N\sqrt{h}}\partial_{\boldsymbol{\beta}\boldsymbol{\varsigma}}\mathcal{L}_N(\boldsymbol{\theta}) & \frac{1}{N}\partial_{\boldsymbol{\varsigma}\boldsymbol{\varsigma}}\mathcal{L}_N(\boldsymbol{\theta}) \end{bmatrix}, \tag{25}$$

$$\mathbf{s}_N := \begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\varsigma}}_N - \boldsymbol{\varsigma}_0) \end{bmatrix}, \qquad \boldsymbol{\lambda}_N := \begin{bmatrix} -\frac{1}{\sqrt{Nh}}\partial_{\boldsymbol{\beta}}\mathcal{L}_N(\boldsymbol{\theta}_0) \\ -\frac{1}{\sqrt{N}}\partial_{\boldsymbol{\varsigma}}\mathcal{L}_N(\boldsymbol{\theta}_0) \end{bmatrix}, \tag{26}$$

and $\mathbf{D}_N := \int_0^1 \mathbf{C}_N(\boldsymbol{\theta}_0 + t(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0))\,\mathrm{d}t$. Then (24) is equivalent to $\mathbf{D}_N\mathbf{s}_N = \boldsymbol{\lambda}_N$. Let

$$\mathbf{C}(\boldsymbol{\theta}_0) := \begin{bmatrix} \mathbf{C}_\beta(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \mathbf{C}_\varsigma(\boldsymbol{\theta}_0) \end{bmatrix}, \tag{27}$$

where

$$\big[\mathbf{C}_\beta(\boldsymbol{\theta}_0)\big]_{i_1,i_2} := \int \big(\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{x})\big)^\top \big(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\big)^{-1}\big(\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{x})\big)\,\mathrm{d}\nu_0(\mathbf{x}), \quad 1 \le i_1, i_2 \le r,$$

$$\big[\mathbf{C}_\varsigma(\boldsymbol{\theta}_0)\big]_{j_1,j_2} := \frac{1}{2}\operatorname{Tr}\big((\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\big), \quad 1 \le j_1, j_2 \le s.$$

Now, we state the theorem for asymptotic normality; the proof is in the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

THEOREM 5.2. *Assume (A1)–(A6). Let* $\mathbf{X}$ *be the solution of* (1), *and* $\widehat{\boldsymbol{\theta}}_N = (\widehat{\boldsymbol{\beta}}_N, \widehat{\boldsymbol{\varsigma}}_N)$ *be the estimator that minimizes either* (22) *or* (23). *If* $\boldsymbol{\theta}_0 \in \Theta$, $\mathbf{C}(\boldsymbol{\theta}_0)$ *is positive definite,* $h \to 0$, $Nh \to \infty$ *and* $Nh^2 \to 0$, *then under* $\mathbb{P}_{\boldsymbol{\theta}_0}$,

$$\begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\varsigma}}_N - \boldsymbol{\varsigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \mathbf{C}^{-1}(\boldsymbol{\theta}_0)\big). \tag{28}$$

The estimator of the diffusion parameter converges faster than the estimator of the drift parameter. Gobet (2002) showed that for a discretely sampled SDE model, the optimal convergence rates for the drift and diffusion parameters are $1/\sqrt{Nh}$ and $1/\sqrt{N}$, respectively. Thus, our estimators reach optimal rates. Moreover, the estimators are asymptotically efficient since $\mathbf{C}$ is the Fisher information matrix for the corresponding continuous-time diffusion (see Kessler (1997), Gobet (2002)). Finally, since the asymptotic correlation is zero between the drift and diffusion estimators, they are asymptotically independent.

6. Simulation study. This section presents the simulation study of the Lorenz system, illustrating the theory and comparing the proposed estimators with other likelihood-based estimators. We briefly recall the estimators, describe the simulation process and the optimization in the programming language `R` (R Core Team (2022)), and present and analyze the results.

6.1. *Estimators used in the study.* The EM transition distribution (16) for the Lorenz system (20) is

$$\begin{bmatrix} X_{t_k} \\ Y_{t_k} \\ Z_{t_k} \end{bmatrix} \Bigg| \begin{bmatrix} X_{t_{k-1}} \\ Y_{t_{k-1}} \\ Z_{t_{k-1}} \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x + hp(y-x) \\ y + h(rx - y - xz) \\ z + h(xy - cz) \end{bmatrix}, \begin{bmatrix} h\sigma_1^2 & 0 & 0 \\ 0 & h\sigma_2^2 & 0 \\ 0 & 0 & h\sigma_3^2 \end{bmatrix} \right).$$

We do not write the closed-form distributions for K (17), LL (18) and HE (19), but we use the corresponding formulas to implement the likelihoods. We implement the two splitting strategies proposed in Section 2.5, leading to four estimators: $\mathrm{LT}_{\mathrm{mix}}$, $\mathrm{LT}_{\mathrm{avg}}$, $\mathrm{S}_{\mathrm{mix}}$ and $\mathrm{S}_{\mathrm{avg}}$. To further speed up computation time, we use the same trick for calculating $\boldsymbol{\Omega}_h$ in (6) as for calculating $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$; see the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

6.2. *Trajectory simulation.* To simulate sample paths, we use the EM discretization with a step size of $h^{\mathrm{sim}} = 0.0001$, which is small enough for the EM discretization to perform well. Then we subsample the trajectory to get a larger time step h, decreasing discretization errors. We perform $M = 1000$ Monte Carlo repetitions.

6.3. *Optimization in* `R`. To optimize the objective functions, we use the `R` package `torch` (Falbel and Luraschi (2022)), which uses AD instead of the traditional finite differentiation used in `optim`. The two main advantages of AD are precision and speed. Finite differentiation is subject to floating point precision errors and is slow in high dimensions (Baydin et al. (2017)). Conversely, AD is exact and fast, and thus used in numerous applications, such as MLE or training neural networks.

We tried all available optimizers in the `torch` package and chose the resilient backpropagation algorithm `optim_rprop` based on Riedmiller and Braun (1992). It performed faster than the rest and was more precise in finding the global minimum. We used the default hyperparameters and set the optimization iterations to 200. We chose the precision of 10^{-5} between the updated and the parameters from the previous iteration as the convergence criteria. For starting values, we used $(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$. All estimators except HE converged after approximately 80 iterations. The HE estimator only converged with the smallest time step, $h = 0.005$, achieving convergence in 43%–72% of cases across various sample sizes N. This probably occurs due to a polynomial approximation of the likelihood that can be unstable at the boundaries, especially for larger h. Incorporating higher-order approximations and adding constraints in the optimization step might improve performance. For further analysis, see the Supplementary Material (Pilipovic, Samson and Ditlevsen (2024)).

6.4. *Comparing criteria.* We compare eight estimators based on their precision and speed. We compute the absolute relative error (ARE) for each component $\hat{\theta}_N^{(i)}$ of the estimator $\hat{\boldsymbol{\theta}}_N$:

$$\mathrm{ARE}\big(\hat{\theta}_N^{(i)}\big) = \frac{1}{M}\sum_{r=1}^{M} \frac{|\hat{\theta}_{N,r}^{(i)} - \theta_{0,r}^{(i)}|}{\theta_{0,r}^{(i)}}.$$

For S and LL, we compare the distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ more closely.

The running times are calculated using the `tictoc` package in `R`, measured from the start of the optimization step until the convergence criterion is met. To avoid the influence of running time outliers, we compute the median over M repetitions.

6.5. *Results.* In Figure 2, AREs are shown on log scale as a function of h. While most estimators work well for a step size no greater than 0.01, only LL, $\mathrm{S}_{\mathrm{mix}}$ and $\mathrm{S}_{\mathrm{avg}}$ perform well for $h = 0.05$. The $\mathrm{LT}_{\mathrm{avg}}$ is not competitive even for $h = 0.005$. The performance of $\mathrm{LT}_{\mathrm{mix}}$ varies, sometimes approaching the performance of K, while other times performing similarly to EM. Thus, $\mathrm{LT}_{\mathrm{mix}}$ is not a good choice for this specific model. The bias of EM starts to show for $h = 0.01$ escalating for $h = 0.05$. The largest bias appears in the diffusion parameters due to the poor approximation of $\boldsymbol{\Omega}_h^{\mathrm{EM}}$. K is less biased than EM except for p and r when $h = 0.05$. The HE estimator converged only for $h = 0.005$. The ARE is calculated

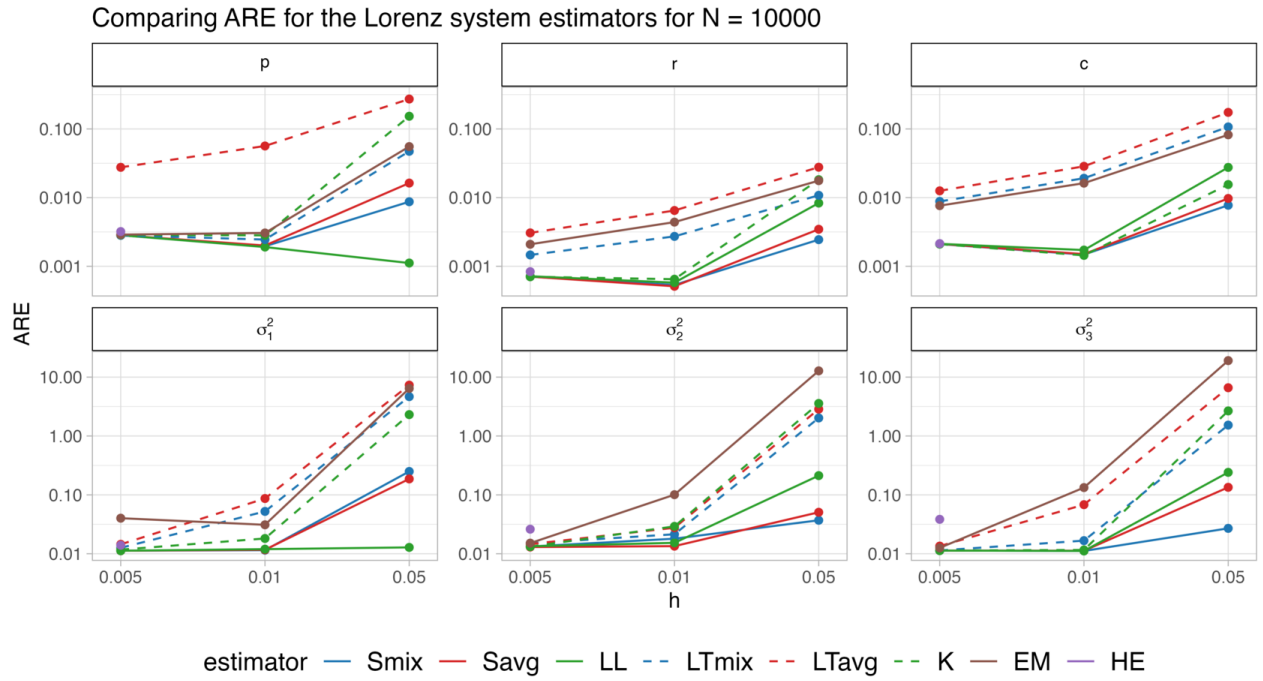

FIG. 2. *Comparing the absolute relative error (ARE) as a function of increasing discretization step h for eight estimators in the stochastic Lorenz system. The sample size is* $N = 10{,}000$. *The y-axis is on log scale. The HE estimator (purple dot) converged only for* $h = 0.005$, *and only for* 60% *of the simulated data sets.*

from the 601 simulations out of a total of 1000 in which convergence was achieved. For these, the performance of HE in estimating drift parameters is comparable to the best estimators. However, the diffusion parameters are not well estimated, with the estimation of σ_3^2 being the least accurate. Drift parameters are generally estimated better for larger h for fixed N due to a longer observation interval $T = Nh$, reflecting the $\sqrt{Nh}$ rate of convergence.

We zoom in on the distributions of $\mathrm{S}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$, LL in Figure 3. We also include HE for $h = 0.005$, based on the 60% converged estimates. For clarity, we removed some outliers for σ_1^2 and σ_2^2. This did not change the shape of the distributions, it only truncated the tails. Estimators $\mathrm{S}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$ and LL perform similarly, especially for the smallest h, where HE performs slightly worse, particularly for p, σ_2^2 and σ_3^2. For $h = 0.05$, the drift parameters are underestimated by approximately 5–10%, while the diffusion parameters are overestimated by up to 20%. Both S estimators performed better than LL, except for p and σ_1^2.

While the LL and S estimators perform similarly in terms of precision, Figure 4 shows the superiority of the S estimators over LL in computational costs. The LL becomes increasingly computationally expensive for increasing N because it calculates N covariance matrices for each parameter value. The next slowest estimators are $\mathrm{S}_{\mathrm{mix}}$ and HE, followed by $\mathrm{LT}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$, K, $\mathrm{LT}_{\mathrm{avg}}$ and, finally, EM is the fastest. The speed of EM is almost constant in N. Additionally, it seems that the running times do not depend on h. Thus, we recommend using the S estimators, especially for large N.

Figures 5 and 6 show that the theoretical results hold for $\mathrm{S}_{\mathrm{mix}}$ and $\mathrm{LT}_{\mathrm{mix}}$. We compare how the distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ change with sample size N and step size h. With increasing N, the variance decreases, whereas the mean does not change. For that, we need smaller h. To obtain negligible bias for $\mathrm{LT}_{\mathrm{mix}}$, we need a step size smaller than $h = 0.005$. However, $\mathrm{S}_{\mathrm{mix}}$ is practically unbiased up to $h = 0.01$. This shows that LT estimators might not be a good choice in practice, while the S estimators are.

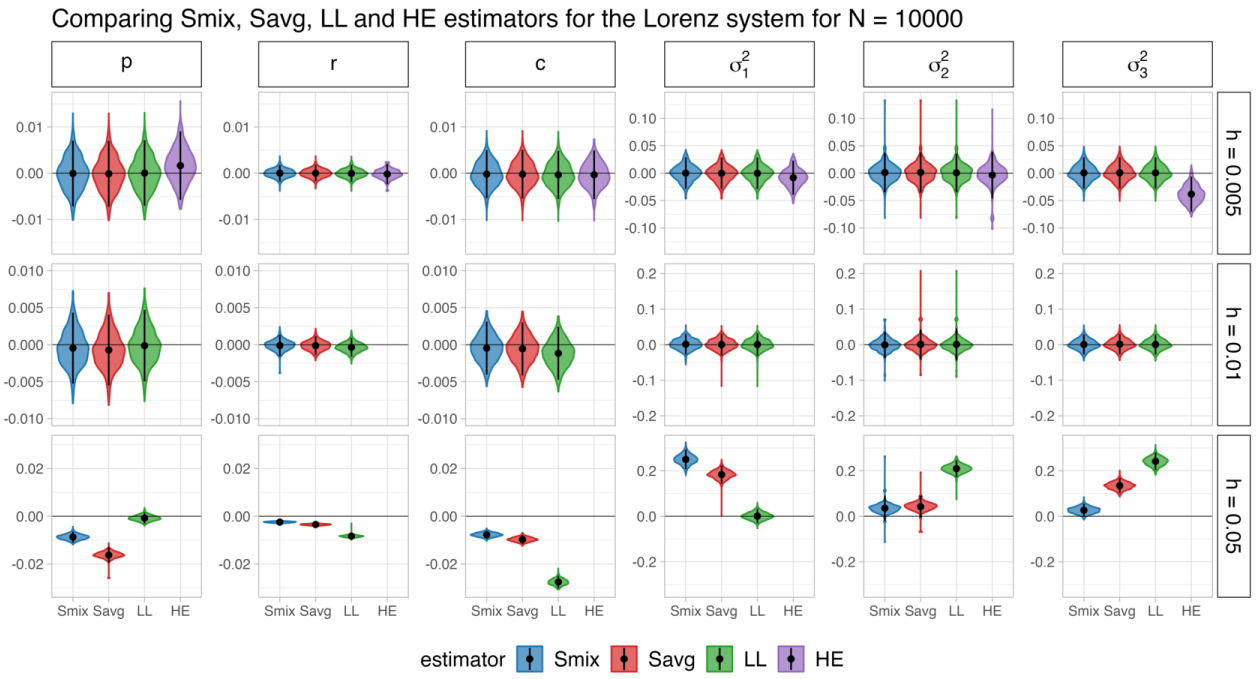

FIG. 3. *Comparing the normalized distributions of* $(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \oslash \boldsymbol{\theta}_0$ *(where* $\oslash$ *is the element-wise division) of the Lorenz system for the* S_{mix}, S_{avg}, *LL and HE estimators for* $N = 10{,}000$. *Each column represents one parameter, and each row represents one value of the discretization step* h. *The black dot with a vertical bar in each violin plot represents the mean and the standard deviation. The HE estimator (purple) converged only for* $h = 0.005$, *and only for* 60% *of the simulated data sets.*

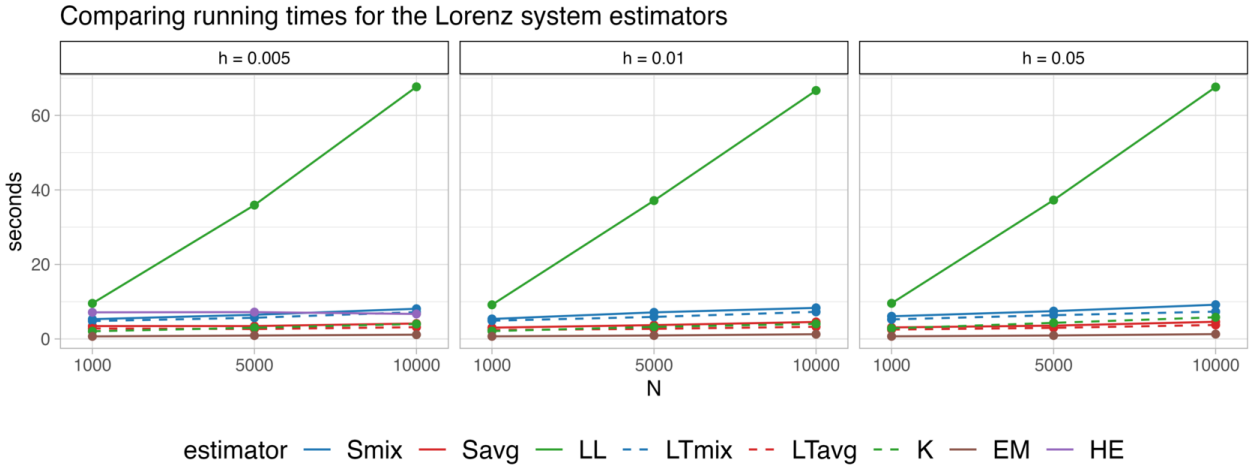

FIG. 4. *Running times as a function of* N *for different estimators of the Lorenz system. Each column shows one value of* h. *On the* x*-axis is the sample size* N, *and on the* y*-axis is the running time in seconds. The HE estimator (purple) achieved convergence only for* $h = 0.005$, *and only in* 43%–72% *of cases across various sample sizes* N.

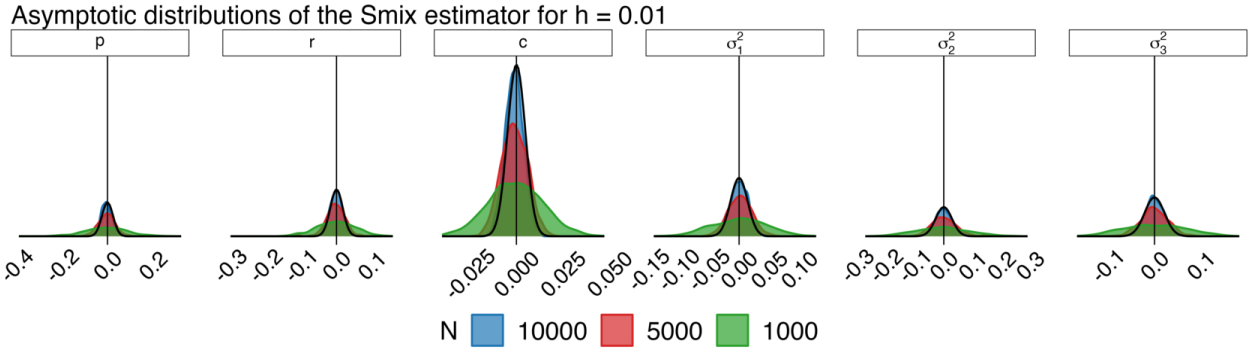

FIG. 5. *Comparing distributions of* $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ *for the* S_{mix} *estimator with theoretical asymptotic distributions (28) for each parameter (columns), for* $h = 0.01$ *and* $N \in \{1000, 5000, 10{,}000\}$ *(colors). The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for* $N = 10{,}000$ *and* $h = 0.01$.

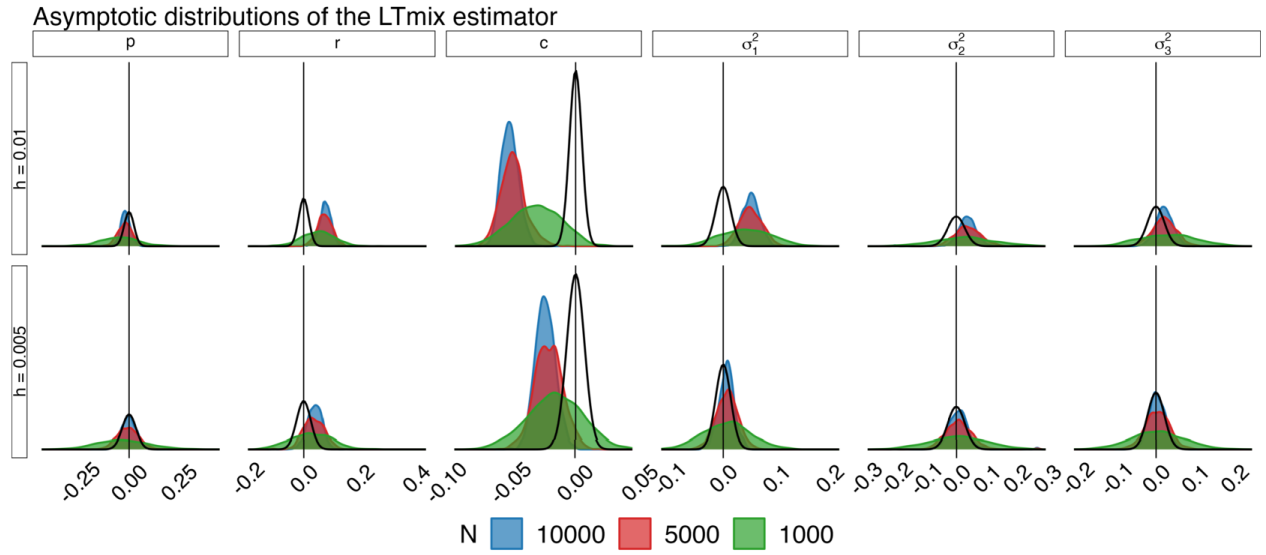

FIG. 6. *Comparing distributions of* $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ *for the* $\mathrm{LT}_{\mathrm{mix}}$ *estimator with theoretical asymptotic distributions* (28) *for each parameter* (*columns*), *for* $h \in \{0.005, 0.01\}$ (*rows*) *and* $N \in \{1000, 5000, 10{,}000\}$ (*colors*). *The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for* $N = 10{,}000$ *and corresponding* h.

The solid black lines in Figures 5 and 6 represent the theoretical asymptotic distributions computed from (28). For the Lorenz system (20), the precision matrix (27) is given by

$$\mathbf{C}(\boldsymbol{\theta}_0) = \operatorname{diag}\left(\int \frac{(y-x)^2}{\sigma_{1,0}^2} \, \mathrm{d}\nu_0(\mathbf{x}), \int \frac{x^2}{\sigma_{2,0}^2} \, \mathrm{d}\nu_0(\mathbf{x}), \int \frac{z^2}{\sigma_{3,0}^2} \, \mathrm{d}\nu_0(\mathbf{x}), \frac{1}{2\sigma_{1,0}^4}, \frac{1}{2\sigma_{2,0}^4}, \frac{1}{2\sigma_{3,0}^4} \right).$$

The integrals are approximated by taking the mean over all data points and all Monte Carlo repetitions.

Some outliers of $\hat{\sigma}_2^2$ are removed from Figures 5 and 6 by truncating the tails.

7. Conclusion. We proposed two new estimators for nonlinear multivariate SDEs. They are based on splitting schemes, a numerical approximation that preserves all important properties of the model. It was known that the LT splitting scheme has L^p convergence rate of order 1. We proved that the same holds for the S splitting. This result was expected because the overall trajectories of the S and LT splittings coincide up to the first $h/2$ and the last $h/2$ move of the flow $\Phi_{h/2}^{[2]}$. Nonetheless, S splitting is more precise in one-step predictions, which is crucial for the estimators because the objective function consists of densities between consecutive data points. Therefore, the obtained S estimator is less biased than the LT.

We proved that both estimators have optimal convergence rates for discrete observations of the SDEs. These rates are $\sqrt{N}$ for the diffusion parameter and $\sqrt{Nh}$ for the drift parameter. We also showed that the asymptotic variance of the estimators is the inverse of the Fisher information for the continuous time model. Thus, the estimators are efficient.

In the simulation study of the stochastic Lorenz system, we show the superior performance of the S estimators. We compared eight estimators based on different discretization schemes. Estimators based on Ozaki's LL and the S splitting schemes demonstrated the highest precision. However, the running time of LL is notably influenced by the sample size N, unlike the S estimator, which experiences a more gradual increase in runtime with larger N. This makes the S estimator more appropriate for large sample sizes. The LT, EM, K and HE estimators perform well for small h, but for larger h the bias increases.

While the proposed estimators are versatile, they come with certain limitations. These include assumptions like additive noise and equidistant observations. However, under specific

conditions, the Lamperti transformation can relax the constraint of additive noise. Equidistant observations can easily be relaxed due to the continuous-time formulation. Furthermore, we assumed that the diffusion parameter $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is invertible. However, there are applications where models with degenerate noise naturally arise, like second-order differential equations.

Acknowledgments. PP is also affiliated with the Bielefeld Graduate School of Economics and Management at the University of Bielefeld in Germany.

We would like to thank three anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the paper. We are thankful to the third reviewer for providing the HE method implementation for the Lorenz system.

Funding. The European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)"; and Novo Nordisk Foundation NNF20OC0062958.

This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

SUPPLEMENTARY MATERIAL

Supplementary article (DOI: 10.1214/24-AOS2371SUPPA; .pdf). The supplementary article (Pilipovic, Samson and Ditlevsen (2024)) contains proofs of results from the main text, auxiliary results, additional discussions and figures.

Computer code (DOI: 10.1214/24-AOS2371SUPPB; .zip). R code that reproduces numerical results for the simulation study in Section 6.

REFERENCES

ABDULLE, A., VILMART, G. and ZYGALAKIS, K. C. (2015). Long time accuracy of Lie–Trotter splitting methods for Langevin dynamics. *SIAM J. Numer. Anal.* **53** 1–16. MR3296612 https://doi.org/10.1137/140962644

ABLEIDINGER, M. and BUCKWAR, E. (2016). Splitting integrators for the stochastic Landau–Lifshitz equation. *SIAM J. Sci. Comput.* **38** A1788–A1806. MR3511359 https://doi.org/10.1137/15M103529X

ABLEIDINGER, M., BUCKWAR, E. and HINTERLEITNER, H. (2017). A stochastic version of the Jansen and Rit neural mass model: Analysis and numerics. *J. Math. Neurosci.* **7** Paper No. 8, 35. MR3683994 https://doi.org/10.1186/s13408-017-0046-4

AÏT-SAHALIA, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica* **70** 223–262. MR1926260 https://doi.org/10.1111/1468-0262.00274

AÏT-SAHALIA, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Ann. Statist.* **36** 906–937. MR2396819 https://doi.org/10.1214/009053607000000622

ALAMO, A. and SANZ-SERNA, J. M. (2016). A technique for studying strong and weak local errors of splitting stochastic integrators. *SIAM J. Numer. Anal.* **54** 3239–3257. MR3570281 https://doi.org/10.1137/16M1058765

ALYUSHINA, L. A. (1987). Euler polygonal lines for Itô equations with monotone coefficients. *Teor. Veroyatn. Primen.* **32** 367–373. MR0902767

ANN, N., PEBRIANTI, D., ABAS, M. and BAYUAJI, L. (2022). Parameter estimation of Lorenz attractor: A combined deep neural network and K-means clustering approach. In *Recent Trends in Mechatronics Towards Industry* 4.0. *Lecture Notes in Electrical Engineering* **730** 321–331. Springer, Singapore.

ARNST, M., LOUPPE, G., VAN HULLE, R., GILLET, L., BUREAU, F. and DENOËL, V. (2022). A hybrid stochastic model and its Bayesian identification for infectious disease screening in a university campus with application to massive COVID-19 screening at the University of Liège. *Math. Biosci.* **347** Paper No. 108805, 14. MR4403088 https://doi.org/10.1016/j.mbs.2022.108805

BARBU, V. (1988). A product formula approach to nonlinear optimal control problems. *SIAM J. Control Optim.* **26** 497–520. MR0937669 https://doi.org/10.1137/0326030

BAYDIN, A. M. G., PEARLMUTTER, B. A., RADUL, A. A. and SISKIND, J. M. (2017). Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res.* **18** Paper No. 153, 43. MR3800512

BENSOUSSAN, A., GLOWINSKI, R. and RĂŞCANU, A. (1992). Approximation of some stochastic differential equations by the splitting up method. *Appl. Math. Optim.* **25** 81–106. MR1133253 https://doi.org/10.1007/BF01184157

BIBBY, B. M. and SØRENSEN, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1** 17–39. MR1354454 https://doi.org/10.2307/3318679

BLANES, S., CASAS, F. and MURUA, A. (2008). Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Mat. Apl.* 45, 89–145. MR2477860

BOU-RABEE, N. and OWHADI, H. (2010). Long-run accuracy of variational integrators in the stochastic context. *SIAM J. Numer. Anal.* **48** 278–297. MR2608370 https://doi.org/10.1137/090758842

BRÉHIER, C-E. and GOUDENÉGE, L. (2019). Analysis of some splitting schemes for the stochastic Allen-Cahn equation. *Discrete Contin. Dyn. Syst. Ser. B* **24** 4169–4190. MR3986273 https://doi.org/10.3934/dcdsb.2019077

BUCKWAR, E., SAMSON, A., TAMBORRINO, M. and TUBIKANEC, I. (2022). A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh–Nagumo model. *Appl. Numer. Math.* **179** 191–220. MR4422320 https://doi.org/10.1016/j.apnum.2022.04.018

BUCKWAR, E., TAMBORRINO, M. and TUBIKANEC, I. (2020). Spectral density-based and measure-preserving ABC for partially observed diffusion processes. An illustration on Hamiltonian SDEs. *Stat. Comput.* **30** 627–648. MR4065223 https://doi.org/10.1007/s11222-019-09909-6

CHANG, J. and CHEN, S. X. (2011). On the approximate maximum likelihood estimation for diffusion processes. *Ann. Statist.* **39** 2820–2851. MR3012393 https://doi.org/10.1214/11-AOS922

CHOI, S. (2013). Closed-form likelihood expansions for multivariate time-inhomogeneous diffusions. *J. Econometrics* **174** 45–65. MR3045019 https://doi.org/10.1016/j.jeconom.2011.12.007

CHOI, S. (2015). Explicit form of approximate transition probability density functions of diffusion processes. *J. Econometrics* **187** 57–73. MR3347294 https://doi.org/10.1016/j.jeconom.2015.02.003

CHOPIN, N. and PAPASPILIOPOULOS, O. (2020). *An Introduction to Sequential Monte Carlo*. *Springer Series in Statistics*. Springer, Cham. MR4215639 https://doi.org/10.1007/978-3-030-47845-2

DACUNHA-CASTELLE, D. and FLORENS-ZMIROU, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19** 263–284. MR0872464 https://doi.org/10.1080/17442508608833428

DIPPLE, S., CHOUDHARY, A., FLAMINO, J., SZYMANSKI, B. and KORNISS, G. (2020). Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities. *Appl. Netw. Sci.* **5**. https://doi.org/10.1007/s41109-020-00259-1

DITLEVSEN, P. and DITLEVSEN, S. (2023). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nat. Commun.* **14** 4254. https://doi.org/10.1038/s41467-023-39810-w

DITLEVSEN, S. and SAMSON, A. (2019). Hypoelliptic diffusions: Filtering and inference from complete and partial observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 361–384. MR3928146

DITLEVSEN, S. and SØRENSEN, M. (2004). Inference for observations of integrated diffusion processes. *Scand. J. Stat.* **31** 417–429. MR2087834 https://doi.org/10.1111/j.1467-9469.2004.02_023.x

DITLEVSEN, S., TAMBORRINO, M. and TUBIKANEC, I. (2023). Network inference in a stochastic multi-population neural mass model via approximate Bayesian computation. Available at arXiv:2306.15787.

DOHNAL, G. (1987). On estimating the diffusion coefficient. *J. Appl. Probab.* **24** 105–114. MR0876173 https://doi.org/10.2307/3214063

DUBOIS, P., GOMEZ, T., PLANCKAERT, L. and PERRET, L. (2020). Data-driven predictions of the Lorenz system. *Phys. D* **408** 132495, 10. MR4087348 https://doi.org/10.1016/j.physd.2020.132495

FALBEL, D. and LURASCHI, J. (2022). torch: Tensors and neural networks with 'GPU' acceleration. Available at https://torch.mlverse.org/docs, https://github.com/mlverse/torch.

FLORENS-ZMIROU, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics* **20** 547–557. MR1047222 https://doi.org/10.1080/02331888908802205

FORMAN, J. L. and SØRENSEN, M. (2008). The Pearson diffusions: A class of statistically tractable diffusion processes. *Scand. J. Stat.* **35** 438–465. MR2446729 https://doi.org/10.1111/j.1467-9469.2007.00592.x

FUCHS, C. (2013). *Inference for Diffusion Processes*: *With Applications in Life Sciences*. Springer, Heidelberg. With a foreword by Ludwig Fahrmeir. MR3015023 https://doi.org/10.1007/978-3-642-25969-2

GENON-CATALOT, V. and JACOD, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **29** 119–151. MR1204521

GLOAGUEN, P., ETIENNE, M.-P. and LE CORFF, S. (2018). Stochastic differential equation based on a multimodal potential to model movement data in ecology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 599–619. MR3787968 https://doi.org/10.1111/rssc.12251

GLOTER, A. (2006). Parameter estimation for a discretely observed integrated diffusion process. *Scand. J. Stat.* **33** 83–104. MR2255111 https://doi.org/10.1111/j.1467-9469.2006.00465.x

GLOTER, A. and YOSHIDA, N. (2021a). Adaptive estimation for degenerate diffusion processes. *Electron. J. Stat.* **15** 1424–1472. MR4255288 https://doi.org/10.1214/20-ejs1777

GLOTER, A. and YOSHIDA, N. (2021b). Adaptive estimation for degenerate diffusion processes. *Electron. J. Stat.* **15** 1424–1472. MR4255288 https://doi.org/10.1214/20-ejs1777

GOBET, E. (2002). LAN property for ergodic diffusions with discrete observations. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 711–737. MR1931584 https://doi.org/10.1016/S0246-0203(02)01107-X

GU, W., WU, H. and XUE, H. (2020). Parameter estimation for multivariate nonlinear stochastic differential equation models: A comparison study. In *Statistical Modeling for Biological Systems*: *In Memory of Andrei Yakovlev* 245–258. Springer, Cham. https://doi.org/10.1007/978-3-030-34675-1_13

HAIRER, E., NØRSETT, S. P. and WANNER, G. (1993). *Solving Ordinary Differential Equations*. *I* : *Nonstiff Problems*, 2nd ed. *Springer Series in Computational Mathematics* **8**. Springer, Berlin. MR1227985

HILBORN, R. C. (1994). *Chaos and Nonlinear Dynamics*: *An Introduction for Scientists and Engineers*: *An Introduction for Scientists and Engineers*. Oxford Univ. Press, New York. MR1263025

HOPKINS, W. E. JR. and WONG, W. S. (1986). Lie–Trotter product formulas for nonlinear filtering. *Stochastics* **17** 313–337. MR0854651 https://doi.org/10.1080/17442508608833395

HUMPHRIES, A. R. and STUART, A. M. (1994). Runge–Kutta methods for dissipative and gradient dynamical systems. *SIAM J. Numer. Anal.* **31** 1452–1485. MR1293524 https://doi.org/10.1137/0731075

HUMPHRIES, A. R. and STUART, A. M. (2002). Deterministic and random dynamical systems: Theory and numerics. In *Modern Methods in Scientific Computing and Applications* (*Montréal*, *QC*, 2001). *NATO Sci. Ser. II Math. Phys. Chem.* **75** 211–254. Kluwer Academic, Dordrecht. MR2004356

HURN, A. S., JEISMAN, J. I. and LINDSAY, K. A. (2007). Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *J. Financ. Econ.* **5** 390–455.

HUTZENTHALER, M., JENTZEN, A. and KLOEDEN, P. E. (2011). Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **467** 1563–1576. MR2795791 https://doi.org/10.1098/rspa.2010.0348

IGUCHI, Y., BESKOS, A. and GRAHAM, M. M. (2022). Parameter estimation with increased precision for elliptic and hypo-elliptic diffusions. Available at arXiv:2211.16384.

JENSEN, B. and POULSEN, R. (2002). Transition densities of diffusion processes: Numerical comparison of approximation techniques. *J. Deriv.* **9** 18–32.

JIMENEZ, J. C., MORA, C. and SELVA, M. (2017). A weak local linearization scheme for stochastic differential equations with multiplicative noise. *J. Comput. Appl. Math.* **313** 202–217. MR3573236 https://doi.org/10.1016/j.cam.2016.09.013

JIMENEZ, J. C., SHOJI, I. and OZAKI, T. (1999). Simulation of stochastic differential equations through the local linearization method. A comparative study. *J. Stat. Phys.* **94** 587–602. MR1675365 https://doi.org/10.1023/A:1004504506041

KAREEM, A. M. and AL-AZZAWI, S. N. (2021). A stochastic differential equations model for internal COVID-19 dynamics. *J. Phys., Conf. Ser.* **1818** 012121. https://doi.org/10.1088/1742-6596/1818/1/012121

KELLER, H. (1996). Attractors and bifurcations of the stochastic Lorenz system Technical report. Institut für Dynamische syteme, Universität Bremen.

KESSLER, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Stat.* **24** 211–229. MR1455868 https://doi.org/10.1111/1467-9469.00059

KLOEDEN, P. E. and PLATEN, E. (1992). *Numerical Solution of Stochastic Differential Equations*. *Applications of Mathematics* (*New York*) **23**. Springer, Berlin. MR1214374 https://doi.org/10.1007/978-3-662-12616-5

KRYLOV, N. V. (1990). A simple proof of the existence of a solution to the Itô equation with monotone coefficients. *Teor. Veroyatn. Primen.* **35** 576–580. MR1091217 https://doi.org/10.1137/1135082

LAZZÚS, J. A., RIVERA, M. and LÓPEZ-CARABALLO, C. H. (2016). Parameter estimation of Lorenz chaotic system using a hybrid swarm intelligence algorithm. *Phys. Lett. A* **380** 1164–1171. MR3457318 https://doi.org/10.1016/j.physleta.2016.01.040

LEIMKUHLER, B. and MATTHEWS, C. (2015). *Molecular Dynamics*: *With Deterministic and Stochastic Numerical Methods*. *Interdisciplinary Applied Mathematics* **39**. Springer, Cham. MR3362507

LI, C. (2013). Maximum-likelihood estimation for diffusion processes via closed-form density expansions. *Ann. Statist.* **41** 1350–1380. MR3113814 https://doi.org/10.1214/13-AOS1118

LÓPEZ-PÉREZ, A., FEBRERO-BANDE, A. and GONZÁLEZ-MANTEIGAV, W. (2021). Parametric estimation of diffusion processes: A review and comparative study. *Mathematics* **9** 859. https://doi.org/10.3390/math9080859

LORENZ, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** 130–141. MR4021434 https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

MAO, X. (2007). *Stochastic Differential Equations and Applications*. Elsevier, Amsterdam.

MCLACHLAN, R. I. and QUISPEL, G. R. W. (2002). Splitting methods. *Acta Numer.* **11** 341–434. MR2009376 https://doi.org/10.1017/S0962492902000053

MICHELOT, T., GLENNIE, R., HARRIS, C. and THOMAS, L. (2021). Varying-coefficient stochastic differential equations with applications in ecology. *J. Agric. Biol. Environ. Stat.* **26** 446–463. MR4292797 https://doi.org/10.1007/s13253-021-00450-6

MICHELOT, T., GLOAGUEN, P., BLACKWELL, P. and ETIENNE, M.-P. (2019). The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods Ecol. Evol.* **10**.

MILSTEIN, G. N. (1987). A theorem on the order of convergence of mean-square approximations of solutions of systems of stochastic differential equations. *Teor. Veroyatn. Primen.* **32** 809–811. MR0927268

MILSTEIN, G. N. and TRETYAKOV, M. V. (2003). Quasi-symplectic methods for Langevin-type equations. *IMA J. Numer. Anal.* **23** 593–626. MR2011342 https://doi.org/10.1093/imanum/23.4.593

MISAWA, T. (2001). A Lie algebraic approach to numerical integration of stochastic differential equations. *SIAM J. Sci. Comput.* **23** 866–890. MR1860968 https://doi.org/10.1137/S106482750037024X

OZAKI, T. (1985). Statistical identification of storage models with application to stochastic hydrology. *J. Amer. Water Resour. Assoc.* **21** 663–675.

OZAKI, T. (1992). A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statist. Sinica* **2** 113–135. MR1152300

OZAKI, T., JIMENEZ, J. C. and HAGGAN-OZAKI, V. (2000). The role of the likelihood function in the estimation of chaos models. *J. Time Series Anal.* **21** 363–387. MR1787661 https://doi.org/10.1111/1467-9892.00189

PICCHINI, U. and DITLEVSEN, S. (2011). Practical estimation of high dimensional stochastic differential mixed-effects models. *Comput. Statist. Data Anal.* **55** 1426–1444. MR2741425 https://doi.org/10.1016/j.csda.2010.10.003

PILIPOVIC, P., SAMSON, A. and DITLEVSEN, S. (2024). Supplement to "Parameter estimation in non-linear multivariate stochastic differential equations based on splitting schemes." https://doi.org/10.1214/24-AOS2371SUPPA, https://doi.org/10.1214/24-AOS2371SUPPB

RIEDMILLER, M. and BRAUN, H. (1992). RPROP—a fast adaptive learning algorithm. Technical report. Proc. of ISCIS VII, Universitat.

SHOJI, I. (1998). Approximation of continuous time stochastic processes by a local linearization method. *Math. Comp.* **67** 287–298. MR1432134 https://doi.org/10.1090/S0025-5718-98-00888-6

SHOJI, I. (2011). A note on convergence rate of a linearization method for the discretization of stochastic differential equations. *Commun. Nonlinear Sci. Numer. Simul.* **16** 2667–2671. MR2772282 https://doi.org/10.1016/j.cnsns.2010.09.008

SHOJI, I. and OZAKI, T. (1998). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stoch. Anal. Appl.* **16** 733–752. MR1632562 https://doi.org/10.1080/07362999808809559

SØRENSEN, M. (2012). Estimating functions for diffusion-type processes. In *Statistical Methods for Stochastic Differential Equations* 1 1–97. CRC Press, Boca Raton. https://doi.org/10.1201/b12126-2

SØRENSEN, M. and UCHIDA, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9** 1051–1069. MR2046817 https://doi.org/10.3150/bj/1072215200

TABOR, M. (1989). *Chaos and Integrability in Nonlinear Dynamics*: *An Introduction*. *A Wiley-Interscience Publication*. Wiley, New York. MR1007309

TIAN, Y. and FAN, M. (2020). Nonlinear integral inequality with power and its application in delay integro-differential equations. *Adv. Difference Equ.* Paper No. 142, 11. MR4085951 https://doi.org/10.1186/s13662-020-02596-y

TRETYAKOV, M. V. and ZHANG, Z. (2013). A fundamental mean-square convergence theorem for SDEs with locally Lipschitz coefficients and its applications. *SIAM J. Numer. Anal.* **51** 3135–3162. MR3129758 https://doi.org/10.1137/120902318

UCHIDA, M. and YOSHIDA, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Process. Appl.* **122** 2885–2924. MR2931346 https://doi.org/10.1016/j.spa.2012.04.001

VAN LOAN, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Trans. Automat. Control* **23** 395–404. MR0494865 https://doi.org/10.1109/TAC.1978.1101743

VATIWUTIPONG, P. and PHEWCHEAN, N. (2019). Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Adv. Difference Equ.* Paper No. 276, 7. MR3978552 https://doi.org/10.1186/s13662-019-2214-1

YANG, N., CHEN, N. and WAN, X. (2019). A new delta expansion for multivariate diffusions via the Itô–Taylor expansion. *J. Econometrics* **209** 256–288. MR3944752 https://doi.org/10.1016/j.jeconom.2019.01.003

ZHUANG, L., CAO, L., WU, Y., ZHONG, Y., ZHANGZHONG, L., ZHENG, W. and WANG, L. (2020). Parameter estimation of Lorenz chaotic system based on a hybrid Jaya–Powell algorithm. *IEEE Access* **8** 20514–20522.

R CORE TEAM (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

II Parameter Estimation in Nonlinear Multivariate Second-order SDEs with Additive Noise

This chapter contains the following paper:

- [Pilipovic et al., 2024b] P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting for parametric inference in second-order stochastic differential equations, 2024b.

  Paper status: Submitted.

In this paper, we generalize the splitting-scheme parameter estimation to second-order SDEs. This class of models is interesting to analyze because they are typically converted into first-order systems by introducing an auxiliary velocity variable, creating a first-order hypoelliptic system. Moreover, the velocity variable is not observed, making these models ideal for investigating different methodologies for both hypoelliptic systems and partial observations.

To motivate this type of model, we describe paleoclimate data from the Greenland ice core in Section 2 and attempt to model it using the Kramers oscillator. The Kramers oscillator is a physical model that describes the motion of a particle in a double-well potential with friction. It is used to reflect the switching between metastable states during glacial periods. This application highlights the real-world relevance and potential impact of our approach.

To address hypoellipticity, we propose an estimator based on the Strang splitting scheme. Given that the velocity variable is unobserved in real-world data, we adapt the Strang-based estimator to function under partial observation scenarios. This adaptation leads to the development of two types of estimators: one that uses the full likelihood based on the full hypoelliptic SDE and another that relies on the marginal probability derived from the velocity coordinate. These estimators are not only theoretically robust but also practical, being easy to implement and computationally fast. In Section 3, we prove the main results and highlight the differences in the asymptotic variance of the diffusion estimator, depending on the type of likelihood used.

STRANG SPLITTING FOR PARAMETRIC INFERENCE IN SECOND-ORDER STOCHASTIC DIFFERENTIAL EQUATIONS

A PREPRINT

Predrag Pilipovic
Department of Mathematical Sciences
University of Copenhagen
2100 Copenhagen, Denmark
predrag@math.ku.dk
Bielefeld Graduate School of Economics and Management
University of Bielefeld
33501 Bielefeld, Germany
predrag.pilipovic@uni-bielefeld.de

Adeline Samson
Univ. Grenoble Alpes
CNRS, Grenoble INP, LJK
38000 Grenoble, France
adeline.leclercq-samson@univ-grenoble-alpes.fr

Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen
2100 Copenhagen, Denmark
susanne@math.ku.dk

ABSTRACT

We address parameter estimation in second-order stochastic differential equations (SDEs), prevalent in physics, biology, and ecology. Second-order SDE is converted to a first-order system by introducing an auxiliary velocity variable raising two main challenges. First, the system is hypoelliptic since the noise affects only the velocity, making the Euler-Maruyama estimator ill-conditioned. To overcome that, we propose an estimator based on the Strang splitting scheme. Second, since the velocity is rarely observed we adjust the estimator for partial observations. We present four estimators for complete and partial observations, using full likelihood or only velocity marginal likelihood. These estimators are intuitive, easy to implement, and computationally fast, and we prove their consistency and asymptotic normality. Our analysis demonstrates that using full likelihood with complete observations reduces the asymptotic variance of the diffusion estimator. With partial observations, the asymptotic variance increases due to information loss but remains unaffected by the likelihood choice. However, a numerical study on the Kramers oscillator reveals that using marginal likelihood for partial observations yields less biased estimators. We apply our approach to paleoclimate data from the Greenland ice core and fit it to the Kramers oscillator model, capturing transitions between metastable states reflecting observed climatic conditions during glacial eras.

Keywords Second-order stochastic differential equations, Hypoellipticity, Partial observations, Strang splitting estimator, Greenland ice core data, Kramers oscillator

1 Introduction

Second-order stochastic differential equations (SDEs) are an effective instrument for modeling complex systems showcasing both deterministic and stochastic dynamics, which incorporate the second derivative of a variable - the acceleration. These models are extensively applied in many fields, including physics [Rosenblum and Pikovsky, 2003], molecular dynamics [Leimkuhler and Matthews, 2015], ecology [Johnson et al., 2008, Michelot and Blackwell, 2019], paleoclimate research [Ditlevsen et al., 2002], and neuroscience [Ziv et al., 1994, Jansen and Rit, 1995].

The general form of a second-order SDE in Langevin form is given as follows:

$$\ddot{\mathbf{X}}_t = \mathbf{F}(\mathbf{X}_t, \dot{\mathbf{X}}_t, \boldsymbol{\beta}) + \boldsymbol{\Sigma}\boldsymbol{\xi}_t. \tag{1}$$

Here, $\mathbf{X}_t \in \mathbb{R}^d$ denotes the variable of interest, the dot indicates derivative with respect to time t, drift $\mathbf{F}$ represents the deterministic force, and $\boldsymbol{\xi}_t$ is a white noise representing the system's random perturbations around the deterministic force. We assume that $\boldsymbol{\Sigma}$ is constant, that is the noise is additive.

The main goal of this study is to estimate parameters in second-order SDEs. We first reformulate the d-dimensional second-order SDE (1) into a $2d$-dimensional SDE in Itô's form. We define an auxiliary velocity variable, and express the second-order SDE in terms of its position $\mathbf{X}_t$ and velocity $\mathbf{V}_t$:

$$\begin{aligned} \mathrm{d}\mathbf{X}_t &= \mathbf{V}_t\,\mathrm{d}t, & \mathbf{X}_0 &= \mathbf{x}_0, \\ \mathrm{d}\mathbf{V}_t &= \mathbf{F}\left(\mathbf{X}_t, \mathbf{V}_t; \boldsymbol{\beta}\right)\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, & \mathbf{V}_0 &= \mathbf{v}_0, \end{aligned} \tag{2}$$

where $\mathbf{W}_t$ is a standard Wiener process. We refer to $\mathbf{X}_t$ and $\mathbf{V}_t$ as the smooth and rough coordinates, respectively.

A specific example of model (2) is $\mathbf{F}(\mathbf{x}, \mathbf{v}) = -\mathbf{c}(\mathbf{x}, \mathbf{v})\mathbf{v} - \nabla\mathbf{U}(\mathbf{x})$, for some function $\mathbf{c}(\cdot)$ and potential $\mathbf{U}(\cdot)$. Then, model (2) is called a stochastic damping Hamiltonian system. This system describes the motion of a particle subjected to potential, dissipative, and random forces [Wu, 2001]. An example of a stochastic damping Hamiltonian system is the Kramers oscillator introduced in Section 2.1.

Let $\mathbf{Y}_t = (\mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top$, $\widetilde{\mathbf{F}}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) = (\mathbf{v}^\top, \mathbf{F}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta})^\top)^\top$ and $\widetilde{\boldsymbol{\Sigma}} = (\mathbf{0}^\top, \boldsymbol{\Sigma}^\top)^\top$. Then (2) is formulated as

$$\mathrm{d}\mathbf{Y}_t = \widetilde{\mathbf{F}}\left(\mathbf{Y}_t; \boldsymbol{\beta}\right)\mathrm{d}t + \widetilde{\boldsymbol{\Sigma}}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{Y}_0 = \mathbf{y}_0. \tag{3}$$

The notation $\widetilde{\ }$ over an object indicates that it is associated with process $\mathbf{Y}_t$. Specifically, the object is of dimension $2d$ or $2d \times 2d$.

When it exists, the unique solution of (3) is called a diffusion or diffusion process. System (3) is usually not fully observed since the velocity $\mathbf{V}_t$ is not observable. Thus, our primary objective is to estimate the underlying drift parameter $\boldsymbol{\beta}$ and the diffusion parameter $\boldsymbol{\Sigma}$, based on discrete observations of either $\mathbf{Y}_t$ (referred to as complete observation case), or only $\mathbf{X}_t$ (referred to as partial observation case). Diffusion $\mathbf{Y}_t$ is said to be hypoelliptic since the matrix

$$\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^\top = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \end{bmatrix} \tag{4}$$

is not of full rank, while $\mathbf{Y}_t$ admits a smooth density. Thus, (2) is a subclass of a larger class of hypoelliptic diffusions.

Parametric estimation for hypoelliptic diffusions is an active area of research. Ditlevsen and Sørensen [2004] studied discretely observed integrated diffusion processes. They proposed to use prediction-based estimating functions, which are suitable for non-Markovian processes and which do not require access to the unobserved component. They proved consistency and asymptotic normality of the estimators for $N \to \infty$, but without any requirements on the sampling interval h. Certain moment conditions are needed to obtain results for fixed h, which are often difficult to fulfill for nonlinear drift functions. The estimator was applied to paleoclimate data in Ditlevsen et al. [2002], similar to the data we analyze in Section 5.

Gloter [2006] also focused on parametric estimation for discretely observed integrated diffusion processes, introducing a contrast function using the Euler-Maruyama discretization. He studied the asymptotic properties as the sampling interval $h \to 0$ and the sample size $N \to \infty$, under the so-called rapidly increasing experimental design $Nh \to \infty$ and $Nh^2 \to 0$. To address the ill-conditioned contrast from the Euler-Maruyama discretization, he suggested using only the rough equations of the SDE. He proposed to recover the unobserved integrated component through the finite difference approximation $(\mathbf{X}_{t_{k+1}} - \mathbf{X}_{t_k})/h$. This approximation makes the estimator biased and requires a correction factor of 3/2 in one of the terms of the contrast function for partial observations. Consequently, the correction increases the asymptotic variance of the estimator of the diffusion parameter. Samson and Thieullen [2012] expanded the ideas of [Gloter, 2006] and proved the results of [Gloter, 2006] in more general models. Similar to [Gloter, 2006], their focus was on contrasts using the Euler-Maruyama discretization limited to only the rough equations.

Pokern et al. [2009] proposed an Itô-Taylor expansion, adding a noise term of order $h^{3/2}$ to the smooth component in the numerical scheme. They argued against the use of finite differences for approximating unobserved components. Instead, he suggested using the Itô-Taylor expansion leading to non-degenerate conditionally Gaussian approximations of the transition density and using Markov Chain Monte Carlo (MCMC) Gibbs samplers for conditionally imputing missing components based on the observations. They found out that this approach resulted in a biased estimator of the drift parameter of the rough component.

Ditlevsen and Samson [2019] focused on both filtering and inference methods for complete and partial observations. They proposed a contrast estimator based on the strong order 1.5 scheme [Kloeden and Platen, 1992], which incorporates noise of order $h^{3/2}$ into the smooth component, similar to [Pokern et al., 2009]. Moreover, they retained terms of order h^2 in the mean, which removed the bias in the drift parameters noted in [Pokern et al., 2009]. They proved consistency and asymptotic normality under complete observations, with the standard rapidly increasing experimental design $Nh \to \infty$ and $Nh^2 \to 0$. They adopted an unconventional approach by using two separate contrast functions, resulting in marginal asymptotic results rather than a joint central limit theorem. The model was limited to a scalar smooth component and a diagonal diffusion coefficient matrix for the rough component.

Melnykova [2020] developed a contrast estimator using local linearization (LL) [Ozaki, 1985, Shoji and Ozaki, 1998, Ozaki et al., 2000] and compared it to the least-squares estimator. She employed local linearization of the drift function, providing a non-degenerate conditional Gaussian discretization scheme, enabling the construction of a contrast estimator that achieves asymptotic normality under the standard conditions $Nh \to \infty$ and $Nh^2 \to 0$. She proved a joint central limit theorem, bypassing the need for two separate contrasts as in Ditlevsen and Samson [2019]. The models in Ditlevsen and Samson [2019] and Melnykova [2020] allow for parameters in the smooth component of the drift, in contrast to models based on second-order differential equations.

Recent work by Gloter and Yoshida [2020, 2021] introduced adaptive and non-adaptive methods in hypoelliptic diffusion models, proving asymptotic normality in the complete observation regime. In line with this work, we briefly review their non-adaptive estimator. It is based on a higher-order Itô-Taylor expansion that introduces additional Gaussian noise onto the smooth coordinates, accompanied by an appropriate higher-order mean approximation of the rough coordinates. The resulting estimator was later termed the local Gaussian (LG), which should be differentiated from LL. The LG estimator can be viewed as an extension of the estimator proposed in Ditlevsen and Samson [2019], with fewer restrictions on the class of models. Gloter and Yoshida [2020, 2021] found that using the full SDE to create a contrast reduces the asymptotic variance of the estimator of the diffusion parameter compared to methods using only rough coordinates in the case of complete observations.

The most recent contributions are Iguchi et al. [2023a,b], Iguchi and Beskos [2023], building on the foundation of the LG estimator and focusing on high-frequency regimes addressing limitations in earlier methods. Iguchi et al. [2023b] presented a new closed-form contrast estimator for hypoelliptic SDEs (denoted as Hypo-I) based on Edgeworth-type density expansion and Malliavin calculus that achieves asymptotic normality under the less restrictive condition of $Nh^3 \to 0$. Iguchi et al. [2023a] focused on a highly degenerate class of SDEs (denoted as Hypo-II) where smooth coordinates split into further sub-groups and proposed estimators for both complete and partial observation settings. Iguchi and Beskos [2023] further refined the conditions for estimators asymptotic normality for both Hypo-I and Hypo-II under a weak design $Nh^p \to 0$, for $p \geq 2$.

The existing methods are generally based on approximations with varying degrees of refinements to correct for possible nonlinearities. This implies that they quickly degrade for highly nonlinear models if the step size is increased. In particular, this is the case for Hamiltonian systems. Instead, we propose to use splitting schemes, more precisely the Strang splitting scheme.

Splitting schemes are established techniques initially developed for solving ordinary differential equations (ODEs) and have proven to be effective also for SDEs [Ableidinger et al., 2017, Buckwar et al., 2022, Pilipovic et al., 2024]. These schemes yield accurate results in many practical applications since they incorporate nonlinearities in their construction. This makes them particularly suitable for second-order SDEs, where they have been widely used. Early work in dissipative particle dynamics [Shardlow, 2003, Serrano et al., 2006], applications to molecular dynamics [Vanden-Eijnden and Ciccotti, 2006, Melchionna, 2007, Leimkuhler and Matthews, 2015] and studies on internal particles [Pavliotis et al., 2009] all highlight the scheme's versatility. Burrage et al. [2007], Bou-Rabee and Owhadi [2010], and Abdulle et al. [2015] focused on the long-run statistical properties such as invariant measures. Bou-Rabee [2017], Bréhier and Goudenège [2019] and Adams et al. [2022] used splitting schemes for stochastic partial differential equations (SPDEs).

Despite the extensive use of splitting schemes in different areas, statistical applications have been lacking. We have recently proposed statistical estimators for elliptic SDEs [Pilipovic et al., 2024]. The straightforward and intuitive schemes lead to robust, easy-to-implement estimators, offering an advantage over more numerically intensive and less user-friendly state-of-the-art methods. We use the Strang splitting scheme to approximate the transition density between two consecutive observations and derive the pseudo-likelihood function since the exact likelihood function is often unknown or intractable. Then, to estimate parameters, we employ maximum likelihood estimation (MLE). However, two specific statistical problems arise due to hypoellipticity and partial observations.

First, hypoellipticity leads to degenerate Euler-Maruyama transition schemes, which can be addressed by constructing the pseudo-likelihood solely from the rough equations of the SDE, referred to as the rough likelihood hereafter. The

Strang splitting technique enables the estimator to incorporate both smooth and rough components (referred to as the full likelihood). It is also possible to construct Strang splitting estimators using only the rough likelihood, raising the question of which estimator performs better. Our results are in line with Gloter and Yoshida [2020, 2021] in the complete observation setting, where we find that using the full likelihood reduces the asymptotic variance of the diffusion estimator. We found the same results in the simulation study for the LL estimator proposed by Melnykova [2020].

Second, we suggest to treat the unobserved velocity by approximating it using finite difference methods. While Gloter [2006] and Samson and Thieullen [2012] exclusively use forward differences, we investigate also central and backward differences. The forward difference approach leads to a biased estimator unless it is corrected. One of the main contributions of this work is finding suitable corrections of the pseudo-likelihoods for different finite difference approximations such that the Strang estimators are asymptotically unbiased. This also ensures consistency of the diffusion parameter estimator, at the cost of increasing its asymptotic variance.

When only partial observations are available, we explore the impact of using the full likelihood versus the rough likelihood and how different finite differentiation approximations influence the parametric inference. We find that the choice of likelihood does not affect the asymptotic variance of the estimator. However, our simulation study on the Kramers oscillator suggests that using the full likelihood in finite sample setups introduce more bias than using only the rough marginal likelihood, which is the opposite of the complete observation setting. Finally, we analyze a paleoclimate ice core dataset from Greenland using a second-order SDE.

The main contributions of this paper are:

1. We extend the Strang splitting estimator of [Pilipovic et al., 2024] to hypoelliptic models given by second-order SDEs, including appropriate correction factors to obtain consistency.
2. When complete observations are available, we show that the asymptotic variance of the estimator of the diffusion parameter is smaller when maximizing the full likelihood. In contrast, for partial observations, we show that the asymptotic variance remains unchanged regardless of using the full or marginal likelihood of the rough coordinates.
3. We discuss the influence on the statistical properties of using the forward difference approximation for imputing the unobserved velocity variables compared to using the backward or the central difference.
4. We evaluate the performance of the estimators through a simulation study of a second-order SDE, the Kramers oscillator. Additionally, we show numerically in a finite sample study that the marginal likelihood for partial observations is more favorable than the full likelihood.
5. We fit the Kramers oscillator to a paleoclimate ice core dataset from Greenland and estimate the average time needed to pass between two metastable states.

The structure of the paper is as follows. In Section 2, we introduce the class of SDE models, define hypoellipticity, introduce the Kramers oscillator, and explain the Strang splitting scheme and its associated estimators. The asymptotic properties of the estimator are established in Section 3. The theoretical results are illustrated in a simulation study on the Kramers Oscillator in Section 4. Section 5 illustrates our methodology on the Greenland ice core data, while the technical results and the proofs of the main theorems and properties are in Section 6 and Supplementary Material S1, respectively.

Notation. We use capital bold letters for random vectors, vector-valued functions, and matrices, while lowercase bold letters denote deterministic vectors. $\|\cdot\|$ denotes both the L^2 vector norm in $\mathbb{R}^d$. Superscript (i) on a vector denotes the i-th component, while on a matrix it denotes the i-th column. Double subscript ij on a matrix denotes the component in the i-th row and j-th column. The transpose is denoted by $\top$. Operator $\mathrm{Tr}(\cdot)$ returns the trace of a matrix and $\det(\cdot)$ the determinant. $\mathbf{I}_d$ denotes the d-dimensional identity matrix, while $\mathbf{0}_{d\times d}$ is a d-dimensional zero square matrix. We denote by $[a_i]_{i=1}^d$ a vector with coordinates a_i, and by $[b_{ij}]_{i,j=1}^d$ a matrix with coordinates b_{ij}, for $i,j=1,\dots,d$. For a real-valued function $g:\mathbb{R}^d\to\mathbb{R}$, $\partial_{x^{(i)}}g(\mathbf{x})$ denotes the partial derivative with respect to $x^{(i)}$ and $\partial^2_{x^{(i)}x^{(j)}}g(\mathbf{x})$ denotes the second partial derivative with respect to $x^{(i)}$ and $x^{(j)}$. The nabla operator $\nabla_{\mathbf{x}}$ denotes the gradient vector of g with respect of $\mathbf{x}$, that is, $\nabla_{\mathbf{x}}g(\mathbf{x})=[\partial_{x^{(i)}}g(\mathbf{x})]_{i=1}^d$. $\mathbb{H}$ denotes the Hessian matrix of function g, $\mathbb{H}_g(\mathbf{x})=[\partial_{x^{(i)}x^{(j)}}g(\mathbf{x})]_{i,j=1}^d$. For a vector-valued function $\mathbf{F}:\mathbb{R}^d\to\mathbb{R}^d$, the differential operator $D_{\mathbf{x}}$ denotes the Jacobian matrix $D_{\mathbf{x}}\mathbf{F}(\mathbf{x})=[\partial_{x^{(i)}}F^{(j)}(\mathbf{x})]_{i,j=1}^d$. Let $\mathbf{R}$ represent a vector (or a matrix) valued function defined on $(0,1)\times\mathbb{R}^d$ (or $(0,1)\times\mathbb{R}^{d\times d}$), such that, for some constant C, $\|\mathbf{R}(a,\mathbf{x})\|<aC(1+\|\mathbf{x}\|)^C$ for all $a,\mathbf{x}$. When denoted by R, it refers to a scalar function. For an open set A, the bar $\overline{A}$ indicates closure. We write $\xrightarrow{\mathbb{P}}$ for convergence in probability $\mathbb{P}$.

2 Problem setup

Let $\mathbf{Y} = (\mathbf{Y}_t)_{t\geq 0}$ in (3) be defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\boldsymbol{\theta}})$ with a complete right-continuous filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$, and let the d-dimensional Wiener process $\mathbf{W} = (\mathbf{W}_t)_{t\geq 0}$ be adapted to $\mathcal{F}_t$. The probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$. Rewrite equation (3) as follows:

$$\mathrm{d}\mathbf{Y}_t = \widetilde{\mathbf{A}}(\boldsymbol{\beta})(\mathbf{Y}_t - \widetilde{\mathbf{b}}(\boldsymbol{\beta}))\,\mathrm{d}t + \widetilde{\mathbf{N}}\left(\mathbf{Y}_t; \boldsymbol{\beta}\right)\mathrm{d}t + \widetilde{\boldsymbol{\Sigma}}\,\mathrm{d}\mathbf{W}_t, \tag{5}$$

where

$$\widetilde{\mathbf{A}}(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{A}_{\mathbf{x}}(\boldsymbol{\beta}) & \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta}) \end{bmatrix}, \quad \widetilde{\mathbf{b}}(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{b}(\boldsymbol{\beta}) \\ \mathbf{0}_d \end{bmatrix}, \quad \widetilde{\mathbf{N}}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{0}_d \\ \mathbf{N}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) \end{bmatrix}. \tag{6}$$

Function $\mathbf{F}$ in (2) is thus split as $\mathbf{F}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) = \mathbf{A}_{\mathbf{x}}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{b}(\boldsymbol{\beta})) + \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})\mathbf{v} + \mathbf{N}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta})$.

Let $\overline{\Theta}_\beta \times \overline{\Theta}_\Sigma = \overline{\Theta}$ denote the closure of the parameter space with Θ_β and Θ_Σ being two convex open bounded subsets of $\mathbb{R}^r$ and $\mathbb{R}^{d\times d}$, respectively. The function $\mathbf{N} : \mathbb{R}^{2d} \times \overline{\Theta}_\beta \to \mathbb{R}^d$ is assumed locally Lipschitz; functions $\mathbf{A}_{\mathbf{x}}$ and $\mathbf{A}_{\mathbf{v}}$ are defined on $\overline{\Theta}_\beta$ and take values in $\mathbb{R}^{d\times d}$; and the parameter matrix $\boldsymbol{\Sigma}$ takes values in $\mathbb{R}^{d\times d}$. The matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is assumed to be positive definite, shaping the variance of the rough coordinates. As any square root of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ induces the same distribution, $\boldsymbol{\Sigma}$ is identifiable only up to equivalence classes. Hence, estimation of the parameter $\boldsymbol{\Sigma}$ means estimation of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. The drift function $\widetilde{\mathbf{F}}$ in (3) is divided into a linear part given by the matrix $\widetilde{\mathbf{A}}$ and a nonlinear part given by $\widetilde{\mathbf{N}}$.

The true value of the parameter is denoted by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$, and we assume that $\boldsymbol{\theta}_0 \in \Theta$. When referring to the true parameters, we write $\mathbf{A}_{\mathbf{x},0}$, $\mathbf{A}_{\mathbf{v},0}$, $\mathbf{b}_0$, $\mathbf{N}_0(\mathbf{x})$, $\mathbf{F}_0(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$ instead of $\mathbf{A}_{\mathbf{x}}(\boldsymbol{\beta}_0)$, $\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta}_0)$, $\mathbf{b}(\boldsymbol{\beta}_0)$, $\mathbf{N}(\mathbf{x}; \boldsymbol{\beta}_0)$, $\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_0)$ and $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top$, respectively. We write $\mathbf{A}_{\mathbf{x}}$, $\mathbf{A}_{\mathbf{v}}$, $\mathbf{b}$, $\mathbf{N}(\mathbf{x})$, $\mathbf{F}(\mathbf{x})$, and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ for any parameter $\boldsymbol{\theta}$.

2.1 Example: The Kramers oscillator

The abrupt temperature changes during the ice ages, known as the Dansgaard–Oeschger (DO) events, are essential elements for understanding the climate [Dansgaard et al., 1993]. These events occurred during the last glacial era spanning approximately the period from 115,000 to 12,000 years before present and are characterized by rapid warming phases followed by gradual cooling periods, revealing colder (stadial) and warmer (interstadial) climate states [Rasmussen et al., 2014].

To analyze the DO events in Section 5, we propose a stochastic model of the escape dynamics in metastable systems, the Kramers oscillator [Kramers, 1940], originally formulated to model the escape rate of Brownian particles from potential wells. The escape rate is related to the mean first passage time — the time needed for a particle to exceed the potential's local maximum for the first time, starting at a neighboring local minimum. This rate depends on variables such as the damping coefficient, noise intensity, temperature, and specific potential features, including the barrier's height and curvature at the minima and maxima. We apply this framework to quantify the rate of climate transitions between stadial and interstadial periods. This provides an estimate on the probability distribution of the ocurrence of DO events, contributing to our understanding of the global climate system.

Following Arnold and Imkeller [2000], we introduce the Kramers oscillator as the stochastic Duffing oscillator - an example of a second-order SDE and a stochastic damping Hamiltonian system. The Duffing oscillator [Duffing, 1918] is a forced nonlinear oscillator, featuring a cubic stiffness term. The governing equation is given by:

$$\ddot{x}_t + \eta\dot{x}_t + \frac{\mathrm{d}}{\mathrm{d}x}U(x_t) = f(t), \quad \text{where} \quad U(x) = -a\frac{x^2}{2} + b\frac{x^4}{4}, \quad \text{with} \quad a, b > 0, \quad \eta \geq 0. \tag{7}$$

The parameter η in (7) indicates the damping level, a regulates the linear stiffness, and b determines the nonlinear component of the restoring force. In the special case where $b = 0$, the equation simplifies to a damped harmonic oscillator. Function f represents the driving force and is usually set to $f(t) = \eta\cos(\omega t)$, which introduces deterministic chaos [Korsch and Jodl, 1999].

When the driving force is $f(t) = \sqrt{2\eta\mathrm{T}}\xi(t)$, where $\xi(t)$ is white noise, equation (7) characterizes the stochastic movement of a particle within a bistable potential well, interpreting $\mathrm{T} > 0$ as the temperature of a heat bath. Setting $\sigma = \sqrt{2\eta\mathrm{T}}$, equation (7) can be reformulated as an Itô SDE for variables X_t and $V_t = \dot{X}_t$, expressed as:

$$\begin{aligned} \mathrm{d}X_t &= V_t\,\mathrm{d}t, \\ \mathrm{d}V_t &= \left(-\eta V_t - \frac{\mathrm{d}}{\mathrm{d}x}U(X_t)\right)\mathrm{d}t + \sigma\,\mathrm{d}W_t, \end{aligned} \tag{8}$$

where W_t denotes a standard Wiener process. The parameter set of SDE (8) is $\boldsymbol{\theta} = \{\eta, a, b, \sigma^2\}$.

The existence and uniqueness of the invariant measure $\nu_0(\mathrm{d}x, \mathrm{d}y)$ of (8) is proved in Theorem 3 in [Arnold and Imkeller, 2000]. The invariant measure ν_0 is linked to the invariant density π_0 through $\nu_0(\mathrm{d}x, \mathrm{d}y) = \pi_0(x, v)\,\mathrm{d}x\,\mathrm{d}y$. Here we write $\pi_0(x, v)$ instead of $\pi(x, v; \boldsymbol{\theta}_0)$, and $\pi(x, v)$ instead of $\pi(x, v; \boldsymbol{\theta})$. The Fokker-Plank equation for π is given by

$$-v\frac{\partial}{\partial x}\pi(x,v) + \eta\pi(x,v) + \eta v\frac{\partial}{\partial v}\pi(x,v) + \frac{\mathrm{d}}{\mathrm{d}x}U(x)\frac{\partial}{\partial v}\pi(x,v) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial v^2}\pi(x,v) = 0. \tag{9}$$

The invariant density that solves the Fokker-Plank equation is:

$$\pi(x,v) = C\exp\left(-\frac{2\eta}{\sigma^2}U(x)\right)\exp\left(-\frac{\eta}{\sigma^2}v^2\right), \tag{10}$$

where C is the normalizing constant.

The marginal invariant probability of V_t is thus Gaussian with zero mean and variance $\sigma^2/(2\eta)$. The marginal invariant probability of X_t is bimodal driven by the potential $U(x)$:

$$\pi(x) = C\exp\left(-\frac{2\eta}{\sigma^2}U(x)\right). \tag{11}$$

At steady state, for a particle moving in any potential $U(x)$ and driven by random Gaussian noise, the position x and velocity v are independent of each other. This is reflected by the decomposition of the joint density $\pi(x, v)$ into $\pi(x)\pi(v)$.

Fokker-Plank equation (9) can also be used to derive the mean first passage time τ which is inversely related to Kramers' escape rate κ [Kramers, 1940]:

$$\tau = \frac{1}{\kappa} \approx \frac{2\pi}{\left(\sqrt{1+\frac{\eta^2}{4\omega^2}} - \frac{\eta}{2\omega}\right)\Omega}\exp\left(\frac{\Delta U}{\mathrm{T}}\right),$$

where $x_{\mathrm{barrier}} = 0$ is the local maximum of $U(x)$ and $x_{\mathrm{well}} = \pm\sqrt{a/b}$ are the local minima, $\omega = \sqrt{|U''(x_{\mathrm{barrier}})|} = \sqrt{a}$, $\Omega = \sqrt{U''(x_{\mathrm{well}})} = \sqrt{2a}$, and $\Delta U = U(x_{\mathrm{barrier}}) - U(x_{\mathrm{well}}) = a^2/4b$, . The formula is derived assuming strong friction, or an over-damped system ($\eta \gg \omega$), and a small parameter $\mathrm{T}/\Delta U \ll 1$, indicating sufficiently deep potential wells. For the potential defined in (7), the mean waiting time τ is then approximated by

$$\tau \approx \frac{\sqrt{2}\pi}{\sqrt{a+\frac{\eta^2}{4}} - \frac{\eta}{2}}\exp\left(\frac{a^2\eta}{2b\sigma^2}\right). \tag{12}$$

2.2 Hypoellipticity

The SDE (5) is said to be hypoelliptic if its quadratic diffusion matrix $\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^\top$ is not of full rank, while its solutions admit a smooth transition density with respect to the Lebesgue measure. According to Hörmander's theorem [Nualart, 2006], this is fulfilled if the SDE in its Stratonovich form satisfies the weak Hörmander condition. Since $\boldsymbol{\Sigma}$ does not depend on $\mathbf{y}$, the Itô and Stratonovich forms coincide.

We begin by recalling the concept of Lie brackets: for smooth vector fields $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$, the i-th component of the Lie bracket, $[\boldsymbol{f}, \boldsymbol{g}]^{(i)}$, is defined as

$$[\boldsymbol{f}, \boldsymbol{g}]^{(i)} \coloneqq D_{\mathbf{y}}^\top g^{(i)}(\mathbf{y})\boldsymbol{f}(\mathbf{y}) - D_{\mathbf{y}}^\top f^{(i)}(\mathbf{y})\boldsymbol{g}(\mathbf{y}).$$

We define the set $\mathcal{H}$ of vector fields by initially including $\widetilde{\boldsymbol{\Sigma}}^{(i)}$, $i = 1, 2, ..., 2d$, and then recursively adding Lie brackets

$$H \in \mathcal{H} \Rightarrow [\widetilde{\mathbf{F}}, H], [\widetilde{\boldsymbol{\Sigma}}^{(1)}, H], \ldots, [\widetilde{\boldsymbol{\Sigma}}^{(2d)}, H] \in \mathcal{H}.$$

The weak Hörmander condition is met if the vectors in $\mathcal{H}$ span $\mathbb{R}^{2d}$ at every point $\mathbf{y} \in \mathbb{R}^{2d}$. The initial vectors span $\{(\mathbf{0}, \mathbf{v}) \in \mathbb{R}^{2d} \mid \mathbf{v} \in \mathbb{R}^d\}$, a d-dimensional subspace. We therefore need to verify the existence of some $H \in \mathcal{H}$ with a non-zero first element. The first iteration of the system yields

$$[\widetilde{\mathbf{F}}, \widetilde{\boldsymbol{\Sigma}}^{(i)}]^{(1)} = -\boldsymbol{\Sigma}^{(i)},$$
$$[\widetilde{\boldsymbol{\Sigma}}^{(i)}, \widetilde{\boldsymbol{\Sigma}}^{(j)}]^{(1)} = \mathbf{0},$$

for $i, j = 1, 2, ..., 2d$. The first equation is non-zero, as are all subsequent iterations. Thus, the second-order SDE defined in (5) is always hypoelliptic.

2.3 Assumptions

The following assumptions are a generalization of those presented in [Pilipovic et al., 2024].

Let $T > 0$ be the length of the observed time interval. We assume that (5) has a unique strong solution $\mathbf{Y} = \{\mathbf{Y}_t \mid t \in [0, T]\}$, adapted to $\mathcal{F} = \{\mathcal{F}_t \mid t \in [0, T]\}$, which follows from the following first two assumptions (Theorem 2 in Alyushina [1988], Theorem 1 in Krylov [1991], Theorem 3.5 in Mao [2007]). We need the last three assumptions to prove the properties of the estimators.

(A1) Function $\mathbf{N}$ is twice continuously differentiable with respect to both $\mathbf{y}$ and $\boldsymbol{\theta}$, i.e., $\mathbf{N} \in C^2$. Moreover, it is globally one-sided Lipschitz continuous with respect to $\mathbf{y}$ on $\mathbb{R}^{2d} \times \overline{\Theta}_\beta$. That is, there exists a constant $C > 0$ such that for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{2d}$,

$$(\mathbf{y}_1 - \mathbf{y}_2)^\top (\mathbf{N}(\mathbf{y}_1; \boldsymbol{\beta}) - \mathbf{N}(\mathbf{y}_2; \boldsymbol{\beta})) \leq C \|\mathbf{y}_1 - \mathbf{y}_2\|^2.$$

(A2) Function $\mathbf{N}$ exhibits at most polynomial growth in $\mathbf{y}$, uniformly in $\boldsymbol{\theta}$. Specifically, there exist constants $C > 0$ and $\chi \geq 1$ such that for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{2d}$,

$$\|\mathbf{N}(\mathbf{y}_1; \boldsymbol{\beta}) - \mathbf{N}(\mathbf{y}_2; \boldsymbol{\beta})\|^2 \leq C \left(1 + \|\mathbf{y}_1\|^{2\chi-2} + \|\mathbf{y}_2\|^{2\chi-2}\right) \|\mathbf{y}_1 - \mathbf{y}_2\|^2.$$

Additionally, its derivatives exhibit polynomial growth in $\mathbf{y}$, uniformly in $\boldsymbol{\theta}$.

(A3) The solution $\mathbf{Y}$ to SDE (5) has invariant probability $\nu_0(\mathrm{d}\mathbf{y})$.

(A4) $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is invertible on $\overline{\Theta}_\Sigma$.

(A5) $\boldsymbol{\beta}$ is identifiable, that is, if $\mathbf{F}(\mathbf{y}, \boldsymbol{\beta}_1) = \mathbf{F}(\mathbf{y}, \boldsymbol{\beta}_2)$ for all $\mathbf{y} \in \mathbb{R}^{2d}$, then $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

Assumption (A1) ensures finiteness of the moments of the solution $\mathbf{X}$ [Tretyakov and Zhang, 2013], i.e.,

$$\mathbb{E}[\sup_{t \in [0,T]} \|\mathbf{Y}_t\|^{2p}] < C(1 + \|\mathbf{y}_0\|^{2p}), \qquad \forall\, p \geq 1. \tag{13}$$

Assumption (A3) is necessary for the ergodic theorem to ensure convergence in distribution. Assumption (A4) ensures that the model (5) is hypoelliptic. Assumption (A5) ensures the identifiability of the drift parameter.

2.4 Strang splitting scheme

Consider the following splitting of (5):

$$\mathrm{d}\mathbf{Y}_t^{[1]} = \widetilde{\mathbf{A}}(\mathbf{Y}_t^{[1]} - \widetilde{\mathbf{b}})\,\mathrm{d}t + \widetilde{\boldsymbol{\Sigma}}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{Y}_0^{[1]} = \mathbf{y}_0, \tag{14}$$

$$\mathrm{d}\mathbf{Y}_t^{[2]} = \widetilde{\mathbf{N}}(\mathbf{Y}_t^{[2]})\,\mathrm{d}t, \qquad \mathbf{Y}_0^{[2]} = \mathbf{y}_0. \tag{15}$$

There are no assumptions on the choice of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{b}}$, and thus the nonlinear function $\widetilde{\mathbf{N}}$. Indeed, we show that the asymptotic results hold for any choice of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{b}}$ in both the complete and the partial observation settings. This extends the results in Pilipovic et al. [2024], where it is shown to hold in the elliptic complete observation case, as well. While asymptotic results are invariant to the choice of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{b}}$, finite sample properties of the scheme and the corresponding estimators are very different, and it is important to choose the splitting wisely. Intuitively, when the process is close to a fixed point of the drift, the linear dynamics are dominating, whereas far from the fixed points, the nonlinearities might be dominating. If the drift has a fixed point $\mathbf{y}^\star$, we therefore suggest setting $\widetilde{\mathbf{A}} = D_{\mathbf{y}}\widetilde{\mathbf{F}}(\mathbf{y}^\star)$ and $\widetilde{\mathbf{b}} = \mathbf{y}^\star$. This choice is confirmed in simulations (for more details see Pilipovic et al. [2024]).

Solution of SDE (14) is an Ornstein–Uhlenbeck (OU) process given by the following h-flow:

$$\mathbf{Y}_{t_k}^{[1]} = \Phi_h^{[1]}(\mathbf{Y}_{t_{k-1}}^{[1]}) = \widetilde{\boldsymbol{\mu}}_h(\mathbf{Y}_{t_{k-1}}^{[1]}; \boldsymbol{\beta}) + \widetilde{\boldsymbol{\varepsilon}}_{h,k}, \tag{16}$$

$$\widetilde{\boldsymbol{\mu}}_h(\mathbf{y}; \boldsymbol{\beta}) \coloneqq e^{\widetilde{\mathbf{A}}h}(\mathbf{y} - \widetilde{\mathbf{b}}) + \widetilde{\mathbf{b}}, \tag{17}$$

$$\widetilde{\boldsymbol{\Omega}}_h = \int_0^h e^{\widetilde{\mathbf{A}}(h-u)} \widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^\top e^{\widetilde{\mathbf{A}}^\top(h-u)}\,\mathrm{d}u, \tag{18}$$

where $\widetilde{\boldsymbol{\varepsilon}}_{h,k} \overset{i.i.d}{\sim} \mathcal{N}_{2d}(\mathbf{0}, \widetilde{\boldsymbol{\Omega}}_h)$ for $k = 1, \ldots, N$. It is useful to rewrite $\widetilde{\boldsymbol{\Omega}}_h$ in the following block matrix form,

$$\widetilde{\boldsymbol{\Omega}}_h = \begin{bmatrix} \boldsymbol{\Omega}_h^{[\mathrm{SS}]} & \boldsymbol{\Omega}_h^{[\mathrm{SR}]} \\ \boldsymbol{\Omega}_h^{[\mathrm{RS}]} & \boldsymbol{\Omega}_h^{[\mathrm{RR}]} \end{bmatrix}, \tag{19}$$

where S in the superscript stands for smooth and R stands for rough. The Schur complement of $\widetilde{\boldsymbol{\Omega}}_h$ with respect to $\boldsymbol{\Omega}_h^{[\mathrm{RR}]}$ and the determinant of $\widetilde{\boldsymbol{\Omega}}_h$ are given by:

$$\boldsymbol{\Omega}_h^{[\mathrm{S|R}]} := \boldsymbol{\Omega}_h^{[\mathrm{SS}]} - \boldsymbol{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\Omega}_h^{[\mathrm{RR}]})^{-1}\boldsymbol{\Omega}_h^{[\mathrm{RS}]}, \qquad \det \widetilde{\boldsymbol{\Omega}}_h = \det \boldsymbol{\Omega}_h^{[\mathrm{RR}]} \det \boldsymbol{\Omega}_h^{[\mathrm{S|R}]}.$$

Assumptions (A1)-(A2) ensure the existence and uniqueness of the solution of (15) (Theorem 1.2.17 in Humphries and Stuart [2002]). Thus, there exists a unique function $\widetilde{\boldsymbol{f}}_h : \mathbb{R}^{2d} \times \Theta_\beta \to \mathbb{R}^{2d}$, for $h \geq 0$, such that

$$\mathbf{Y}_{t_k}^{[2]} = \Phi_h^{[2]}(\mathbf{Y}_{t_{k-1}}^{[2]}) = \widetilde{\boldsymbol{f}}_h(\mathbf{Y}_{t_{k-1}}^{[2]}; \boldsymbol{\beta}). \tag{20}$$

For all $\boldsymbol{\beta} \in \Theta_\beta$, the h-flow $\widetilde{\boldsymbol{f}}_h$ fulfills the following semi-group properties:

$$\widetilde{\boldsymbol{f}}_0(\mathbf{y}; \boldsymbol{\beta}) = \mathbf{y}, \qquad \widetilde{\boldsymbol{f}}_{t+s}(\mathbf{y}; \boldsymbol{\beta}) = \widetilde{\boldsymbol{f}}_t(\widetilde{\boldsymbol{f}}_s(\mathbf{y}; \boldsymbol{\beta}); \boldsymbol{\beta}), \;\; t, s \geq 0.$$

For $\mathbf{y} = (\mathbf{x}^\top, \mathbf{v}^\top)^\top$, we have:

$$\widetilde{\boldsymbol{f}}_h(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{f}_h(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) \end{bmatrix}, \tag{21}$$

where $\boldsymbol{f}_h(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta})$ is the solution of the ODE with vector field $\mathbf{N}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta})$.

We introduce another assumption needed to define the pseudo-likelihood based on the splitting scheme.

(A6) Inverse function $\widetilde{\boldsymbol{f}}_h^{-1}(\mathbf{y}; \boldsymbol{\beta})$ is defined asymptotically for all $\mathbf{y} \in \mathbb{R}^{2d}$ and all $\boldsymbol{\beta} \in \Theta_\beta$, when $h \to 0$.

Then, the inverse of $\tilde{\boldsymbol{f}}_h$ can be decomposed as:

$$\widetilde{\boldsymbol{f}}_h^{-1}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{f}_h^{\star -1}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta}) \end{bmatrix}, \tag{22}$$

where $\boldsymbol{f}_h^{\star -1}(\mathbf{x}, \mathbf{v}; \boldsymbol{\beta})$ is the rough part of the inverse of $\widetilde{\boldsymbol{f}}_h^{-1}$. It does not equal $\boldsymbol{f}_h^{-1}$ since the inverse does not propagate through coordinates when $\boldsymbol{f}_h$ depends on $\mathbf{x}$.

We are now ready to define the Strang splitting scheme for model (5).

Definition 2.1 *(Strang splitting) Let Assumptions (A1)-(A2) hold. The Strang approximation of the solution of* (5) *is given by:*

$$\Phi_h^{[\mathrm{str}]}(\mathbf{Y}_{t_{k-1}}^{[\mathrm{str}]}) = (\Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]})(\mathbf{Y}_{t_{k-1}}^{[\mathrm{str}]}) = \widetilde{\boldsymbol{f}}_{h/2}(\widetilde{\boldsymbol{\mu}}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{Y}_{t_{k-1}}^{[\mathrm{str}]})) + \widetilde{\boldsymbol{\varepsilon}}_{h,k}). \tag{23}$$

Remark 1 *The order of composition in the splitting schemes is not unique. Changing the order in the Strang splitting leads to a sum of 2 independent random variables, one Gaussian and one non-Gaussian, whose likelihood is not trivial. Thus, we only use the splitting* (23)*.*

2.5 Strang splitting estimators

In this section, we introduce four estimators, all based on the Strang splitting scheme. We distinguish between estimators based on complete observations (denoted by C when both $\mathbf{X}$ and $\mathbf{V}$ are observed) and partial observations (denoted by P when only $\mathbf{X}$ is observed). In applications, we typically only have access to partial observations, however, the full observation estimator is used as a building block for the partial observation case. Additionally, we distinguish the estimators based on the type of likelihood function employed. These are the full likelihood (denoted by F) and the marginal likelihood of the rough component (denoted by R). We furthermore use the conditional likelihood based on the smooth component given the rough part (denoted by S | R) to decompose the full likelihood.

2.5.1 Complete observations

Assume we observe the complete sample $\mathbf{Y}_{0:t_N} := (\mathbf{Y}_{t_k})_{k=1}^N$ from (5) at time steps $0 = t_0 < t_1 < ... < t_N = T$. For notational simplicity, we assume equidistant step size $h = t_k - t_{k-1}$. Strang splitting scheme (23) is a nonlinear transformation of a Gaussian random variable $\widetilde{\boldsymbol{\mu}}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{Y}_{t_{k-1}}^{[\mathrm{str}]})) + \widetilde{\boldsymbol{\varepsilon}}_{h,k}$. We define:

$$\widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta}) \coloneqq \widetilde{\boldsymbol{f}}_{h/2}^{-1}(\mathbf{Y}_{t_k}; \boldsymbol{\beta}) - \widetilde{\boldsymbol{\mu}}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{Y}_{t_{k-1}}; \boldsymbol{\beta}); \boldsymbol{\beta}), \tag{24}$$

and apply change of variables to get:

$$p(\mathbf{y}_{t_k} \mid \mathbf{y}_{t_{k-1}}) = p_{\mathcal{N}(\mathbf{0},\widetilde{\mathbf{\Omega}}_h)}(\widetilde{\mathbf{z}}_{k,k-1} \mid \mathbf{y}_{t_{k-1}})|\det D_{\mathbf{y}} \widetilde{\boldsymbol{f}}_{h/2}^{-1}(\mathbf{y}_{t_k})|.$$

Using $-\log|\det D_{\mathbf{y}} \widetilde{\boldsymbol{f}}_{h/2}^{-1}(\mathbf{y};\boldsymbol{\beta})| = \log|\det D_{\mathbf{y}} \widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y};\boldsymbol{\beta})|$ and $\det D_{\mathbf{y}} \widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y};\boldsymbol{\beta}) = \det D_{\mathbf{v}} \boldsymbol{f}_{h/2}(\mathbf{y};\boldsymbol{\beta})$, together with the Markov property of $\mathbf{Y}_{0:t_N}$, we get the following objective function based on the full log-likelihood:

$$\mathcal{L}^{[\mathrm{CF}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) \coloneqq \sum_{k=1}^{N} \left(\log\det \widetilde{\mathbf{\Omega}}_h(\boldsymbol{\theta}) + \widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta})^\top \widetilde{\mathbf{\Omega}}_h(\boldsymbol{\theta})^{-1} \widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta}) + 2\log|\det D_{\mathbf{v}} \boldsymbol{f}_{h/2}(\mathbf{Y}_{t_k};\boldsymbol{\beta})|\right). \quad (25)$$

Now, split $\widetilde{\mathbf{Z}}_{k,k-1}$ from (24) into the smooth and rough parts $\widetilde{\mathbf{Z}}_{k,k-1} = ((\mathbf{Z}^{[\mathrm{S}]}_{k,k-1})^\top, (\mathbf{Z}^{[\mathrm{R}]}_{k,k-1})^\top)^\top$ defined as:

$$\mathbf{Z}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) \coloneqq [\widetilde{Z}^{(i)}_{k,k-1}(\boldsymbol{\beta})]_{i=1}^{d} = \mathbf{X}_{t_k} - \boldsymbol{\mu}_h^{[\mathrm{S}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}), \quad (26)$$

$$\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) \coloneqq [\widetilde{Z}^{(i)}_{k,k-1}(\boldsymbol{\beta})]_{i=d+1}^{2d} = \boldsymbol{f}^{\star-1}_{h/2}(\mathbf{Y}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h^{[\mathrm{R}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}), \quad (27)$$

where

$$\boldsymbol{\mu}_h^{[\mathrm{S}]}(\mathbf{y};\boldsymbol{\beta}) \coloneqq [\widetilde{\mu}_h^{(i)}(\mathbf{y};\boldsymbol{\beta})]_{i=1}^{d}, \qquad \boldsymbol{\mu}_h^{[\mathrm{R}]}(\mathbf{y};\boldsymbol{\beta}) \coloneqq [\widetilde{\mu}_h^{(i)}(\mathbf{y};\boldsymbol{\beta})]_{i=d+1}^{2d}. \quad (28)$$

We also define the following sequence of vectors

$$\mathbf{Z}^{[\mathrm{S|R}]}_{k,k-1}(\boldsymbol{\beta}) \coloneqq \mathbf{Z}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) - \mathbf{\Omega}_h^{[\mathrm{SR}]}(\mathbf{\Omega}_h^{[\mathrm{RR}]})^{-1}\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}). \quad (29)$$

The formula for jointly normal distributions yields:

$$p_{\mathcal{N}(\mathbf{0},\widetilde{\mathbf{\Omega}}_h)}(\widetilde{\mathbf{z}}_{k,k-1} \mid \mathbf{y}_{t_{k-1}}) = p_{\mathcal{N}(\mathbf{0},\mathbf{\Omega}_h^{[\mathrm{RR}]})}(\mathbf{z}^{[\mathrm{R}]}_{k,k-1} \mid \mathbf{y}_{t_{k-1}}) \cdot p_{\mathcal{N}(\mathbf{\Omega}_h^{[\mathrm{SR}]}(\mathbf{\Omega}_h^{[\mathrm{RR}]})^{-1}\mathbf{z}^{[\mathrm{R}]}_{k,k-1},\mathbf{\Omega}_h^{[\mathrm{S|R}]})}(\mathbf{z}^{[\mathrm{S}]}_{k,k-1} \mid \mathbf{z}^{[\mathrm{R}]}_{k,k-1}, \mathbf{y}_{t_{k-1}}).$$

This leads to dividing the full log-likelihood $\mathcal{L}^{[\mathrm{CF}]}$ into a sum of the marginal log-likelihood $\mathcal{L}^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})$ and the smooth-given-rough log-likelihood $\mathcal{L}^{[\mathrm{CS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})$:

$$\mathcal{L}^{[\mathrm{CF}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = \mathcal{L}^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) + \mathcal{L}^{[\mathrm{CS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}),$$

where

$$\mathcal{L}^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) \coloneqq \sum_{k=1}^{N} \Bigg(\log\det \mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}) + \mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta})^\top \mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1}\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) + 2\log\left|\det D_{\mathbf{v}} \boldsymbol{f}_{h/2}(\mathbf{Y}_{t_k};\boldsymbol{\beta})\right| \Bigg), \quad (30)$$

$$\mathcal{L}^{[\mathrm{CS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) \coloneqq \sum_{k=1}^{N} \left(\log\det \mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta}) + \mathbf{Z}^{[\mathrm{S|R}]}_{k,k-1}(\boldsymbol{\beta})^\top \mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})^{-1}\mathbf{Z}^{[\mathrm{S|R}]}_{k,k-1}(\boldsymbol{\beta})\right). \quad (31)$$

The terms containing the drift parameter in $\mathcal{L}^{[\mathrm{CR}]}$ in (30) are of order $h^{1/2}$, as in the elliptic case, whereas the terms containing the drift parameter in $\mathcal{L}^{[\mathrm{CS|R}]}$ in (31) are of order $h^{3/2}$. Consequently, under a rapidly increasing experimental design where $Nh \to \infty$ and $Nh^2 \to 0$, the objective function (31) is degenerate for estimating the drift parameter. However, it contributes to the estimation of the diffusion parameter when the full objective function (25) is used. We show in later sections that employing (25) results in a lower asymptotic variance for the diffusion parameter making it more efficient in complete observation scenarios.

The estimators based on complete observations are then defined as:

$$\hat{\boldsymbol{\theta}}_N^{[\mathrm{obj}]} \coloneqq \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{[\mathrm{obj}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}), \quad \mathrm{obj} \in \{[\mathrm{CF}], [\mathrm{CR}]\}. \quad (32)$$

Although the full objective function is based on twice as many equations as the marginal likelihood, its implementation complexity, speed, and memory requirements are similar to the marginal objective function. Therefore, if the complete observations are available, we recommend using the objective function (25) based on the full likelihood.

2.5.2 Partial observations

Assume we only observe the smooth coordinates $\mathbf{X}_{0:t_N} \coloneqq (\mathbf{X}_{t_k})_{k=0}^N$. The observed process $\mathbf{X}_t$ alone is not a Markov process, although the complete process $\mathbf{Y}_t$ is. To approximate $\mathbf{V}_{t_k}$, we define the backward difference process:

$$\Delta_h \mathbf{X}_{t_k} \coloneqq \frac{\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}}}{h}. \tag{33}$$

From SDE (2) it follows that

$$\Delta_h \mathbf{X}_{t_k} = \frac{1}{h} \int_{t_{k-1}}^{t_k} \mathbf{V}_t \, \mathrm{d}t. \tag{34}$$

We propose to approximate $\mathbf{V}_{t_k}$ using $\Delta_h \mathbf{X}_{t_k}$ by any of the three approaches:

1. Backward difference approximation: $\mathbf{V}_{t_k} \approx \Delta_h \mathbf{X}_{t_k}$;
2. Forward difference approximation: $\mathbf{V}_{t_k} \approx \Delta_h \mathbf{X}_{t_{k+1}}$;
3. Central difference approximation: $\mathbf{V}_{t_k} \approx \frac{\Delta_h \mathbf{X}_{t_k} + \Delta_h \mathbf{X}_{t_{k+1}}}{2}$.

The forward difference approximation performs best in our simulation study, which is also the approximation method employed in Gloter [2006] and Samson and Thieullen [2012].

In the field of numerical approximations of ODEs, backward and forward finite differences have the same order of convergence, whereas the central difference has a higher convergence rate. However, the diffusion parameter estimator based on the central difference $(\mathbf{X}_{t_{k+1}} - \mathbf{X}_{t_{k-1}})/2h$ is less suitable because this approximation skips a data point and thus increases the estimator's variance. For further discussion, see Remark 6.

Thus, we focus exclusively on forward differences, following Gloter [2006], Samson and Thieullen [2012], and all proofs are done for this approximation. Similar results also hold for the backward difference, with some adjustments needed in the conditional moments due to filtration issues.

We start by approximating $\widetilde{\mathbf{Z}}$ for the case of partial observations denoted by $\widetilde{\overline{\mathbf{Z}}}$:

$$\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta}) \coloneqq \widetilde{\boldsymbol{f}}_{h/2}^{-1}(\mathbf{X}_{t_k}, \Delta_h \mathbf{X}_{t_{k+1}}; \boldsymbol{\beta}) - \widetilde{\boldsymbol{\mu}}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{X}_{t_{k-1}}, \Delta_h \mathbf{X}_{t_k}; \boldsymbol{\beta}); \boldsymbol{\beta}). \tag{35}$$

The smooth and rough parts of $\widetilde{\overline{\mathbf{Z}}}$ are thus equal to:

$$\overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) \coloneqq \mathbf{X}_{t_k} - \boldsymbol{\mu}_h^{[\mathrm{S}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{X}_{t_{k-1}}, \Delta_h \mathbf{X}_{t_k}; \boldsymbol{\beta}); \boldsymbol{\beta}), \tag{36}$$

$$\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) \coloneqq \boldsymbol{f}_{h/2}^{\star -1}(\mathbf{X}_{t_k}, \Delta_h \mathbf{X}_{t_{k+1}}; \boldsymbol{\beta}) - \boldsymbol{\mu}_h^{[\mathrm{R}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{X}_{t_{k-1}}, \Delta_h \mathbf{X}_{t_k}; \boldsymbol{\beta}); \boldsymbol{\beta}), \tag{37}$$

and

$$\overline{\mathbf{Z}}^{[\mathrm{S|R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) \coloneqq \overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) - \boldsymbol{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\Omega}_h^{[\mathrm{RR}]})^{-1} \overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}). \tag{38}$$

Compared to $\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}$ in (27), $\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}$ in (37) depends on three consecutive data points, with the additional point $\mathbf{X}_{t_{k+1}}$ entering through $\Delta_h \mathbf{X}_{t_{k+1}}$. Furthermore, $\mathbf{X}_{t_k}$ enters both $\boldsymbol{f}_{h/2}^{\star -1}$ and $\widetilde{\boldsymbol{\mu}}_h^{[\mathrm{R}]}$, rending them coupled. This coupling has a significant influence on later derivations of the estimator's asymptotic properties, in contrast to the elliptic case where the derivations simplify.

While it might seem straightforward to incorporate $\widetilde{\overline{\mathbf{Z}}}$, $\overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}$ and $\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k,k-1}$ into the objective functions (25), (30) and (31), it introduces bias in the estimators of the diffusion parameters, as also discussed in [Gloter, 2006, Samson and Thieullen, 2012]. The bias arises because $\mathbf{X}_{t_k}$ enters in both $\boldsymbol{f}_{h/2}^{\star -1}$ and $\widetilde{\boldsymbol{\mu}}_h^{[\mathrm{R}]}$, and the covariances of $\widetilde{\overline{\mathbf{Z}}}$, $\overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}$, and $\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k,k-1}$ differ from their complete observation counterparts. To eliminate this bias, Gloter [2006], Samson and Thieullen [2012] applied a correction of $2/3$ multiplied to $\log\det$ of the covariance term in the objective functions, which is $\log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ in the Euler-Maruyama discretization. We also need appropriate corrections to our objective functions (25), (30) and (31), however, caution is necessary because $\log\det \widetilde{\boldsymbol{\Omega}}_h(\boldsymbol{\theta})$ depends on both drift and diffusion parameters. To counterbalance this, we also incorporate an adjustment to h in $\boldsymbol{\Omega}_h$. Moreover, we add the term $4 \log |\det D_{\mathbf{v}} \boldsymbol{f}_{h/2}|$ to objective function (31) to obtain consistency of the drift estimator under partial observations. The detailed derivation of these correction factors will be elaborated in the following sections.

We thus propose the following objective functions:

$$\mathcal{L}^{[\mathrm{PF}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) \coloneqq \frac{4}{3}(N-2)\log\det\widetilde{\boldsymbol{\Omega}}_{3h/4}(\boldsymbol{\theta}) \tag{39}$$
$$+\sum_{k=1}^{N-1}\left(\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta})^{\top}\widetilde{\boldsymbol{\Omega}}_h(\boldsymbol{\theta})^{-1}\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta})+6\log|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_{k+1}};\boldsymbol{\beta})|\right),$$

$$\mathcal{L}^{[\mathrm{PR}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) \coloneqq \frac{2}{3}(N-2)\log\det\boldsymbol{\Omega}^{[\mathrm{RR}]}_{3h/2}(\boldsymbol{\theta}) \tag{40}$$
$$+\sum_{k=1}^{N-1}\left(\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta})^{\top}\boldsymbol{\Omega}^{[\mathrm{RR}]}_h(\boldsymbol{\theta})^{-1}\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta})+2\log\left|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_{k+1}};\boldsymbol{\beta}\right)\right|\right),$$

$$\mathcal{L}^{[\mathrm{PS|R}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) \coloneqq 2(N-2)\log\det\boldsymbol{\Omega}^{[\mathrm{S|R}]}_h(\boldsymbol{\theta}) \tag{41}$$
$$+\sum_{k=1}^{N-1}\left(\overline{\mathbf{Z}}^{[\mathrm{S|R}]}_{k+1,k,k-1}(\boldsymbol{\beta})^{\top}\boldsymbol{\Omega}^{[\mathrm{S|R}]}_h(\boldsymbol{\theta})^{-1}\overline{\mathbf{Z}}^{[\mathrm{S|R}]}_{k+1,k,k-1}(\boldsymbol{\beta})+4\log|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_{k+1}};\boldsymbol{\beta})|\right). \tag{42}$$

Remark 2 *Due to the correction factors in the objective functions, we now have that*

$$\mathcal{L}^{[\mathrm{PF}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) \neq \mathcal{L}^{[\mathrm{PR}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) + \mathcal{L}^{[\mathrm{PS|R}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}). \tag{43}$$

However, when expanding the objective functions (39)-(41) *using Taylor series to the lowest necessary order in h, their approximations will satisfy equality in* (43)*, as shown in Section 6.*

Remark 3 *Adding the extra term $4\log|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}|$ in* (41) *is necessary to keep the consistency of the drift parameter. However, this term is not initially present in objective function* (31)*, making this correction somehow artificial. This can potentially make the objective function further from the true log-likelihood.*

The estimators based on the partial sample are then defined as:

$$\hat{\boldsymbol{\theta}}^{[\mathrm{obj}]}_N \coloneqq \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{[\mathrm{obj}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}), \quad \mathrm{obj}\in\{[\mathrm{PF}],[\mathrm{PR}]\}. \tag{44}$$

In the partial observation case, the asymptotic variances of the diffusion estimators are identical whether using (39) or (40), in contrast to the complete observation scenario. This variance is shown to be $9/4$ times higher than the variance of the estimator $\hat{\boldsymbol{\theta}}^{[\mathrm{CF}]}_N$, and $9/8$ times higher than that of the estimator based on the marginal likelihood $\hat{\boldsymbol{\theta}}^{[\mathrm{CR}]}_N$.

The numerical study in Section 4 shows that the estimator based on the marginal objective function (40) is less biased than the one based on the full objective function (39) in finite sample scenarios with partial observations. A potential reason for this is discussed in Remark 3. Therefore, we recommend using the objective function (40) for partial observations.

3 Main results

This section states the two main results – consistency and asymptotic normality of all four proposed estimators. The key ideas for proofs are presented in Supplementary Materials S1.

First, we state the consistency of the estimators in both complete and partial observation cases. Let $\mathcal{L}^{[\mathrm{obj}]}_N$ be one of the objective functions (25), (30), (39) or (40) and $\widehat{\boldsymbol{\theta}}^{[\mathrm{obj}]}_N$ the corresponding estimator. Thus,

$$\mathrm{obj}\in\{[\mathrm{CF}],[\mathrm{CR}],[\mathrm{PF}],[\mathrm{PR}]\}.$$

We use superscript $[\mathrm{C}\cdot]$ to refer to any objective function in the complete observation case. Likewise, $[\cdot\mathrm{R}]$ stands for an objective function based on the rough marginal likelihood either in the complete or the partial observation case.

Theorem 3.1 (Consistency of the estimators) *Assume (A1)-(A6), $h\to 0$, and $Nh\to\infty$. Then under the complete observation or partial observation case, it holds:*

$$\widehat{\boldsymbol{\beta}}^{[\mathrm{obj}]}_N \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \boldsymbol{\beta}_0, \qquad \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}^{[\mathrm{obj}]}_N \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}_0.$$

Remark 4 *We split the full objective function* (25) *into the sum of the rough marginal likelihood* (30) *and the conditional smooth-given-rough likelihood* (31)*. Even if* (31) *cannot identify the drift parameter $\boldsymbol{\beta}$, it is an important intermediate step in understanding the full objective function* (25)*. This can be seen in the proof of Theorem 3.1, where we first establish consistency of the diffusion estimator with a convergence rate of $\sqrt{N}$, which is faster than $\sqrt{Nh}$, the convergence rate of the drift estimators. Then, under complete observations, we show that*

$$\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta},\boldsymbol{\sigma}_0)-\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma}_0)) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \int (\mathbf{F}_0(\mathbf{y})-\mathbf{F}(\mathbf{y}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{y})-\mathbf{F}(\mathbf{y}))\,\mathrm{d}\nu_0(\mathbf{y}). \tag{45}$$

The right-hand side of (45) *is non-negative, with a unique zero for $\mathbf{F}=\mathbf{F}_0$. Conversely, for objective function* (31)*, it holds:*

$$\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{CS|R}]}(\boldsymbol{\beta},\boldsymbol{\sigma})-\mathcal{L}_N^{[\mathrm{CS|R}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0. \tag{46}$$

Hence, (46) *does not have a unique minimum, making the drift parameter unidentifiable. Similar conclusions are drawn in the partial observation case.*

Now, we state the asymptotic normality of the estimator. First, we need some preliminaries. Let $\rho>0$ and $\mathcal{B}_\rho(\boldsymbol{\theta}_0)=\{\boldsymbol{\theta}\in\Theta \mid \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\rho\}$ be a ball around $\boldsymbol{\theta}_0$. Since $\boldsymbol{\theta}_0\in\Theta$, for sufficiently small $\rho>0$, $\mathcal{B}_\rho(\boldsymbol{\theta}_0)\in\Theta$. For $\hat{\boldsymbol{\theta}}_N^{[\mathrm{obj}]}\in\mathcal{B}_\rho(\boldsymbol{\theta}_0)$, the mean value theorem yields:

$$\left(\int_0^1 \mathbb{H}_{\mathcal{L}_N^{[\mathrm{obj}]}}(\boldsymbol{\theta}_0+t(\hat{\boldsymbol{\theta}}_N^{[\mathrm{obj}]}-\boldsymbol{\theta}_0))\,\mathrm{d}t\right)(\hat{\boldsymbol{\theta}}_N^{[\mathrm{obj}]}-\boldsymbol{\theta}_0)=-\nabla_{\boldsymbol{\theta}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0). \tag{47}$$

Define:

$$\mathbf{C}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}) \coloneqq \begin{bmatrix} \left[\frac{1}{Nh}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta})\right]_{i_1,i_2=1}^{r} & \left[\frac{1}{N\sqrt{h}}\partial^2_{\beta^{(i)}\sigma^{(j)}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta})\right]_{i=1,j=1}^{r,s} \\ \left[\frac{1}{N\sqrt{h}}\partial^2_{\sigma^{(j)}\beta^{(i)}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta})\right]_{i=1,j=1}^{r,s} & \left[\frac{1}{N}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta})\right]_{j_1,j_2=1}^{s} \end{bmatrix}, \tag{48}$$

$$\mathbf{s}_N^{[\mathrm{obj}]} \coloneqq \begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N^{[\mathrm{obj}]}-\boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\sigma}}_N^{[\mathrm{obj}]}-\boldsymbol{\sigma}_0) \end{bmatrix}, \qquad \boldsymbol{\lambda}_N^{[\mathrm{obj}]} \coloneqq \begin{bmatrix} -\frac{1}{\sqrt{Nh}}\nabla_{\boldsymbol{\beta}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0) \\ -\frac{1}{\sqrt{N}}\nabla_{\boldsymbol{\sigma}}\mathcal{L}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0) \end{bmatrix}, \tag{49}$$

and $\mathbf{D}_N^{[\mathrm{obj}]} \coloneqq \int_0^1 \mathbf{C}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0+t(\hat{\boldsymbol{\theta}}_N^{[\mathrm{obj}]}-\boldsymbol{\theta}_0))\,\mathrm{d}t$. Then, (47) is equivalent to $\mathbf{D}_N^{[\mathrm{obj}]}\mathbf{s}_N^{[\mathrm{obj}]}=\boldsymbol{\lambda}_N^{[\mathrm{obj}]}$. Let:

$$[\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)]_{i_1,i_2} \coloneqq \int (\partial_{\beta^{(i_1)}}\mathbf{F}_0(\mathbf{y}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\beta^{(i_2)}}\mathbf{F}_0(\mathbf{y}))\,\mathrm{d}\nu_0(\mathbf{y}),\ 1\le i_1,i_2\le r, \tag{50}$$

$$[\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)]_{j_1,j_2} \coloneqq \mathrm{Tr}((\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}),\ 1\le j_1,j_2\le s. \tag{51}$$

Theorem 3.2 *Let assumptions (A1)-(A6) hold, and let $h\to 0$, $Nh\to\infty$, and $Nh^2\to 0$. Then under complete observations, it holds:*

$$\begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N^{[\mathrm{CR}]}-\boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\sigma}}_N^{[\mathrm{CR}]}-\boldsymbol{\sigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & 2\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)^{-1} \end{bmatrix}\right),$$

$$\begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N^{[\mathrm{CF}]}-\boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\sigma}}_N^{[\mathrm{CF}]}-\boldsymbol{\sigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)^{-1} \end{bmatrix}\right),$$

under $\mathbb{P}_{\boldsymbol{\theta}_0}$. If only partial observations are available and the unobserved coordinates are approximated using the forward or backward differences, then

$$\begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N^{[\mathrm{PR}]}-\boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\sigma}}_N^{[\mathrm{PR}]}-\boldsymbol{\sigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \frac{9}{4}\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)^{-1} \end{bmatrix}\right),$$

$$\begin{bmatrix} \sqrt{Nh}(\hat{\boldsymbol{\beta}}_N^{[\mathrm{PF}]}-\boldsymbol{\beta}_0) \\ \sqrt{N}(\hat{\boldsymbol{\sigma}}_N^{[\mathrm{PF}]}-\boldsymbol{\sigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \frac{9}{4}\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)^{-1} \end{bmatrix}\right),$$

under $\mathbb{P}_{\boldsymbol{\theta}_0}$.

Here, we only outline the proof. According to Theorem 1 in Kessler [1997] or Theorem 1 in Sørensen and Uchida [2003], Lemmas 3.3 and 3.4 below are enough for establishing asymptotic normality of $\hat{\boldsymbol{\theta}}_N$. For more details, see proof of Theorem 1 in Sørensen and Uchida [2003].

Lemma 3.3 *Let $\mathbf{C}_N(\boldsymbol{\theta}_0)$ be defined in* (48)*. For $h \to 0$ and $Nh \to \infty$, it holds:*

$$\mathbf{C}_N^{[\mathrm{CR}]}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \begin{bmatrix} 2\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}, \qquad \mathbf{C}_N^{[\mathrm{PR}]}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \begin{bmatrix} 2\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \frac{2}{3}\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix},$$
$$\mathbf{C}_N^{[\mathrm{CF}]}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \begin{bmatrix} 2\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & 2\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}, \qquad \mathbf{C}_N^{[\mathrm{PF}]}(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \begin{bmatrix} 2\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \frac{8}{3}\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}.$$

Moreover, let ρ_N be a sequence such that $\rho_N \to 0$, then in all cases it holds:

$$\sup_{\|\boldsymbol{\theta}\| \le \rho_N} \|\mathbf{C}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0 + \boldsymbol{\theta}) - \mathbf{C}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0)\| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

Lemma 3.4 *Let $\boldsymbol{\lambda}_N$ be defined* (49)*. For $h \to 0$, $Nh \to \infty$ and $Nh^2 \to 0$, it holds:*

$$\boldsymbol{\lambda}_N^{[\mathrm{CR}]} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 4\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & 2\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}\right), \qquad \boldsymbol{\lambda}_N^{[\mathrm{PR}]} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 4\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}\right),$$
$$\boldsymbol{\lambda}_N^{[\mathrm{CF}]} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 4\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & 4\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}\right), \qquad \boldsymbol{\lambda}_N^{[\mathrm{PF}]} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 4\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & 16\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0) \end{bmatrix}\right),$$

under $\mathbb{P}_{\boldsymbol{\theta}_0}$.

Now, the two previous lemmas suggest

$$\mathbf{s}_N^{[\mathrm{obj}]} = (\mathbf{D}_n^{[\mathrm{obj}]})^{-1}\boldsymbol{\lambda}_N^{[\mathrm{obj}]} \xrightarrow{d} \mathbf{C}_N^{[\mathrm{obj}]}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\lambda}_N^{[\mathrm{obj}]}.$$

The previous line is not completely formal, but it gives the intuition. For more details on formally deriving the result, see Section 7.4 in Pilipovic et al. [2024] or proof of Theorem 1 in Sørensen and Uchida [2003].

4 Simulation study

This Section illustrates the simulation study of the Kramers oscillator (8), demonstrating the theoretical aspects and comparing our proposed estimators against estimators based on the EM and LL approximations. We chose to compare our proposed estimators to these two, because the EM estimator is routinely used in applications, and the LL estimator has shown to be one of the best state-of-the-art methods, see Pilipovic et al. [2024] for the elliptic case. The true parameters are set to $\eta_0 = 6.5, a_0 = 1, b_0 = 0.6$ and $\sigma_0^2 = 0.1$. We outline the estimators specifically designed for the Kramers oscillator, explain the simulation procedure, describe the optimization implemented in the `R` programming language R Core Team [2022], and then present and interpret the results.

4.1 Estimators used in the study

For the Kramers oscillator (8), the EM transition distribution is:

$$\begin{bmatrix} X_{t_k} \\ V_{t_k} \end{bmatrix} \mid \begin{bmatrix} X_{t_{k-1}} \\ V_{t_{k-1}} \end{bmatrix} = \begin{bmatrix} x \\ v \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x + hv \\ v + h\left(-\eta v + ax - bx^3\right) \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & h\sigma^2 \end{bmatrix}\right).$$

The ill-conditioned variance of this discretization restricts us to an estimator that only uses the marginal likelihood of the rough coordinate. The estimator for complete observations directly follows from the Gaussian distribution. The estimator for partial observations is defined as [Samson and Thieullen, 2012]:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{EM}}^{[\mathrm{PR}]} = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{2}{3}(N-3)\log\sigma^2 + \frac{1}{h\sigma^2}\sum_{k=1}^{N-2}(\Delta_h X_{t_{k+1}} - \Delta_h X_{t_k} - h(-\eta\Delta_h X_{t_{k-1}} + aX_{t_{k-1}} - bX_{t_{k-1}}^3))^2 \right\}.$$

To our knowledge, the LL estimator has not previously been applied to partial observations. Given the similar theoretical and computational performance of the Strang and LL discretizations, we suggest (without formal proof) to adjust the LL objective functions with the same correction factors as used in the Strang approach. The numerical evidence indicates

that the LL estimator has the same asymptotic properties as those proved for the Strang estimator. We omit the definition of the LL estimator due to its complexity (see Melnykova [2020], Pilipovic et al. [2024] and accompanying code).

To define S estimators based on the Strang splitting scheme, we first split SDE (8) as follows:

$$\mathrm{d}\begin{bmatrix} X_t \\ V_t \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -2a & -\eta \end{bmatrix}}_{\mathbf{A}} \left(\begin{bmatrix} X_t \\ V_t \end{bmatrix} - \underbrace{\begin{bmatrix} x_{\pm}^{\star} \\ 0 \end{bmatrix}}_{\mathbf{b}} \right) \mathrm{d}t + \underbrace{\begin{bmatrix} 0 \\ aX_t - bX_t^3 + 2a(X_t - x_{\pm}^{\star}) \end{bmatrix}}_{\mathbf{N}(X_t, V_t)} \mathrm{d}t + \begin{bmatrix} 0 \\ \sigma \end{bmatrix} \mathrm{d}W_t,$$

where $x_{\pm}^{\star} = \pm\sqrt{a/b}$ are the two stable points of the dynamics. Since there are two stable points, we suggest splitting with $x_{+}^{\star}$, when $X_t > 0$, and $x_{-}^{\star}$, when $X_t < 0$. This splitting follows the guidelines from [Pilipovic et al., 2024]. Note that the nonlinear ODE driven by $\mathbf{N}(x, v)$ has a trivial solution where x is a constant. To obtain Strang estimators, we plug in the corresponding components in the objective functions (25), (30), (39) and (40).

4.2 Trajectory simulation

We simulate a sample path using the EM discretization with a step size of $h^{\mathrm{sim}} = 0.0001$ to ensure good performance. To reduce discretization errors, we sub-sample from the path at wider intervals to get time step $h = 0.1$. The path has $N = 5000$ data points. We repeat the simulations to obtain 250 data sets.

4.3 Optimization in `R`

For optimizing the objective functions, we proceed as in Pilipovic et al. [2024] using the `R` package `torch` [Falbel and Luraschi, 2022], which allows automatic differentiation. The optimization employs the resilient backpropagation algorithm, `optim_rprop`. We use the default hyperparameters and limit the number of optimization iterations to 2000. The convergence criterion is set to a precision of 10^{-5} for the difference between estimators in consecutive iterations. The initial parameter values are set to $(-0.1, -0.1, 0.1, 0.1)$.

4.4 Results

The results of the simulation study are presented in Figure 1. Figure 1A) presents the distributions of the normalized estimators in the complete and partial observation cases. The S and LL estimators exhibit nearly identical performance, particularly in the complete observation scenario. In contrast, the EM method displays significant underperformance and notable bias. The variances of the S and LL rough-likelihood estimators of σ^2 are higher compared to those derived from the full likelihood, aligning with theoretical expectations. Interestingly, in the partial observation scenario, Figure 1A) reveals that estimators employing the full likelihood display greater finite sample bias compared to those based on the rough likelihood. Possible reasons for this bias are discussed in Remark 3. However, it is noteworthy that this bias is eliminated for smaller time steps, e.g. $h = 0.0001$ (not shown), thus confirming the theoretical asymptotic results. This observation suggests that the rough likelihood is preferable under partial observations due to its lower bias. Backward finite difference approximations of the velocity variables perform similarly to the forward differences and are therefore excluded from the figure for clarity.

We closely examine the variances of the S estimators of σ^2 in Figure 1B). The LL estimators are omitted due to their similarity to the S estimators, and because the computation times for the LL estimators are prohibitive. To align more closely with the asymptotic predictions, we opt for $h = 0.02$ and conduct 1000 simulations. Additionally, we set $\sigma_0^2 = 100$ to test different noise levels. Atop each empirical distribution, we overlay theoretical normal densities that match the variances as per Theorem 3.2. The theoretical variance is derived from $C_{\sigma^2}(\boldsymbol{\theta}_0)$ in (51), which for the Kramers oscillator in (8) is:

$$C_{\sigma^2}(\boldsymbol{\theta}_0) = \frac{1}{\sigma_0^4}. \tag{52}$$

Figure 1 illustrates that the lowest variance of the diffusion estimator is observed when using the full likelihood with complete observations. The second lowest variance is achieved using the rough likelihood with complete observations. The largest variance is observed in the partial observation case; however, it remains independent of whether the full or rough likelihood is used. Once again, we observe that using the full likelihood introduces additional finite sample bias.

In Figure 1C), we compare running times calculated using the `tictoc` package in `R`. Running times are measured from the start of the optimization step until convergence. The figure depicts the median over 250 repetitions to mitigate the influence of outliers. The EM method is notably the fastest; however, the S estimators exhibit only slightly slower performance. The LL estimators are 10-100 times slower than the S estimators, depending on whether complete or partial observations are used and whether the full or rough likelihood is employed.

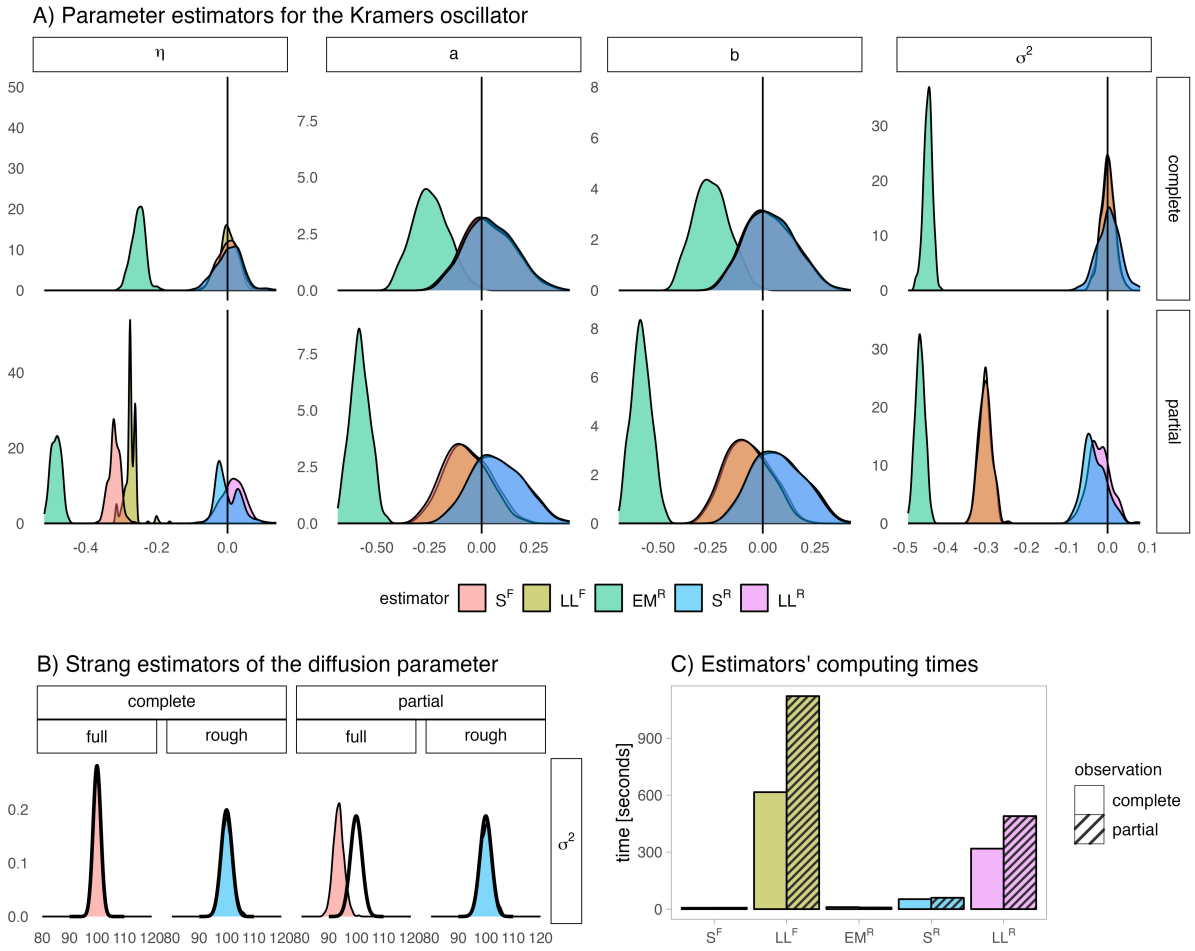

Figure 1: Parameter estimates in a simulation study for the Kramers oscillator, eq. (8). The color code remains consistent across all three figures. **A)** Normalized distributions of parameter estimation errors $(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \oslash \boldsymbol{\theta}_0$ in both complete and partial observation cases, based on 250 simulated data sets with $h = 0.1$ and $N = 5000$. Each column corresponds to a different parameter, while the color indicates the type of estimator. Estimators are distinguished by superscripted objective functions (F for full and R for rough). **B)** Distribution of $\widehat{\sigma}^2_N$ estimators based on 1000 simulations with $h = 0.02$ and $N = 5000$ across different observation settings (complete or partial) and likelihood choices (full or rough) using the Strang splitting scheme. The true value of σ^2 is set to $\sigma^2_0 = 100$. Theoretical normal densities are overlaid for comparison. Theoretical variances are calculated based on $C_{\sigma^2}(\boldsymbol{\theta}_0)$, eq. (52). **C)** Median computing time in seconds for one estimation of various estimators based on 250 simulations with $h = 0.1$ and $N = 5000$. Shaded color patterns represent times in the partial observation case, while no color pattern indicates times in the complete observation case.

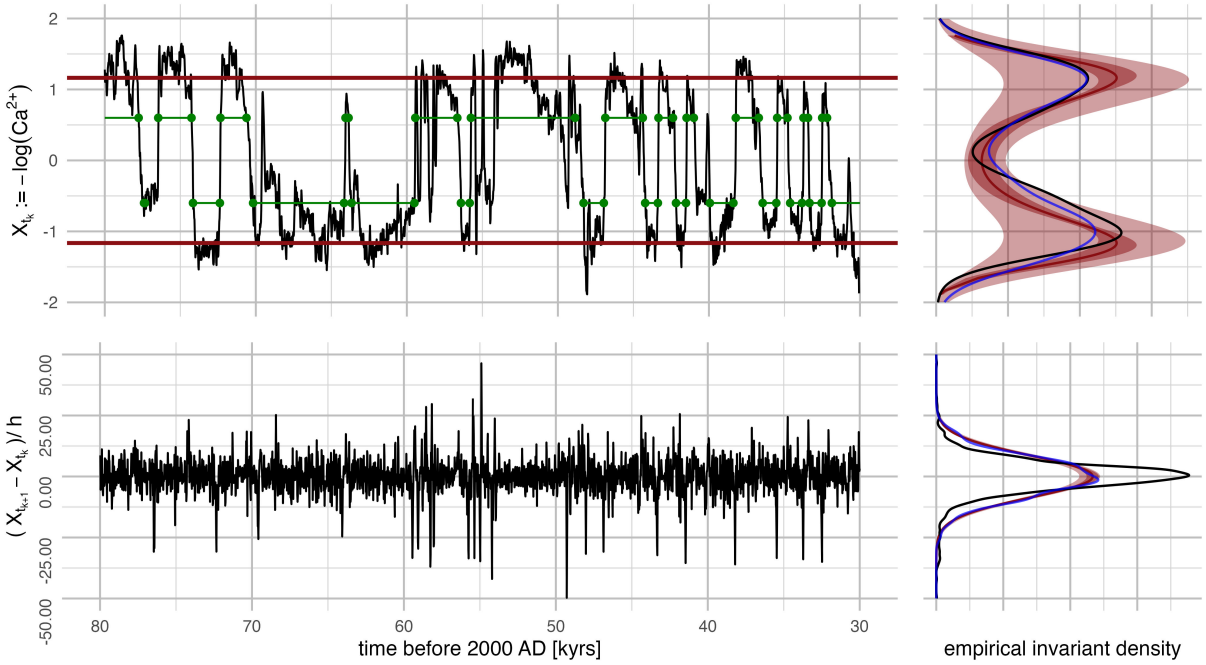

Figure 2: Ice core data from Greenland. Left: Trajectories over time (in kilo years) of the centered negative logarithm of the Ca^{2+} measurements (top) and forward difference approximations of its rate of change (bottom). The two vertical dark red lines represent the estimated stable equilibria of the double-well potential function. Green points denote up- and down-crossings of level ± 0.6, conditioned on having crossed the other level. Green vertical lines indicate empirical estimates of occupancy in either of the two metastable states. **Right:** Empirical densities (black) alongside estimated invariant densities with confidence intervals (dark red), prediction intervals (light red), and the empirical density of a simulated sample from the estimated model (blue).

5 Application to Greenland Ice Core Data

During the last glacial period, significant climatic shifts known as Dansgaard-Oeschger (DO) events have been documented in paleoclimatic records [Dansgaard et al., 1993]. Proxy data from Greenland ice cores, particularly stable water isotope composition (δ^{18}O) and calcium ion concentrations (Ca^{2+}), offer valuable insights into these past climate variations [Boers et al., 2017, 2018, Boers, 2018, Ditlevsen et al., 2002, Lohmann and Ditlevsen, 2019, Hassanibesheli et al., 2020].

The δ^{18}O ratio, reflecting the relative abundance of ^{18}O and ^{16}O isotopes in ice, serves as a proxy for paleotemperatures during snow deposition. Conversely, calcium ions, originating from dust deposition, exhibit a strong negative correlation with δ^{18}O, with higher calcium ion levels indicating colder conditions. Here, we prioritize Ca^{2+} time series due to its finer temporal resolution.

In Greenland ice core records, the DO events manifest as abrupt transitions from colder climates (stadials) to approximately 10 degrees warmer climates (interstadials) within a few decades. Although the waiting times between state switches last a couple of thousand years, their spacing exhibits significant variability. The underlying mechanisms driving these changes remain largely elusive, prompting discussions on whether they follow cyclic patterns, result from external forcing, or emerge from noise-induced processes [Boers, 2018, Ditlevsen et al., 2007]. We aim to determine if the observed data can be explained by noise-induced transitions of the Kramers oscillator.

The measurements were conducted at the summit of the Greenland ice sheet as part of the Greenland Icecore Project (GRIP) [Anklin et al., 1993, Andersen et al., 2004]. Originally, the data were sampled at 5 cm intervals, resulting in a non-equidistant time series due to ice compression at greater depths, where 5 cm of ice core spans longer time periods. For our analysis, we use a version of the data transformed into a uniformly spaced series through 20-year binning and averaging. This transformation simplifies the analysis and highlights significant climatic trends. The dataset is available in the supplementary material of [Rasmussen et al., 2014, Seierstad et al., 2014].

To address the large amplitudes and negative correlation with temperature, we transform the data to minus the logarithm of Ca^{2+}, where higher values of the transformed variable indicate warmer climates at the time of snow deposition. Additionally, we center the transformed measurements around zero. With the 20-year binning, to obtain one point per 20 years, we average across the bins, resulting in a time step of $h = 0.02\text{kyr}$ ($1\text{kyr} = 1000$ years). Additionally, we addressed a few missing values using the `na.approx` function from the `zoo` package. Following the approach of Hassanibesheli et al. [2020], we analyze a subset of the data with a sufficiently good signal-to-noise ratio. Hassanibesheli et al. [2020] examined the data from 30 to 60kyr before present. Here, we extend the analysis to cover 30kyr to 80kyr, resulting in a time interval of $T = 50\text{kyr}$ and a sample size of $N = 2500$. We approximate the velocity of the transformed Ca^{2+} by the forward difference method. The trajectories and empirical invariant distributions are illustrated in Figure 2.

We fit the Kramers oscillator to the $-\log Ca^{2+}$ time series and estimate parameters using the Strang estimator. Following Theorem 3.2, we compute $\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)$ from (50). Applying the invariant density $\pi_0(x, v)$ from (10), which decouples into $\pi_0(x)$ (11) and a Gaussian zero-mean and $\sigma_0^2/(2\eta_0)$ variance, leads us to:

$$\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0) = \begin{bmatrix} \frac{1}{2\eta_0} & 0 & 0 \\ 0 & \frac{1}{\sigma_0^2}\int_{-\infty}^{\infty} x^2\pi_0(x)\,\mathrm{d}x & -\frac{1}{\sigma_0^2}\int_{-\infty}^{\infty} x^4\pi_0(x)\,\mathrm{d}x \\ 0 & -\frac{1}{\sigma_0^2}\int_{-\infty}^{\infty} x^4\pi_0(x)\,\mathrm{d}x & \frac{1}{\sigma_0^2}\int_{-\infty}^{\infty} x^6\pi_0(x)\,\mathrm{d}x \end{bmatrix}. \tag{53}$$

Thus, to obtain 95% confidence intervals (CI) for the estimated parameters, we plug $\widehat{\boldsymbol{\theta}}_N$ into (52) and (53). The estimators and confidence intervals are shown in Table 1. We also calculate the expected waiting time τ, eq. (12), of crossing from one state to another, and its confidence interval using the Delta Method.

Parameter	Estimate	95% CI
η	62.5	$59.4 - 65.6$
a	296.7	$293.6 - 299.8$
b	219.1	$156.4 - 281.7$
σ^2	9125	$8589 - 9662$
τ	3.97	$3.00 - 4.94$

Table 1: Estimated parameters of the Kramers oscillator from Greenland ice core data.

The model fit is assessed in the right panels of Figure 2. Here, we present the empirical distributions of the two coordinates along with the fitted theoretical invariant distribution and a 95% confidence interval. Additionally, a prediction interval for the distribution is provided by simulating 1000 datasets from the fitted model, matching the size of the empirical data. We estimate the empirical distributions for each simulated dataset and construct a 95% prediction interval using the pointwise 2.5th and 97.5th percentiles of these estimates. A single example trace is included in blue. While the fitted distribution for $-\log Ca^{2+}$ appears to fit well, even with this symmetric model, the velocity variables are not adequately captured. This discrepancy is likely due to the presence of extreme values in the data that are not effectively accounted for by additive Gaussian noise. Consequently, the model compensates by estimating a large variance.

We estimate the waiting time between metastable states to be approximately 4000 years. However, this approximation relies on certain assumptions, namely $62.5 \approx \eta \gg \sqrt{a} \approx 17.2$ and $73 \approx \sigma^2/2\eta \ll a^2/4b \approx 100$. Thus, the accuracy of the approximation may not be highly accurate.

Defining the current state of the process is not straightforward. One method involves identifying successive up- and down-crossings of predefined thresholds within the smoothed data. However, the estimated occupancy time in each state depends on the level of smoothing applied and the distance of crossing thresholds from zero. Using a smoothing technique involving running averages within windows of 11 data points (equivalent to 220 years) and detecting down- and up-crossings of levels ± 0.6, we find an average occupancy time of 4058 years in stadial states and 3550 years in interstadial states. Nevertheless, the actual occupancy times exhibit significant variability, ranging from 60 to 6900 years, with the central 50% of values falling between 665 and 2115 years. This classification of states is depicted in green in Figure 2. Overall, the estimated mean occupancy time inferred from the Kramers oscillator appears reasonable.

6 Technical results

In this Section, we present all the necessary technical properties that are used to derive the main results of the paper.

We start by expanding $\widetilde{\mathbf{\Omega}}_h$ and its block components $\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1}, \mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})^{-1}, \log\det\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}), \log\det\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})$ and $\log|\det D\boldsymbol{f}_{h/2}(\mathbf{y};\boldsymbol{\beta})|$ when h goes to zero. Then, we expand $\widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta})$ and $\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta})$ around $\mathbf{Y}_{t_{k-1}}$ when h goes to zero. The main tools used are Itô's lemma, Taylor expansions, and Fubini's theorem. The final result is stated in Propositions 6.6 and 6.7. The approximations depend on the drift function $\mathbf{F}$, the nonlinear part $\mathbf{N}$, and some correlated sequences of Gaussian random variables. Finally, we obtain approximations of the objective functions (25), (30), (31) and (39) - (41). Proofs of all the stated propositions and lemmas in this section are in Supplementary Material S1.

6.1 Covariance matrix $\widetilde{\mathbf{\Omega}}_h$

The covariance matrix $\widetilde{\mathbf{\Omega}}_h$ is approximated by:

$$\begin{aligned}\widetilde{\mathbf{\Omega}}_h &= \int_0^h e^{\widetilde{\mathbf{A}}(h-u)}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top e^{\widetilde{\mathbf{A}}^\top(h-u)}\,\mathrm{d}u \\ &= h\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top + \frac{h^2}{2}(\widetilde{\mathbf{A}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top + \widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top\widetilde{\mathbf{A}}^\top) + \frac{h^3}{6}(\widetilde{\mathbf{A}}^2\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top + 2\widetilde{\mathbf{A}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top\widetilde{\mathbf{A}}^\top + \widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top(\widetilde{\mathbf{A}}^2)^\top) \\ &\quad + \frac{h^4}{24}(\widetilde{\mathbf{A}}^3\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top + 3\widetilde{\mathbf{A}}^2\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top\widetilde{\mathbf{A}}^\top + 3\widetilde{\mathbf{A}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top(\widetilde{\mathbf{A}}^2)^\top + \widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^\top(\widetilde{\mathbf{A}}^3)^\top) + \mathbf{R}(h^5,\mathbf{y}_0). \qquad (54)\end{aligned}$$

The following lemma approximates each block of $\widetilde{\mathbf{\Omega}}_h$ up to the first two leading orders of h. The result follows directly from equations (4), (6), and (54).

Lemma 6.1 *The covariance matrix $\widetilde{\mathbf{\Omega}}_h$ defined in* (54)-(19) *approximates block-wise as:*

$$\begin{aligned}\mathbf{\Omega}_h^{[\mathrm{SS}]}(\boldsymbol{\theta}) &= \frac{h^3}{3}\mathbf{\Sigma}\mathbf{\Sigma}^\top + \frac{h^4}{8}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\mathbf{\Sigma}\mathbf{\Sigma}^\top + \mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top) + \mathbf{R}(h^5,\mathbf{y}_0),\\ \mathbf{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\theta}) &= \frac{h^2}{2}\mathbf{\Sigma}\mathbf{\Sigma}^\top + \frac{h^3}{6}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\mathbf{\Sigma}\mathbf{\Sigma}^\top + 2\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top) + \mathbf{R}(h^4,\mathbf{y}_0),\\ \mathbf{\Omega}_h^{[\mathrm{RS}]}(\boldsymbol{\theta}) &= \frac{h^2}{2}\mathbf{\Sigma}\mathbf{\Sigma}^\top + \frac{h^3}{6}(2\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\mathbf{\Sigma}\mathbf{\Sigma}^\top + \mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top) + \mathbf{R}(h^4,\mathbf{y}_0),\\ \mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}) &= h\mathbf{\Sigma}\mathbf{\Sigma}^\top + \frac{h^2}{2}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\mathbf{\Sigma}\mathbf{\Sigma}^\top + \mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top) + \mathbf{R}(h^3,\mathbf{y}_0).\end{aligned}$$

Building on Lemma 6.1, we calculate products, inverses, and logarithms of the components of $\widetilde{\mathbf{\Omega}}_h$ in the following lemma.

Lemma 6.2 *For the covariance matrix $\widetilde{\mathbf{\Omega}}_h$ defined in* (54) *it holds:*

(i) $\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1} = \frac{1}{h}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1} - \frac{1}{2}((\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}\mathbf{A}_\mathbf{v}(\boldsymbol{\beta}) + \mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}) + \mathbf{R}(h,\mathbf{y}_0)$;

(ii) $\mathbf{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\theta})\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1} = \frac{h}{2}\mathbf{I} - \frac{h^2}{12}(\mathbf{A}_\mathbf{v} - \mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}) + \mathbf{R}(h^3,\mathbf{y}_0)$;

(iii) $\mathbf{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\theta})\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1}\mathbf{\Omega}_h^{[\mathrm{RS}]}(\boldsymbol{\theta}) = \frac{h^3}{4}\mathbf{\Sigma}\mathbf{\Sigma}^\top + \frac{h^4}{8}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\mathbf{\Sigma}\mathbf{\Sigma}^\top + \mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top) + \mathbf{R}(h^5,\mathbf{y}_0)$;

(iv) $\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta}) = \frac{h^3}{12}\mathbf{\Sigma}\mathbf{\Sigma}^\top + \mathbf{R}(h^5,\mathbf{y}_0)$;

(v) $\log\det\mathbf{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}) = d\log h + \log\det\mathbf{\Sigma}\mathbf{\Sigma}^\top + h\,\mathrm{Tr}\,\mathbf{A}_\mathbf{v}(\boldsymbol{\beta}) + R(h^2,\mathbf{y}_0)$;

(vi) $\log\det\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta}) = 3d\log h + \log\det\mathbf{\Sigma}\mathbf{\Sigma}^\top + R(h^2,\mathbf{y}_0)$;

(vii) $\log\det\widetilde{\mathbf{\Omega}}_h(\boldsymbol{\theta}) = 4d\log h + 2\log\det\mathbf{\Sigma}\mathbf{\Sigma}^\top + h\,\mathrm{Tr}\,\mathbf{A}_\mathbf{v}(\boldsymbol{\beta}) + R(h^2,\mathbf{y}_0)$.

Remark 5 *We adjusted the objective functions for partial observations using the term $c\log\det\mathbf{\Omega}_{h/c}^{[\cdot]}$, where c is a correction constant. This adjustment keeps the term $h\,\mathrm{Tr}\,\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})$ in (v)-(vii) constant, not affecting the asymptotic distribution of the drift parameter. There is no h^4-term in $\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})$ which simplifies the approximation of $\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})^{-1}$ and $\log\det\mathbf{\Omega}_h^{[\mathrm{S|R}]}(\boldsymbol{\theta})$. Consequently, this makes* (41) *a bad choice for estimating the drift parameter.*

6.2 Nonlinear solution $\widetilde{\boldsymbol{f}}_h$

We now state a useful proposition for the nonlinear solution $\widetilde{\boldsymbol{f}}_h$ (Section 1.8 in [Hairer et al., 1993]).

Proposition 6.3 *Let Assumptions (A1), (A2) and (A6) hold. When $h \to 0$, the h-flow of* (15) *approximates as:*

$$\widetilde{\boldsymbol{f}}_h(\mathbf{y}) = \mathbf{y} + h\widetilde{\mathbf{N}}(\mathbf{y}) + \frac{h^2}{2}(D_{\mathbf{y}}\widetilde{\mathbf{N}}(\mathbf{y}))\widetilde{\mathbf{N}}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}), \tag{55}$$

$$\widetilde{\boldsymbol{f}}_h^{-1}(\mathbf{y}) = \mathbf{y} - h\widetilde{\mathbf{N}}(\mathbf{y}) + \frac{h^2}{2}(D_{\mathbf{y}}\widetilde{\mathbf{N}}(\mathbf{y}))\widetilde{\mathbf{N}}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}). \tag{56}$$

Applying the previous proposition on (21) and (22), we get:

$$\boldsymbol{f}_h(\mathbf{y}) = \mathbf{v} + h\mathbf{N}(\mathbf{y}) + \frac{h^2}{2}(D_{\mathbf{v}}\mathbf{N}(\mathbf{y}))\mathbf{N}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}), \tag{57}$$

$$\boldsymbol{f}_h^{\star -1}(\mathbf{y}) = \mathbf{v} - h\mathbf{N}(\mathbf{y}) + \frac{h^2}{2}(D_{\mathbf{v}}\mathbf{N}(\mathbf{y}))\mathbf{N}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}). \tag{58}$$

The following lemma approximates $\log|\det D\boldsymbol{f}_{h/2}(\mathbf{y};\boldsymbol{\beta})|$ in the objective functions and connects it with Lemma 6.2.

Lemma 6.4 *Let $\widetilde{\boldsymbol{f}}_h$ be the function defined in* (21)*. It holds:*

$$2\log|\det D\boldsymbol{f}_{h/2}(\mathbf{Y}_{t_k};\boldsymbol{\beta})| = h\,\mathrm{Tr}\,D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}) + R(h^{3/2}, \mathbf{Y}_{t_{k-1}}),$$
$$2\log|\det D\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k}, \Delta_h\mathbf{X}_{t_{k+1}};\boldsymbol{\beta})| = h\,\mathrm{Tr}\,D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}) + R(h^{3/2}, \mathbf{Y}_{t_{k-1}}).$$

An immediate consequence of the previous lemma and that $D_{\mathbf{v}}\mathbf{F}(\mathbf{y};\boldsymbol{\beta}) = \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta}) + D_{\mathbf{v}}\mathbf{N}(\mathbf{y};\boldsymbol{\beta})$ is

$$\log\det\boldsymbol{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}) + 2\log|\det D\boldsymbol{f}_{h/2}(\mathbf{Y}_{t_k};\boldsymbol{\beta})| = \log\det h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + h\,\mathrm{Tr}\,D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}) + R(h^{3/2}, \mathbf{Y}_{t_{k-1}}).$$

The same equality holds when $\mathbf{Y}_{t_k}$ is approximated by $(\mathbf{X}_{t_k}, \Delta_h\mathbf{X}_{t_{k+1}})$. The following lemma expands function $\boldsymbol{\mu}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y}))$ up to the highest necessary order of h.

Lemma 6.5 *For the functions $\widetilde{\boldsymbol{f}}_h$ in* (21) *and $\widetilde{\boldsymbol{\mu}}_h$ in* (28)*, it holds*

$$\boldsymbol{\mu}_h^{[\mathrm{S}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y})) = \mathbf{x} + h\mathbf{v} + \frac{h^2}{2}\mathbf{F}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}), \tag{59}$$

$$\boldsymbol{\mu}_h^{[\mathrm{R}]}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y})) = \mathbf{v} + h(\mathbf{F}(\mathbf{y}) - \frac{1}{2}\mathbf{N}(\mathbf{y})) + \mathbf{R}(h^2, \mathbf{y}). \tag{60}$$

6.3 Random variables $\widetilde{\mathbf{Z}}_{k,k-1}$ and $\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}$

To approximate the random variables $\mathbf{Z}_{k,k-1}^{[\mathrm{S}]}(\boldsymbol{\beta})$, $\mathbf{Z}_{k,k-1}^{[\mathrm{R}]}(\boldsymbol{\beta})$, $\overline{\mathbf{Z}}_{k,k-1}^{[\mathrm{S}]}(\boldsymbol{\beta})$, and $\overline{\mathbf{Z}}_{k+1,k,k-1}^{[\mathrm{R}]}(\boldsymbol{\beta})$ around $\mathbf{Y}_{t_{k-1}}$, we start by defining the following random sequences:

$$\boldsymbol{\eta}_{k-1} \coloneqq \frac{1}{h^{1/2}}\int_{t_{k-1}}^{t_k} \mathrm{d}\mathbf{W}_t, \tag{61}$$

$$\boldsymbol{\xi}_{k-1} \coloneqq \frac{1}{h^{3/2}}\int_{t_{k-1}}^{t_k} (t - t_{k-1})\,\mathrm{d}\mathbf{W}_t, \qquad \boldsymbol{\xi}'_k \coloneqq \frac{1}{h^{3/2}}\int_{t_k}^{t_{k+1}} (t_{k+1} - t)\,\mathrm{d}\mathbf{W}_t, \tag{62}$$

$$\boldsymbol{\zeta}_{k-1} \coloneqq \frac{1}{h^{5/2}}\int_{t_{k-1}}^{t_k} (t - t_{k-1})^2\,\mathrm{d}\mathbf{W}_t, \qquad \boldsymbol{\zeta}'_k \coloneqq \frac{1}{h^{5/2}}\int_{t_k}^{t_{k+1}} (t_{k+1} - t)^2\,\mathrm{d}\mathbf{W}_t. \tag{63}$$

The random variables (61)-(63) are Gaussian with mean zero. Moreover, at time t_k they are $\mathcal{F}_{t_{k+1}}$ measurable and independent of $\mathcal{F}_{t_k}$. The following linear combinations of (61)-(63) appear in the expansions in the partial observation case:

$$\mathbf{U}_{k,k-1} \coloneqq \boldsymbol{\xi}'_k + \boldsymbol{\xi}_{k-1}, \tag{64}$$

$$\mathbf{Q}_{k,k-1} \coloneqq \boldsymbol{\zeta}'_k + 2\boldsymbol{\eta}_{k-1} - \boldsymbol{\zeta}_{k-1}. \tag{65}$$

It is not hard to check that $\boldsymbol{\xi}'_k + \boldsymbol{\eta}_{k-1} - \boldsymbol{\xi}'_{k-1} = \mathbf{U}_{k,k-1}$. This alternative representation of $\mathbf{U}_{k,k-1}$ will be used later in proofs.

The Itô isometry yields:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}] = \mathbf{I}, \qquad \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\eta}_{k-1}\boldsymbol{\xi}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}] = \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\eta}_{k-1}\boldsymbol{\xi}_{k-1}'^\top \mid \mathcal{F}_{t_{k-1}}] = \frac{1}{2}\mathbf{I}, \tag{66}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}_{k-1}'^\top \mid \mathcal{F}_{t_{k-1}}] = \frac{1}{6}\mathbf{I}, \qquad \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}] = \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}'_k\boldsymbol{\xi}_k'^\top \mid \mathcal{F}_{t_{k-1}}] = \frac{1}{3}\mathbf{I}, \tag{67}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top \mid \mathcal{F}_{t_{k-1}}] = \frac{2}{3}\mathbf{I}, \qquad \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{U}_{k,k-1}(\mathbf{U}_{k,k-1} + 2\boldsymbol{\xi}'_{k-1})^\top \mid \mathcal{F}_{t_{k-1}}] = \mathbf{I}. \tag{68}$$

The covariances of other combinations of the random variables (61)-(63) are not needed for the proofs. However, to derive asymptotic properties, we need some fourth moments calculated in Supplementary Materials S1.

The following two propositions are the last building blocks for approximating the objective functions (30)-(31) and (40)-(41).

Proposition 6.6 *The random variables $\widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta})$ in* (24) *and $\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta})$ in* (35) *are approximated by:*

$$\begin{aligned}
\mathbf{Z}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) &= h^{3/2}\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \frac{h^2}{2}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) + \frac{h^{5/2}}{2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\zeta}'_{k-1} + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}),\\
\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) &= h^{1/2}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1}\\
&\quad + h^{3/2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}),\\
\overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) &= -\frac{h^2}{2}\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \frac{h^{5/2}}{2}D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}),\\
\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= h^{1/2}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}\\
&\quad - h^{3/2}D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \frac{h^{3/2}}{2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\mathbf{Q}_{k,k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Remark 6 *Proposition 6.6 yield*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta})\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta})^\top \mid \mathbf{Y}_{t_{k-1}}] &= h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}) = \boldsymbol{\Omega}_h^{[\mathrm{RR}]} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}),\\
\mathbb{E}_{\boldsymbol{\theta}_0}[\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta})\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta})^\top \mid \mathbf{Y}_{t_{k-1}}] &= \frac{2}{3}h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}) = \frac{2}{3}\boldsymbol{\Omega}_h^{[\mathrm{RR}]} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Thus, the correction factor $2/3$ in (40) *compensates for the underestimation of the covariance of $\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta})$. Similarly, it can be shown that the same underestimation happens when using the backward difference. On the other hand, when using the central difference, it can be shown that*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\overline{\mathbf{Z}}^{[\mathrm{R}],central}_{k+1,k,k-1}(\boldsymbol{\beta})\overline{\mathbf{Z}}^{[\mathrm{R}],central}_{k+1,k,k-1}(\boldsymbol{\beta})^\top \mid \mathbf{Y}_{t_{k-1}}] = \frac{5}{12}h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}),$$

which is a larger deviation from $\boldsymbol{\Omega}_h^{[\mathrm{RR}]}$, yielding a larger correcting factor and larger asymptotic variance of the diffusion parameter estimator.

Proposition 6.7 *Let $\widetilde{\mathbf{Z}}_{k,k-1}(\boldsymbol{\beta})$ and $\widetilde{\overline{\mathbf{Z}}}_{k+1,k,k-1}(\boldsymbol{\beta})$ be defined in* (24) *and* (35)*, respectively. Then,*

$$\begin{aligned}
\mathbf{Z}^{[\mathrm{S|R}]}_{k,k-1}(\boldsymbol{\beta}) &= -\frac{h^{3/2}}{2}\boldsymbol{\Sigma}_0(\boldsymbol{\eta}_{k-1} - 2\boldsymbol{\xi}'_{k-1}) + \frac{h^{5/2}}{12}(\mathbf{A}_{\mathbf{v}} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1}\\
&\quad + \frac{h^{5/2}}{4}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} - \frac{h^{5/2}}{2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0(\boldsymbol{\xi}'_{k-1} - \boldsymbol{\zeta}'_{k-1}) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}),\\
\overline{\mathbf{Z}}^{[\mathrm{S|R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= -\frac{h^{3/2}}{2}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} - \frac{h^2}{2}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \frac{h^{5/2}}{12}(\mathbf{A}_{\mathbf{v}} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}\\
&\quad + \frac{h^{5/2}}{4}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} - \frac{h^{5/2}}{4}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\mathbf{Q}_{k,k-1} + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

6.4 Objective functions

Starting with the complete observation case, we approximate objective functions (30) and (31) up to order $R(h^{3/2}, \mathbf{Y}_{t_{k-1}})$ to prove the asymptotic properties of the estimators $\hat{\boldsymbol{\theta}}_N^{[\mathrm{CR}]}$ and $\hat{\boldsymbol{\theta}}_N^{[\mathrm{CS|R}]}$. After omitting the terms of order $R(h, \mathbf{Y}_{t_{k-1}})$ that do not depend on $\boldsymbol{\beta}$, we obtain the following approximations:

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) &= (N-1)\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} \\ &+ 2\sqrt{h}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})) \\ &+ h\sum_{k=1}^{N}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})) \\ &- h\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} + h\sum_{k=1}^{N}\operatorname{Tr} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta}),\end{aligned} \tag{69}$$

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{CS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) &= (N-1)\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + 3\sum_{k=1}^{N}(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1}) \\ &- 3h\sum_{k=1}^{N}(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} \\ &- h\sum_{k=1}^{N}(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1}\end{aligned} \tag{70}$$

$$\mathcal{L}_N^{[\mathrm{CF}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = \mathcal{L}_N^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) + \mathcal{L}_N^{[\mathrm{CS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\,. \tag{71}$$

The two last sums in (70) converge to zero because $\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})\boldsymbol{\eta}_{k-1}^\top|\mathcal{F}_{t_{k-1}}] = \mathbf{0}$. Moreover, (70) lacks the quadratic form of $\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}_0(\mathbf{Y}_{t_{k-1}})$, that is crucial for the asymptotic variance of the drift estimator. This implies that the objective function $\mathcal{L}_N^{[\mathrm{CS|R}]}$ is not suitable for estimating the drift parameter. Conversely, (70) provides a correct and consistent estimator of the diffusion parameter, indicating that the full objective function (the sum of $\mathcal{L}_N^{[\mathrm{CR}]}$ and $\mathcal{L}_N^{[\mathrm{CS|R}]}$) consistently estimates $\boldsymbol{\theta}$.

Similarly, the approximated objective functions in the partial observation case are:

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{PR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) &= \frac{2}{3}(N-2)\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} \\ &+ 2\sqrt{h}\sum_{k=1}^{N}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})) \\ &+ h\sum_{k=1}^{N-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})) \\ &- h\sum_{k=1}^{N-1}(\mathbf{U}_{k,k-1}+2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} + h\sum_{k=1}^{N-1}\operatorname{Tr} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta}),\end{aligned} \tag{72}$$

$$\begin{aligned}\mathcal{L}_N^{[\mathrm{PS|R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) &= 2(N-2)\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + 3\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} \\ &+ 6\sqrt{h}\sum_{k=1}^{N}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0) \\ &- 3h\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} + 2h\sum_{k=1}^{N-1}\operatorname{Tr} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_k};\boldsymbol{\beta}),\end{aligned} \tag{73}$$

$$\mathcal{L}_N^{[\mathrm{PF}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right)=\mathcal{L}_N^{[\mathrm{PR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right)+\mathcal{L}_N^{[\mathrm{PS|R}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right). \tag{74}$$

This time, the term with $\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}$ vanishes because

$$\mathrm{Tr}(\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}))=0$$

due to the symmetry of the matrices and the trace cyclic property.

Even though the partial observation objective function $\mathcal{L}^{[\mathrm{PR}]}\left(\mathbf{X}_{0:t_N};\boldsymbol{\theta}\right)$ (40) depends only on $\mathbf{X}_{0:t_N}$, we could approximate it with $\mathcal{L}_N^{[\mathrm{PR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right)$ (72). This is useful for proving the asymptotic normality of the estimator since its asymptotic distribution will depend on the invariant probability ν_0 defined for the solution $\mathbf{Y}$.

The absence of the quadratic form $\mathbf{F}(\mathbf{Y}_{t_{k-1}})-\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})$ in (73) indicates that $\mathcal{L}_N^{[\mathrm{PS|R}]}$ is not suitable for estimating the drift parameter. Additionally, the penultimate term in (73) does not vanish, needing an additional correction term of $2h\sum_{k=1}^{N-1}\mathrm{Tr}\,D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_k};\boldsymbol{\beta})$ for consistency. This correction is represented as $4\log|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}|$ in (41). Notably, this term is absent in the complete objective function (31), making this adjustment somewhat artificial and could potentially deviate further from the true log-likelihood. Consequently, the objective function based on the full likelihood (39) inherits this characteristic from (73), suggesting that in the partial observation scenario, using only the rough likelihood (72) may be more appropriate.

7 Conclusion

Many fundamental laws of physics and chemistry are formulated as second-order differential equations, a model class important for understanding complex dynamical systems in various fields such as biology and economics. The extension of these deterministic models to stochastic second-order differential equations represents a natural generalization, allowing for the incorporation of uncertainties and variability inherent in real-world systems. However, robust statistical methods for analyzing data generated from such stochastic models have been lacking, presenting a significant challenge due to the inherent degeneracy of the noise and partial observation.

In this study, we propose estimating model parameters using a recently developed methodology of Strang splitting estimator for SDEs. This estimator has demonstrated finite sample efficiency with relatively large sample time steps, particularly in handling highly nonlinear models. We adjust the estimator to the partial observation setting and employ either the full likelihood or only the marginal likelihood based on the rough coordinates. For all four obtained estimators, we establish the consistency and asymptotic normality.

The application of the Strang estimator to a historical paleoclimate dataset obtained from ice cores in Greenland has yielded valuable insights and analytical tools for comprehending abrupt climate shifts throughout history. Specifically, we employed the stochastic Duffing oscillator, also known as the Kramers oscillator, to analyze the data.

While our focus in this paper has been primarily confined to second-order SDEs with no parameters in the smooth components, we are confident that our findings can be extended to encompass models featuring parameters in the drift of the smooth coordinates. This opens up directions for further exploration and application of our methodology to a broader range of complex dynamical systems, promising deeper insights into their behavior and underlying mechanisms.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)"; and Novo Nordisk Foundation NNF20OC0062958.

References

A. Abdulle, G. Vilmart, and K. C. Zygalakis. Long Time Accuracy of Lie–Trotter Splitting Methods for Langevin Dynamics. *SIAM Journal on Numerical Analysis*, 53(1):1–16, 2015.

M. Ableidinger, E. Buckwar, and H. Hinterleitner. A stochastic version of the jansen and rit neural mass model: Analysis and numerics. *Journal of Mathematical Neuroscience*, 7, 2017. ISSN 2190-8567.

D. Adams, M. H. Duong, and G. dos Reis. Operator-splitting schemes for degenerate, non-local, conservative-dissipative systems. *Discrete and Continuous Dynamical Systems*, 42(11):5453–5486, 2022.

L. A. Alyushina. Euler Polygonal Lines for Itô Equations with Monotone Coefficients. *Theory of Probability & Its Applications*, 32(2):340–345, 1988.

K. Andersen, N. Azuma, J. Barnola, M. Bigler, P. Biscaye, N. Caillon, J. Chappellaz, H. Clausen, D. Dahl-Jensen, H. Fischer, J. Flückiger, D. Fritzsche, Y. Fujii, K. Goto-Azuma, K. Grønvold, N. Gundestrup, M. Hansson, C. Huber, C. Hvidberg, and J. White. High-resolution record of northern hemisphere climate extending into the last interglacial period. *Nature*, 431:147–51, 10 2004.

M. Anklin, J. M. Barnola, J. Beer, T. Blunier, J. Chappellaz, H. B. Clausen, D. Dahljensen, W. Dansgaard, M. Deangelis, R. Delmas, P. Duval, M. Fratta, A. Fuchs, K. Fuhrer, N. Gundestrup, C. Hammer, P. Iversen, S. Johnsen, J. Jouzel, and E. W. Wolff. Climate instability during the last interglacial period recorded in the grip ice core. *Nature*, 364: 203–207, 07 1993.

L. Arnold and P. Imkeller. The Kramers Oscillator Revisited. In J. A. Freund and T. Pöschel, editors, *Stochastic Processes in Physics, Chemistry, and Biology*, pages 280–291. Springer Berlin Heidelberg, 2000.

N. Boers. Early-warning signals for Dansgaard-Oeschger events in a high-resolution ice core record. *Nature Communications*, 9, 07 2018.

N. Boers, M. D. Chekroun, H. Liu, D. Kondrashov, D.-D. Rousseau, A. Svensson, M. Bigler, and M. Ghil. Inverse stochastic–dynamic models for high-resolution Greenland ice core records. *Earth System Dynamics*, 8(4):1171–1190, 2017.

N. Boers, M. Ghil, and D.-D. Rousseau. Ocean circulation, ice shelf, and sea ice interactions explain Dansgaard–Oeschger cycles. *Proceedings of the National Academy of Sciences*, 115:E11005–E11014, 11 2018.

G. W. Bohrnstedt and A. S. Goldberger. On the Exact Covariance of Products of Random Variables. *Journal of the American Statistical Association*, 64(328):1439–1442, 1969.

N. Bou-Rabee. Cayley splitting for second-order langevin stochastic partial differential equations. *arXiv: Probability*, 2017. URL `https://api.semanticscholar.org/CorpusID:119132768`.

N. Bou-Rabee and H. Owhadi. Long-Run Accuracy of Variational Integrators in the Stochastic Context. *SIAM Journal on Numerical Analysis*, 48(1):278–297, Jan. 2010.

C.-E. Bréhier and L. Goudenège. Analysis of some splitting schemes for the stochastic Allen-Cahn equation. *Discrete and Continuous Dynamical Systems - B*, 24(8):4169–4190, 2019.

E. Buckwar, A. Samson, M. Tamborrino, and I. Tubikanec. A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *Applied Numerical Mathematics*, 179:191–220, 2022.

K. Burrage, I. Lenane, and G. Lythe. Numerical Methods for Second-Order Stochastic Differential Equations. *SIAM Journal on Scientific Computing*, 29(1):245–264, 2007.

I. Crimaldi and L. Pratelli. Convergence results for multivariate martingales. *Stochastic Processes and their Applications*, 115(4):571–577, 2005.

W. Dansgaard, S. Johnsen, and H. e. a. Clausen. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, 364:218–220, 1993.

P. D. Ditlevsen, S. Ditlevsen, and K. K. Andersen. The fast climate fluctuations during the stadial and interstadial climate states. *Annals of Glaciology*, 35:457–462, 2002.

P. D. Ditlevsen, K. K. Andersen, and A. Svensson. The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle. *Climate of the Past*, 3(1):129–134, 2007.

S. Ditlevsen and A. Samson. Hypoelliptic diffusions: filtering and inference from complete and partial observations. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 81(2):361–384, 2019.

S. Ditlevsen and M. Sørensen. Inference for observations of integrated diffusion processes. *Scandinavian Journal of Statistics*, 31(3):417–429, 2004.

G. Duffing. *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. Vieweg, 1918.

D. Falbel and J. Luraschi. *torch: Tensors and Neural Networks with 'GPU' Acceleration*, 2022.

V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'I.H.P. Probabilités et statistiques*, 29(1):119–151, 1993.

A. Gloter. Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM: Probability and Statistics*, 4:205–227, 2000.

A. Gloter. Parameter Estimation for a Discretely Observed Integrated Diffusion Process. *Scandinavian Journal of Statistics*, 33(1):83–104, 2006.

A. Gloter and N. Yoshida. Adaptive and non-adaptive estimation for degenerate diffusion processes, 2020.

A. Gloter and N. Yoshida. Adaptive estimation for degenerate diffusion processes. *Electronic Journal of Statistics*, 15 (1):1424 – 1472, 2021.

E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg, 1993.

F. Hassanibesheli, N. Boers, and J. Kurths. Reconstructing complex system dynamics from time series: a method comparison. *New journal of physics*, 2020.

A. R. Humphries and A. M. Stuart. *Deterministic and random dynamical systems: theory and numerics*, pages 211–254. Springer Netherlands, Dordrecht, 2002.

Y. Iguchi and A. Beskos. Parameter inference for hypo-elliptic diffusions under a weak design condition, 2023.

Y. Iguchi, A. Beskos, and M. Graham. Parameter inference for degenerate diffusion processes, 2023a.

Y. Iguchi, A. Beskos, and M. M. Graham. Parameter Estimation with Increased Precision for Elliptic and Hypo-elliptic Diffusions, 2023b.

B. H. Jansen and V. G. Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366, 1995.

D. S. Johnson, J. M. London, M.-A. Lea, and J. W. Durban. Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5):1208–1215, 2008.

M. Kessler. Estimation of an Ergodic Diffusion from Discrete Observations. *Scandinavian Journal of Statistics*, 24(2): 211–229, 1997.

P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1992.

H. Korsch and H. Jodl. *Chaos: A Program Collection for the PC*. Springer, 1999.

H. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304, 1940.

N. V. Krylov. A Simple Proof of the Existence of a Solution of Itô's Equation with Monotone Coefficients. *Theory of Probability & Its Applications*, 35(3):583–587, 1991.

B. Leimkuhler and C. Matthews. Molecular dynamics. *Interdisciplinary applied mathematics*, 36, 2015.

J. Lohmann and P. Ditlevsen. A consistent statistical model selection for abrupt glacial climate changes. *Climate Dynamics*, 52, 06 2019.

X. Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

S. Melchionna. Design of quasisymplectic propagators for Langevin dynamics. *The Journal of Chemical Physics*, 127 (4):044108, 07 2007.

A. Melnykova. Parametric inference for hypoelliptic ergodic diffusions with full observations, 2020.

T. Michelot and P. G. Blackwell. State-switching continuous-time correlated random walks. *Methods in Ecology and Evolution*, 10(5):637–649, 2019.

D. Nualart. *The Malliavin Calculus and Related Topics*. Probability and Its Applications. Springer Berlin Heidelberg, 2006. ISBN 9783540283294.

T. Ozaki. Statistical Identification of Storage Models with Application to Stochastic Hydrology. *Journal of The American Water Resources Association*, 21:663–675, 1985.

T. Ozaki, J. C. Jimenez, and V. Haggan-Ozaki. The Role of the Likelihood Function in the Estimation of Chaos Models. *Journal of Time Series Analysis*, 21(4):363–387, 2000.

G. Pavliotis, A. Stuart, and K. Zygalakis. Calculating effective diffusivities in the limit of vanishing molecular diffusion. *Journal of Computational Physics*, 228(4):1030–1055, 2009.

P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *arXiv preprint arXiv:2211.11884*, 2024. To appear in The Annals of Statistics.

Y. Pokern, A. M. Stuart, and P. Wiberg. Parameter Estimation for Partially Observed Hypoelliptic Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(1):49–73, 01 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

S. O. Rasmussen, M. Bigler, S. P. Blockley, T. Blunier, S. L. Buchardt, H. B. Clausen, I. Cvijanovic, D. Dahl-Jensen, S. J. Johnsen, H. Fischer, V. Gkinis, M. Guillevic, W. Z. Hoek, J. J. Lowe, J. B. Pedro, T. Popp, I. K. Seierstad, J. P. Steffensen, A. M. Svensson, P. Vallelonga, B. M. Vinther, M. J. Walker, J. J. Wheatley, and M. Winstrup. A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: refining and extending the INTIMATE event stratigraphy. *Quaternary Science Reviews*, 106:14–28, 2014.

M. Rosenblum and A. Pikovsky. Synchronization: from pendulum clocks to chaotic lasers and chemical oscillators. *Contemporary Physics*, 44(5):401–416, 2003.

A. Samson and M. Thieullen. Contrast estimator for completely or partially observed hypoelliptic diffusion. *Stochastic Processes and their Applications*, 122(7):2521–2552, 2012.

I. Seierstad, P. Abbott, M. Bigler, T. Blunier, A. Bourne, E. Brook, S. L. Buchardt, C. Buizert, H. Clausen, E. Cook, D. Dahl-Jensen, S. Davies, M. Guillevic, S. Johnsen, D. Pedersen, T. Popp, S. Rasmussen, J. Severinghaus, A. Svensson, and B. Vinther. Consistently dated records from the Greenland GRIP, GISP2 and NGRIP ice cores for the past 104 ka reveal regional millennial-scale δ18O gradients with possible Heinrich event imprint. *Quaternary Science Reviews*, 106, 11 2014.

M. Serrano, G. De Fabritiis, P. Español, and P. Coveney. A stochastic Trotter integration scheme for dissipative particle dynamics. *Mathematics and Computers in Simulation*, 72(2):190–194, 2006.

T. Shardlow. Splitting for dissipative particle dynamics. *SIAM Journal on Scientific Computing*, 24(4):1267–1282, Dec. 2003.

I. Shoji and T. Ozaki. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4):733–752, 1998.

M. Sørensen and M. Uchida. Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli*, 9 (6):1051 – 1069, 2003. ISSN 1350-7265.

M. V. Tretyakov and Z. Zhang. A Fundamental Mean-Square Convergence Theorem for SDEs with Locally Lipschitz Coefficients and Its Applications. *SIAM Journal on Numerical Analysis*, 51(6):3135–3162, 2013.

E. Vanden-Eijnden and G. Ciccotti. Second-order integrators for Langevin equations with holonomic constraints. *Chemical Physics Letters - CHEM PHYS LETT*, 429, 09 2006.

L. Wu. Large and moderate deviations and exponential convergence for stochastic damping hamiltonian systems. *Stochastic Processes and their Applications*, 91(2):205–238, 2001.

N. Yoshida. Asymptotic behavior of M-estimator and related random field for diffusion process. *Annals of the Institute of Statistical Mathematics*, 42(2):221–251, June 1990.

I. Ziv, D. Baxter, and J. Byrne. Simulator for neural networks and action-potentials - description and application. *Journal of Neurophysiology*, 71(1):294–308, 1994.

III Parameter Estimation in Nonlinear Multivariate SDEs with Pearson-type Noise

This chapter contains the following paper:

- P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting parameter estimator for nonlinear multivariate stochastic differential equations with Pearson-type multiplicative noise, 2024.

  Paper status: working paper.

This paper introduces a new model class called multivariate Pearson diffusions. This class features a linear drift and a quadratic function of the state vector in the squared diffusion matrix. It can be viewed as a generalization of univariate Pearson diffusion, where the noise structure allows for the explicit derivation of the first two moments. Additionally, it can be seen as a generalization of a multivariate affine diffusion, which is defined with a linear drift and a linear function of the state vector in the squared diffusion matrix. We derive a closed-form expression for the mean and covariance matrix for multivariate Pearson diffusions using a theorem for computing integrals involving matrix exponentials (Theorem 2).

We also propose a splitting scheme for a nonlinear process with Pearson-type noise—a process that shares the same diffusion structure as a multivariate Pearson diffusion but has a nonlinear drift. We suggest splitting this nonlinear drift into linear and nonlinear components and then obtaining a pseudo-likelihood from the Strang splitting scheme. Unlike in the previous two papers, the linear sub-SDE in this context does not yield an Ornstein-Uhlenbeck (OU) process due to the presence of multiplicative noise. To solve for the multivariate Pearson diffusion, we recommend approximating the transition density as Gaussian, with the correct first two moments.

The paper also discusses two existing models from the literature that fit within the framework of nonlinear multivariate SDEs with Pearson-type noise. The first model is a coupled multivariate Wright-Fisher diffusion, used in genetic research to describe allele frequencies across multiple loci. The second model is the stochastic SIR model, which describes the spread of disease within a population. This model can be considered a special case of a generalized Lotka-Volterra model, which also falls under the category of nonlinear multivariate SDEs with Pearson-type noise.

Finally, we introduce a new model called the student Kramers oscillator, a generalization of the Kramers oscillator. We prove the existence and uniqueness of the solution to

the governing SDE and use this model in a simulation study to illustrate the performance of the Strang splitting estimator. The study demonstrates that the new estimator provides more accurate estimates of diffusion parameters than any other method evaluated.

STRANG SPLITTING PARAMETER ESTIMATOR FOR NONLINEAR MULTIVARIATE STOCHASTIC DIFFERENTIAL EQUATIONS WITH PEARSON-TYPE MULTIPLICATIVE NOISE

A PREPRINT

Predrag Pilipovic
Department of Mathematical Sciences
University of Copenhagen
2100 Copenhagen, Denmark
predrag@math.ku.dk
Bielefeld Graduate School of Economics and Management
University of Bielefeld
33501 Bielefeld, Germany
predrag.pilipovic@uni-bielefeld.de

Adeline Samson
Univ. Grenoble Alpes
CNRS, Grenoble INP, LJK
38000 Grenoble, France
adeline.leclercq-samson@univ-grenoble-alpes.fr

Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen
2100 Copenhagen, Denmark
susanne@math.ku.dk

ABSTRACT

This paper extends the one-dimensional Pearson diffusion framework to multivariate models where the squared diffusion coefficient is a quadratic function of the state $\mathbf{X}_t$. It generalizes the multivariate affine diffusion models, allowing for explicit computation of first and second moments despite the unknown transition densities. We further allow the drift to be nonlinear, thus introducing multivariate nonlinear diffusions with Pearson-type multiplicative noise. For example, we propose the Student Kramers oscillator, which is the Kramers oscillator with a Pearson-type noise, which has the student's t-distribution as an invariant distribution, allowing for heavier tails than the usual additive noise. We prove the existence and uniqueness of Student Kramers oscillator and the existence of its invariant measure. We propose a novel approach for parameter estimation based on Strang splitting combined with Gaussian transition density approximation. We start by splitting the nonlinear diffusion into a linear multivariate Pearson diffusion and a nonlinear ordinary differential equation (ODE). We solve the nonlinear ODE and approximate the flow of the multivariate Pearson diffusion using Gaussian approximation, where the first two moments are exact. The Strang splitting approximation and the corresponding estimator are obtained by composing the solutions of the split subsystems. A case study using the Student Kramers oscillator demonstrates that our estimator performs comparably to the Euler-Maruyama, Kessler, and local linearization estimators. Specifically, the results are similar for drift parameters, while the proposed estimator outperforms the others for diffusion parameters.

Keywords Gaussian approximation, Multivariate Pearson diffusion, Nonlinear drift, Strang estimator, Strang splitting scheme

1 Introduction

Pearson diffusions, introduced as a versatile class of tractable one-dimensional diffusion models, have found extensive use in various fields due to their rich statistical properties and ease of parameter estimation [Forman and Sørensen,

2008]. These models are particularly advantageous because moments and conditional moments are explicitly available, making them an attractive choice for statistical inference. For instance, the Ornstein-Uhlenbeck (OU) and the Cox-Ingersoll-Ross (CIR) processes, special cases of Pearson diffusions, are often used in practical applications due to their mathematical tractability. Pearson diffusions, with their mean-reverting linear drift and diffusion coefficient defined as a second-order polynomial, encompass a wide range of stationary distributions, namely the family of Pearson distributions, including light-tailed and heavy-tailed distributions and with different state spaces.

The estimation of parameters in Pearson diffusions is facilitated by the use of optimal martingale estimating functions using exact moments, found by use of eigenfunctions of the infinitesimal generator [Bibby and Sørensen, 1995, Kessler and Sørensen, 1999, Forman and Sørensen, 2008]. This method simplifies the estimation process and ensures consistent estimators in low-frequency sampling scenarios, often encountered in empirical data. Other estimation methods, such as the generalized method of moments, quasi-likelihood, and non-linear weighted least squares, are also applicable to Pearson diffusions, highlighting their flexibility and ease of implementation.

Despite the tractability and versatility of Pearson diffusions, there has been limited work on extending these models to higher dimensions. Leonenko and Phillips [2012] employed a spectral high-order approximation of the Fokker–Planck equations for Pearson diffusions. They suggested extending their method to the multivariate case defined by a quadratic form without a linear term and intercept, similar to equation (3), but did not pursue this further. Thus, a significant gap remains in the literature concerning the generalization of these models to multivariate settings. Only the OU process is straightforwardly extended to arbitrary dimensions, following a multivariate Gaussian distribution. A generalization is desired to model more complex dynamical systems with diverse types of stationary distributions shaped by the diffusion function.

A Pearson diffusion [Forman and Sørensen, 2008] is a solution to a one-dimensional stochastic differential equation (SDE) of the form

$$\mathrm{d}X_t = a(X_t - b)\,\mathrm{d}t + \sqrt{\alpha X_t^2 + \beta X_t + \gamma}\,\mathrm{d}W_t, \tag{1}$$

where $a < 0$, and α, β, and γ are such that the square root is well-defined when X_t is in the state space. The parameters of (1) are $\boldsymbol{\theta} = \{a, b, \alpha, \beta, \gamma\}$, where $-a > 0$ determines the speed of mean reversion, b is the mean of the invariant distribution, and α, β, and γ shape the state space and the invariant distribution.

Pearson diffusions can be classified into six cases based on the form of the squared diffusion coefficient $\sigma^2(x) = \alpha x^2 + \beta x + \gamma$. Each case presents specific conditions for the existence and uniqueness of ergodic solutions and corresponding invariant distributions. This classification is based on equivalence classes because the Pearson class of diffusions is closed under translations and scale transformations. Specifically, if $(X_t)_{t\geq 0}$ is an ergodic Pearson diffusion, so is $(\tilde{X}_t)_{t\geq 0}$ where $\tilde{X}_t = \psi X_t + \phi$. Up to translation and scale transformations, the ergodic Pearson diffusions are classified into six forms [Forman and Sørensen, 2008].

- For $\sigma^2(x) = \gamma$, it is an OU process defined on the entire real line. The unique ergodic solution exists for all $b \in \mathbb{R}$, with the invariant distribution being normal with mean b and variance $\gamma/(-2a)$.
- For $\sigma^2(x) = \beta x$, it is a CIR process, also called a square-root process. The process is defined on the positive half-line $(0, \infty)$. A unique ergodic solution exists for $-2ab \geq \beta$. The invariant distribution is the gamma distribution with scale parameter $\beta/(-2a)$ and shape parameter $-2ab/\beta$. For $0 < -2ab < \beta$, the boundary at zero can be reached with positive probability, but with an instantaneous reflecting boundary, the process remains stationary with the same gamma invariant distribution.
- For $\sigma^2(x) = \alpha x^2$, it is a geometric Brownian motion type process, also called a GARCH diffusion. It is defined on the positive half-line $(0, \infty)$. A unique ergodic solution exists for all $\alpha > 0$ and $b > 0$. The invariant distribution is an inverse gamma distribution with shape parameter $1 - 2a/\alpha$ and scale parameter $\alpha/(-2ab)$. The variance only exists for $\alpha < -2a$.
- For $\sigma^2(x) = \alpha(x^2 + 1)$, the diffusion is defined on the entire real line. A unique ergodic solution exists for all $\alpha > 0$ and $b \in \mathbb{R}$. If $b = 0$, the invariant distribution is a scaled Student's t-distribution with $1 - 2a/\alpha$ degrees of freedom. For $b \neq 0$, the invariant distribution is a skew t-distribution.
- For $\sigma^2(x) = \alpha x(x + 1)$, the process is defined on the positive half-line $(0, \infty)$. A unique ergodic solution exists for all $\alpha > 0$ and $b \geq \alpha/(-2a)$. The invariant distribution is a scaled F-distribution with $-4ab/\alpha$ and $2(1 - 2a/\alpha)$ degrees of freedom. If $0 < b < \alpha/(-2a)$, the boundary at zero can be reached, but with an instantaneous reflecting boundary, the process remains stationary with the same F-distribution.
- For $\sigma^2(x) = \alpha x(x - 1)$, the process is a Jacobi diffusion, defined on the interval $(0, 1)$. A unique ergodic solution exists for $\alpha < 0$ and b such that $\min(b, 1 - b) \geq \alpha/(2a)$. The invariant distribution is a Beta distribution with shape parameters $2ab/\alpha$ and $2a(1 - b)/\alpha$. If $0 < b < \alpha/(2a)$, the boundary at zero can be reached, and similar remarks apply to the boundary at one when $0 < 1 - b < \alpha/(2a)$.

This classification illustrates the diverse applications of Pearson diffusions, allowing for explicit computation of invariant distributions and parameter conditions across different state spaces and diffusion forms. An important feature of Pearson diffusions is the ability to find explicit expressions for the marginal and conditional moments. The k-th absolute moment, $\mathbb{E}|X_t|^k$, is finite if and only if $\alpha/(-2a) < 1/(k-1)$. This implies that all moments exist if $\alpha \leq 0$. However, for $\alpha > 0$, only moments satisfying $k < -2a/\alpha + 1$ exist [Forman and Sørensen, 2008].

Generalizing Pearson diffusions to multivariate settings retains univariate Pearson diffusions' tractability and statistical convenience, such as deriving explicit forms for the first two moments and conditional moments. This generalization allows for a specific Pearson-type invariant distribution for each coordinate of a high-dimensional process. For example, a process could be defined in a d-dimensional hypercube $[0,1]^d$, with each coordinate having a Beta invariant distribution. Alternatively, different combinations of Pearson-type noises could appear in various coordinates of the quadratic diffusion matrix. Moreover, the quadratic diffusion matrix may be hypoelliptic, meaning it can be singular while the diffusion still admits a smooth density. Furthermore, allowing the drift to be nonlinear opens new directions for modeling complex systems in fields such as finance, biology, and physics, capturing dependencies and interactions between multiple variables and providing a more realistic representation of underlying processes.

The motivation for this study is to develop efficient and accurate methods for parameter estimation in nonlinear SDEs with Pearson-type multiplicative noise. Traditional estimation techniques often fall short in these scenarios due to the complexity of the noise structure and nonlinearity in the drift function. Our goal is to create a robust framework for parameter estimation that can be applied to a wide range of models, including hypoelliptic diffusions. Recent advances have highlighted the potential of pseudo-likelihood approaches, which approximate the likelihood, simplifying the estimation process while maintaining accuracy. By building on these advances, we aim to provide a more effective and versatile method for parameter estimation in complex systems.

Parameter estimation for SDEs with non-constant or multiplicative noise has been extensively studied. These models are characterized by the complexity added by the noise term, which depends on the state of the process itself. This dependency complicates the estimation process and has led to the development of various specialized methods. Much work has focused on methods that transform the SDE into another SDE with additive noise. For a review of likelihood-based parameter estimation of SDEs with additive noise, see [Pilipovic et al., 2024a,b] and references therein. These methods can be applied if the Lamperti transform is available.

Martingale estimating functions have emerged as a powerful tool for parameter estimation in SDEs. They exploit the underlying process's martingale property to construct consistent and asymptotically normal estimators. Kessler and Sørensen [1999] introduced optimal martingale estimating functions based on eigenfunctions of the generator for Pearson diffusions. This method simplifies estimation and ensures high efficiency, particularly in high-frequency sampling scenarios often encountered in financial data. Sørensen [2008] further demonstrated that these optimal estimating functions compare to maximum likelihood estimation under high-frequency asymptotics, offering a simpler alternative.

Quasi-likelihood methods, which approximate the likelihood using a tractable form, have also been applied to SDEs with non-constant noise. Kessler [1997] proposed an estimator, often referred to as the Kessler (K) estimator, which approximates the unknown transition density of a diffusion process with a Gaussian density. This approximation is achieved using the true conditional mean and covariance, or approximations derived from the infinitesimal generator of the diffusion process.

Building on Kessler's foundational work, Uchida and Yoshida [2012] extended the Kessler estimator to the setting of multivariate elliptic diffusions. They developed an adaptive-type contrast estimator that also achieves a central limit theorem under the same design condition proposed by Kessler, specifically $Nh^p \to 0$, for $p \geq 2$.

More recently, Iguchi and Beskos [2023] further refined the conditions under which the K estimator achieves asymptotic normality. They focused on hypoelliptic SDEs, which pose additional challenges due to the degeneracy of the diffusion coefficient. They proposed a modified estimator that remains consistent and asymptotically normal under a weaker design condition, specifically $Nh^p \to 0$ for $p \geq 2$. This advancement addresses some of the limitations of previous methods and enhances the robustness of the K-type estimators in more general settings.

Gloter [2006] developed a contrast function using the Euler-Maruyama (EM) discretization for integrated diffusion processes, focusing on the asymptotic properties as the sampling interval approaches zero and the sample size approaches infinity. By addressing the ill-conditioned nature of the contrast from the EM discretization, Gloter [2006] suggested using rough equations of the SDE and recovering the unobserved components through finite difference approximations. However, this approach introduced bias and required correction factors, affecting the estimator's variance.

Like Kessler [1997], Hurn et al. [2013] developed a quasi-maximum likelihood procedure for parameter estimation in multi-dimensional diffusions, where the transition density is approximated by a multivariate Gaussian distribution.

Unlike Kessler [1997], Hurn et al. [2013] focused specifically on systems with affine drift and diffusion functions, where both the drift and quadratic diffusion matrix are linear functions of the state variable. For such affine models, they derive closed-form expressions for the first and second moments of the true transition density, which are exact due to the affine structure. This explicit calculation of moments enables them to construct a Gaussian approximation that provides consistent parameter estimates even when the true transition density is misspecified. This approach ensures robustness in parameter estimation. For non-affine models, Hurn et al. [2013] showed that numerical methods can still accurately compute the required moments of the transition distribution, ensuring that Gaussian approximations achieve high computational precision in integral evaluations.

Recent advances in this field have continued to build on these foundational methods. For instance, Gloter and Yoshida [2020, 2021] introduced adaptive and non-adaptive methods for hypoelliptic diffusion models, demonstrating asymptotic normality in complete observation regimes. They used higher-order Itô-Taylor expansions to introduce additional Gaussian noise into the smooth coordinates, accompanied by higher-order mean approximations for the rough coordinates. These contributions have refined the conditions for the estimators' asymptotic normality, extending these methods' applicability to a broader class of models.

In this paper, we extend the one-dimensional Pearson diffusion to a multivariate setting, where the quadratic diffusion matrix is a quadratic function of the state variable. This generalization builds on the concept of affine diffusions as used in [Hurn et al., 2013], where first and second moments can be explicitly computed despite the unknown transition density. While Hurn et al. [2013] focused on the linear case, encompassing certain types of Pearson diffusions such as the OU and CIR processes, our approach fully generalizes to all Pearson diffusions. We further extend this framework from linear multivariate Pearson diffusions to nonlinear drifts. To illustrate our approach, we define the Student Kramers oscillator and prove that its solution exists, is unique, and possesses an invariant measure. We propose a novel method for parameter estimation in nonlinear multivariate Pearson diffusions that combines Strang splitting with Gaussian transition density approximation. Specifically, we split the nonlinear drift into a multivariate Pearson diffusion and a nonlinear ODE, then approximate the transition density of the linear component as Gaussian with the exact mean and covariance matrix. We solve the nonlinear ODE and compose the solutions of the split subsystems to achieve the Strang splitting approximation, upon which our Strang (S) estimator is based. We demonstrate its performance through a detailed study of the Student Kramers oscillator.

The main contributions of this paper are:

1. We propose a new class of models based on multivariate Pearson diffusions, extending the univariate Pearson diffusions to more complex systems with tractable properties.
2. We explicitly compute the first and second moments for the proposed multivariate Pearson diffusion models, ensuring accurate and efficient parameter estimation.
3. We introduce the Student Kramers oscillator as a specific example of a nonlinear hypoelliptic multivariate Pearson diffusion, demonstrating that its solution exists, is unique and admits an invariant density. This example is also used for the simulation study.
4. We develop a new parameter estimation method based on a splitting scheme for nonlinear SDEs with a multivariate Pearson-type diffusion matrix. Specifically, we split the nonlinear drift into a multivariate Pearson diffusion and a nonlinear ODE. The solution of multivariate Pearson diffusion is approximated by Gaussian transition density with the exact first two moments.
5. We conduct a simulation study including three other methods where our estimator performs similarly to the others for the drift parameters but outperforms them for the diffusion parameters. The local linearization (LL) estimator is designed only for additive noise SDEs. Nevertheless, we implemented it using the Lamperti transform.

The structure of the paper is as follows. Section 2 presents the problem setup, illustrated with the coupled Wright-Fisher diffusion, the stochastic SIR model, and the newly introduced Student Kramers oscillator. Section 3 introduces the class of multivariate Pearson diffusion models and their theoretical properties and explicitly computes the first and second moments. Section 4 describes the new parameter estimation method and briefly recalls the EM, K, and LL estimators, including applying the Kessler method for Gaussian approximation. In Section 5, we conduct a simulation study to evaluate the performance of the new estimator against existing methods. Finally, Section 6 provides concluding remarks and potential directions for future research.

Notation. We use capital bold letters for random vectors, vector-valued functions, and matrices, while lowercase bold letters denote deterministic vectors. $\|\cdot\|$ denotes both the L^2 vector norm in $\mathbb{R}^d$. Superscript (i) on a vector denotes the i-th component. Double subscript ij on a matrix denotes the component in the i-th row and j-th column. The transpose is denoted by $\top$. Operator $\mathrm{Tr}(\cdot)$ returns the trace, $\det(\cdot)$ the determinant and vec vectorization of a matrix. The

Kronecker product and the sum of two matrices are $\otimes$ and $\oplus$, respectively. $\mathbf{I}_d$ denotes the d-dimensional identity matrix, while $\mathbf{0}_{d\times d}$ is a d-dimensional zero square matrix. We denote by $[a_i]_{i=1}^d$ a vector with coordinates a_i, and by $[b_{ij}]_{i,j=1}^d$ a matrix with coordinates b_{ij}, for $i,j=1,\dots,d$. For a real-valued function $g:\mathbb{R}^d\to\mathbb{R}$, $\partial_{x^{(i)}}g(\mathbf{x})$ denotes the partial derivative with respect to $x^{(i)}$ and $\partial^2_{x^{(i)}x^{(j)}}g(\mathbf{x})$ denotes the second partial derivative with respect to $x^{(i)}$ and $x^{(j)}$. The nabla operator $\nabla_{\mathbf{x}}$ denotes the gradient vector of g with respect of $\mathbf{x}$, that is, $\nabla_{\mathbf{x}}g(\mathbf{x})=[\partial_{x^{(i)}}g(\mathbf{x})]_{i=1}^d$. For a vector-valued function $\mathbf{F}:\mathbb{R}^d\to\mathbb{R}^d$, the differential operator $D_{\mathbf{x}}$ denotes the Jacobian matrix $D_{\mathbf{x}}\mathbf{F}(\mathbf{x})=[\partial_{x^{(i)}}F^{(j)}(\mathbf{x})]_{i,j=1}^d$. Let $\mathbf{R}$ represent a vector (or a matrix) valued function defined on $(0,1)\times\mathbb{R}^d$ (or $(0,1)\times\mathbb{R}^{d\times d}$), such that, for some constant C, $\|\mathbf{R}(a,\mathbf{x})\|<aC(1+\|\mathbf{x}\|)^C$ for all $a,\mathbf{x}$. When denoted by R, it refers to a scalar function. For an open set A, the bar $\overline{A}$ indicates closure. δ_{ij} is the Kronecker delta function that equals one for $i=j$ and zero otherwise. The indicator function is denoted as $\mathbb{1}$.

2 Problem setup

We assume that the following SDE

$$\mathrm{d}\mathbf{X}_t=\mathbf{F}(\mathbf{X}_t;\boldsymbol{\theta}^{(1)})\,\mathrm{d}t+\boldsymbol{\Sigma}(\mathbf{X}_t;\boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t \tag{2}$$

has a unique strong solution $\mathbf{X}_t\in\mathcal{X}\subset\mathbb{R}^d$ defined on a complete probability space $(\Omega,\mathcal{F},\mathbb{P}_{\boldsymbol{\theta}})$ with a complete right-continuous filtration $(\mathcal{F}_t)_{t\geq 0}$, where $\mathbf{W}=(\mathbf{W}_t)_{t\geq 0}$ is a d-dimensional Wiener process adapted to $\mathcal{F}_t$. The probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by the parameter $\boldsymbol{\theta}=(\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)})$. Moreover, we assume that the squared diffusion function $\boldsymbol{\Sigma}(\mathbf{X}_t;\boldsymbol{\theta}^{(2)})$ has the following form

$$[\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})^\top]_{ij}=\mathbf{x}^\top\boldsymbol{\alpha}^{ij}\mathbf{x}+\mathbf{x}^\top\boldsymbol{\beta}^{ij}+\gamma^{ij},\qquad i,j=1,2,...,d, \tag{3}$$

where $\boldsymbol{\alpha}^{ij}\in\mathbb{R}^{d\times d},\boldsymbol{\beta}^{ij}\in\mathbb{R}^d,\gamma^{ij}\in\mathbb{R}$ are known or unknown parameters of the diffusion function such that $\boldsymbol{\alpha}^{ij}$ are symmetric, and $\boldsymbol{\alpha}^{ij}=\boldsymbol{\alpha}^{ji}$, $\boldsymbol{\beta}^{ij}=\boldsymbol{\beta}^{ji}$, $\gamma^{ij}=\gamma^{ji}$, for all $i,j=1,2,\dots,d$. The following section provides three examples of SDEs that fit this setup.

Remark 1 *Due to symmetry, each matrix $\boldsymbol{\alpha}^{ij}$ has $d(d+1)/2$ parameters. Therefore, $[\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})^\top]_{ij}$ has $(d+1)(d+2)/2$ parameters. Since also $\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})^\top$ is symmetric, it has $d(d+1)^2(d+2)/4$ parameters, which is of order d^4. This setup thus assumes an over-parameterized quadratic diffusion matrix where each component depends on all quadratic combinations of the state $\mathbf{X}_t$. In most applications, the quadratic diffusion matrix will be diagonal or nearly diagonal, as will $\boldsymbol{\alpha}^{ij}$, where most quadratic interactions between the state $\mathbf{X}_t$ are unnecessary. Moreover, most diffusion parameters will be constants, i.e., known parameters. Consequently, $\boldsymbol{\alpha}^{ij}$, $\boldsymbol{\beta}^{ij}$, and γ^{ij} will generally be sparse with few non-zero elements or with a few unknown parameters to be estimated. However, we do not impose sparsity conditions to maximize generality and flexibility and work under an over-parameterized setup.*

Rewrite (2) as follows

$$\mathrm{d}\mathbf{X}_t=\mathbf{A}(\boldsymbol{\theta}^{(1)})(\mathbf{X}_t-\mathbf{b}(\boldsymbol{\theta}^{(1)}))\,\mathrm{d}t+\mathbf{N}(\mathbf{X}_t;\boldsymbol{\theta}^{(1)})\,\mathrm{d}t+\boldsymbol{\Sigma}(\mathbf{X}_t,\boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t,\qquad \mathbf{X}_0=\mathbf{x}_0, \tag{4}$$

such that $\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})=\mathbf{A}(\boldsymbol{\theta}^{(1)})(\mathbf{x}-\mathbf{b}(\boldsymbol{\theta}^{(1)}))+\mathbf{N}\left(\mathbf{x};\boldsymbol{\theta}^{(1)}\right)$. Let $\overline{\Theta}=\overline{\Theta}_{\theta_1}\times\overline{\Theta}_{\theta_2}$ be the parameter space with Θ_{θ_1} and Θ_{θ_2} being two open convex bounded subsets of $\mathbb{R}^r$ and $\mathbb{R}^{d\times d}$, respectively.

Functions $\mathbf{F},\mathbf{N}:\mathbb{R}^d\times\overline{\Theta}_{\theta_1}\to\mathbb{R}^d$ are assumed locally Lipschitz, and $\mathbf{A}$, $\mathbf{b}$ are defined on $\overline{\Theta}_{\theta_1}$ and take values in $\mathbb{R}^{d\times d}$ and $\mathbb{R}^d$, respectively. Matrix $\boldsymbol{\Sigma}$ takes values in $\mathbb{R}^{d\times m}$. The matrix function $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top:\mathbb{R}^d\times\overline{\Theta}_{\theta_2}\to\mathbb{R}^{d\times d}$ is assumed to be positive semidefinite. From here after, we write $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x};\boldsymbol{\theta}^{(2)})$ instead of $\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}^{(2)})^\top$. Since any square root of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x};\boldsymbol{\theta}^{(2)})$ induces the same distribution, $\boldsymbol{\Sigma}(\mathbf{x})$ is only identifiable up to equivalence classes. Thus, we work only with $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x};\boldsymbol{\theta}^{(2)})$ and from now on, when we refer to $\boldsymbol{\theta}^{(2)}$ we mean parameters from $\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\}$ from equation (3). The drift function $\mathbf{F}$ in (2) is split up into a linear part given by matrix $\mathbf{A}$ and vector $\mathbf{b}$ and a nonlinear part given by $\mathbf{N}$. This splitting is essential for defining the splitting schemes and the objective functions for estimating $\boldsymbol{\theta}$.

We denote the true parameter value by $\boldsymbol{\theta}_0=(\boldsymbol{\theta}_0^{(1)},\boldsymbol{\theta}_0^{(2)})$ and assume that $\boldsymbol{\theta}_0\in\Theta$. Sometimes we write $\mathbf{A}_0$, $\mathbf{b}_0$, $\mathbf{N}_0(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\mathbf{x})$ instead of $\mathbf{A}(\boldsymbol{\theta}_0^{(1)})$, $\mathbf{b}(\boldsymbol{\theta}_0^{(1)})$, $\mathbf{N}(\mathbf{x};\boldsymbol{\theta}_0^{(1)})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}(\mathbf{x};\boldsymbol{\theta}_0^{(2)})^\top$, when referring to the true parameters. We write $\mathbf{A}$, $\mathbf{b}$, $\mathbf{N}(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})$ for any parameter $\boldsymbol{\theta}$. Sometimes, we suppress the parameter to simplify notation, e.g., $\mathbb{E}$ implicitly refers to $\mathbb{E}_{\boldsymbol{\theta}}$.

2.1 Examples

This section provides three examples of SDEs that fit the described framework. Two of these SDEs are standard models from the literature, and the third one is a novel extension of the Kramers oscillator, which we denote as the Student Kramers oscillator. For each of the three examples, we explicitly find diffusion parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$.

The first model is the coupled multivariate Wright-Fisher diffusion, which describes the evolution of allele frequencies at multiple loci with Pearson-type multiplicative noise corresponding to Jacobi diffusions. This implies that the multivariate Pearson diffusion with this noise type would have an invariant generalized multivariate beta distribution. The second example is the Stochastic SIR model, a widely used epidemiological model that describes the spread of infectious diseases within a population. It characterizes the transitions between susceptible, infectious, and recovered individuals. The third example is the Student Kramer oscillator, a hypoelliptic model for processes with two quasi-stationary states.

2.1.1 Coupled multivariate Wright-Fisher diffusion

The coupled Wright-Fisher diffusion [Aurell et al., 2019] is a multidimensional model that describes the evolution of genetic frequencies across multiple loci and alleles. It is formulated as the strong solution to a system of SDEs, with the drift terms incorporating interactions between loci to account for inter-locus selection [Favero et al., 2021]. The model presented in this section relies on both papers [Aurell et al., 2019] and [Favero et al., 2021].

Let L be the number of loci, and let M_l be the number of alleles at locus l for $l = 1, 2, \ldots, L$. The set of allele frequencies at time t at each locus is denoted by the $(\sum_{l=1}^{L} M_l)$-dimensional frequency vector $\mathbf{X}_t = (\boldsymbol{X}_t^{(1)\top}, \boldsymbol{X}_t^{(2)\top}, \ldots, \boldsymbol{X}_t^{(L)\top})^\top$. Vector $\boldsymbol{X}_t^{(l)} = (X_t^{(l1)}, X_t^{(l2)}, \ldots, X_t^{(lM_l)})^\top$ represents the allele frequencies at locus l at time t, that is

$$X_t^{li} \geq 0 \qquad \text{and} \qquad \sum_{j=1}^{M_l} X_t^{lj} = 1, \tag{5}$$

for all $i = 1, 2, \ldots, M_l$ and $l = 1, 2, \ldots L$. The dynamics of $\mathbf{X}_t$ are governed by the following SDE

$$\mathrm{d}\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t)\,\mathrm{d}t + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\nabla V(\mathbf{X}_t)\,\mathrm{d}t + \boldsymbol{\Sigma}(\mathbf{X}_t)\,\mathrm{d}\mathbf{W}_t, \tag{6}$$

where $\boldsymbol{\mu}(\mathbf{X})$ is the mutation drift vector, $\boldsymbol{\Sigma}(\mathbf{X})$ is the diffusion matrix, and $V(\mathbf{X})$ is a potential function encoding selection and interaction effects.

The diffusion matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X})$ for the multivariate Wright-Fisher process is block-diagonal, reflecting the independence of loci

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}) = \begin{bmatrix} \boldsymbol{\Sigma}\boldsymbol{\Sigma}_1^\top(\boldsymbol{X}^{(1)}) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^\top(\boldsymbol{X}^{(2)}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}\boldsymbol{\Sigma}_L^\top(\boldsymbol{X}^{(L)}) \end{bmatrix}, \tag{7}$$

where each block $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_l^\top(\boldsymbol{X}^{(l)})$ for locus l is given by

$$[\boldsymbol{\Sigma}\boldsymbol{\Sigma}_l^\top(\boldsymbol{X}^{(l)})]_{ij} = \begin{cases} X^{(li)}(1 - X^{(li)}) & \text{if } i = j \\ -X^{(li)}X^{(lj)} & \text{if } i \neq j \end{cases}; \quad i, j = 1, \ldots, M_l \tag{8}$$

Equivalently, this can be written as

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}_l^\top(\boldsymbol{X}^{(l)}) = \mathrm{diag}(\boldsymbol{X}^{(l)}) - \boldsymbol{X}^{(l)}\boldsymbol{X}^{(l)\top}. \tag{9}$$

The drift parameter is a sum of two functions, $\boldsymbol{\mu}(\mathbf{X})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X})\nabla V(\mathbf{X})$. The first function, $\boldsymbol{\mu}$, represents the mutation dynamics. It is assumed that mutations occur independently at each locus. Specifically, for the l-th locus, the mutation rate is θ_l, and the mutation probability matrix is

$$\boldsymbol{P}^{(l)} = \begin{bmatrix} p_{11}^{(l)} & \cdots & p_{1M_l}^{(l)} \\ \vdots & \ddots & \vdots \\ p_{M_l1}^{(l)} & \cdots & p_{M_lM_l}^{(l)} \end{bmatrix}, \qquad \sum_{j=1}^{M_l} p_{ij}^{(l)} = 1, \quad i = 1, 2, \ldots, M_l. \tag{10}$$

The transition rates of mutations from allele i to allele j at locus l are given by $u_{ij}^{(l)} = \frac{\theta_l}{2} p_{ij}^{(l)}$. Following the standard Wright-Fisher model with parent-dependent mutations, the i-th component of function $\boldsymbol{\mu}^{(l)}$ is

$$\mu^{(li)}(\mathbf{X}^{(l)}) = \sum_{j=1}^{M_l} (u_{ji}^{(l)} X^{(lj)} - u_{ij}^{(l)} X^{(li)}). \tag{11}$$

The second term in the drift, $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X})\nabla V(\mathbf{X})$, represents selection for the current allele type at the current locus. The fitness potential $V(\mathbf{X}_t)$ is explicitly constructed such that the coupled Wright-Fisher diffusion (6) treats the effects of selection and mutation as independent mechanisms, ignoring their cross-effects. This means that the term $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X})\nabla V(\mathbf{X})$ includes at most pairwise interactions between different loci and their allele types. Thus, $V(\mathbf{X})$ is given by

$$V(\mathbf{X}) = \mathbf{X}^\top\mathbf{h} + \frac{1}{2}\mathbf{X}^\top\mathbf{J}\mathbf{X} \tag{12}$$

and represents the selection and interaction effects, where $\mathbf{h}$ is the locus selection parameter, and $\mathbf{J}$ is the interaction matrix, a symmetric block matrix with the blocks on the diagonal equal to zero matrices. The gradient of $V(\mathbf{X})$ is:

$$\nabla V(\mathbf{X}) = \mathbf{h} + \mathbf{J}\mathbf{X}. \tag{13}$$

Both components in the drift are structured to operate on individual loci without interaction.

The Wright-Fisher diffusion model described by (6) is a specific case of the general SDE presented in (2). Specifically, the squared diffusion matrix (8) for each locus l conforms to the structure given in (3). The matrix $\boldsymbol{\alpha}^{(l)ij}$ is of size $M_l \times M_l$ and is entirely composed of zeros except at the positions $(m,n) \in \{(i,j),(j,i)\}$, which equal $-1/2$ and positions $(m,n) = \{(i,i)\}$ which equal -1. This can be expressed more compactly as

$$\boldsymbol{\alpha}^{(l)ij} = [-\frac{1}{2}(\delta_{mi}\delta_{nj} + \delta_{ni}\delta_{mj})]_{m,n=1}^{M_l}. \tag{14}$$

The vector $\boldsymbol{\beta}^{(l)ij}$ is zero everywhere except at the position $m = i = j$, where it takes the value 1. This can be formally written as

$$\boldsymbol{\beta}^{(l)ij} = [\delta_{mi}\delta_{mj}]_{m=1}^{M_l}. \tag{15}$$

Finally, the parameter γ^{ij} is always zero for this model.

Note that there are no parameters to estimate in the diffusion matrix. However, the model can be extended by adding a parameter in front of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_l^\top$.

2.2 Stochastic SIR model

The Stochastic Susceptible-Infectious-Recovered (SIR) model is used in epidemiology to describe the spread of a disease within a population. The model divides the population into three categories: susceptible (S), infectious (I), and recovered (R). In the Stochastic SIR model, the transitions between categories are governed by probabilistic events, reflecting the inherent randomness in disease transmission and recovery processes.

The following SDE describes the stochastic SIR model:

$$\mathrm{d}\begin{bmatrix} S_t \\ I_t \end{bmatrix} = \begin{bmatrix} -\alpha S_t I_t \\ \alpha S_t I_t - \beta I_t \end{bmatrix}\mathrm{d}t + \frac{1}{\sqrt{N}}\begin{bmatrix} \sqrt{\alpha S_t I_t} & 0 \\ -\sqrt{\alpha S_t I_t} & \sqrt{\beta I_t} \end{bmatrix}\mathrm{d}\mathbf{W}_t, \tag{16}$$

where S_t, I_t are the fraction of susceptible and infectious individuals, respectively, at time t in a population of size N, α is the transmission rate, β is the recovery rate, and $\mathbf{W}_t$ is a two-dimensional Wiener process representing the stochastic noise. The fraction of recovered is given by $R_t = 1 - S_t - I_t$.

This model also fits the setup in (2) since the square diffusion matrix is a second-order polynomial of the state vector, that is,

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(s,i) = \frac{1}{N}\begin{bmatrix} \alpha si & -\alpha si \\ -\alpha si & \alpha si + \beta i \end{bmatrix}. \tag{17}$$

Thus, the matrices $\boldsymbol{\alpha}^{ij}$ from (3) are

$$\boldsymbol{\alpha}^{11} = \boldsymbol{\alpha}^{22} = \frac{\alpha}{N}\begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}, \qquad \boldsymbol{\alpha}^{12} = \boldsymbol{\alpha}^{21} = \frac{\alpha}{N}\begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix}, \tag{18}$$

and $\boldsymbol{\beta}^{ij}$ are

$$\boldsymbol{\beta}^{11} = \boldsymbol{\beta}^{12} = \boldsymbol{\beta}^{21} = \mathbf{0}, \qquad \boldsymbol{\beta}^{22} = \frac{\beta}{N}\begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{19}$$

All γ^{ij} are zeros.

2.2.1 The Student Kramers oscillator

In Pilipovic et al. [2024b], we analyzed the Kramers oscillator with additive noise

$$\begin{aligned} \mathrm{d}X_t &= V_t \,\mathrm{d}t, \\ \mathrm{d}V_t &= \left(-\eta V_t + aX_t^3 + cX_t\right) \mathrm{d}t + \sigma \,\mathrm{d}W_t, \end{aligned} \tag{20}$$

where $a < 0$ and $c, \sigma > 0$, and $\eta \geq 0$. This oscillator is a second-order SDE and a stochastic damping Hamiltonian system. The Kramers oscillator (20) characterizes the stochastic movement of a particle within a bistable potential

$$U(x) = -a\frac{x^4}{4} - c\frac{x^2}{2}. \tag{21}$$

The parameter η in (20) indicates the damping level, c regulates the linear stiffness, and a determines the nonlinear component of the restoring force. When $a = 0$, the equation simplifies to a damped harmonic oscillator.

In Pilipovic et al. [2024b], we fitted the Kramers oscillator (20) to Greenland Ice Core data [Rasmussen et al., 2014] to understand the abrupt temperature changes during the ice ages, known as the Dansgaard–Oeschger (DO) events. We found that this model only partially fits the data. Precisely, the velocity variable V_t did not adequately capture the spread in the observed velocity. We conjecture that this discrepancy is likely due to extreme values in the data that are not well accounted for by additive Gaussian noise. This paper proposes generalizing the model (20) to allow for more heavy-tailed intrinsic noise.

Motivated by these heavy-tailed patterns, we extend the diffusion function to yield a stationary distribution that would follow a Student's t-distribution if the drift function were linear, in contrast to the Gaussian distribution for additive noise. Specifically, the invariant distribution of Pearson diffusion (1) when $\alpha > 0$, $\beta^2 - 4\alpha\gamma \leq 0$, $\alpha < 2a$ is a generalized Student's t-distribution

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \qquad x \in \mathbb{R}, \tag{22}$$

where $\Gamma(\cdot)$ is the gamma function and

$$\nu = -\frac{2a}{\alpha} + 1, \quad \mu = -\frac{\beta}{2\alpha}, \quad \sigma^2 = \frac{\gamma}{\alpha\nu} - \mu^2.$$

The generalized Student t-distribution is peaked around the location parameter μ. If $\mu \neq 0$, it is a skewed distribution. The peak width is determined by the scale parameter $\sigma > 0$. The shape parameter, or degrees of freedom, ν, influences the heaviness of the tails.

We also generalize the potential function $U(x)$ (21) to allow for asymmetries between the two modes of the distribution. The generalized potential function is

$$U(x) = -a\frac{x^4}{4} - b\frac{x^3}{3} - c\frac{x^2}{2} - dx. \tag{23}$$

Finally, we define the Student Kramers oscillator (Figure 1) as a solution to

$$\begin{aligned} \mathrm{d}X_t &= V_t \,\mathrm{d}t, \\ \mathrm{d}V_t &= (-\eta V_t - U'(X_t)) \,\mathrm{d}t + \sqrt{\alpha V_t^2 + \beta V_t + \gamma} \,\mathrm{d}W_t, \end{aligned} \tag{24}$$

where $\eta \geq 0$, $a < 0$, $\alpha > 0$, $\beta^2 - 4\alpha\gamma < 0$, $\alpha < 2\eta$, and X_t, V_t are defined on $\mathbb{R}$. These conditions are necessary to have a non-exploding solution with Student-type noise. We denote any SDE for an SDE with Student-type noise if the corresponding SDE with linear drift is a Student Pearson diffusion (see also Forman and Sørensen [2008], Leonenko and Phillips [2012]). The condition $\alpha < 2\eta$ is sufficient for an invariant density to exist. Thus, the unknown parameters are

$$\boldsymbol{\theta}^{(1)} = \{\eta, a, b, c, d\}, \text{ and } \boldsymbol{\theta}^{(2)} = \{\alpha, \beta, \gamma\}. \tag{25}$$

The squared diffusion function of SDE (24) is

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(x, v) = \begin{bmatrix} 0 & 0 \\ 0 & \alpha v^2 + \beta v + \gamma \end{bmatrix}, \tag{26}$$

so it aligns with the structure given in (3). Namely, all $\boldsymbol{\alpha}^{11}, \boldsymbol{\alpha}^{12}, \boldsymbol{\alpha}^{21}$ are zero matrices in $\mathbb{R}^{2\times 2}$, all $\boldsymbol{\beta}^{11}, \boldsymbol{\beta}^{12}, \boldsymbol{\beta}^{21}$ are zero vectors in $\mathbb{R}^2$, and all $\gamma^{11}, \gamma^{12}, \gamma^{21} = 0$. Moreover,

$$\boldsymbol{\alpha}^{22} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha \end{bmatrix}, \qquad \boldsymbol{\beta}^{22} = \begin{bmatrix} 0 \\ \beta \end{bmatrix}, \qquad \gamma^{22} = \gamma. \tag{27}$$

Now, we state and prove the following theorem about the solution of SDE (24).

Theorem 2.1 *A unique, strong solution exists to the Student Kramers SDE* (24)*, when* $\eta \geq 0$*,* $a < 0$*,* $\alpha > 0$*,* $\beta^2 - 4\alpha\gamma < 0$*. An invariant probability measure exists for the system when* $\alpha < 2\eta$*.*

Proof To prove the existence and uniqueness of the solution to (24), we show that the diffusion function

$$\mathbf{\Sigma}(x, v) = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{\alpha v^2 + \beta v + \gamma} \end{bmatrix}$$

is Lipschitz and that the drift function

$$\mathbf{F}(x, v) = \begin{bmatrix} v \\ -\eta v + ax^3 + bx^2 + cx + d \end{bmatrix}$$

is one-sided Lipschitz and of at most polynomial growth (see Assumptions (A2) and (A3) in Section 2.3).

To see that $\mathbf{\Sigma}(x, v)$ is Lipschitz, we show that the derivative of $\sigma(v) = \sqrt{\alpha v^2 + \beta v + \gamma}$ is bounded for all $v \in \mathbb{R}$. We have

$$|\sigma'(v)| = \frac{|2\alpha v + \beta|}{2\sqrt{\alpha v^2 + \beta v + \gamma}}.$$

If $v = 0$, then $|\sigma'(v)| = \frac{|\beta|}{2\sqrt{\gamma}}$. Otherwise, for $|v| > c_1 > 0$ and $\sqrt{1 + \beta/(\alpha v) + \gamma/(\alpha v^2)} > c_2 > 0$, we have

$$|\sigma'(v)| \leq \frac{|\alpha v|}{\sqrt{\alpha v^2 + \beta v + \gamma}} + \frac{|\beta|}{2\sqrt{\alpha v^2 + \beta v + \gamma}} < \frac{\sqrt{\alpha}}{c_2} + \frac{|\beta|}{2\sqrt{\alpha} c_1 c_2}.$$

To show that $\mathbf{F}$ is one-sided Lipschitz, we start by computing the Jacobian matrix of $\mathbf{F}$ as

$$D\mathbf{F}(x, v) = \begin{bmatrix} 0 & 1 \\ 3ax^2 + 2bx + c & -\eta \end{bmatrix}. \tag{28}$$

Since $a < 0$ and $\eta > 0$, all components of $D\mathbf{F}(x, v)$ are upper bounded, so $\mathbf{F}$ is one-sided Lipschitz. By construction, $\mathbf{F}$ has polynomial growth. Thus, a unique, strong solution exists to SDE (24).

Next, we prove that SDE (24) is hypoelliptic using Hörmander's condition [Hörmander, 1967]. To apply this condition, we first write SDE (24) in Stratonovich form:

$$\begin{aligned} \mathrm{d}X_t &= V_t \, \mathrm{d}t, \\ \mathrm{d}V_t &= \left(-\eta V_t + aX_t^3 + bX_t^2 + cX_t + d - \frac{1}{2}\left(\alpha V_t + \frac{1}{2}\beta \right) \right) \mathrm{d}t + \sqrt{\alpha V_t^2 + \beta V_t + \gamma} \circ \mathrm{d}W_t. \end{aligned} \tag{29}$$

Now, the associated drift and diffusion vector fields are

$$\mathbf{V}_0(x, v) = \begin{bmatrix} v \\ \left(-\eta - \frac{1}{2}\alpha\right) v + ax^3 + bx^2 + cx + d - \frac{1}{4}\beta \end{bmatrix}, \quad \mathbf{V}_1(x, v) = \begin{bmatrix} 0 \\ \sqrt{\alpha v^2 + \beta v + \gamma} \end{bmatrix}. \tag{30}$$

We recall that the Lie bracket of smooth vector fields $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$, defined as

$$[\boldsymbol{f}, \boldsymbol{g}] \coloneqq D\boldsymbol{g}(\mathbf{x})\boldsymbol{f}(\mathbf{x}) - D\boldsymbol{f}(\mathbf{x})\boldsymbol{g}(\mathbf{x}),$$

is used to verify Hörmander's condition.

We define the set $\mathcal{H}$ of vector fields iteratively as

$$\begin{aligned} \mathbf{V}_1 &\in \mathcal{H}, \\ \mathbf{H} \in \mathcal{H} \Rightarrow [\mathbf{V}_0, \mathbf{H}], [\mathbf{V}_1, \mathbf{H}] &\in \mathcal{H}. \end{aligned}$$

The weak Hörmander condition is met if the vectors in $\mathcal{H}$ span $\mathbb{R}^d$ at every point $\mathbf{x} \in \mathbb{R}^d$. Initially, vector $\mathbf{V}_1$ spans $\{(0, y) \in \mathbb{R}^2 \mid y \in \mathbb{R}\}$, a one-dimensional subspace. Therefore, we need to verify the existence of some $\mathbf{H} \in \mathcal{H}$ with a non-zero first element. It is easy to see that

$$\begin{aligned} [\mathbf{V}_0, \mathbf{V}_1]^{(1)} &= -\sigma(v) < 0, \quad \forall v \in \mathbb{R}, \\ [\mathbf{V}_1, \mathbf{V}_0]^{(1)} &= \sigma(v) \geq 0, \quad \forall v \in \mathbb{R}. \end{aligned}$$

Thus, the Student Kramers SDE defined in (24) is hypoelliptic, meaning it admits a smooth transition density.

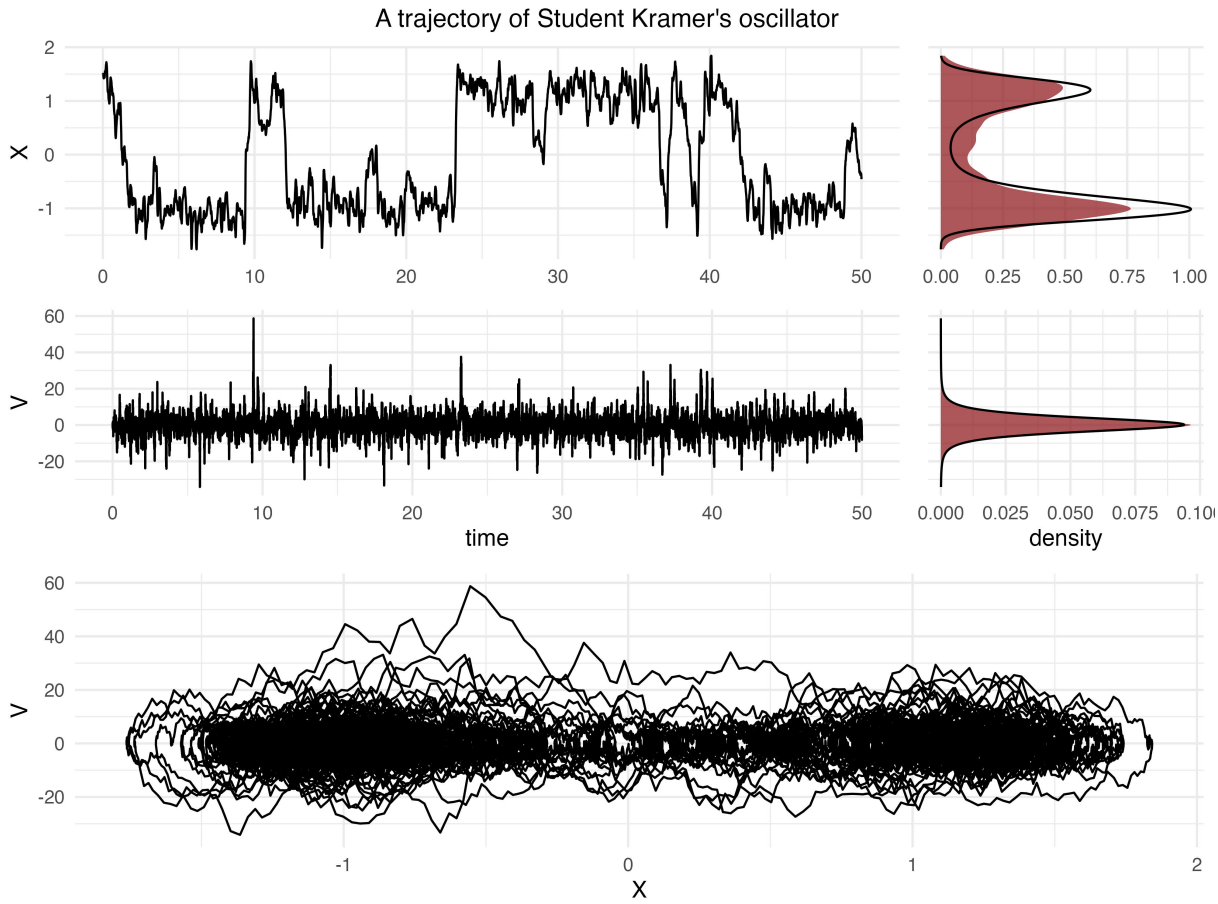

Figure 1: A trajectory of Student Kramers oscillator (24) simulated with the Milstein scheme with $h = 0.001$ and $N = 50000$, and with true parameter $\eta_0 = 30$, $a_0 = -125$, $b_0 = 40$, $c_0 = 150$, $d_0 = -20$, $\alpha_0 = 20$, $\beta_0 = -8$, and $\gamma_0 = 1280.8$. The first and second rows show the evolution of the individual components X_t and V_t, respectively. The last row shows the evolution of pair (X_t, V_t). The first two rows also depict the empirical invariant densities of X_t and V_t in red, overlined by the approximated theoretical invariant densities.

Since SDE (24) is hypoelliptic and the drift and diffusion functions with their corresponding derivatives grow at most polynomially, SDE (24) has the strong Feller property (see, for example, Theorem 1.2 in Hairer and Pillai [2011]). Then, we use Theorem 4.5 from Meyn and Tweedie [1993] to prove that SDE (24) has an invariant measure. We choose $V(x, v) = \frac{1}{2}v^2 + U(x)$ as a Lyapunov function. Since $a < 0$, $\lim_{\|\mathbf{x}\| \to \infty} V(\mathbf{x}) = +\infty$.

For SDE (24), the infinitesimal generator is

$$\mathbb{L}\phi(x, v) = v\frac{\partial \phi}{\partial x} + (-\eta v - U'(x))\frac{\partial \phi}{\partial v} + \frac{1}{2}(\alpha v^2 + \beta v + \gamma)\frac{\partial^2 \phi}{\partial v^2}.$$

Then,

$$\mathbb{L}V(x, v) = \left(\frac{\alpha}{2} - \eta\right) v^2 + \frac{\beta}{2}v + \frac{\gamma}{2}. \tag{31}$$

Since $\alpha < 2\eta$, we can find a compact set $K \subset \mathbb{R}^2$ and constants $c_1 > 0, c_2 \in \mathbb{R}$, such that

$$\mathbb{L}V(\mathbf{x}) \leq -c_1\|\mathbf{x}\|^2 + c_2\mathbb{1}\{\mathbf{x} \in K\}. \tag{32}$$

According to Theorem 4.5 in Meyn and Tweedie [1993], an invariant measure π for SDE (24) exists.

2.3 Assumptions

The main assumption is that SDE (4) has a unique strong solution $\mathbf{X} = (\mathbf{X}_t)_{t\in[0,T]}$, adapted to $(\mathcal{F}_t)_{t\in[0,T]}$. We have two alternative assumptions on the diffusion matrix $\mathbf{\Sigma}$ defined in (3) and the SDE (4), determining whether the process

is elliptic (the diffusion matrix is of full rank) or hypoelliptic (the diffusion matrix is of reduced rank and the solution of (4) admits a smooth density).

(A1) $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ defined in (3) is positive definite on $\mathcal{X} \times \overline{\Theta}_{\theta_2}$.

(A1') $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ defined in (3) is positive semidefinite on $\mathcal{X} \times \overline{\Theta}_{\theta_2}$ and SDE (2) is hypoelliptic.

The matrix $\boldsymbol{\Sigma}$ and thus $\overline{\Theta}_{\theta_2}$ define the state space. For example, for the scalar Pearson diffusion (1), the state space can only be the entire real line, $\mathcal{X} = \mathbb{R}$, if $\alpha = \beta = 0$ (normal distribution) or $\alpha \geq 0$ and the polynomial has no real roots (t-distribution, possibly skewed). Here, we give a general framework assuming that the state and parameter space are known and that the diffusion has a strong and unique solution. From a statistical point of view, $\mathcal{X}$ is known, so assumptions (A1) or (A1') define the parameter space $\overline{\Theta}_{\theta_2}$.

We extend the first two assumptions in Pilipovic et al. [2024a], which are special cases of assumptions in Tretyakov and Zhang [2013]. Now, following Tretyakov and Zhang [2013], the following two assumptions ensure that a unique, strong solution exists on the state space $\mathcal{X}$. These two assumptions are not necessary, but they are sufficient for a unique solution to exist. We choose these assumptions as a natural continuation of our previous papers [Pilipovic et al., 2024a,b].

(A2) Function $\mathbf{N}$ is twice continuously differentiable with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$. Additionally, for all $\boldsymbol{\theta} \in \overline{\Theta}$, for a sufficiently large $p \geq 1$, there is a constant $C_{\boldsymbol{\theta}} > 0$ such that:

$$(\mathbf{x}-\mathbf{y})^\top (\mathbf{N}(\mathbf{x};\boldsymbol{\theta}^{(1)})-\mathbf{N}(\mathbf{y};\boldsymbol{\theta}^{(1)}))+\frac{2p-1}{2}\sum_{i=1}^{d}\|\boldsymbol{\Sigma}_i(\mathbf{x};\boldsymbol{\theta}^{(2)})-\boldsymbol{\Sigma}_i(\mathbf{y};\boldsymbol{\theta}^{(2)})\|^2 \leq C_{\boldsymbol{\theta}}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{X}.$$

(A3) Function $\mathbf{N}$ grows at most polynomially in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}^{(1)}$, i.e., there exist constants $C_{\boldsymbol{\theta}^{(1)}} > 0$ and $p \geq 1$ such that:

$$\|\mathbf{N}(\mathbf{x};\boldsymbol{\theta}^{(1)}) - \mathbf{N}(\mathbf{y};\boldsymbol{\theta}^{(1)})\|^2 \leq C_{\boldsymbol{\theta}^{(1)}}(1+\|\mathbf{x}\|^{2p-2}+\|\mathbf{y}\|^{2p-2})\|\mathbf{x}-\mathbf{y}\|^2, \qquad \forall \mathbf{x},\mathbf{y} \in \mathcal{X}.$$

Additionally, its derivatives are of polynomial growth in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}^{(1)}$.

We need the following two assumptions to ensure the ergodicity and identifiability of the parameters.

(A4) The solution $\mathbf{X}$ of SDE (2) has invariant probability measure $\nu_0(\mathrm{d}\mathbf{x})$.

(A5) Functions $\mathbf{F}$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ are identifiable in $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, respectively. That is, if $\mathbf{F}(\mathbf{x},\boldsymbol{\theta}^{(1)}) = \mathbf{F}(\mathbf{x},\boldsymbol{\theta}^{(1)}_\star)$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x},\boldsymbol{\theta}^{(2)}) = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x},\boldsymbol{\theta}^{(2)}_\star)$ for all $\mathbf{x} \in \mathbb{R}^d$, then $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(1)}_\star$ and $\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(2)}_\star$.

We assume discrete observations $(\mathbf{X}_{t_k})_{k=0}^N \equiv \mathbf{X}_{0:t_N}$ of SDE (4) at time steps $0 = t_0 < t_1 < \cdots < t_N = T$. For notational simplicity, we assume equidistant step size $h = t_k - t_{k-1}$.

3 Multivariate Pearson diffusions

In this section, we give details on the extension of Pearson diffusions to a multivariate setting. While this introduces additional complexity, we retain the desirable property of knowing first and second moments explicitly. A multivariate Pearson diffusion is described by the SDE (4) with $\mathbf{N} = \mathbf{0}$, that is

$$\mathbf{F}(\mathbf{x}) = \mathbf{A}(\mathbf{x}-\mathbf{b}), \tag{33}$$

$$[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{ij} = \mathbf{x}^\top\boldsymbol{\alpha}^{ij}\mathbf{x} + \mathbf{x}^\top\boldsymbol{\beta}^{ij} + \gamma^{ij}, \qquad i,j = 1,2,...,d, \tag{34}$$

where it is impossible to represent $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})$ in matrix form without using tensors, so we vectorize $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})$.

In the following, we use standard properties of the vec operator and the Kronecker product $\otimes$ (for more details see Magnus [2019]). We start by noticing that

$$\begin{aligned}
[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{ij} &= \operatorname{vec}(\mathbf{x}^\top\boldsymbol{\alpha}^{ij}\mathbf{x}) + \mathbf{x}^\top\boldsymbol{\beta}^{ij} + \gamma^{ij} \\
&= \operatorname{vec}(\mathbf{x}^\top \otimes \mathbf{x}^\top)\operatorname{vec}(\boldsymbol{\alpha}^{ij}) + \boldsymbol{\beta}^{ij\top}\mathbf{x} + \gamma_{ij} \\
&= ((\mathbf{x}\otimes\mathbf{x})^\top \operatorname{vec}(\boldsymbol{\alpha}^{ij}))^\top + \boldsymbol{\beta}^{ij\top}\mathbf{x} + \gamma_{ij} \\
&= \operatorname{vec}(\boldsymbol{\alpha}^{ij})^\top \operatorname{vec}(\mathbf{x}\mathbf{x}^\top) + \boldsymbol{\beta}^{ij\top}\mathbf{x} + \gamma_{ij}.
\end{aligned}$$

Then, we express the vectorization of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})$ as

$$\operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})) = \begin{bmatrix} [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{11} \\ \vdots \\ [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{1d} \\ \vdots \\ [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{dd} \end{bmatrix} = \begin{bmatrix} \operatorname{vec}(\boldsymbol{\alpha}^{11})^\top \\ \vdots \\ \operatorname{vec}(\boldsymbol{\alpha}^{1d})^\top \\ \vdots \\ \operatorname{vec}(\boldsymbol{\alpha}^{dd})^\top \end{bmatrix} \operatorname{vec}(\mathbf{x}\mathbf{x}^\top) + \begin{bmatrix} \boldsymbol{\beta}^{11\top} \\ \vdots \\ \boldsymbol{\beta}^{1d\top} \\ \vdots \\ \boldsymbol{\beta}^{dd\top} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \gamma^{11} \\ \vdots \\ \gamma^{1d} \\ \vdots \\ \gamma^{dd} \end{bmatrix} = \check{\boldsymbol{\alpha}} \operatorname{vec}(\mathbf{x}\mathbf{x}^\top) + \check{\boldsymbol{\beta}}\mathbf{x} + \check{\boldsymbol{\gamma}}, \tag{35}$$

where we defined

$$\begin{aligned} \check{\boldsymbol{\alpha}} &\coloneqq [\operatorname{vec}(\boldsymbol{\alpha}^{11})^\top, \dots, \operatorname{vec}(\boldsymbol{\alpha}^{1d})^\top, \dots, \operatorname{vec}(\boldsymbol{\alpha}^{dd})^\top]^\top \in \mathbb{R}^{d^2 \times d^2}, \\ \check{\boldsymbol{\beta}} &\coloneqq [\boldsymbol{\beta}^{11\top}, \dots, \boldsymbol{\beta}^{1d\top}, \dots, \boldsymbol{\beta}^{dd\top}]^\top \in \mathbb{R}^{d^2 \times d}, \\ \check{\boldsymbol{\gamma}} &\coloneqq (\gamma^{11\top}, \dots, \gamma^{1d\top}, \dots, \gamma^{dd\top})^\top = \operatorname{vec}(\boldsymbol{\gamma})^\top \in \mathbb{R}^{d^2}. \end{aligned}$$

Now, we focus on computing the first two moments of SDE (2). If $\boldsymbol{\phi}(t, \mathbf{x})$ is a sufficiently smooth function, then from the Itô formula the expected value of $\boldsymbol{\phi}(t, \mathbf{X}_t)$ evolves according to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\boldsymbol{\phi}(t, \mathbf{X}_t)] = \mathbb{E}\left[\frac{\partial \boldsymbol{\phi}(t, \mathbf{X}_t)}{\partial t}\right] + \sum_{i=1}^{d} \mathbb{E}\left[\frac{\partial \boldsymbol{\phi}(t, \mathbf{X}_t)}{\partial x^{(i)}} F^{(i)}(\mathbf{X}_t)\right] + \frac{1}{2}\sum_{i,j=1}^{d} \mathbb{E}\left[\frac{\partial^2 \boldsymbol{\phi}(t, \mathbf{X}_t)}{\partial x^{(i)} \partial x^{(j)}} [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)]_{ij}\right]. \tag{36}$$

To find the mean vector $\mathbf{m}(t) = \mathbb{E}[\mathbf{X}_t]$, we set $\boldsymbol{\phi}(t, \mathbf{x}) = \mathbf{x}$ and obtain

$$\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} = \mathbb{E}[\mathbf{F}(\mathbf{X}_t)] = \mathbf{A}(\mathbf{m}(t) - \mathbf{b}). \tag{37}$$

The solution of the linear ODE (37) is

$$\mathbf{m}(t) = \exp(\mathbf{A}t)(\mathbf{m}(0) - \mathbf{b}) + \mathbf{b}, \tag{38}$$

where $\mathbf{m}(0)$ is a given initial condition.

For the covariance matrix $\mathbf{C}(t) = \mathbb{E}[(\mathbf{X}_t - \mathbf{m}(t))(\mathbf{X}_t - \mathbf{m}(t))^\top]$, setting $\boldsymbol{\phi}(t, \mathbf{x}) = \mathbf{x}\mathbf{x}^\top - \mathbf{m}(t)\mathbf{m}(t)^\top$, we derive:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} &= \mathbb{E}[\mathbf{F}(\mathbf{X}_t)(\mathbf{X}_t - \mathbf{m}(t))^\top] + \mathbb{E}[(\mathbf{X}_t - \mathbf{m}(t))\mathbf{F}(\mathbf{X}_t)^\top] + \mathbb{E}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)] && (39) \\ &= \mathbf{A}\mathbf{C}(t) + \mathbf{C}(t)\mathbf{A}^\top + \mathbb{E}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)]. && (40) \end{aligned}$$

The previous equation is a matrix differential equation with a form of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)$ that can not be explicitly written without using tensors. To solve it, we transform it into a linear ordinary differential equation. Using the linearity of differentiation, vectorization, and expectation operators, we obtain

$$\begin{aligned} \frac{\mathrm{d}\operatorname{vec}(\mathbf{C}(t))}{\mathrm{d}t} &= \operatorname{vec}(\mathbf{A}\mathbf{C}(t)) + \operatorname{vec}(\mathbf{C}(t)\mathbf{A}^\top) + \mathbb{E}[\operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t))] \\ &= (\mathbf{I} \otimes \mathbf{A}) \operatorname{vec}(\mathbf{C}(t)) + (\mathbf{A} \otimes \mathbf{I}) \operatorname{vec}(\mathbf{C}(t)) + \check{\boldsymbol{\alpha}} \operatorname{vec}(\mathbb{E}[\mathbf{X}_t \mathbf{X}_t^\top]) + \check{\boldsymbol{\beta}} \operatorname{vec}(\mathbb{E}[\mathbf{X}_t]) + \check{\boldsymbol{\gamma}} \\ &= (\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}}) \operatorname{vec}(\mathbf{C}(t)) + \check{\boldsymbol{\alpha}} \operatorname{vec}(\mathbf{m}(t)\mathbf{m}(t)^\top) + \check{\boldsymbol{\beta}}\mathbf{m}(t) + \check{\boldsymbol{\gamma}} \\ &= (\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}}) \operatorname{vec}(\mathbf{C}(t)) + \operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{m}(t))). \end{aligned}$$

The solution to this linear ODE, given initial condition $\mathbf{C}(0)$, is

$$\operatorname{vec}(\mathbf{C}(t)) = \exp(\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}}) \operatorname{vec}(\mathbf{C}(0)) + \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t - s)) \operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{m}(s)))\, \mathrm{d}s. \tag{41}$$

At least two methods exist to evaluate the integral in Eq. (41). First, we can expand $\operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{m}(s)))$ in a Taylor series around $s \to 0$ and apply Theorem 1 from Carbonell et al. [2008]. Second, we can substitute $\mathbf{m}(s)$ from Eq. (38) into Eq. (41) and directly use Theorem 1 from Van Loan [1978]. Here, we opt for the second approach. We begin with

$$\int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t - s)) \operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{m}(s)))\, \mathrm{d}s$$

$$
\begin{aligned}
&= \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))(\check{\boldsymbol{\alpha}} \operatorname{vec}(\mathbf{m}(s)\mathbf{m}(s)^\top) + \check{\boldsymbol{\beta}}\mathbf{m}(s) + \check{\boldsymbol{\gamma}})\,\mathrm{d}s \\
&= \mathbf{I}_1(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}((\mathbf{m}(0) - \mathbf{b})(\mathbf{m}(0) - \mathbf{b})^\top) + \mathbf{I}_2(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}((\mathbf{m}(0) - \mathbf{b})\mathbf{b}^\top) \\
&+ \mathbf{I}_3(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}(\mathbf{b}(\mathbf{m}(0) - \mathbf{b})^\top) + \mathbf{I}_4(t, \mathbf{A}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\beta}})(\mathbf{m}(0) - \mathbf{b}) + \mathbf{I}_5(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{b})),
\end{aligned} \tag{42}
$$

where we define

$$
\mathbf{I}_1(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \coloneqq \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))\check{\boldsymbol{\alpha}} \exp((\mathbf{A} \oplus \mathbf{A})s)\,\mathrm{d}s \in \mathbb{R}^{d^2 \times d^2}, \tag{43}
$$

$$
\mathbf{I}_2(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \coloneqq \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))\check{\boldsymbol{\alpha}} \exp((\mathbf{I} \oplus \mathbf{A})s)\,\mathrm{d}s \in \mathbb{R}^{d^2 \times d^2}, \tag{44}
$$

$$
\mathbf{I}_3(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \coloneqq \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))\check{\boldsymbol{\alpha}} \exp((\mathbf{A} \oplus \mathbf{I})s)\,\mathrm{d}s \in \mathbb{R}^{d^2 \times d^2}, \tag{45}
$$

$$
\mathbf{I}_4(t, \mathbf{A}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\beta}}) \coloneqq \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))\check{\boldsymbol{\beta}} \exp(\mathbf{A}s)\,\mathrm{d}s \in \mathbb{R}^{d^2 \times d}, \tag{46}
$$

$$
\mathbf{I}_5(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \coloneqq \int_0^t \exp((\mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}})(t-s))\,\mathrm{d}s \in \mathbb{R}^{d^2 \times d^2}. \tag{47}
$$

Each of these integrals (43)-(47) can be evaluated using Theorem 1 from Van Loan [1978]. For instance, if we define the following block matrix

$$
\mathbf{M}_1(\mathbf{A}, \check{\boldsymbol{\alpha}}) = \begin{bmatrix} \mathbf{A} \oplus \mathbf{A} + \check{\boldsymbol{\alpha}} & \check{\boldsymbol{\alpha}} \\ \mathbf{0} & \mathbf{A} \oplus \mathbf{A} \end{bmatrix}, \tag{48}
$$

then,

$$
\exp(\mathbf{M}_1(\mathbf{A}, \check{\boldsymbol{\alpha}})t) = \begin{bmatrix} \star & \mathbf{I}_1(t, \mathbf{A}, \check{\boldsymbol{\alpha}}) \\ \mathbf{0} & \star \end{bmatrix}, \tag{49}
$$

where $\star$ denotes a matrix of no particular interest. Similarly, we can compute the other integrals $\mathbf{I}_2$-$\mathbf{I}_5$.

Consequently, we are now equipped to explicitly compute $\operatorname{vec}(\mathbf{C}(t))$, from which, by reshaping, we can easily derive $\mathbf{C}(t)$. It is important to note that while this algorithm explicitly calculates the covariance matrix of a multivariate Pearson diffusion, it may be computationally intensive for large dimensions d due to the need to compute matrices of size d^2. This can be optimized by employing symmetric vectorization svec (cf. [de Klerk, 2002]) or half-vectorization vech (cf. [Magnus, 2019]) instead of the vec operator.

3.1 Strang splitting scheme

Consider the following splitting of (4)

$$
\mathrm{d}\mathbf{X}_t^{[1]} = \mathbf{A}(\mathbf{X}_t^{[1]} - \mathbf{b})\,\mathrm{d}t + \boldsymbol{\Sigma}(\mathbf{X}_t; \boldsymbol{\theta}^{(2)})\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0^{[1]} = \mathbf{x}_0, \tag{50}
$$

$$
\mathrm{d}\mathbf{X}_t^{[2]} = \mathbf{N}(\mathbf{X}_t^{[2]}; \boldsymbol{\theta}^{(1)})\,\mathrm{d}t, \qquad \mathbf{X}_0^{[2]} = \mathbf{x}_0. \tag{51}
$$

Equation (50) describes a multivariate Pearson diffusion whose solution cannot be explicitly obtained in general. However, we can approximate it by approximating the transition density as Gaussian as suggested by Kessler [1997]. Kessler [1997] proposed approximating the transition density with a Gaussian distribution, matching the SDE's mean and variance. Generally, these two moments cannot be obtained without knowing the transition density, but they can be approximated using the infinitesimal generator (for more details, see Kessler [1997]). For the multivariate Pearson diffusions, mean and covariance can be calculated exactly as presented in the previous section, enabling us to approximate the solution to (50) as

$$
\mathbf{X}_{t_k}^{[1]} = \Psi_h^{[1]}(\mathbf{X}_{t_{k-1}}^{[1]}) = \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}^{[1]}; \boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_h(\mathbf{X}_{t_{k-1}}^{[1]}; \boldsymbol{\theta}), \tag{52}
$$

where $\boldsymbol{\xi}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\theta}) \overset{i.i.d}{\sim} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\theta}))$ for $k = 1, \ldots, N$. We can find the function $\boldsymbol{\mu}_h$ by taking the conditional expectation of $\mathbf{X}_{t_k}^{[1]}$ given $\mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}$, which is equivalent to setting $\mathbf{m}(0) = \mathbf{x}$ in (38). Thus, we obtain that

$$
\boldsymbol{\mu}_h(\mathbf{x}; \boldsymbol{\theta}^{(1)}) = \mathbf{m}(h) = \exp(\mathbf{A}h)(\mathbf{x} - \mathbf{b}) + \mathbf{b}. \tag{53}
$$

Similarly, conditioning on $\mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}$ is equivalent to setting $\mathbf{C}(0) = \mathbf{0}$ in (41), since there is no randomness. Then, we find the covariance $\boldsymbol{\Omega}_h(\mathbf{x}; \boldsymbol{\theta})$ of $\boldsymbol{\xi}_h(\mathbf{x})$ from (42) as

$$
\operatorname{vec}(\boldsymbol{\Omega}_h(\mathbf{x}; \boldsymbol{\theta})) = \operatorname{vec}(\mathbf{C}(h)) = \mathbf{I}_1(h, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}((\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^\top) + \mathbf{I}_2(h, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}((\mathbf{x} - \mathbf{b})\mathbf{b}^\top)
$$

$$+ \mathbf{I}_3(h, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}(\mathbf{b}(\mathbf{x}-\mathbf{b})^\top) + \mathbf{I}_4(h, \mathbf{A}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\beta}})(\mathbf{x}-\mathbf{b}) + \mathbf{I}_5(h, \mathbf{A}, \check{\boldsymbol{\alpha}}) \operatorname{vec}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{b})). \tag{54}$$

Assumptions (A2) and (A3) ensure the existence and uniqueness of the solution of (51) (Theorem 1.2.17 in Humphries and Stuart [2002]). Thus, there exists a unique function $\boldsymbol{f}_h : \mathbb{R}^d \times \overline{\Theta}_{\theta^{(1)}} \to \mathbb{R}^d$, for $h \geq 0$, such that

$$\mathbf{X}^{[2]}_{t_k} = \Phi^{[2]}_h(\mathbf{X}^{[2]}_{t_{k-1}}) = \boldsymbol{f}_h(\mathbf{X}^{[2]}_{t_{k-1}}; \boldsymbol{\theta}^{(1)}). \tag{55}$$

Ideally, we would like to find an explicit function $\boldsymbol{f}_h$ in (55). Still, sometimes we need to approximate it using standard numerical tools, like Runge-Kutta algorithms. Here, we assume $\boldsymbol{f}_h$ is readily available, but all the following results also hold if we approximate $\boldsymbol{f}_h$ up to order h^2 (see Proposition 2.2 in Pilipovic et al. [2024a]).

For all $\boldsymbol{\theta}^{(1)} \in \overline{\Theta}_{\theta^{(1)}}$, the time flow $\boldsymbol{f}_h$ fulfills the following semi-group properties:

$$\boldsymbol{f}_0(\mathbf{x}; \boldsymbol{\theta}^{(1)}) = \mathbf{x}, \qquad \boldsymbol{f}_{t+s}(\mathbf{x}; \boldsymbol{\theta}^{(1)}) = \boldsymbol{f}_t(\boldsymbol{f}_s(\mathbf{x}; \boldsymbol{\theta}^{(1)}); \boldsymbol{\theta}^{(1)}), \;\; t, s \geq 0. \tag{56}$$

However, to define the estimator, we need the backward flow $\boldsymbol{f}_{-h} = \boldsymbol{f}_h^{-1}$, which might not be defined for all $h \geq 0, \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta}^{(1)} \in \overline{\Theta}_{\theta^{(1)}}$. We, therefore, introduce the following and last assumption.

(A6) There exists $h_0 > 0$ such that function $\boldsymbol{f}_h^{-1}(\mathbf{x}; \boldsymbol{\theta}^{(1)})$ is defined for all $h \in [0, h_0), \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta}^{(1)} \in \overline{\Theta}_{\theta^{(1)}}$.

Now, we are ready to define the Strang splitting approximation of SDE (4).

Definition 3.1 *Let Assumptions (A1)-(A3) hold. The Strang splitting approximation of the solution of* (4) *is given by:*

$$\mathbf{X}^{[\mathrm{S}]}_{t_k} := \Phi^{[\mathrm{S}]}_h(\mathbf{X}^{[\mathrm{S}]}_{t_{k-1}}) = (\Phi^{[2]}_{h/2} \circ \Psi^{[1]}_h \circ \Phi^{[2]}_{h/2})(\mathbf{X}^{[\mathrm{S}]}_{t_{k-1}}) = \boldsymbol{f}_{h/2}(\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}^{[\mathrm{S}]}_{t_{k-1}})) + \boldsymbol{\xi}_{h,k}(\boldsymbol{f}_{h/2}(\mathbf{X}^{[\mathrm{S}]}_{t_{k-1}}))). \tag{57}$$

3.2 Taylor expanding $\boldsymbol{\Omega}_h$

It may be useful to approximate $\boldsymbol{\Omega}_h$ for future proofs. However, it is important to note that we use the exact expressions in the implementation of the estimators, with the approximations serving only as a tool for potential theoretical proofs.

To Taylor expand $\boldsymbol{\Omega}_h(\mathbf{x}; \boldsymbol{\theta})$ (54), we do not expand every integral (43)-(47) separately. Instead, we focus on the covariance ODE from (39).

Recall that the solution of (50) is approximated such that its density is Gaussian

$$p^{[1]}(\mathbf{x}, t) = \mathcal{N}(\mathbf{x}; \mathbf{m}(t), \mathbf{C}(t)). \tag{58}$$

Recall the covariance ODE from (39)

$$\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = \mathbf{A}\mathbf{C}(t) + \mathbf{C}(t)\mathbf{A}^\top + \mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right], \tag{59}$$

where $\mathbb{E}^{[1]}$ is the expectation corresponding to probability density (58). Moreover, we denote by $\mathbb{L}^{[1]}$ the infinitesimal generator of SDE (50), that is, for sufficiently smooth function $\boldsymbol{\phi}$

$$\mathbb{L}^{[1]}\boldsymbol{\phi}(\mathbf{x}) = D_{\mathbf{x}}\boldsymbol{\phi}(\mathbf{x})\mathbf{A}(\mathbf{x}-\mathbf{b}) + \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{i,j}\boldsymbol{\phi}(\mathbf{x})[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})]_{ij}.$$

Now, we can Taylor expand $\mathbf{C}(t)$ around $t = 0$

$$\mathbf{C}(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbf{C}(0). \tag{60}$$

Taking the derivative of ODE (59) with respect to t and using the definition of the infinitesimal generator $\mathbb{L}^{[1]}$, $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}^{[1]}[\boldsymbol{\phi}(\mathbf{X}_t)] = \mathbb{E}^{[1]}\left[\mathbb{L}^{[1]}\boldsymbol{\phi}(\mathbf{X}_t)\right]$, we compute

$$\begin{aligned}\frac{\mathrm{d}^2\mathbf{C}(t)}{\mathrm{d}t^2} &= \mathbf{A}^2\mathbf{C}(t) + 2\mathbf{A}\mathbf{C}(t)\mathbf{A}^\top + \mathbf{C}(t)(\mathbf{A}^\top)^2 \\ &\quad + \mathbf{A}\mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right] + \mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right]\mathbf{A}^\top + \mathbb{E}^{[1]}\left[\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right].\end{aligned} \tag{61}$$

Taking another derivative of (61) with respect to t yields

$$\frac{\mathrm{d}^3\mathbf{C}(t)}{\mathrm{d}t^3} = \mathbf{A}^3\mathbf{C}(t) + 3\mathbf{A}^2\mathbf{C}(t)\mathbf{A}^\top + 3\mathbf{A}\mathbf{C}(t)(\mathbf{A}^\top)^2 + \mathbf{C}(t)(\mathbf{A}^\top)^3$$

$$+ \mathbf{A}^2\mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right] + 2\mathbf{A}\mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right]\mathbf{A}^\top + \mathbb{E}^{[1]}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right](\mathbf{A}^\top)^2 + \mathbb{E}^{[1]}\left[\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right]$$
$$+ \mathbf{A}\mathbb{E}^{[1]}\left[\left[\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right] + \mathbb{E}^{[1]}\left[\left[\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right]\mathbf{A}^\top + \mathbb{E}^{[1]}\left[(\mathbb{L}^{[1]})^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{X}_t)\right].\right.\right.$$

Evaluating at time $t = 0$, we use that there is no randomness in the initial condition so that $\mathbf{C}(0) = \mathbf{0}$. Conditioning on $\mathbf{X}_0 = \mathbf{x}$, we get

$$\frac{\mathrm{d}\mathbf{C}(0)}{\mathrm{d}t} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}), \tag{62}$$
$$\frac{\mathrm{d}^2\mathbf{C}(0)}{\mathrm{d}t^2} = \mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}), \tag{63}$$
$$\frac{\mathrm{d}^3\mathbf{C}(0)}{\mathrm{d}t^3} = \mathbf{A}^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + 2\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^{2\top} \tag{64}$$
$$+ \mathbf{A}\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + \mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \mathbb{L}^{[1]2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}). \tag{65}$$

Finally, plugging back the previous results in (60), we get

$$\begin{aligned}\boldsymbol{\Omega}_h(\mathbf{x};\boldsymbol{\theta}) = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) &+ \frac{h^2}{2}\left(\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\right)\\ &+ \frac{h^3}{6}\left(\mathbf{A}^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + 2\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^{2\top}\right)\\ &+ \frac{h^3}{6}\left(\mathbf{A}\mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}) + \mathbb{L}^{[1]}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\mathbf{A}^\top + \mathbb{L}^{[1]2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x})\right) + \mathbf{R}(h^4,\mathbf{x}).\end{aligned} \tag{66}$$

4 Estimators

This section introduces the new S estimator, given a sample $\mathbf{X}_{0:t_N}$. Subsequently, we provide a brief overview of the EM, K, and LL estimators. The LL estimator is defined only for SDEs with constant diffusion coefficients. However, it is sometimes possible to transform the SDE using the Lamperti transform to obtain a new SDE with a constant diffusion coefficient. These SDEs are referred to as reducible diffusions [Aït-Sahalia, 2008]. We, therefore, present the LL estimator here because the Student Kramers oscillator treated in our simulation study in Section 5 is reducible, and the LL estimator can thus also be used.

4.1 Strang splitting estimator

The S splitting (57) is a nonlinear transformation of the Gaussian random variable

$$\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta}^{(1)});\boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,k}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta}^{(1)});\boldsymbol{\theta}).$$

We define

$$\mathbf{Z}_{t_k}(\boldsymbol{\theta}^{(1)}) \coloneqq \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\theta}^{(1)}) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta}^{(1)});\boldsymbol{\theta}^{(1)}) \tag{67}$$

and apply a change of variables to derive the following objective function

$$\begin{aligned}\mathcal{L}^{[\mathrm{S}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) = &\sum_{k=1}^{N}\left(\log(\det\boldsymbol{\Omega}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta}^{(1)});\boldsymbol{\theta})) + \mathbf{Z}_{t_k}(\boldsymbol{\theta}^{(1)})^\top\boldsymbol{\Omega}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\theta}^{(1)});\boldsymbol{\theta})^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\theta}^{(1)})\right)\\ &+ 2\sum_{k=1}^{N}\log|\det D\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k};\boldsymbol{\theta}^{(1)})|.\end{aligned} \tag{68}$$

The S estimator is then defined as

$$\widehat{\boldsymbol{\theta}}_N^{[\mathrm{S}]} \coloneqq \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{[\mathrm{S}]}\left(\mathbf{X}_{0:t_N};\boldsymbol{\theta}\right). \tag{69}$$

4.2 Euler-Maruyama estimator

The EM method uses a first-order Taylor expansion of the SDE (2)

$$\mathbf{X}_{t_k}^{[\mathrm{EM}]} \coloneqq \mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]} + h\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]};\boldsymbol{\theta}^{(1)}) + \boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]}, \tag{70}$$

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]} \overset{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}, h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ [Kloeden and Platen, 1992]. The transition density $p^{[\mathrm{EM}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ is Gaussian, so the pseudo-likelihood follows trivially.

4.3 Kessler estimator

The K estimator assumes Gaussian transition densities $p^{[\mathrm{K}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}}; \boldsymbol{\theta})$ with the true mean and covariance of the solution $\mathbf{X}$ [Kessler, 1997], that is, the density of SDE (2) is approximated by

$$p(\mathbf{x}, t) \approx \mathcal{N}(\mathbf{x}; \mathbf{m}(t), \mathbf{C}(t)),$$

where $\mathbf{m}(t)$ and $\mathbf{C}(t)$ are the mean and covariance of the solution of SDE (2). When the moments are unknown, they are approximated using the infinitesimal generator $\mathbb{L}$, that is, for sufficiently smooth functions $\boldsymbol{\phi}$

$$\mathbb{E}[\boldsymbol{\phi}(\mathbf{X}_{t+h}) \mid \mathbf{X}_t = \mathbf{x}] = \sum_{n=0}^{\infty} \frac{h^n}{n!} \mathbb{L}^n \boldsymbol{\phi}(\mathbf{x}). \tag{71}$$

By setting $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$ and $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top$ in (71), we obtain approximations of the first and second moments of SDE (2), respectively. Then, we find the covariance matrix indirectly by computing

$$\mathbb{E}[\mathbf{X}_{t+h}\mathbf{X}_{t+h}^\top \mid \mathbf{X}_t = \mathbf{x}] - \mathbb{E}[\mathbf{X}_{t+h} \mid \mathbf{X}_t = \mathbf{x}]\mathbb{E}[\mathbf{X}_{t+h} \mid \mathbf{X}_t = \mathbf{x}]^\top.$$

Then, the K second-order approximation becomes

$$\mathbf{X}_{t_k}^{[\mathrm{K}]} \coloneqq \boldsymbol{\mu}_h^{[\mathrm{K}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}; \boldsymbol{\theta}) + \boldsymbol{\xi}_h^{[\mathrm{K}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}; \boldsymbol{\theta}), \tag{72}$$

where $\boldsymbol{\xi}_h^{[\mathrm{K}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}; \boldsymbol{\theta}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h^{[\mathrm{K}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K}]}; \boldsymbol{\theta}))$, and

$$\boldsymbol{\mu}_h^{[\mathrm{K}]}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{x} + h\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)}) + \frac{h^2}{2}\left(D\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)})\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)}) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 \mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)})}{\partial x^{(i)} \partial x^{(j)}} [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}, \boldsymbol{\theta}^{(2)})]_{ij} \right),$$

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{K}]}(\mathbf{x}; \boldsymbol{\theta}) = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)}) + \frac{h^2}{2}\Bigg(& D\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)}) + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)}) D^\top \mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)}) \\ & + \sum_{i=1}^{d} \frac{\partial \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)})}{\partial x^{(i)}} F^{(i)}(\mathbf{x}; \boldsymbol{\theta}^{(1)}) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)})}{\partial x^{(i)} \partial x^{(j)}} [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\mathbf{x}; \boldsymbol{\theta}^{(2)})]_{ij} \Bigg).\end{aligned}$$

While this method is straightforward to implement, a few problems exist in practice.

First, obtaining closed-form formulas for $\boldsymbol{\mu}_h^{[\mathrm{K}]}$ and $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ becomes more complex for higher-order approximations. Thus, second-order approximation is the most commonly used in practice. However, for hypoelliptic systems, we need the third-order approximation of $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$. We do not provide a general formula for the third-order approximation using (65). Still, we calculate $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ up to order h^3 for Student Kramers oscillator in Section 5.

Second, $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ does not need to be positive definite. To avoid this problem, Kessler [1997] suggested Taylor expanding $\log\det \boldsymbol{\Omega}_h^{[\mathrm{K}]}$ and $(\boldsymbol{\Omega}_h^{[\mathrm{K}]})^{-1}$ in h.

4.4 Ozaki's local linearization estimator

Ozaki's LL method assumes an SDE with additive noise, that is

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}(\mathbf{X}_t; \boldsymbol{\theta}^{(1)})\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t. \tag{73}$$

If the starting SDE (2) is reducible, that is, if there is a bijective transformation between (2) and (73), then we can use the LL estimator. For SDE (73), the diffusion parameter $\boldsymbol{\theta}^{(2)}$ is the whole matrix $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. Here, we briefly present the estimator for SDE (73).

First, we approximate the drift of (2) between consecutive observations by a linear function and then we find the closed-form solution of the resulting linear SDE (see, [Jimenez et al., 1999]). The approximation becomes

$$\mathbf{X}_{t_k}^{[\mathrm{LL}]} \coloneqq \boldsymbol{\mu}_h^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}; \boldsymbol{\theta}) + \boldsymbol{\xi}_h^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}; \boldsymbol{\theta}), \tag{74}$$

where $\boldsymbol{\xi}_h^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}; \boldsymbol{\theta}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}; \boldsymbol{\theta}))$, and

$$\boldsymbol{\Omega}_h^{[\mathrm{LL}]}(\mathbf{x}; \boldsymbol{\theta}) \coloneqq \int_0^h e^{D\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)})(h-u)} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{D\mathbf{F}(\mathbf{x}; \boldsymbol{\theta}^{(1)})^\top (h-u)}\,\mathrm{d}u,$$

$$\boldsymbol{\mu}_h^{[\mathrm{LL}]}(\mathbf{x};\boldsymbol{\theta}) \coloneqq \mathbf{x} + \mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)}))\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)}) + (h\mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})) - \mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})))\mathbf{M}(\mathbf{x};\boldsymbol{\theta}),$$

$$\mathbf{R}_{h,i}(D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})) \coloneqq \int_0^h \exp(D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})u)u^i \,\mathrm{d}u, \qquad i = 0, 1,$$

$$\mathbf{M}(\mathbf{x};\boldsymbol{\theta}) \coloneqq \frac{1}{2}(\operatorname{Tr}\mathbf{H}_1(\mathbf{x};\boldsymbol{\theta}), \ldots, \operatorname{Tr}\mathbf{H}_d(\mathbf{x};\boldsymbol{\theta}))^\top, \quad \mathbf{H}_k(\mathbf{x};\boldsymbol{\theta}) \coloneqq \left[\frac{\partial^2 F^{(k)}(\mathbf{x};\boldsymbol{\theta}^{(1)})}{\partial x^{(i)}\partial x^{(j)}}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{ij}\right]_{i,j=1}^{d}.$$

We can efficiently compute $\mathbf{R}_{h,i}$ and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ using formulas from [Van Loan, 1978]. For more details, see [Pilipovic et al., 2024a]. Thus, $p^{[\mathrm{LL}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ is Gaussian and standard likelihood inference applies.

While this method usually performs best in numerical studies, it is the slowest due to $e^{D\mathbf{F}(\mathbf{x};\boldsymbol{\theta}^{(1)})h}$ in $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$. Moreover, it can only be applied to reducible diffusions; even if possible, reducing a diffusion can be complicated.

5 Simulation study

In this section, we conduct a simulation study of the student Kramers oscillator (24), demonstrating its theoretical aspects and comparing our proposed estimators against those based on the EM, K, and LL approximations. The choice of these three methods for comparison is motivated by their widespread use and established performance: the EM estimator is commonly applied in practice, our proposed method generalizes the K approximation, and the LL estimator is recognized as a state-of-the-art method as shown in Pilipovic et al. [2024a,b].

The true parameters for the simulation are set as follows: $\eta_0 = 30$, $a_0 = -125$, $b_0 = 40$, $c_0 = 150$, $d_0 = -20$, $\alpha_0 = 20$, $\beta_0 = -8$, and $\gamma_0 = 1280.8$. The squared diffusion coefficient is thus $(\Sigma(x,v))_{22}^2 = 20(v-0.2)^2 + 1280$ and strictly positive for all $v \in \mathbb{R}$.

We begin by outlining the estimators tailored for the Student Kramers oscillator. We then detail the simulation procedure and describe the optimization process implemented in the `R` programming language R Core Team [2022]. Finally, we present and interpret the results of our study.

5.1 Strang splitting estimator

To define S estimators based on the Strang splitting scheme, we first split SDE (24) as follows

$$\mathrm{d}\begin{bmatrix}X_t\\V_t\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 1\\ c & -\eta\end{bmatrix}}_{\mathbf{A}}\left(\begin{bmatrix}X_t\\V_t\end{bmatrix} - \underbrace{\begin{bmatrix}-d/c\\0\end{bmatrix}}_{\mathbf{b}}\right)\mathrm{d}t + \underbrace{\begin{bmatrix}0\\ aX_t^3 + bX_t^2\end{bmatrix}}_{\mathbf{N}(X_t,V_t)}\mathrm{d}t + \begin{bmatrix}0 & 0\\ 0 & \sqrt{\alpha V_t^2 + \beta V_t + \gamma}\end{bmatrix}\mathrm{d}W_t.$$

The nonlinear ODE driven by $\mathbf{N}(x,v)$ has a trivial solution where x is a constant. We incorporate these components into the objective function (68) to obtain the S estimator.

5.2 Euler-Maruyama estimator

For the Kramers oscillator (24), the EM transition distribution is given by

$$\begin{bmatrix}X_{t_k}\\V_{t_k}\end{bmatrix} \mid \begin{bmatrix}X_{t_{k-1}}\\V_{t_{k-1}}\end{bmatrix} = \begin{bmatrix}x\\v\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}x + hv\\ v - h(\eta v + U'(x))\end{bmatrix}, \begin{bmatrix}0 & 0\\ 0 & h(\alpha v^2 + \beta v + \gamma)\end{bmatrix}\right).$$

Due to the ill-conditioned variance of this discretization, we use an estimator that relies solely on the marginal likelihood of V_{t_k}

$$\widehat{\boldsymbol{\theta}}_N^{[\mathrm{EM}]} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{N}\left(\log(h(\alpha V_{t_{k-1}}^2 + \beta V_{t_{k-1}} + \gamma)) + \frac{(V_{t_k} - V_{t_{k-1}} + h(\eta V_{t_{k-1}} + U'(X_{t_{k-1}})))^2}{h(\alpha V_{t_{k-1}}^2 + \beta V_{t_{k-1}} + \gamma)}\right).$$

5.3 Kessler estimator

From (72), the transition distribution of the K approximation is

$$\begin{bmatrix}X_{t_k}\\V_{t_k}\end{bmatrix} \mid \begin{bmatrix}X_{t_{k-1}}\\V_{t_{k-1}}\end{bmatrix} = \begin{bmatrix}x\\v\end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}_h^{[\mathrm{K}]}(x,v), \boldsymbol{\Omega}_h^{[\mathrm{K}]}(x,v)\right),$$

where

$$\boldsymbol{\mu}_h^{[\mathrm{K}]}(x,v) = \begin{bmatrix} x + hv - \frac{h^2}{2}(\eta v + U'(x)) \\ v - h(\eta v + U'(x)) + \frac{h^2}{2}(\eta^2 v + \eta U'(x) - U''(x)v) \end{bmatrix}. \tag{75}$$

Here, we cannot use $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ from (72) directly because it is singular. Instead, we use formula (65) to add one more order of h to $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ and obtain

$$\boldsymbol{\Omega}_h^{[\mathrm{K}]}(x,v)^{(1,1)} = \frac{h^3}{3}\left(v\left(v\alpha+\beta\right)+\gamma\right), \tag{76}$$

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{K}]}(x,v)^{(1,2)} &= \frac{h^2}{2}\left(v\left(v\alpha+\beta\right)+\gamma\right) + \frac{h^3}{6}\left(2bvx^2\alpha + 2avx^3\alpha + v^2\alpha^2 + bx^2\beta + ax^3\beta + v\alpha\beta + d\left(2v\alpha+\beta\right)\right) \\ &\quad + \frac{h^3}{6}\left(cx\left(2v\alpha+\beta\right) + \alpha\gamma - \left(5v^2\alpha + 4v\beta + 3\gamma\right)\eta\right),\end{aligned} \tag{77}$$

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{K}]}(x,v)^{(2,2)} &= h\left(v\left(v\alpha+\beta\right)+\gamma\right) + \frac{h^2}{2}\left(2bvx^2\alpha + 2avx^3\alpha + v^2\alpha^2 + bx^2\beta + ax^3\beta + v\alpha\beta + d\left(2v\alpha+\beta\right)\right) \\ &\quad + \frac{h^2}{2}\left(cx\left(2v\alpha+\beta\right) + \alpha\gamma - \left(4v^2\alpha + 3v\beta + 2\gamma\right)\eta\right) \\ &\quad + \frac{h^3}{6}\left(2d^2\alpha + 8bv^2x\alpha + 12av^2x^2\alpha\right) \\ &\quad + \frac{h^3}{6}\left(2b^2x^4\alpha + 4abx^5\alpha + 2a^2x^6\alpha + 2bvx^2\alpha^2 + 2avx^3\alpha^2 + v^2\alpha^3 + 6bvx\beta + 9avx^2\beta + bx^2\alpha\beta\right) \\ &\quad + \frac{h^3}{6}\left(ax^3\alpha\beta + v\alpha^2\beta + d\alpha\left(4x(c+x(b+ax)) + 2v\alpha + \beta\right) + 4bx\gamma + 6ax^2\gamma + \alpha^2\gamma + 2c^2x^2\alpha\right) \\ &\quad + \frac{h^3}{6}\left(-d\left(10v\alpha+3\beta\right)\eta - \left(bx^2\left(10v\alpha+3\beta\right) + ax^3\left(10v\alpha+3\beta\right) + \alpha\left(6v^2\alpha + 5v\beta + 4\gamma\right)\right)\eta\right) \\ &\quad + \frac{h^3}{6}c\left(4v^2\alpha + 4x^3(b+ax)\alpha + 3v\beta + x\alpha\beta + 2\gamma + 2vx\alpha(\alpha - 5\eta) - 3x\beta\eta\right) \\ &\quad + \frac{h^3}{6}\left(12v^2\alpha + 7v\beta + 4\gamma\right)\eta^2.\end{aligned} \tag{78}$$

Although we could trim $\boldsymbol{\Omega}_h^{[\mathrm{K}]}$ to include only the lowest order terms of h, leading to an estimator based on the strong order 1.5 scheme in Ditlevsen and Samson [2019], we use the full approximation given by formulas (76)-(78). The K estimator is defined as

$$\widehat{\boldsymbol{\theta}}_N^{[\mathrm{K}]} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{N} \left(\log\det \boldsymbol{\Omega}_h^{[\mathrm{K}]}(\mathbf{Y}_{t_{k-1}}) + (\mathbf{Y}_{t_k} - \boldsymbol{\mu}_h^{[\mathrm{K}]}(\mathbf{Y}_{t_{k-1}}))^\top \boldsymbol{\Omega}_h^{[\mathrm{K}]}(\mathbf{Y}_{t_{k-1}})^{-1}(\mathbf{Y}_{t_k} - \boldsymbol{\mu}_h^{[\mathrm{K}]}(\mathbf{Y}_{t_{k-1}}))\right),$$

where $Y_{t_k} = (X_{t_k}, V_{t_k})$ for $k = 0, 1, \ldots, N$.

5.4 Ozaki's local linearization estimator

To derive the LL estimator, we first need to transform the SDE (24) to one with a constant diffusion coefficient. We achieve this by applying the Lamperti transform ψ, similar to the approach in Nagahara [1996] for a one-dimensional nonlinear student-type Pearson diffusion. We have

$$U_t = \psi(V_t) = \int^{V_t} \frac{dv}{\sqrt{\alpha v^2 + \beta v + \gamma}} = \frac{1}{\sqrt{\alpha}} \operatorname{arcsinh}\left(\frac{2\alpha V_t + \beta}{\sqrt{4\alpha\gamma - \beta^2}}\right).$$

Since we assume that $\alpha > 0$ and $4\alpha\gamma - \beta^2 \leq 0$, then

$$V_t = \psi^{-1}(U_t) = \frac{\sqrt{4\alpha\gamma - \beta^2}\sinh(\sqrt{\alpha}U_t) - \beta}{2\alpha}.$$

We transform SDE (24) by applying Itô's lemma to $\widetilde{\boldsymbol{\psi}}(X_t, V_t) = (X_t, \psi(V_t))$

$$\mathrm{d}X_t = F^{(1)}(\widetilde{\boldsymbol{\psi}}^{-1}(X_t, U_t))\,\mathrm{d}t,$$

$$\mathrm{d}U_t = \left(\frac{\partial \psi}{\partial v}(\widetilde{\boldsymbol{\psi}}^{-1}(X_t, U_t)) F^{(2)}(\widetilde{\boldsymbol{\psi}}^{-1}(X_t, U_t)) + \frac{1}{2} \frac{\partial^2 \psi}{\partial v^2}(\widetilde{\boldsymbol{\psi}}^{-1}(X_t, U_t) \sigma^2(\psi^{-1}(U_t))) \right) \mathrm{d}t + \mathrm{d}W_t.$$

The transformed SDE becomes

$$\begin{aligned}
\mathrm{d}X_t &= \left(\sqrt{\gamma - \frac{\beta^2}{4\alpha}} \frac{\sinh(\sqrt{\alpha} U_t)}{\sqrt{\alpha}} - \frac{\beta}{2\alpha} \right) \mathrm{d}t, \\
\mathrm{d}U_t &= \left(- \left(\eta + \frac{\alpha}{2} \right) \frac{\tanh(\sqrt{\alpha} U_t)}{\sqrt{\alpha}} + \frac{\frac{\beta}{2\alpha}\eta - U'(X_t)}{\sqrt{\gamma - \frac{\beta^2}{4\alpha}} \cosh(\sqrt{\alpha} U_t)} \right) \mathrm{d}t + \mathrm{d}W_t.
\end{aligned}$$

To implement the LL estimator, we need to find $D\mathbf{F}(x, u)$ for the corresponding drift function $\mathbf{F}$, which is

$$D\mathbf{F}(x,u) = \begin{bmatrix} 0 & \sqrt{\gamma - \frac{\beta^2}{4\alpha}} \cosh(\sqrt{\alpha} U_t) \\ -\frac{U''(x)}{\sqrt{\gamma - \frac{\beta^2}{4\alpha}} \cosh(\sqrt{\alpha} U_t)} & -\frac{\eta + \frac{\alpha}{2}}{\cosh^2(\sqrt{\alpha} U_t)} - \frac{\frac{\beta}{2\alpha}\eta - U'(X_t)}{\sqrt{\gamma - \frac{\beta^2}{4\alpha}} \cosh(\sqrt{\alpha} U_t)} \sqrt{\alpha} \tanh(\sqrt{\alpha} U_t) \end{bmatrix}.$$

Then, the LL estimator for SDE (24) is given by

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_N^{[\mathrm{LL}]} = \arg\min_{\boldsymbol{\theta}} \Bigg\{ & \sum_{k=1}^{N} (\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_k}) - \boldsymbol{\mu}_h^{[\mathrm{LL}]}(\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_{k-1}})))^\top \boldsymbol{\Omega}_h^{[\mathrm{LL}]}(\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_{k-1}}))^{-1} (\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_k}) - \boldsymbol{\mu}_h^{[\mathrm{LL}]}(\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_{k-1}}))) \\
& + \sum_{k=1}^{N} \log\det \boldsymbol{\Omega}_h^{[\mathrm{LL}]}(\widetilde{\boldsymbol{\psi}}(\mathbf{Y}_{t_{k-1}})) + \sum_{k=1}^{N} \log(\psi'(V_{t_k})^2) \Bigg\},
\end{aligned} \tag{79}$$

where $\boldsymbol{\mu}_h^{[\mathrm{LL}]}$ and $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$ are defined in (74). The last term in (79) arises due to the change of variable formula.

5.5 Trajectory simulation

We simulate sample paths using the Milstein discretization scheme [Kloeden and Platen, 1992] with a step size of $h^{\mathrm{sim}} = 0.0001$ to ensure high accuracy. To reduce discretization errors, we sub-sample the path at wider intervals to obtain time steps $h = 0.005$, $h = 0.01$, and $h = 0.02$. We fix the time interval length to $T = 50$ and adjust the sample sizes for each h accordingly. We repeat the simulations to obtain 1000 data sets.

5.6 Optimization in `R`

For optimizing the objective functions, we follow the approach in Pilipovic et al. [2024a] using the `R` package `torch` [Falbel and Luraschi, 2024], which supports automatic differentiation. The optimization employs the resilient backpropagation algorithm, `optim_rprop`. We use the default hyperparameters and limit the optimization iterations to 1000. The convergence criterion is set to a precision of 10^{-5} for the difference between estimators in consecutive iterations. The initial parameter values are set to $(50, -200, 10, 100, 10, 30, -5, 1000)$.

5.7 Results

Figure 2 presents the simulation study results. It shows the distributions of the four normalized estimators EM, K, LL, and S. The time interval T is fixed at 50, and the step size h varies from 0.02 to 0.005, where each h is half the size of the previous. Consequently, the sample size doubles from $N = 2500$ to $N = 10000$ to keep the observation interval constant. Figure 2A presents the estimators of the parameters of the potential $U(x)$, while Figure 2B focuses on the damping and diffusion parameters. All the estimators perform well for the potential parameters, with LL being the best for the largest h. In contrast, all methods except S show significant bias for the damping and diffusion parameters.

6 Conclusion

This paper extends the Pearson diffusion framework to multivariate models with quadratic diffusion coefficients and nonlinear drift. We provided a concrete example with the Student Kramers oscillator, a hypoelliptic two-dimensional model with nonlinear drift and a t-distribution type noise, proving the existence, uniqueness, and invariant measure of its solution. We also briefly described two other examples: a coupled multivariate Wright-Fisher diffusion of

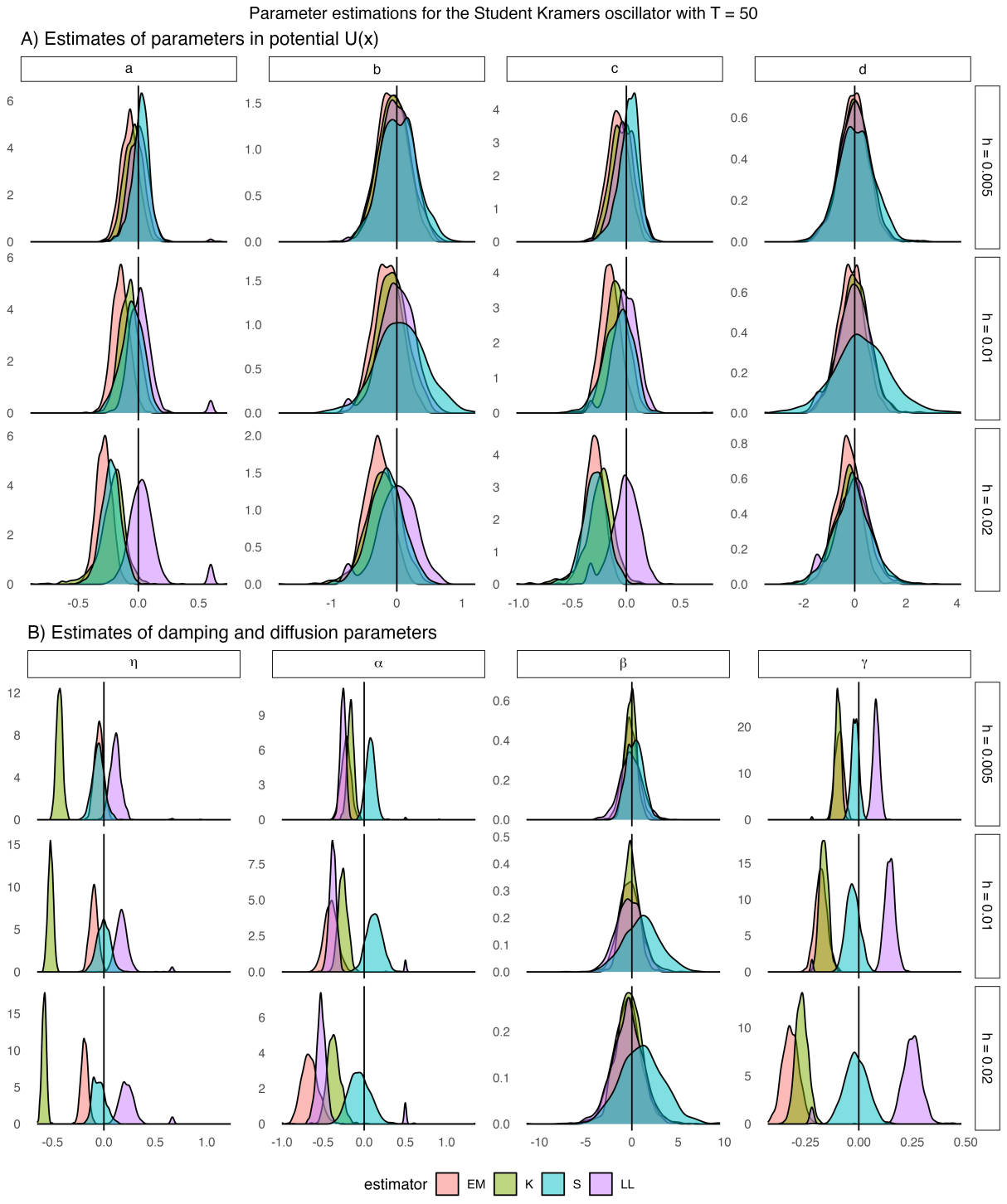

Figure 2: Normalized distributions of parameter estimation errors $(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \oslash \boldsymbol{\theta}_0$ based on 1000 simulated datasets with a fixed time interval of length $T = 50$. Different colors indicate the type of estimator. Each column corresponds to a different parameter, and each row corresponds to a different value of h, and consequently N. **A)** Distributions of the estimators of the potential $U(x)$. **B)** Distributions of the estimators of the damping parameter η and diffusion parameters.

arbitrary dimension (depending on the number of loci and alleles at each locus) and the stochastic SIR model. Through comprehensive simulations, we illustrated the performance of our estimator, which is particularly suitable for diffusion parameters.

This work's implications are useful for the field of stochastic modeling. By broadening the scope of Pearson diffusions to encompass multivariate and nonlinear drift scenarios, we enable more accurate and flexible modeling of complex dynamical systems. Our estimator's ability to handle these scenarios while maintaining computational efficiency positions it as a valuable tool for researchers and practitioners dealing with multivariate data.

The Student Kramers oscillator example showcases the potential of our approach in real-world scenarios, offering insights into the behavior of stochastic systems under various conditions.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)", and from Novo Nordisk Foundation NNF20OC0062958.

References

Y. Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906 – 937, 2008. doi:10.1214/009053607000000622. URL https://doi.org/10.1214/009053607000000622.

E. Aurell, M. Ekeberg, and T. Koski. On a multilocus wright-fisher model with mutation and a svirezhev-shahshahani gradient-like selection dynamics. *arXiv: Probability*, 2019. URL https://api.semanticscholar.org/CorpusID:209370499.

B. M. Bibby and M. Sørensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1 (1/2):17–39, 1995. ISSN 13507265. URL http://www.jstor.org/stable/3318679.

F. Carbonell, J. Jímenez, and L. Pedroso. Computing multiple integrals involving matrix exponentials. *Journal of Computational and Applied Mathematics*, 213:300–305, 03 2008. doi:10.1016/j.cam.2007.01.007.

E. de Klerk. *Aspects of semidefinite programming: Interior point algorithms and selected applications*. Number 65 in Applied optimization, ISSN 1384-6485. Kluwer Academic Publishers, Netherlands, 2002. ISBN 1402005474. Pagination: xvi, 283.

S. Ditlevsen and A. Samson. Hypoelliptic diffusions: filtering and inference from complete and partial observations. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 81(2):361–384, 2019.

D. Falbel and J. Luraschi. *torch: Tensors and Neural Networks with 'GPU' Acceleration*, 2024. URL https://torch.mlverse.org/docs. R package version 0.13.0, https://github.com/mlverse/torch.

M. Favero, H. Hult, and T. Koski. A dual process for the coupled wright-fisher diffusion. *Journal of Mathematical Biology*, 82(6), 2021. doi:10.1007/s00285-021-01555-9. URL https://doi.org/10.1007/s00285-021-01555-9.

J. L. Forman and M. Sørensen. The pearson diffusions: A class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, 35(3):438–465, 2008. doi:https://doi.org/10.1111/j.1467-9469.2007.00592.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2007.00592.x.

A. Gloter. Parameter Estimation for a Discretely Observed Integrated Diffusion Process. *Scandinavian Journal of Statistics*, 33(1):83–104, 2006.

A. Gloter and N. Yoshida. Adaptive and non-adaptive estimation for degenerate diffusion processes, 2020.

A. Gloter and N. Yoshida. Adaptive estimation for degenerate diffusion processes. *Electronic Journal of Statistics*, 15 (1):1424 – 1472, 2021.

M. Hairer and N. Pillai. Regularity of laws and ergodicity of hypoelliptic sdes driven by rough paths. *The Annals of Probability*, 41, 04 2011. doi:10.1214/12-AOP777.

L. Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119(none):147 – 171, 1967. doi:10.1007/BF02392081. URL https://doi.org/10.1007/BF02392081.

A. R. Humphries and A. M. Stuart. *Deterministic and random dynamical systems: theory and numerics*, pages 211–254. Springer Netherlands, Dordrecht, 2002. ISBN 978-94-010-0510-4. URL https://doi.org/10.1007/978-94-010-0510-4_6.

A. Hurn, K. Lindsay, and A. McClelland. A quasi-maximum likelihood method for estimating the parameters of multivariate diffusions. *Journal of Econometrics*, 172(1):106–126, 2013. ISSN 0304-4076. doi:https://doi.org/10.1016/j.jeconom.2012.09.002. URL https://www.sciencedirect.com/science/article/pii/S0304407612002187.

Y. Iguchi and A. Beskos. Parameter inference for hypo-elliptic diffusions under a weak design condition, 2023.

J. Jimenez, I. Shoji, and T. Ozaki. Simulation of Stochastic Differential Equations Through the Local Linearization Method. A Comparative Study. *J. Stat. Phys.*, 94:587–602, 02 1999.

M. Kessler. Estimation of an Ergodic Diffusion from Discrete Observations. *Scand. J. Stat.*, 24(2):211–229, 1997. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616449.

M. Kessler and M. Sørensen. Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299 – 314, 1999.

P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1992. ISBN 9783540540625. doi:10.1007/978-3-662-12616-5. URL https://books.google.dk/books?id=BCvtssom1CMC.

G. Leonenko and T. Phillips. High-order approximation of pearson diffusion processes. *Journal of Computational and Applied Mathematics*, 236(11):2853–2868, 2012. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2012.01.022. URL https://www.sciencedirect.com/science/article/pii/S0377042712000337.

J. Magnus. *Matrix Differential Calculus with Applications in Statistics and Econometrics: Third edition and E-book*. John Wiley, 2019.

S. P. Meyn and R. L. Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993. URL http://www.jstor.org/stable/1427522.

Y. Nagahara. Non-gaussian distribution for stock returns and related stochastic differential equation. *Financial Engineering and the Japanese Markets*, 3(2):121–149, 1996. doi:10.1007/BF00868083. URL https://doi.org/10.1007/BF00868083.

P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics*, 52:842–867, 2024a.

P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting for parametric inference in second-order stochastic differential equations, 2024b. Under review in JRSS - series B.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

S. O. Rasmussen, M. Bigler, S. P. Blockley, T. Blunier, S. L. Buchardt, H. B. Clausen, I. Cvijanovic, D. Dahl-Jensen, S. J. Johnsen, H. Fischer, V. Gkinis, M. Guillevic, W. Z. Hoek, J. J. Lowe, J. B. Pedro, T. Popp, I. K. Seierstad, J. P. Steffensen, A. M. Svensson, P. Vallelonga, B. M. Vinther, M. J. Walker, J. J. Wheatley, and M. Winstrup. A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: refining and extending the INTIMATE event stratigraphy. *Quaternary Science Reviews*, 106:14–28, 2014.

M. Sørensen. Efficient estimation for ergodic diffusions sampled at high frequency. CREATES Research Papers 2007-46, Department of Economics and Business Economics, Aarhus University, Jan. 2008. URL https://ideas.repec.org/p/aah/create/2007-46.html.

M. V. Tretyakov and Z. Zhang. A Fundamental Mean-Square Convergence Theorem for SDEs with Locally Lipschitz Coefficients and Its Applications. *SIAM Journal on Numerical Analysis*, 51(6):3135–3162, 2013. URL https://doi.org/10.1137/120902318.

M. Uchida and N. Yoshida. Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications*, 122(8):2885–2924, 2012.

C. Van Loan. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23(3):395–404, 1978. doi:10.1109/TAC.1978.1101743.

A Appendix to Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations with Additive Noise

SUPPLEMENT TO "PARAMETER ESTIMATION IN NONLINEAR MULTIVARIATE STOCHASTIC DIFFERENTIAL EQUATIONS BASED ON SPLITTING SCHEMES"

BY PREDRAG PILIPOVIC, ADELINE SAMSON AND SUSANNE DITLEVSEN

Section S1 provides proofs for all propositions, lemmas, and theorems. References to equations and sections that do not begin with "S" refer to the main paper. The properties necessary for subsequent proofs are outlined in Section S2. These properties encompass Grönwall's and Rosenthal's inequalities, as well as Central Limit Theorems for a sum of triangular arrays. In Section S3 we discuss in more detail the LL and HE estimators.

If not stated, we assume the parameters are the true ones, and the expectations are taken under the probability measure. Occasionally, we omit explicit parameter notation to enhance clarity. For instance, $\mathbb{E}$ implicitly denotes $\mathbb{E}_{\boldsymbol{\theta}}$.

S1. Proofs. In Section S1.1, we provide the proof for the Lie-Trotter splitting (LT), while Section S1.2 contains the proofs for the Strang splitting (S). Proof of L^p convergence of the splitting scheme is in Section S1.3. The proof of Lemma 4.1 is in Section S1.4. Additionally, the proofs of moment bounds are detailed in Section S1.5. Sections S1.6 and S1.7 present proofs of consistency and asymptotic normality of the estimators, respectively.

S1.1. *Proof for the Lie-Trotter splitting.*

PROOF OF PROPOSITION 3.4. To establish the proposition, we compare the actual first moment of the solution to SDE (1), as obtained from Lemma 2.1, with the moment derived through Taylor expansion of the LT approximation. First, we prove the proposition for LT splitting as defined in the paper. By performing the Taylor expansion of $\mathbb{E}[\Phi_h^{[\mathrm{LT}]}(\mathbf{x})] = \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}((\mathbf{x})) = e^{\mathbf{A}h}\boldsymbol{f}_h(\mathbf{x}) + (\mathbf{I} - e^{\mathbf{A}h})\mathbf{b}$ around $h = 0$, using Proposition 2.2, we arrive at:

$$\text{(S1)} \quad \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}((\mathbf{x})) = \mathbf{x} + h(\mathbf{A}(\mathbf{x}-\mathbf{b}) + \mathbf{N}(\mathbf{x})) + \frac{h^2}{2}(\mathbf{A}^2(\mathbf{x}-\mathbf{b}) + 2\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x})) + \mathbf{R}(h^3, \mathbf{x}).$$

The coefficient of h in (S1) is $\mathbf{F}(\mathbf{x})$, which aligns with the coefficient of h in the theoretical moment of the solution to (1) as provided in Lemma 2.1. However, in Lemma 2.1, $\boldsymbol{\Sigma}$ appears in the coefficient of h^2, while it does not appear in (S1). Consequently, to achieve the order of convergence $\mathbf{R}(h^3, \mathbf{x})$, we need to make the following unrealistic assumption.

$$\text{(SA)} \quad \sum_{i=1}^d \sum_{j=1}^d [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{ij} \partial_{ij}^2 F^{(i)}(\mathbf{x}) = 0, \quad \text{for all } k = 1, \ldots, d.$$

Upon comparing expression (S1) with the true moments of the SDE solution under Assumption (SA), we arrive at $(D\mathbf{F}(\mathbf{x}))\mathbf{N}(\mathbf{x}) = (D\mathbf{N}(\mathbf{x}))\mathbf{F}(\mathbf{x})$ to ensure equality of the coefficient at order h^2. However, the last equation holds true for all $\mathbf{x} \in \mathbb{R}^d$ only when $\mathbf{N}$ is linear. Therefore, achieving the order $\mathbf{R}(h^3, \mathbf{x})$ one-step convergence is feasible only if SDE (1) is linear.

We now aim to show that changing the composition order within the LT does not affect the one-step convergence order. To demonstrate this, we define the reversed LT:

$$\mathbf{X}_{t_k}^{[\mathrm{LT}]^\star} := \Phi_h^{[\mathrm{LT}]^\star}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]^\star}) = (\Phi_h^{[2]} \circ \Phi_h^{[1]})(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]^\star}) = \boldsymbol{f}_h(\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]^\star}) + \boldsymbol{\xi}_{h,k}).$$

We compute $\mathbb{E}[\boldsymbol{f}_h(\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}) + \boldsymbol{\xi}_{h,k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]$, which is equivalent to calculating $\mathbb{E}[\boldsymbol{f}_h(\mathbf{X}_{t_k}^{[1]}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}] = \mathbb{E}[\boldsymbol{f}_h(\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}^{[1]}) + \boldsymbol{\xi}_{h,k}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}]$. The infinitesimal generator $L_{[1]}$ for SDE (3) is defined on the class of sufficiently smooth functions $g : \mathbb{R}^d \to \mathbb{R}$ by $L_{[1]}g(\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_g(\mathbf{x}))$. This yields:

$$\text{(S2)} \qquad \mathbb{E}[g(\mathbf{X}_{t_k}^{[1]}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}] = g(\mathbf{x}) + hL_{[1]}g(\mathbf{x}) + \frac{h^2}{2}L_{[1]}^2 g(\mathbf{x}) + R(h^3, \mathbf{x}).$$

We apply (S2) to $g(\mathbf{x}) = f_h^{(i)}(\mathbf{x})$. For calculating $L_{[1]}f_h^{(i)}(\mathbf{x})$ and $L_{[1]}^2 f_h^{(i)}(\mathbf{x})$, we use the Taylor expansion of $\boldsymbol{f}_h(\mathbf{x})$ around $h = 0$, as provided in Proposition 2.2. The partial derivatives are $\partial_j f_h^{(i)}(\mathbf{x}) = \delta_j^i +$

$h\partial_j N^{(i)}(\mathbf{x}) + R(h^2,\mathbf{x})$ and $\partial^2_{jk} f_h^{(i)}(\mathbf{x}) = h\partial^2_{jk} N^{(i)}(\mathbf{x}) + R(h^2,\mathbf{x})$. Since $L_{[1]} f_h^{(i)}(\mathbf{x})$ is multiplied by h in (S2), we only need to calculate it up to order $R(h,\mathbf{x})$. We have $L_{[1]} f_h^{(i)}(\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + h(\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla N^{(i)}(\mathbf{x}) + \frac{h}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})) + R(h^2,\mathbf{x})$. Similarly, we have $L^2_{[1]} f_h^{(i)}(\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla (\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + R(h,\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \mathbf{A}^{(i)} + R(h,\mathbf{x})$. Thus,

(S3)

$$\begin{aligned}
&\mathbb{E}[f_h^{(i)}(\mathbf{X}^{[1]}_{t_{k-1}}) \mid \mathbf{X}^{[1]}_{t_{k-1}} = \mathbf{x}] = x^{(i)} + hN^{(i)}(\mathbf{x}) + \frac{h^2}{2}(\mathbf{N}(\mathbf{x}))^\top \nabla N^{(i)}(\mathbf{x}) \\
&+ h(\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + h^2(\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla N^{(i)}(\mathbf{x}) + \frac{h^2}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})) + \frac{h^2}{2}(\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \mathbf{A}^{(i)} + R(h^3,\mathbf{x}) \\
&= x^{(i)} + hF^{(i)}(\mathbf{x}) + \frac{h^2}{2}((\mathbf{F}(\mathbf{x}))^\top (\nabla N^{(i)}(\mathbf{x})) + (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla F^{(i)}(\mathbf{x}) + \operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x}))) + R(h^3,\mathbf{x}).
\end{aligned}$$

Using that $F^{(i)}(\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + N^{(i)}(\mathbf{x})$, $\frac{\partial F^{(i)}(\mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A}^{(i)})^\top + \nabla N^{(i)}(\mathbf{x})$ and $\mathbf{H}_{F^{(i)}}(\mathbf{x}) = \mathbf{H}_{N^{(i)}}(\mathbf{x})$, the expectation of the true process rewrites as:

$$\begin{aligned}
\mathbb{E}[X^{(i)}_{t_k} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] &= x^{(i)} + hF^{(i)}(\mathbf{x}) \\
&+ \frac{h^2}{2}((\mathbf{N}(\mathbf{x}))^\top \nabla F^{(i)}(\mathbf{x}) + (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla F^{(i)}(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x}))) + R(h^3,\mathbf{x}).
\end{aligned}$$

The final equation coincides with equation (S3) only up to order $R(h,\mathbf{x})$. Despite the reversed LT has the term with $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ at the order h^2, the coefficients do not match. Thus, to obtain order $R(h^2,\mathbf{x})$, the condition $(\mathbf{N}(\mathbf{x}))^\top \nabla F^{(i)}(\mathbf{x}) - \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})) = (\mathbf{F}(\mathbf{x}))^\top \nabla N^{(i)}(\mathbf{x})$, must hold for all $i = 1,\dots,d$. Given Assumption (SA), the condition for achieving a higher one-step convergence order remains equivalent to the case of the original LT. □

S1.2. *Proof for the Strang Splitting.* We continue employing the Taylor expansion to establish the numerical properties of the S approximation. To begin, we introduce a helpful Lemma S1.1 regarding the approximation of the composition of the mean function $\boldsymbol{\mu}_h$ and the nonlinear solution $\boldsymbol{f}_{h/2}$. Lemma S1.1 expands $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x}))$ around $h = 0$ in various ways, each retaining the crucial terms necessary for the subsequent proofs.

LEMMA S1.1. *For the mean function $\boldsymbol{\mu}_h$ and the nonlinear solution $\boldsymbol{f}_{h/2}$ the following three identities hold:*

1. $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \boldsymbol{f}_{h/2}(\mathbf{x}) + h\mathbf{A}(\mathbf{x}-\mathbf{b}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3,\mathbf{x})$
2. $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3,\mathbf{x})$.
3. $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \mathbf{x} + h\mathbf{A}(\mathbf{x}-\mathbf{b}) + \frac{h}{2}\mathbf{N}(\mathbf{x}) + \frac{h^2}{2}(\mathbf{A}^2(\mathbf{x}-\mathbf{b}) + \mathbf{A}\mathbf{N}(\mathbf{x}) + \frac{1}{4}(D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x})) + \mathbf{R}(h^3,\mathbf{x})$.

PROOF. We prove only the first two identities, as the last one follows the same reasoning. Utilizing the definition of $\boldsymbol{\mu}_h$, its Taylor expansion, and the expansion of $\boldsymbol{f}_{h/2}$, we obtain: $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = (\mathbf{I} + h\mathbf{A} + \frac{h^2}{2}\mathbf{A}^2)(\boldsymbol{f}_{h/2}(\mathbf{x})-\mathbf{b}) + \mathbf{b} + \mathbf{R}(h^3,\mathbf{x}) = \boldsymbol{f}_{h/2}(\mathbf{x}) + h\mathbf{A}(\mathbf{x}-\mathbf{b}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3,\mathbf{x})$, which concludes the first part.

For the second part, Proposition 2.2 gives $\boldsymbol{f}_{h/2}(\mathbf{x}) - \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) = h\mathbf{N}(\mathbf{x}) + \mathbf{R}(h^3,\mathbf{x})$. This leads to: $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3,\mathbf{x})$. □

PROOF OF PROPOSITION 3.6. We begin by introducing a new function of $\mathbf{x}$, arising from the third property of Lemma S1.1:

$$\mathbf{Q}_h(\mathbf{x}) := \frac{h}{2}(2\mathbf{A}(\mathbf{x}-\mathbf{b}) + \mathbf{N}(\mathbf{x})) + \frac{h^2}{8}(4\mathbf{A}^2(\mathbf{x}-\mathbf{b}) + 4\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x})).$$

Then, for a generic random vector $\mathbf{X}$ we use Proposition 2.2 and Lemma S1.1 to write:

$$\boldsymbol{f}_{h/2}(\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X})) + \boldsymbol{\xi}_h) = \boldsymbol{f}_{h/2}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \mathbf{R}(h^3,\mathbf{X}))$$

$$= \mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \frac{h}{2}\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)$$

$$+ \frac{h^2}{8}(D\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h))\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) + \mathbf{R}(h^3, \mathbf{X}). \tag{S4}$$

Consequently, we expand:

$$\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) = \mathbf{N}(\mathbf{X}) + (D\mathbf{N}(\mathbf{X}))(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)$$

$$+ \frac{1}{2}[(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)]_{i=1}^d + \mathbf{R}(h^2, \mathbf{X}). \tag{S5}$$

The term $[\mathbf{Q}_h(\mathbf{X})^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\mathbf{Q}_h(\mathbf{X})]_{i=1}^d$ is $\mathbf{R}(h^2, \mathbf{X})$, while the terms with only one $\boldsymbol{\xi}_h$ have zero means. Thus,

$$\mathbb{E}[\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}] = \mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x}) + \frac{1}{2}[\mathbb{E}[\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h \mid \mathbf{X} = \mathbf{x}]]_{i=1}^d + \mathbf{R}(h^2, \mathbf{x}). \tag{S6}$$

Lastly, we compute:

$$\mathbb{E}[\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathrm{tr}(\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}] = \mathrm{tr}(\mathbf{H}_{N^{(i)}}(\mathbf{X})\mathbb{E}[\boldsymbol{\xi}_h \boldsymbol{\xi}_h^\top])$$

$$= \sum_{j,k=1}^d \partial_{jk}^2 N^{(i)}(\mathbf{x})[\mathrm{var}(\boldsymbol{\xi}_h)]_{jk} = \sum_{j,k=1}^d \partial_{jk}^2 F^{(i)}(\mathbf{x})[\boldsymbol{\Omega}_h]_{jk}.$$

We use the approximation of the variance of the random vector $\boldsymbol{\xi}_h$ to get $\mathbb{E}[\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}] = \mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x}) + \frac{h}{2}[\sum_{j,k=1}^d [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{jk}\partial_{jk}^2 F^{(i)}(\mathbf{x})]_{i=1}^d + \mathbf{R}(h^2, \mathbf{x})$. Taking the expectation of (S4) and incorporating the previous equation completes the proof. □

S1.3. *Proof of L^p convergence of the splitting scheme.* Now, we present the proof of L^p convergence stated in Theorem 3.7.

PROOF OF THEOREM 3.7. We use Theorem 3.3 to prove L^p convergence. It is sufficient to prove the two conditions (1) and (2). To prove condition (1), we need to prove the following property:

$$(\mathbb{E}[\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})\|^p \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])^{\frac{1}{p}} = R(h^{q_2}, \mathbf{x}),$$

where $q_2 = 3/2$. We start with $\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})\|^p = \|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\mathbf{F}(\mathbf{X}_{t_{k-1}}) - \boldsymbol{\xi}_{h,k} + \mathbf{R}(h^{3/2}, \mathbf{X}_{t_{k-1}})\|^p$. For more details on the expansion of $\Phi_h^{[\mathrm{S}]}$, see the previous proof. We approximate $\boldsymbol{\xi}_{h,k} = \int_{t_{k-1}}^{t_k} e^{\mathbf{A}(t_k - s)}\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_s$ by:

$$\boldsymbol{\xi}_{h,k} = \int_{t_{k-1}}^{t_k} (\mathbf{I} + (t_k - s)\mathbf{A})\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_s + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}})$$

$$= \boldsymbol{\Sigma}(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}}) + \mathbf{A}\boldsymbol{\Sigma}\int_{t_{k-1}}^{t_k} (t_k - s)\,\mathrm{d}\mathbf{W}_s + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}}).$$

Using the fact that $\int_{t_{k-1}}^{t_k} (t_k - s)\,\mathrm{d}\mathbf{W}_s \sim \mathcal{N}(\mathbf{0}, \frac{h^3}{3}\mathbf{I})$, we deduce that $\boldsymbol{\xi}_{h,k} = \boldsymbol{\Sigma}(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}}) + \mathbf{R}(h^{3/2}, \mathbf{X}_{t_{k-1}})$. Then, Hölder's inequality yields:

$$\|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\mathbf{F}(\mathbf{X}_{t_{k-1}}) - \boldsymbol{\Sigma}(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}})\|^p$$

$$\leq h^{p-1}\int_{t_{k-1}}^{t_k} \|(\mathbf{F}(\mathbf{X}_s) - \mathbf{F}(\mathbf{X}_{t_{k-1}}))\|^p\,\mathrm{d}s.$$

Assumption (A2), the integral norm inequality, Cauchy-Schwartz, and Hölder's inequalities, together with the mean value theorem yield:

$$\mathbb{E}[\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})\|^p \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]$$

$$\leq C(\mathbb{E}[h^{p-1}\int_{t_{k-1}}^{t_k} \|\mathbf{F}(\mathbf{X}_s) - \mathbf{F}(\mathbf{X}_{t_{k-1}})\|^p\,\mathrm{d}s \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])$$

$$\leq C(h^{p-1}\int_{t_{k-1}}^{t_k}\mathbb{E}[\|\mathbf{X}_s-\mathbf{X}_{t_{k-1}}\|^p\|\int_0^1 D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_s-u(\mathbf{X}_s-\mathbf{X}_{t_{k-1}}))\,\mathrm{d}u\|^p\mid\mathbf{X}_{t_{k-1}}=\mathbf{x}]\,\mathrm{d}s)$$

$$\leq C\Bigg(h^{p-1}\int_{t_{k-1}}^{t_k}(\mathbb{E}[\|\mathbf{X}_s-\mathbf{X}_{t_{k-1}}\|^{2p}\mid\mathbf{X}_{t_{k-1}}=\mathbf{x}])^{\frac{1}{2}}$$

$$(\mathbb{E}[\|\int_0^1 D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_s-u(\mathbf{X}_s-\mathbf{X}_{t_{k-1}}))\,\mathrm{d}u\|^{2p}\mid\mathbf{X}_{t_{k-1}}=\mathbf{x}])^{\frac{1}{2}}\,\mathrm{d}s\Bigg)$$

$$\leq C(h^{p-1}\int_{t_{k-1}}^{t_k}h^{\frac{p}{2}}\,\mathrm{d}s)=R(h^{3p/2},\mathbf{x}).$$

In the last line, we used Lemma 4.1. This proves condition (1) of Theorem 3.3.

Now, we prove condition (2). We use (5) and (11) to write $\mathbf{X}_{t_k}^{[\mathrm{S}]}=\boldsymbol{f}_{h/2}(e^{\mathbf{A}h}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]})-\mathbf{X}_{t_{k-1}}^{[1]})+\mathbf{X}_{t_k}^{[1]})$. Define $\mathbf{R}_{t_k}:=e^{\mathbf{A}h}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k}^{[\mathrm{S}]})-\mathbf{X}_{t_k}^{[1]})$, and use the associativity (9) to get $\mathbf{R}_{t_k}=e^{\mathbf{A}h}(\boldsymbol{f}_h(\mathbf{R}_{t_{k-1}}+\mathbf{X}_{t_k}^{[1]})-\mathbf{X}_{t_k}^{[1]})$. The proof of the boundness of the moments of $\mathbf{R}_{t_k}$ is the same as in Lemma 2 in Buckwar et al. (2022). Finally, we have $\mathbf{X}_{t_k}^{[\mathrm{S}]}=\boldsymbol{f}_{h/2}^{-1}(e^{-\mathbf{A}h}\mathbf{R}_{t_k}+\mathbf{X}_{t_k}^{[1]})$. Since $\boldsymbol{f}_{h/2}^{-1}$ grows polynomially and $\mathbf{X}_{t_k}^{[1]}$ has finite moments, $\mathbf{X}_{t_k}^{[\mathrm{S}]}$ must have finite moments too. This concludes the proof. □

S1.4. *Proof of Lemma 4.1.*

PROOF OF LEMMA 4.1. We first prove (1). In the following, C_1 and C_2 denote constants. We use the triangular inequality and Hölder's inequality to obtain:

$$\|\mathbf{X}_t-\mathbf{X}_{t_{k-1}}\|^p\leq 2^{p-1}(\|\int_{t_{k-1}}^t\mathbf{F}(\mathbf{X}_s;\boldsymbol{\theta})\,\mathrm{d}s\|^p+\|\boldsymbol{\Sigma}(\mathbf{W}_t-\mathbf{W}_{t_{k-1}})\|^p)$$

$$\leq 2^{p-1}((\int_{t_{k-1}}^t C_1(1+\|\mathbf{X}_s\|)^{C_1}\,\mathrm{d}s)^p+\|\boldsymbol{\Sigma}(\mathbf{W}_t-\mathbf{W}_{t_{k-1}})\|^p)$$

$$\leq 2^{p-1}C_1^p(\int_{t_{k-1}}^t(1+\|\mathbf{X}_s-\mathbf{X}_{t_{k-1}}\|+\|\mathbf{X}_{t_{k-1}}\|)^{C_1}\,\mathrm{d}s)^p+2^{p-1}\|\boldsymbol{\Sigma}(\mathbf{W}_t-\mathbf{W}_{t_{k-1}})\|^p$$

$$\leq 2^{C_1+2p-3}C_1^p(t-t_{k-1})^{p-1}(\int_{t_{k-1}}^t\|\mathbf{X}_s-\mathbf{X}_{t_{k-1}}\|^{pC_1}\,\mathrm{d}s+(t-t_{k-1})^p(1+\|\mathbf{X}_{t_{k-1}}\|)^{pC_1})$$

$$+2^{p-1}\|\boldsymbol{\Sigma}(\mathbf{W}_t-\mathbf{W}_{t_{k-1}})\|^p.$$

In the second inequality, we used the polynomial growth (A2) of $\mathbf{F}$. Furthermore, for some constant C_2 that depends on p, we have $\mathbb{E}[\|\boldsymbol{\Sigma}(\mathbf{W}_t-\mathbf{W}_{t_{k-1}})\|^p\mid\mathcal{F}_{t_{k-1}}]=(t-t_{k-1})^{p/2}C_2(p)$. Then, for $h<1$, there exists a constant C_p that depends on p, such that:

$$C_p(t-t_{k-1})^{2p-1}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p}+C_p(t-t_{k-1})^{p/2}\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p}.$$

The last inequality holds because the term of order $p/2$ is dominating when $t-t_{k-1}<1$. Denote $m(t)=\mathbb{E}[\|\mathbf{X}_t-\mathbf{X}_{t_{k-1}}\|^p\mid\mathcal{F}_{t_{k-1}}]$. Then, we have:

$$m(t)\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p}+C_p\int_{t_{k-1}}^t m^{C_1}(s)\,\mathrm{d}s. \tag{S7}$$

Now, we apply the generalized Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020), stated in Section S2) on (S7). Since we consider a super-linear growth, we can assume that there exist $C_1>1$ and $C_p>0$, such that:

$$m(t)\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p}+(\kappa^{1-C_1}(t)-(C_1-1)2^{C_1-1}C_p(t-t_{k-1}))^{\frac{1}{1-C_1}}$$

$$\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p}+C\kappa(t), \tag{S8}$$

where $\kappa(t) = C_p(t - t_{k-1})^{C_1 p/2+1}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p}$. The bound C in inequality (S8) makes sense, because the term:

$$(1 - (C_1 - 1)2^{C_1-1}C_p(t - t_{k-1})\kappa^{\frac{1}{1-C_1}}(t))^{\frac{1}{1-C_1}}$$

is positive by Lemma 2.3 from Tian and Fan (2020). Additionally, the same term reaches its maximum value of 1, for $t = t_{k-1}$. The constant C in (S8) includes some terms that depend on $t - t_{k-1}$. However, these terms will not change the dominating term of $\kappa(t)$ since $h < 1$. Finally, the terms in $\kappa(t)$ are of order $p/2$, thus for large enough constant C_p, it holds $m(t) \leq C_p(t - t_{k-1})^{p/2}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p}$.

To prove (2), we use that g is of polynomial growth:

$$\begin{aligned}\mathbb{E}[|g(\mathbf{X}_t;\boldsymbol{\theta})| \mid \mathcal{F}_{t_{k-1}}] &\leq C_1\mathbb{E}[(1 + \|\mathbf{X}_{t_{k-1}}\| + \|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\|)^{C_1} \mid \mathcal{F}_{t_{k-1}}] \\ &\leq C_2(1 + \|\mathbf{X}_{t_{k-1}}\|^{C_1} + \mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\|^{C_1} \mid \mathcal{F}_{t_{k-1}}]).\end{aligned}$$

Now, we apply the first part of the lemma, to get:

$$\begin{aligned}\mathbb{E}[|g(\mathbf{X}_t;\boldsymbol{\theta})| \mid \mathcal{F}_{t_{k-1}}] &\leq C_2(1 + \|\mathbf{X}_{t_{k-1}}\|^{C_1} + C'_{t-t_{k-1}}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_3}) \\ &\leq C_{t-t_{k-1}}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C}.\end{aligned}$$

That concludes the proof. □

S1.5. *Proofs of the Moment Bounds.* Before proving the moment bounds, we first demonstrate in Lemma S1.2 how the infinitesimal generator L operates on a product of two functions.

LEMMA S1.2. *Let L be the infinitesimal generator defined in the main text of SDE (1). For sufficiently smooth functions $\alpha, \beta : \mathbb{R}^d \to \mathbb{R}$, it holds:*

$$L(\alpha(\mathbf{x})\beta(\mathbf{x})) = \alpha(\mathbf{x})L\beta(\mathbf{x}) + \beta(\mathbf{x})L\alpha(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x}))).$$

PROOF. We use the generator L and the product rule to get:

$$\begin{aligned}L(\alpha(\mathbf{x})\beta(\mathbf{x})) &= \mathbf{F}(\mathbf{x})^\top\alpha(\mathbf{x})\nabla\beta(\mathbf{x}) + \mathbf{F}(\mathbf{x})^\top\beta(\mathbf{x})\nabla\alpha(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\alpha(\mathbf{x})\mathbf{H}_\beta(\mathbf{x}) + \beta(\mathbf{x})\mathbf{H}_\alpha(\mathbf{x}))) \\ &\quad + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x}))) \\ &= \alpha(\mathbf{x})L\beta(\mathbf{x}) + \beta(\mathbf{x})L\alpha(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x}))).\end{aligned}$$

This concludes the proof. □

PROOF OF PROPOSITION 4.3. Proof of (i). Lemma S1.1 yields:

$$\begin{aligned}\mathbb{E}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}})) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] &= \mathbb{E}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) \\ &= \mathbb{E}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] - \boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) - h\mathbf{F}(\mathbf{x}) \\ &\quad - \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3, \mathbf{x}).\end{aligned}$$

Now, we use the infinitesimal generator L to evaluate the expectation in the last line where the generator L is applied to a vector-valued function. We have:

$$\mathbb{E}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] = \boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) + hL\boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) + \frac{h^2}{2}L^2\boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) + \mathbf{R}(h^3, \mathbf{x}).$$

We use $\boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) = \boldsymbol{f}_{-h/2}(\mathbf{x})$ and Proposition 2.2 to get:

$$\begin{aligned}L\boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) &= L\mathbf{x} - \frac{h}{2}L\mathbf{N}(\mathbf{x}) + \mathbf{R}(h^2, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - \frac{h}{2}L\mathbf{N}(\mathbf{x}) + \mathbf{R}(h^2, \mathbf{x}), \\ L^2\boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) &= L\mathbf{A}(\mathbf{x}-\mathbf{b}) + L\mathbf{N}(\mathbf{x}) + \mathbf{R}(h, \mathbf{x}) = \mathbf{A}\mathbf{F}(\mathbf{x}) + L\mathbf{N}(\mathbf{x}) + \mathbf{R}(h, \mathbf{x}).\end{aligned}$$

It follows that $\mathbb{E}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}})) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] = \mathbf{R}(h^3, \mathbf{x})$.

Proof of (ii). In this proof, we distinguish the true parameters $\boldsymbol{\theta}_0$ from a generic parameter $\boldsymbol{\theta}$. We start with the expansions of $\boldsymbol{f}_h^{-1}$ and $\boldsymbol{\mu}_h$:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0))\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] \\
&= \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{X}_{t_k}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] - \frac{h}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] \\
&- \mathbf{x}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] - \frac{h}{2}(2\mathbf{A}^0(\mathbf{x}-\mathbf{b}_0) + \mathbf{N}_0(\mathbf{x}))\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] + \mathbf{R}(h^2, \mathbf{x}) \\
&= \mathbf{x}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + hL_{\boldsymbol{\theta}_0}(\mathbf{x}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top) - \frac{h}{2}\mathbf{N}_0(\mathbf{x})\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top \\
&- \mathbf{x}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top - h\mathbf{x}L_{\boldsymbol{\theta}_0}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top - h\mathbf{A}^0(\mathbf{x}-\mathbf{b}_0)\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top - \frac{h}{2}\mathbf{N}_0(\mathbf{x})\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + \mathbf{R}(h^2, \mathbf{x}) \\
&= hL_{\boldsymbol{\theta}_0}(\mathbf{x}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top) - h\mathbf{x}L_{\boldsymbol{\theta}_0}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top - h\mathbf{F}_0(\mathbf{x})\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + \mathbf{R}(h^2, \mathbf{x}).
\end{aligned}
$$

Lastly, Lemma S1.2 and the definition of $L_{\boldsymbol{\theta}_0}$ yield:

$$
\begin{aligned}
L_{\boldsymbol{\theta}_0}(\mathbf{x}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top) &= \mathbf{x}L_{\boldsymbol{\theta}_0}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + (L_{\boldsymbol{\theta}_0}\mathbf{x})\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + \frac{1}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top \mathbf{g}(\mathbf{x};\boldsymbol{\beta}) + D\mathbf{g}(\mathbf{x};\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) \\
&= \mathbf{x}L_{\boldsymbol{\theta}_0}\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + \mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)\mathbf{g}(\mathbf{x};\boldsymbol{\beta})^\top + \frac{1}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top \mathbf{g}(\mathbf{x};\boldsymbol{\beta}) + D\mathbf{g}(\mathbf{x};\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top).
\end{aligned}
$$

Proof of (iii). We introduce $\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) = \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)$ and use (ii) to show:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0))(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0))^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] \\
&= \frac{h}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top \mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0) + D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) \\
&- \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0)^\top + \mathbf{R}(h^2, \mathbf{x}).
\end{aligned}
$$

The result follows from property (i) and
$D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0) = \mathbf{I} + \mathbf{R}(h, \mathbf{x})$. □

S1.6. *Proof of consistency of the estimator.* The proof of consistency consists in studying the convergence of the objective function that defines the estimators. The objective function $\mathcal{L}_N(\boldsymbol{\beta}, \boldsymbol{\varsigma})$ (23) can be decomposed into sums of martingale triangular arrays. We thus first state a lemma that proves the convergence of each triangular array involved in the objective function. Then, we will focus on the proof of consistency.

LEMMA S1.3. *Let Assumptions (A1)-(A6) hold, and* $\mathbf{X}$ *be the solution of (1). Let* $\mathbf{g}, \mathbf{g}_1, \mathbf{g}_2 : \mathbb{R}^d \times \Theta \times \Theta \to \mathbb{R}^{\mathbb{d}}$ *be differentiable functions with respect to* $\mathbf{x}$ *and* $\boldsymbol{\theta}$*, with derivatives of polynomial growth in* $\mathbf{x}$*, uniformly in* $\boldsymbol{\theta}$*. If* $h \to 0$ *and* $Nh \to \infty$*, then:*

1. $\frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$;
2. $\frac{h}{N}\sum_{k=1}^{N} \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
3. $\frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
4. $\frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
5. $\frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;

6. $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}}$
$$\int \operatorname{Tr}(D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\,\mathrm{d}\nu_0(\mathbf{x});$$
7. $\frac{h}{N}\sum_{k=1}^{N}\mathbf{g}_1(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}_2(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,$

uniformly in $\boldsymbol{\theta}$.

Lemma S1.3 plays a central role in demonstrating the consistency and asymptotic normality of the proposed estimators. The lemma deals with the uniform convergence of multiple triangular arrays, and proving various aspects of it involves a range of technical tools and methods. Different parts of Lemma S1.3 require distinct strategies to establish appropriate bounds, which can be intricate. Once these bounds are established, we leverage the properties discussed in the preceding section.

For instance, when establishing point-wise convergence, we primarily rely on Lemma S2.2. On the other hand, for proving uniform convergence, we utilize both Lemma S2.3 and Lemma S2.4. Throughout the proof of Lemma S1.3, a recurring theme is to interpret quadratic forms as traces and exploit the cyclic property inherent to them. Additionally, we employ fundamental mathematical tools like the mean value theorem, the Cauchy-Schwartz inequality, and Hölder's inequality in various instances.

Furthermore, there are occasions where we require inequality for norms, particularly the Frobenius norm. To address this, we introduce the Frobenius inner product of matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ in $\mathbb{R}^{n\times m}$ as $\langle\mathbf{M}_1,\mathbf{M}_2\rangle_F := \operatorname{Tr}(\mathbf{M}_1^\top\mathbf{M}_2)$. Leveraging Hölder's inequality on Frobenius norm provides us with the following bound for the trace of a matrix product: $\|\operatorname{Tr}(\mathbf{M}_1^\top\mathbf{M}_2)\| \le \|\operatorname{Tr}(\mathbf{M}_1)\|\|\mathbf{M}_2\|$.

PROOF OF LEMMA S1.3. Proof of 1. As previously discussed, we introduce a martingale array that corresponds to the limit outlined in point 1. We then utilize Lemma S2.2 to facilitate our analysis. We denote $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\varsigma}) := \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$. We have:

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\varsigma})\mid\mathbf{X}_{t_{k-1}}] &= \frac{1}{Nh}\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0))\mid\mathbf{X}_{t_{k-1}}]\\ &= \frac{1}{Nh}\sum_{k=1}^{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\mid\mathbf{X}_{t_{k-1}}])\\ &= \frac{1}{Nh}\sum_{k=1}^{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top+\mathbf{R}(h^2,\mathbf{X}_{t_{k-1}})) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top).\end{aligned}$$

To use the result of Lemma S2.2, we need to prove that covariance of $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\varsigma})$ goes to zero. To achieve this, we leverage Corollary 3.8 and recall that if $\boldsymbol{\rho}$ is a Gaussian random vector $\boldsymbol{\rho}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Pi})$, then $\mathbb{E}[(\boldsymbol{\rho}^T\mathbf{M}\boldsymbol{\rho})^2] = 2\operatorname{Tr}((\mathbf{M}\boldsymbol{\Pi})^2)+(\operatorname{Tr}(\mathbf{M}\boldsymbol{\Pi}))^2$. This leads to:

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\varsigma})^2\mid\mathbf{X}_{t_{k-1}}] &= \frac{1}{N^2h^2}\sum_{k=1}^{N}(\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\xi}_{h,k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\xi}_{h,k})^2\mid\mathbf{X}_{t_{k-1}}]+R(h^{3/2},\mathbf{X}_{t_{k-1}}))\\ &= \frac{1}{Nh}\frac{1}{N}\sum_{k=1}^{N}(2\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top)^2+(\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top))^2+R(h^{1/2},\mathbf{X}_{t_{k-1}})) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,\end{aligned}$$

for $Nh\to\infty$, $h\to 0$. Then, by Lemma S2.2 $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$, for $Nh\to\infty$, $h\to 0$. To establish the uniformity of the limits with respect to $\boldsymbol{\varsigma}$, we turn to Lemma S2.3 and introduce sets Θ_{ς_j} such that $\boldsymbol{\varsigma}=(\varsigma_1,\varsigma_2,\ldots,\varsigma_s)\in\Theta_{\varsigma_1}\times\Theta_{\varsigma_2}\times\cdots\times\Theta_{\varsigma_s}=\Theta_\varsigma$. Then it is enough to show that for all $j=1,\ldots,s$, it holds:

$$\sup_{N\in\mathbb{N}}\mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{\varsigma_j\in\Theta_{\varsigma_j}}|\partial_{\varsigma_j}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)|]<\infty. \tag{S9}$$

We use the well-known rule of matrix differentiation $\partial_{\mathbf{X}}(\mathbf{a}^\top \mathbf{X}^{-1}\mathbf{a}) = -\mathbf{X}^{-1}\mathbf{a}\mathbf{a}^\top\mathbf{X}^{-1}$, where $\mathbf{a}$ is a vector and $\mathbf{X}$ is a symmetric matrix, to get:

$$\partial_{x^{(i)}} \operatorname{Tr}(\mathbf{a}^\top \mathbf{C}^{-1}(\mathbf{x})\mathbf{a}) = -\operatorname{Tr}(\mathbf{C}^{-1}(\mathbf{x})\mathbf{a}\mathbf{a}^\top\mathbf{C}^{-1}(\mathbf{x})\partial_{x^{(i)}}\mathbf{C}(\mathbf{x})) = -\operatorname{Tr}(\mathbf{a}\mathbf{a}^\top\mathbf{C}^{-1}(\mathbf{x})(\partial_{x^{(i)}}\mathbf{C}(\mathbf{x}))\mathbf{C}^{-1}(\mathbf{x})).$$

We omit writing $\boldsymbol{\beta}_0$ for ease of notation. Then, by using the trace bound, the norm inequality, and Assumption (A4), we can deduce that:

$$\sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{\varsigma_j\in\Theta_{\varsigma_j}} |\partial_{\varsigma_j} \frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}|] \leq \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \sup_{\varsigma_j\in\Theta_{\varsigma_j}} |\partial_{\varsigma_j} \operatorname{Tr}(\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k})|]$$

$$\leq \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|]$$

$$\leq \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|^2\|\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\|] \leq C \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top)]$$

$$= C \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \mid \mathbf{X}_{t_{k-1}}]] = C \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}}))] < \infty.$$

Proof of 2. We use Lemma 4.2 to deduce:

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \int \mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta})\,\mathrm{d}\nu_0(\mathbf{x}),$$

uniformly in $\boldsymbol{\theta}$, for $Nh \to \infty$, $h \to 0$. Then we use the bound of $\mathbf{g}$ to conclude the proof of 2.

Proof of 3. For $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}) := \frac{1}{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})$, the limit of $\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}]$ rewrites as:

$$\begin{aligned}
\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}] &= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})) \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{1}{N}\sum_{k=1}^{N} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}]) \\
&= \frac{1}{N}\sum_{k=1}^{N} R(h^3, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. Then, we study the limit of $\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}]$:

$$\begin{aligned}
&\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}](\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \\
&= \frac{1}{N}\sum_{k=1}^{N} R(\frac{h}{N}, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. Lemma S2.2 yields that $\frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$, for $Nh \to \infty$, $h \to 0$. To show the uniformity of the limits with respect to $\boldsymbol{\theta}$, we leverage Lemma S2.4. It is sufficient to demonstrate the existence of constants $p \geq l > r + s$ and $C > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ it holds:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})|^p] \leq C, \tag{S10}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} (Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))|^p] \leq C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^l. \tag{S11}$$

We begin by considering equation (S10). Based on the definition of $\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$ and the assumptions made about $\mathbf{N}$, as well as the fact that $h < 1$, there exist constants C_1 and C_2 such that:

$$\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^p \leq \|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}}\|^p + C_1 h^p (1 + \|\mathbf{X}_{t_k}\|)^{C_1} + C_2 h^p (1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_2}, \tag{S12}$$

Then, Lemma 4.1 yields:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^p \mid \mathbf{X}_{t_{k-1}}] \leq C h^{p/2} (1 + \|\mathbf{X}_{t_{k-1}}\|)^C. \tag{S13}$$

Subsequently, we use the norm inequality, (S13) and both statements of Lemma 4.1 to get:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})|^p] &\leq N^{p-1} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[|Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})|^p] \\
&= \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbb{E}_{\boldsymbol{\theta}_0}[|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})|^p \mid \mathbf{X}_{t_{k-1}}]] \\
&\leq \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^p \mid \mathbf{X}_{t_{k-1}}]\|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|^p \|\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})\|^p] \leq \frac{1}{N} \cdot N \cdot C.
\end{aligned} \tag{S14}$$

This completes the proof of (S10). Now, we focus on (S11). We use the triangular inequality and the Hölder's inequality to derive:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} (Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))|^p] \\
&\leq \frac{2^{p-1}}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}(\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0) - \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0))|^p] \qquad \text{(S15)} \\
&+ \frac{2^{p-1}}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top ((\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1} - (\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1})\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0)|^p]. \qquad \text{(S16)}
\end{aligned}$$

First, we study sum (S15). We use the mean value theorem and the triangular inequalities to get:

$$\begin{aligned}
&\frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}(\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0) - \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0))|^p] \\
&\leq \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^p \mid \mathbf{X}_{t_{k-1}}]\|(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}\|^p \|\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0) - \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0)\|^p] \\
&\leq \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[C_p(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^p \| \int_0^1 D_\beta \mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2),\boldsymbol{\beta}_0)\,\mathrm{d}t\|^p] \\
&\leq C\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^p. \qquad \text{(S17)}
\end{aligned}$$

To bound sum (S16), we introduce the following multivariate matrix-valued function $\mathbf{G}(\boldsymbol{\varsigma}) \coloneqq (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}$. Then, we use the inequality between the operator 2-norm and Frobenius norm, and the definition of the Frobenius norm to get:

$$\|\mathbf{G}(\boldsymbol{\varsigma}_1) - \mathbf{G}(\boldsymbol{\varsigma}_2)\| \leq (\sum_{i,j=1}^{d} \|G_{ij}(\boldsymbol{\varsigma}_1) - G_{ij}(\boldsymbol{\varsigma}_2)\|^2)^{\frac{1}{2}}.$$

Now, apply the mean value theorem on each G_{ij} and Assumption (A4) to get:

$$\|\mathbf{G}(\boldsymbol{\varsigma}_1) - \mathbf{G}(\boldsymbol{\varsigma}_2)\| \leq (\sum_{i,j=1}^{d} \|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\|^2 \| \int_0^t \nabla_{\boldsymbol{\varsigma}} G_{ij}(\boldsymbol{\varsigma}_2 + t(\boldsymbol{\varsigma}_1 - \boldsymbol{\sigma_2}))\, \mathrm{d}t\|^2)^{\frac{1}{2}} \leq C\|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\|.$$

Finally, combining the previous results, we conclude that:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N}(Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2))|^p] &\leq C(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^p + \|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\|^p) \\ &\leq C(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\boldsymbol{\varsigma}_1 - \boldsymbol{\varsigma}_2\|^2)^{p/2} = C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^p,\end{aligned}$$

for $p \geq 2$. This concludes the proof of 3.

Proof of 4. For $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$, we repeat the same derivations as in the proof of 3. to show that the limit of $\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}]$ satisfies:

$$\begin{aligned}&\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}] \\ &= \frac{1}{Nh}\sum_{k=1}^{N} \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}]) = \frac{1}{N}\sum_{k=1}^{N} R(h^2, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,\end{aligned}$$

for $h \to 0$. Similarly we deduce that:

$$\begin{aligned}&\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}] \\ &= \frac{1}{N^2h^2}\sum_{k=1}^{N} \mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}](\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}) \\ &= \frac{1}{N}\sum_{k=1}^{N} R(\frac{1}{Nh}, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,\end{aligned} \tag{S18}$$

for $Nh \to \infty$. To prove uniform convergence, we use Lemma S2.4 along with Rosenthal's inequality from Theorem S2.5, resulting in:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})|^p] \leq C(\mathbb{E}[(\sum_{k=1}^{N} \mathbb{E}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}])^{p/2}] + \sum_{k=1}^{N} \mathbb{E}[|Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})|^p]).$$

The first term is bounded because of (S18). To bound the second term on the right-hand side, we use (S14). Then, for $Nh \to \infty$ and $h \to 0$ and $p > 2$ it holds:

$$\sum_{k=1}^{N} \mathbb{E}[|Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})|^p] \leq \frac{1}{(Nh)^p} \cdot Nh^{p/2} \cdot C = \frac{1}{(Nh)^{p-1}} \cdot h^{p/2-1} \cdot C \leq C.$$

To conclude the proof of uniform convergence, we once again apply Rosenthal's inequality to get:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N}(Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1)-Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))|^p]$$

$$\leq C\mathbb{E}[(\sum_{k=1}^{N}\mathbb{E}[(Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1)-Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))^2\mid\mathbf{X}_{t_{k-1}}])^{p/2}]+C\sum_{k=1}^{N}\mathbb{E}[|(Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1)-Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))|^p]. \tag{S19}$$

To bound the first term in (S19), we follow the reasoning from (S17) and start with:

$$\begin{aligned}&\mathbb{E}[(Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1)-Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))^2\mid\mathbf{X}_{t_{k-1}}]\\&\leq 2\mathbb{E}_{\boldsymbol{\theta}_0}[(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}(\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0)-\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0)))^2\mid\mathbf{X}_{t_{k-1}}]\\&+2\mathbb{E}_{\boldsymbol{\theta}_0}[(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top((\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}-(\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1})\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0))^2\mid\mathbf{X}_{t_{k-1}}].\end{aligned}$$

Then, the rest is the same. Similarly, to bound the second term in (S19), we repeat derivations from (S17) to get:

$$\sum_{k=1}^{N}\mathbb{E}[|(Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1)-Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2))|^p]\leq\frac{1}{(Nh)^p}\cdot Nh^{p/2}\cdot C\cdot\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^p\leq C\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^p,$$

Finally, (S18) and conclusions after (S17) complete the proof of 4.

Proof of 5. We introduce $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}):=\frac{1}{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})$. Proposition 4.3 yields that $\mathbb{E}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top\mid\mathbf{X}_{t_{k-1}}]=\mathbf{R}(h,\mathbf{X}_{t_{k-1}})$. Then, we conclude that $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mid\mathbf{X}_{t_{k-1}}]\to 0$ in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh\to\infty$, $h\to 0$. Moreover, to prove the convergence of $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2\mid\mathbf{X}_{t_{k-1}}]$, it is enough to bound $\frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}[\mathrm{Tr}((\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}))^2)\mid\mathbf{X}_{t_{k-1}}]$. Hölder's inequality, together with Cauchy-Schwartz inequality, Lemma 4.1 and (S13), yield:

$$\begin{aligned}&\frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}[\mathrm{Tr}((\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}))^2)\mid\mathbf{X}_{t_{k-1}}]\\&\leq\frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^2\|\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})\|^2\mid\mathbf{X}_{t_{k-1}}]\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|\\&\leq\frac{C}{N^2}\sum_{k=1}^{N}(\mathbb{E}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^4\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}[\|\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})\|^4\mid\mathbf{X}_{t_{k-1}}])^{\frac{1}{2}}=\frac{1}{N}\sum_{k=1}^{N}R(h/N,\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,\end{aligned} \tag{S20}$$

for $Nh\to\infty$, $h\to 0$. To prove the uniform convergence, we use Lemma S2.4. Again, it is enough to prove (S10) and (S11). Repeating the same steps as in the proof of (S14) leads to (S10). Similarly, to prove (S11) we repeat the same steps as in (S17) using Hölder's inequality, Cauchy-Schwartz inequality, and Lemma 4.1 with (S13).

Proof of 6. We introduce $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}):=\frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})$ and study $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mid\mathbf{X}_{t_{k-1}}]$. Proposition 4.3 yields:

$$\begin{aligned}&\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mid\mathbf{X}_{t_{k-1}}]=\frac{1}{Nh}\sum_{k=1}^{N}\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top\mid\mathbf{X}_{t_{k-1}}])\\&=\frac{1}{2N}\sum_{k=1}^{N}\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})+D\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top+\mathbf{R}(h,\mathbf{X}_{t_{k-1}})))\\&\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}\int\mathrm{Tr}(D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\,\mathrm{d}\nu_0(\mathbf{x}),\end{aligned}$$

for $Nh\to\infty$, $h\to 0$. On the other hand, $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2\mid\mathbf{X}_{t_{k-1}}]=\frac{1}{N}\sum_{k=1}^{N}R(\frac{1}{Nh},\mathbf{X}_{t_{k-1}})\to 0$, in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh\to\infty$, $h\to 0$, which follows from derivations in (S20). To prove uniform convergence, we repeat the same approach as in the previous two proofs.

Proof of 7. First, we use the fact that $\mathbb{E}[\boldsymbol{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] = \boldsymbol{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta}) + \mathbf{R}(h,\mathbf{x})$, for a generic function $\boldsymbol{g}$. Then, for $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}) \coloneqq \frac{h}{N}\mathbf{g}_1(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}_2(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})$ it follows

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0, \qquad \sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}] \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

Again, the proofs of (S10) and (S11) are the same as in property 3, with a distinction of rewriting:

$$\begin{aligned}
&\mathbf{g}_1(\boldsymbol{\beta}_1)^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_1) - \mathbf{g}_1(\boldsymbol{\beta}_2)^\top(\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_2)\\
&\quad= (\mathbf{g}_1(\boldsymbol{\beta}_1) - \mathbf{g}_1(\boldsymbol{\beta}_2))^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_1) + \mathbf{g}_1(\boldsymbol{\beta}_2)^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}(\mathbf{g}_2(\boldsymbol{\beta}_1) - \mathbf{g}_2(\boldsymbol{\beta}_2))\\
&\quad+ \mathbf{g}_1(\boldsymbol{\beta}_2)^\top((\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1} - (\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1})\mathbf{g}_2(\boldsymbol{\beta}_2).
\end{aligned}$$

□

PROOF OF THEOREM 5.1. To establish consistency, we follow the proof of Theorem 1 in Kessler (1997) and study the limit of $\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})$ from (23), rescaled by the correct rate of convergence. More precisely, the consistency of the diffusion parameter is proved by studying the limit of $\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})$, while the consistency of the drift parameter is proved by studying the limit of $\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0,\boldsymbol{\varsigma}))$. We start with the consistency of the diffusion parameter $\boldsymbol{\varsigma}$. We need to prove that:

$$\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) \to \log(\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)) + \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) =: G_1(\boldsymbol{\varsigma},\boldsymbol{\varsigma}_0), \tag{S21}$$

in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh\to\infty$, $h\to 0$, uniformly in $\boldsymbol{\theta}$. To study the limit, we first decompose $\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})$ as follows:

$$\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) = \log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + T_1 + T_2 + T_3 + 2(T_4+T_5+T_6) + R(h,\mathbf{x}_0). \tag{S22}$$

The terms $T_1,\dots,T_6$ are derived from the quadratic form in (23) by adding and subtracting the corresponding terms with $\boldsymbol{\beta}_0$, followed by rearrangements, resulting in the following expressions:

$$\begin{aligned}
T_1 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0),\\
T_2 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}) - \boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}_0))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}) - \boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}_0)),\\
T_3 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}(\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta})),\\
T_4 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta})),\\
T_5 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}) - \boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}_0))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta})),\\
T_6 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}) - \boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}_0))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0).
\end{aligned}$$

Previously, we defined $\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}) \coloneqq \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta})$ and $\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}) \coloneqq \boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta})$. These terms will also play a significant role in proving the asymptotic normality.

The first term of (S22) is a constant. Properties 1, 2, 3, 5, and 7 from Lemma S1.3 give the following limits $T_1 \to \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$ and for $l = 2,3,...,6$, $T_l \to 0$, uniformly in $\boldsymbol{\theta}$. The convergence in probability is equivalent

to the existence of a subsequence converging almost surely. Thus, the convergence in (S21) is almost sure for a subsequence $(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l})$. This implies:

$$\widehat{\boldsymbol{\varsigma}}_{N_l} \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}-a.s.} \boldsymbol{\varsigma}_\infty.$$

The compactness of $\overline{\Theta}$ implies that $(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l})$ converges to a limit $(\boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty)$ almost surely. By continuity of the mapping $\boldsymbol{\varsigma} \mapsto G_1(\boldsymbol{\varsigma}, \boldsymbol{\varsigma}_0)$ we have $\frac{1}{N_l}\mathcal{L}_{N_l}^{[\mathrm{S}]}(\hat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l}) \to G_1(\boldsymbol{\varsigma}_\infty^\top, \boldsymbol{\varsigma}_0)$, in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh \to \infty$, $h \to 0$, uniformly in $\boldsymbol{\theta}$. By the definition of the estimator, $G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) \leq G_1(\boldsymbol{\varsigma}_0, \boldsymbol{\varsigma}_0)$. We also have:

$$\begin{aligned}
&G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) \geq G_1(\boldsymbol{\varsigma}_0, \boldsymbol{\varsigma}_0) \\
&\Leftrightarrow \log(\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top)) + \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) \geq \log(\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)) + \mathrm{Tr}(\mathbf{I}_d) \\
&\Leftrightarrow \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) - \log(\det((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)) \geq d \\
&\Leftrightarrow \sum_{i=1}^d \lambda_i - \log\prod_{i=1}^d \lambda_i \geq \sum_{i=1}^d 1 \Leftrightarrow \sum_{i=1}^d (\lambda_i - 1 - \log\lambda_i) \geq 0,
\end{aligned}$$

where λ_i are the eigenvalues of $(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$, which is a positive definite matrix. The last inequality follows since for any positive x, $\log x \leq x - 1$. Thus, $G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) = G_1(\boldsymbol{\varsigma}_0.\boldsymbol{\varsigma}_0)$. Then, all the eigenvalues λ_i must be equal to 1, hence, $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$. We proved that a convergent subsequence of $\widehat{\boldsymbol{\varsigma}}_N$ tends to $\boldsymbol{\varsigma}_0$ almost surely. From there, the consistency of the estimator of the diffusion coefficient follows.

We now focus on the consistency of the drift parameter. The objective is to prove that the following limit in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh \to \infty$, $h \to 0$, uniformly with respect to $\boldsymbol{\theta}$:

$$\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0, \boldsymbol{\varsigma})) \to G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}), \tag{S23}$$

where:

$$\begin{aligned}
G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}) \coloneqq &\int (\mathbf{F}_0(\mathbf{x}) - \mathbf{F}(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{x}) - \mathbf{F}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}) \\
&+ \int \mathrm{Tr}(D(\mathbf{F}_0(\mathbf{x}) - \mathbf{F}(\mathbf{x}))(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} - \mathbf{I}))\,\mathrm{d}\nu_0(\mathbf{x}).
\end{aligned}$$

To prove it, we decompose $\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}))$ as follows:

$$\begin{aligned}
\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0, \boldsymbol{\varsigma})) &= \mathrm{Tr}(\mathbf{A}(\boldsymbol{\beta}) - \mathbf{A}(\boldsymbol{\beta}_0)) + \frac{1}{h}(T_2 + T_3 + 2(T_4 + T_5 + T_6)) \\
&+ \frac{1}{Nh}\sum_{k=1}^N (\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) - \mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta})) \\
&+ \frac{1}{N}\sum_{k=1}^N \mathrm{Tr}\, D(\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}) - \mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)) + R(h, \mathbf{x}_0).
\end{aligned} \tag{S24}$$

The term $\frac{1}{Nh}\sum_{k=1}^N (\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) - \mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta}))$ converges to $\mathrm{Tr}(\mathbf{A}(\boldsymbol{\beta_0}) - \mathbf{A}(\boldsymbol{\beta}))$, which thus cancels out with the first term in (34). Lemma 4.2 provides the uniform convergence of $\frac{1}{h}T_2$ with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\frac{1}{h}T_2 &= \frac{1}{4N}\sum_{k=1}^N (\mathbf{N}_0(\mathbf{X}_{t_k}) - \mathbf{N}(\mathbf{X}_{t_k}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{X}_{t_k}) - \mathbf{N}(\mathbf{X}_{t_k})) + R(h, \mathbf{x}_0) \\
&\to \frac{1}{4}\int (\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}).
\end{aligned}$$

The limit of $\frac{1}{h}T_3$ computes analogously. To prove $\frac{1}{h}T_4 \to 0$, we use Lemma 9 in Genon-Catalot and Jacob (1993) and Property 4 from Lemma S1.3. Lemma 4.2 yields:

$$\frac{1}{h}T_5 \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{1}{4}\int (\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}) + \frac{1}{2}\int (\mathbf{A}_0(\mathbf{x}-\mathbf{b}_0) - \mathbf{A}(\mathbf{x}-\mathbf{b}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}).$$

Finally, $\frac{1}{h}T_6 \to \frac{1}{2}\int \mathrm{Tr}(D(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))^\top \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\,\mathrm{d}\nu_0(\mathbf{x})$ uniformly in $\boldsymbol{\theta}$, by Property 6 of Lemma S1.3. Lemma 4.2 gives:

$$\frac{1}{N}\sum_{k=1}^{N} \mathrm{Tr}\, D(\mathbf{N}(\mathbf{X}_{t_k}) - \mathbf{N}_0(\mathbf{X}_{t_k})) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \int \mathrm{Tr}\, D(\mathbf{N}(\mathbf{x}) - \mathbf{N}_0(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}),$$

uniformly in $\boldsymbol{\theta}$. This proves (S23). Then, there exists a subsequence N_l such that $(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l})$ converges to a limit $(\boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty)$, almost surely. By continuity of the mapping $(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \mapsto G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma})$, for $N_l h \to \infty$, $h \to 0$, we have the following convergence in $\mathbb{P}_{\boldsymbol{\theta}_0}$:

$$\frac{1}{N_l h}(\mathcal{L}_{N_l}^{[\mathrm{S}]}(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l}) - \mathcal{L}_{N_l}^{[\mathrm{S}]}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\varsigma}}_{N_l})) \to G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty).$$

Then, $G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty) \geq 0$ since $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$. On the other hand, by the definition of the estimator $\mathcal{L}_{N_l}^{[\mathrm{S}]}(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l}) - \mathcal{L}_{N_l}^{[\mathrm{S}]}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\varsigma}}_{N_l}) \leq 0$. Thus, the identifiability assumption (A5) concludes the proof for the S estimator.

To prove the same statement for the LT estimator, the representation of the objective function (S22) has to be adapted. In the LT case, this representation is straightforward. There is no extra logarithmic term and only three instead of six auxiliary T terms are used. This is due to the Gaussian transition density in the LT approximation. □

S1.7. *Proof of asymptotic normality of the estimator.* In this section, we distinguish between the true parameter $\boldsymbol{\theta}_0$ and a generic parameter $\boldsymbol{\theta}$.

PROOF OF THEOREM 5.2. According to Theorem 1 in Kessler (1997) or Theorem 1 in Sørensen and Uchida (2003), Lemmas S1.4 and S1.5 below are enough for establishing the asymptotic normality of $\hat{\boldsymbol{\theta}}_N$. Here, we only present the outline of the proof. For more details, see proof of Theorem 1 in Sørensen and Uchida (2003).

LEMMA S1.4. *Let $\mathbf{C}_N(\boldsymbol{\theta}_0)$ and $\mathbf{C}(\boldsymbol{\theta}_0)$ be as defined in (25) and (27), respectively. If $h \to 0$, $Nh \to \infty$, and $\rho_N \to 0$, then:*

$$\mathbf{C}_N(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 2\mathbf{C}(\boldsymbol{\theta}_0), \qquad \sup_{\|\boldsymbol{\theta}\|\leq\rho_N} \|\mathbf{C}_N(\boldsymbol{\theta}_0+\boldsymbol{\theta}) - \mathbf{C}_N(\boldsymbol{\theta}_0)\| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

LEMMA S1.5. *Let $\boldsymbol{\lambda}_N$ be as defined (26).If $h \to 0$, $Nh \to \infty$ and $Nh^2 \to 0$, then:*

$$\boldsymbol{\lambda}_N \xrightarrow{d} \mathcal{N}(\mathbf{0}, 4\mathbf{C}(\boldsymbol{\theta}_0)),$$

under $\mathbb{P}_{\boldsymbol{\theta}_0}$.

Lemma S1.4 states that $\mathbf{C}_N(\boldsymbol{\theta}_0)$ approaches $2\mathbf{C}(\boldsymbol{\theta}_0)$ as $h \to 0$ and $Nh \to \infty$. Moreover, the difference between $\mathbf{C}_N(\boldsymbol{\theta}_0+\boldsymbol{\theta})$ and $\mathbf{C}_N(\boldsymbol{\theta}_0)$ approaches zero when $\boldsymbol{\theta}$ approaches $\boldsymbol{\theta}_0$, within a distance specified by balls $\mathcal{B}_{\rho_N}(\boldsymbol{\theta}_0)$, where $\rho_N \to 0$. To ensure the asymptotic normality of $\hat{\boldsymbol{\theta}}_N$, Lemma S1.4 is employed to restrict the term $\|\mathbf{D}_N - \mathbf{C}_N(\boldsymbol{\theta}_0)\|$ when $\hat{\boldsymbol{\theta}}_N \in \Theta \cap \mathcal{B}_{\rho_N}(\boldsymbol{\theta}_0)$ as follows:

$$\|\mathbf{D}_N - \mathbf{C}_N(\boldsymbol{\theta}_0)\|\mathbb{1}_{\{\hat{\boldsymbol{\theta}}_N \in \Theta\cap\mathcal{B}_{\rho_N}(\boldsymbol{\theta}_0)\}} \leqslant \sup_{\boldsymbol{\theta}\in\mathcal{B}_{\rho_N}(\boldsymbol{\theta}_0)} \|\mathbf{C}_N(\boldsymbol{\theta}) - \mathbf{C}_N(\boldsymbol{\theta}_0)\| \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$$

Applying again Lemma S1.4 on the previous line, we get $\mathbf{D}_N \to 2\mathbf{C}(\boldsymbol{\theta}_0)$ in $\mathbb{P}_{\boldsymbol{\theta}_0}$, as $h \to 0$ and $Nh \to \infty$.

Lemma S1.5 establishes the convergence in distribution of $\boldsymbol{\lambda}_N$ to $\mathcal{N}(\mathbf{0}, 4\mathbf{C}(\boldsymbol{\theta}_0))$, under $\mathbb{P}_{\boldsymbol{\theta}_0}$, as $h \to 0$ and $Nh \to \infty$. This result provides the groundwork for the asymptotic normality of $\hat{\boldsymbol{\theta}}_N$. Indeed, consider the set $\mathcal{D}_N$ composed of instances where $\mathbf{D}_N$ is invertible. The probability, under $\boldsymbol{\theta}_0$, of $\mathcal{D}_N$ occurring approaches 1, as $h \to 0$ and $Nh \to \infty$. This implies that $\mathbf{D}_N$ is almost surely invertible in this limit. Furthermore, we define $\mathcal{E}_N$ as the intersection of $\{\hat{\boldsymbol{\theta}}_N \in \Theta\}$ and $\mathcal{D}_N$. Then, it can be shown that $\mathbb{1}_{\mathcal{E}_N} \to 1$ in $\mathbb{P}_{\boldsymbol{\theta}_0}$ when $h \to 0$ and $Nh \to \infty$. For $\mathbf{E}_N := \mathbf{D}_N$ on $\mathcal{E}_N$, we have $\mathbf{E}_N \to 2\mathbf{C}(\boldsymbol{\theta}_0)$ in $\mathbb{P}_{\boldsymbol{\theta}_0}$ as $h \to 0$ and $Nh \to \infty$. Given that $\mathbf{s}_N \mathbb{1}_{\mathcal{E}_N} = \mathbf{E}_N^{-1}\mathbf{D}_N \mathbf{s}_N \mathbb{1}_{\mathcal{E}_N} = \mathbf{E}_N^{-1}\boldsymbol{\lambda}_N \mathbb{1}_{\mathcal{E}_N}$ and according to Lemma S1.5, $\mathbf{s}_N \mathbb{1}_{\mathcal{E}_N} \to \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_0)^{-1})$ in distribution as $h \to 0$, $Nh \to \infty$ and $Nh^2 \to 0$.

In conclusion, under $\mathbb{P}_{\boldsymbol{\theta}_0}$, as $h \to 0$, $Nh \to \infty$ and $Nh^2 \to 0$, $\mathbf{s}_N \mathbb{1}_{\mathcal{E}_N}$ is shown to converge in distribution to $\mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_0)^{-1})$. The asymptotic normality for $\hat{\boldsymbol{\theta}}_N$ is, thus, confirmed due to the convergence of $\mathbb{1}_{\mathcal{E}_N} \to 1$. □

PROOF OF LEMMA S1.4. To prove the first part of the lemma, we aim to represent $\mathbf{C}_N(\boldsymbol{\theta}_0)$ from the objective function (14). In doing so, we again employ the approximation (23), focusing solely on the terms that do not converge to zero as $Nh \to \infty$ and $h \to 0$. We start as in the approximation (34) and compute the corresponding derivatives to obtain the first block matrix of $\mathbf{C}_N$ (25). We begin with $\partial_{\beta_{i_1}\beta_{i_2}} \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma})$:

$$\begin{aligned}
\frac{1}{Nh}\partial_{\beta_{i_1}\beta_{i_2}} \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) &= \partial_{\beta_{i_1}\beta_{i_2}} \operatorname{Tr} \mathbf{A}(\boldsymbol{\beta}) + \frac{1}{N}\sum_{k=1}^{N} \partial_{\beta_{i_1}\beta_{i_2}} \operatorname{Tr} D\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}) \\
&+ \partial_{\beta_{i_1}\beta_{i_2}} \frac{1}{h}\Big(T_2(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}) + T_3(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}) + 2(T_4(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}) + T_5(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}) + T_6(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}))\Big) \\
&- \frac{1}{Nh}\sum_{k=1}^{N} \partial_{\beta_{i_1}\beta_{i_2}} (\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{A}(\boldsymbol{\beta}) \mathbf{Z}_{t_k}(\boldsymbol{\beta}))) + R(h, \mathbf{x}_0).
\end{aligned}$$

To determine the convergence of each of the previous terms, we use the definitions of the sums T_is and approximate each T_i using Proposition 2.2 and the Taylor expansion of the function $\boldsymbol{\mu}_h$. As we apply the derivatives $\partial\beta_{i_1}\beta_{i_2}$, the order of h in each sum increases since terms of order $R(1, \mathbf{x}_0)$ are constant with respect to $\boldsymbol{\beta}$. Finally, when evaluating $\frac{1}{Nh}\partial_{\beta_{i_1}\beta_{i_2}} \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, numerous terms will cancel out due to differences of the type $\mathbf{g}(\boldsymbol{\beta}_0; \mathbf{X}_{t_k}, \mathbf{X}_{t_{k-1}}) - \mathbf{g}(\boldsymbol{\beta}; \mathbf{X}_{t_k}, \mathbf{X}_{t_{k-1}})$. Using the results from Lemma S1.3 and the proof of Theorem 5.1, we get the following limits:

$$\begin{aligned}
&\partial_{\beta_{i_1}\beta_{i_2}} \frac{1}{h} T_2(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{1}{2}\int (\partial_{\beta_{i_1}} \mathbf{N}_0(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1} \partial_{\beta_{i_2}} \mathbf{N}_0(\mathbf{x})\, \mathrm{d}\nu_0(\mathbf{x}), \\
&\partial_{\beta_{i_1}\beta_{i_2}} \frac{1}{h} T_3(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \\
&\quad \frac{1}{2}\int (\partial_{\beta_{i_1}} \mathbf{N}_0(\mathbf{x}) + 2\partial_{\beta_{i_1}} \mathbf{A}_0(\mathbf{x}-\mathbf{b}_0))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1} (\partial_{\beta_{i_2}} \mathbf{N}_0(\mathbf{x}) + 2\partial_{\beta_{i_2}} \mathbf{A}_0(\mathbf{x}-\mathbf{b}_0))\, \mathrm{d}\nu_0(\mathbf{x}), \\
&\partial_{\beta_{i_1}\beta_{i_2}} \frac{1}{h} T_5(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{1}{2}\int (\partial_{\beta_{i_1}} \mathbf{F}_0(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1} \partial_{\beta_{i_2}} \mathbf{N}_0(\mathbf{x})\, \mathrm{d}\nu_0(\mathbf{x}) \\
&\qquad\qquad + \frac{1}{2}\int (\partial_{\beta_{i_2}} \mathbf{A}_0(\mathbf{x}-\mathbf{b}_0))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1} \partial_{\beta_{i_1}} \mathbf{N}_0(\mathbf{x})\, \mathrm{d}\nu_0(\mathbf{x}), \\
&\partial_{\beta_{i_1}\beta_{i_2}} \frac{1}{h} T_6(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} -\frac{1}{2}\int \operatorname{Tr}(D\partial_{\beta_{i_1}\beta_{i_2}} \mathbf{N}_0(\mathbf{x}))\, \mathrm{d}\nu_0(\mathbf{x}),
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. Since $\frac{1}{h}T_4 \to 0$, the partial derivatives go to zero too. From Lemma 4.2, for $Nh \to \infty$, $h \to 0$, we have:

$$\frac{1}{N}\sum_{k=1}^{N} \partial_{\beta_{i_1}\beta_{i_2}} \operatorname{Tr} D\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \int \operatorname{Tr}(D\partial_{\beta_{i_1}\beta_{i_2}} \mathbf{N}_0(\mathbf{x}))\, \mathrm{d}\nu_0(\mathbf{x}).$$

Term $\frac{1}{Nh}\sum_{k=1}^{N}\partial_{\beta_{i_1}\beta_{i_2}}(\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta}))$, evaluated in $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, has only one term of order h: $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}\beta_{i_2}}\mathbf{A}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$, which converges to $\partial_{\beta_{i_1}\beta_{i_2}}\operatorname{Tr}\mathbf{A}(\boldsymbol{\beta}_0)$ (Property 1 Lemma S1.3).

Thus, $\frac{1}{Nh}\partial_{\beta_{i_1}\beta_{i_2}}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}_0)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\to 2\int(\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{x}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{x})\,\mathrm{d}\nu_0(\mathbf{x})$, in $\mathbb{P}_{\boldsymbol{\theta}_0}$ for $Nh\to\infty$, $h\to 0$.

Now, we prove $\frac{1}{N\sqrt{h}}\partial_{\beta\varsigma}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}\to 0$, in $\mathbb{P}_{\boldsymbol{\theta}_0}$ for $Nh\to\infty$, $h\to 0$. For a constant C_h, depending on h, $l=2,3,...,6$, and generic functions $\mathbf{g},\mathbf{g}_1$, the following term is at most of order $R(h,\mathbf{x}_0)$:

$$\partial_{\beta_i}T_l(\boldsymbol{\beta},\boldsymbol{\varsigma})=C_h\sum_{k=1}^{N}(\mathbf{g}(\boldsymbol{\beta}_0;\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}})-\mathbf{g}(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}_1(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}),$$

Then, term $\partial_{\beta\varsigma}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})$ still contains $\mathbf{g}(\boldsymbol{\beta}_0;\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}})-\mathbf{g}(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}})$ which is 0 for $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. Moreover, the term $\frac{1}{N}\sum_{k=1}^{N}\partial_{\boldsymbol{\beta}\boldsymbol{\varsigma}}(\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta}))$ is at most of order $R(h,\mathbf{x}_0)$. Thus, $\frac{1}{N\sqrt{h}}\partial_{\beta\varsigma}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}=0$.

Finally, we compute $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})$. As before, it holds $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}T_l(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}\to 0$, for $l=2,3,...,6$. Similarly, we see that $\frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\partial_{\varsigma_{j_1}\varsigma_{j_2}}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$ is at most of order $R(h,\mathbf{x}_0)$. So, we need to compute the following second derivatives $\partial_{\varsigma_{j_1}\varsigma_{j_2}}\log(\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ and $\partial_{\varsigma_{j_1}\varsigma_{j_2}}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$. The first one yields:

$$\begin{aligned}&\partial_{\varsigma_{j_1}\varsigma_{j_2}}\log(\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\\&=\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)-\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top).\end{aligned}$$

On the other hand, we have:

$$\begin{aligned}&\partial_{\varsigma_{j_1}\varsigma_{j_2}}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\\&=-\frac{1}{Nh}\sum_{k=1}^{N}\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_1}\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\\&+\frac{1}{Nh}\sum_{k=1}^{N}\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\\&+\frac{1}{Nh}\sum_{k=1}^{N}\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}).\end{aligned}$$

Then, from Property 1 of Lemma S1.3, we get:

$$\begin{aligned}&\partial_{\varsigma_{j_1}\varsigma_{j_2}}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\Big|_{\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}\\&\xrightarrow[\substack{Nh\to\infty\\h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}}2\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)-\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top).\end{aligned}$$

Thus, $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}\to\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$. Since all the limits used in this proof are uniform in $\boldsymbol{\theta}$, the first part of the lemma is proved. The second part is trivial, because all limits are continuous in $\boldsymbol{\theta}$. □

PROOF OF LEMMA S1.5. First, we compute the first derivatives. We start with:

$$\partial_{\beta_i}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})=-2\sum_{k=1}^{N}\operatorname{Tr}(D\boldsymbol{f}_{h/2,k}(\boldsymbol{\beta})D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}))$$

$$+\frac{2}{h}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^{\top}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}(\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\partial_{\beta_i}\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta})).$$

The first derivative with respect to $\boldsymbol{\varsigma}$ is:

$$\begin{aligned}\partial_{\varsigma_j}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) &= N\partial_{\varsigma_j}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})\\ &\quad+\frac{1}{h}\partial_{\varsigma_j}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^{\top}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))\\ &=-\frac{1}{h}\sum_{k=1}^{N}\Bigg(\operatorname{Tr}\Big((\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^{\top}\\ &\qquad(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}\Big)+\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})\Bigg).\end{aligned}$$

Define:

$$\begin{aligned}\eta_{N,k}^{(i)}(\boldsymbol{\theta}) &:= \frac{2}{\sqrt{Nh}}\operatorname{Tr}(D\boldsymbol{f}_{h/2,k}(\boldsymbol{\beta})D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}))\\ &\quad-\frac{2}{\sqrt{Nh}h}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^{\top}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}\partial_{\beta_i}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))\end{aligned}\tag{S25}$$

$$\begin{aligned}\zeta_{N,k}^{(j)}(\boldsymbol{\theta}) &:= \frac{1}{\sqrt{Nh}}\operatorname{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta})^{\top}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1})\\ &\quad-\frac{1}{\sqrt{N}}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top})^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}),\end{aligned}\tag{S26}$$

and rewrite $\boldsymbol{\lambda}_N$ as $\boldsymbol{\lambda}_N=\sum_{k=1}^{N}[\eta_{N,k}^{(1)}(\boldsymbol{\theta}_0),\ldots,\eta_{N,k}^{(r)}(\boldsymbol{\theta}_0),\zeta_{N,k}^{(1)}(\boldsymbol{\theta}_0),\ldots,\zeta_{N,k}^{(s)}(\boldsymbol{\theta}_0)]^{\top}$. Now, by Proposition 3.1 from Crimaldi and Pratelli (2005), it is sufficient to prove Lemma S1.6. □

LEMMA S1.6. *Let $\eta_{N,k}^{(i)}(\boldsymbol{\theta})$ and $\zeta_{N,k}^{(j)}(\boldsymbol{\theta})$ be defined as in (S25) and (S26), respectively. If $h\to 0$, $Nh\to\infty$, and $Nh^2\to 0$, then for and all $i,i_1,i_2=1,2,...,r$, and $j,j_1,j_2=1,2,...,s$, it holds:*

(i) $\mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{1\le k\le N}|\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)|]\longrightarrow 0$, *and* $\mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{1\le k\le N}|\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)|]\longrightarrow 0$;
(ii) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$, *and* $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(iii) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(iv) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(v) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(vi) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}4[\mathbf{C}_{\beta}(\boldsymbol{\theta}_0)]_{i_1i_2}$;
(vii) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}4[\mathbf{C}_{\varsigma}(\boldsymbol{\theta}_0)]_{j_1j_2}$;
(viii) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(ix) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[(\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0))^2\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(x) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[(\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0))^2\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$;
(xi) $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[(\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)^2)\mid\mathbf{X}_{t_{k-1}}]\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0$.

PROOF OF LEMMA S1.6. The proof of Lemma S1.6 is technical and involves bounding the sums of triangular arrays in such a way that the bound converges to zero in probability $\mathbb{P}_{\boldsymbol{\theta}_0}$ as $h\to 0$, $Nh\to\infty$, and $Nh^2\to 0$. Unlike in the previous proof, this time we do not require uniform convergence.

We begin by expanding $\eta_k^{(i)}$ to differentiate between terms that vanish and those that do not in the limits:

$$\begin{aligned}
\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0) &= \frac{2}{\sqrt{Nh}}\operatorname{Tr}((\mathbf{I}+\frac{h}{2}D\mathbf{N}_0(\mathbf{X}_{t_k}))(-\frac{h}{2}D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})))\\
&\quad-\frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(-\frac{h}{2}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})+\frac{h^2}{8}\partial_{\beta_i}(D\mathbf{N}_0(\mathbf{X}_{t_k}))\mathbf{N}_0(\mathbf{X}_{t_k}))\\
&\quad+\frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0)+R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}})\\
&=-\sqrt{\frac{h}{N}}\operatorname{Tr}(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}))+\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\\
&\quad-\frac{1}{4}\sqrt{\frac{h}{N}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}(D\mathbf{N}_0(\mathbf{X}_{t_k}))\mathbf{N}_0(\mathbf{X}_{t_k})\\
&\quad+\frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0)+R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}}).
\end{aligned}\tag{S27}$$

Proof of (i). Let us begin by examining the limit of the expectation of $\sup_{1\le k\le N}|\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)|$. In equation (S27), all the involved functions are bounded, and the term with the largest order is $R(\sqrt{Nh},\mathbf{X}_{t_{k-1}})$ because $\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0)$ is $\mathbf{R}(h,\mathbf{X}_{t_{k-1}})$. The remaining terms converge to zero. Moreover, terms with coefficients $\frac{1}{\sqrt{Nh}}$ take the form $\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{g}$, where $\mathbf{g}$ is a vector-valued function of either $\mathbf{X}_{t_{k-1}}$ or $\mathbf{X}_{t_k}$. Their expected values are bounded by $R(h,\mathbf{X}_{t_{k-1}})$ at most. Thus, the dominant order becomes $R(\sqrt{h/N},\mathbf{X}_{t_{k-1}})$, which indeed converges to zero.

We proceed to analyze the limit of the expectation of $\sup_{1\le k\le N}|\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)|$. The leading term in $\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)$, as defined in the paper, has an order $R(1/\sqrt{Nh^2},\mathbf{X}_{t_{k-1}})$. Upon calculating its expected value, we obtain an order of $R(h,\mathbf{X}_{t_{k-1}})$. This concludes the proof of (i).

To establish limits (ii)-(v), we need to calculate the expectations of $\eta_{N,k}^{(i)}$ and $\zeta_{N,k}^{(i)}$. By analyzing (S27), we can deduce that $\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}})$, since Proposition 4.3 gives:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\mid\mathbf{X}_{t_{k-1}}]=\sqrt{\frac{h}{N}}\operatorname{Tr}(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}))+R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}}),$$

Similarly, from:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1})\mid\mathbf{X}_{t_{k-1}}]=h\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)+R(h^2,\mathbf{X}_{t_{k-1}})$$

we conclude that $\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(h/\sqrt{N},\mathbf{X}_{t_{k-1}})$. Then, combining the previous, we get:

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(\sqrt{Nh^3},\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,$$

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(\sqrt{Nh^2},\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,$$

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(h^3,\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,$$

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(h^2,\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,$$

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(h^{5/2},\mathbf{X}_{t_{k-1}})\xrightarrow[N\to\infty]{\mathbb{P}_{\boldsymbol{\theta}_0}}0.$$

Now, we prove limit (vi). Here, we focus on the terms of order $1/\sqrt{Nh}$ in $\eta^{(i)}_{N,k}$ which are the only ones that will not converge to zero when multiplying $\eta^{(i_1)}_{N,k}$ and $\eta^{(i_2)}_{N,k}$:

$$\begin{aligned}
\eta^{(i)}_{N,k}(\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}) \\
&+ \frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0) + R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}}) \\
&= \frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}) + \frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}(\mathbf{N}_0(\mathbf{X}_{t_{k-1}}) \\
&+ 2\mathbf{A}_0(\mathbf{X}_{t_{k-1}}-\mathbf{b}_0)) + R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}}) \\
&= \frac{2}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{F}_0(\mathbf{X}_{t_{k-1}}) + \frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^i_{k,k-1}(\boldsymbol{\beta}_0) + R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}}),
\end{aligned}$$

In the previous calculations, we introduced a new notation $\boldsymbol{\psi}^i_{k,k-1}(\boldsymbol{\beta}_0) \coloneqq \partial_{\beta_i}(\mathbf{N}_0(\mathbf{X}_{t_k}) - \mathbf{N}_0(\mathbf{X}_{t_{k-1}}))$. Now, we consider the product $\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0)$ and again focus only on the terms with coefficient $1/Nh$:

$$\begin{aligned}
\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0) &= \frac{4}{Nh}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{2}{Nh}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{2}{Nh}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\boldsymbol{\psi}^{i_2}_{k,k-1}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{1}{Nh}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\boldsymbol{\psi}^{i_2}_{k,k-1}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} + R(1/N,\mathbf{X}_{t_{k-1}}).
\end{aligned}$$

In the previous equation, we must show that the sum of expectations of all the terms except the first converges to zero. We only prove this for the second row; the rest follows analogously. Due to the definition of $\boldsymbol{\psi}^i$, it is clear that $\mathbb{E}_0[\|\boldsymbol{\psi}^i_{k,k-1}(\boldsymbol{\beta}_0)\|^p \mid \mathbf{X}_{t_{k-1}}] = \mathbf{R}(h,\mathbf{X}_{t_{k-1}})$, for all $p \geq 1$. Then, we use property (S13) to obtain:

$$\begin{aligned}
&\frac{1}{Nh}|\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}]| \\
&\leq \frac{1}{Nh}|\operatorname{Tr}(\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1})|\|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\|\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\|\|\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\| \mid \mathbf{X}_{t_{k-1}}] \\
&\leq \frac{C}{Nh}(\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\|^2 \mid \mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\|\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\|^2 \mid \mathbf{X}_{t_{k-1}}])^{\frac{1}{2}} \\
&= \frac{1}{Nh}(R(h^2,\mathbf{X}_{t_{k-1}})R(h,\mathbf{X}_{t_{k-1}}))^{\frac{1}{2}} = R(\sqrt{h}/N,\mathbf{X}_{t_{k-1}}).
\end{aligned}$$

Finally, we use Lemma 4.2 to get:

$$\begin{aligned}
&\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{4}{Nh}\sum_{k=1}^N(\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}] + R(h^{3/2},\mathbf{X}_{t_{k-1}})) \\
&= \frac{4}{N}\sum_{k=1}^N(\operatorname{Tr}(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)) + R(\sqrt{h},\mathbf{X}_{t_{k-1}})) \xrightarrow[N\to\infty]{\mathbb{P}_{\boldsymbol{\theta}_0}} 4[\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)]_{i_1i_2}.
\end{aligned}$$

To prove (vii) we use Corollary 3.8:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j_1)}_{N,k}(\boldsymbol{\theta}_0)\zeta^{(j_2)}_{N,k}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}] \\
&- \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) \\
&= \frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{h,k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}_{h,k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}] \\
&- \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) + R(\sqrt{h}/N, \mathbf{X}_{t_{k-1}}).
\end{aligned}$$

Now, we use the expectation of a product of two quadratic forms of normally distributed random vectors (see for example Section 2 in Kumar (1973)) to get:

$$\begin{aligned}
&\frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{h,k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}_{h,k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}] \\
&= \frac{2}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_1}}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_2}}) + \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_1}})\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_2}}).
\end{aligned}$$

This proves (vii). We omit the proofs of (viii)-(xi) since they follow the same pattern. Namely, we find the leading term and ensure it goes to zero. For the expectations of squares, we can apply the same approach with a product of two quadratic forms. □

S2. Auxiliary properties. In this section, we revisit crucial properties essential for establishing the consistency and asymptotic normality of the proposed estimators. To begin, we invoke Lemma 2.3 from Tian and Fan (2020) as Lemma S2.1, which was used in proving Lemma 4.1. This lemma offers a generalization of the Grönwall's inequality.

Furthermore, Lemma 9 in Genon-Catalot and Jacob (1993) provides conditions for the convergence of a sum of a triangular array and is recalled as Lemma S2.2.

Lemmas S2.3 and S2.4 give sufficient conditions for uniform convergence. The former is sourced from Proposition A1 in Gloter (2006), while the latter comes from Lemma 3.1 from Yoshida (1990). On occasions, Lemma S2.3 might not suffice, warranting the use of Lemma S2.4. Theorem S2.5 is a helpful tool for assessing the conditions of these two lemmas is the Rosenthal's inequality for martingales (Theorem 2.12 in Hall and Heyde (1980)).

Lastly, Theorem S2.6 presents a special case of the central limit theorem for multivariate martingale triangular arrays (Proposition 3.1 from Crimaldi and Pratelli (2005)). This theorem is pivotal for proving the asymptotic normality of the proposed estimators.

LEMMA S2.1 (Generalized Grönwall's inequality, Lemma 2.3 in Tian and Fan (2020)). *Let $p > 1$ and $b > 0$ be constants, and let $a : (0, +\infty) \to (0, +\infty)$ be a continuous function. If*

$$u(t) \le a(t) + b\int_0^t u^p(s)\,\mathrm{d}s,$$

then $u(t) \le a(t) + (\kappa^{1-p}(t) - (p-1)2^{p-1}bt)^{\frac{1}{1-p}}$ and $\kappa^{1-p}(t) > (p-1)2^{p-1}bt$, where

$$\kappa(t) \coloneqq 2^{p-1}b\int_0^t a^p(s)\,\mathrm{d}s. \tag{S28}$$

LEMMA S2.2 (Lemma 9 in Genon-Catalot and Jacob (1993)). *Let $(X_k^N)_{N\in\mathbb{N},1\le k\le N}$ be a triangular array with each row N adapted to a filtration $(\mathcal{G}_k^N)_{1\le k\le N}$, and let U be a random variable. If*

$$\sum_{k=1}^N \mathbb{E}[X_k^N \mid \mathcal{G}_{k-1}^N] \xrightarrow[N\to\infty]{\mathbb{P}} U, \qquad \sum_{k=1}^N \mathbb{E}[(X_k^N)^2 \mid \mathcal{G}_{k-1}^N] \xrightarrow[N\to\infty]{\mathbb{P}} 0,$$

then $\sum_{k=1}^N X_k^N \xrightarrow[N\to\infty]{\mathbb{P}} U$.

LEMMA S2.3 (Proposition A1 in Gloter (2006)). *Let $S_N(\omega, \boldsymbol{\theta})$ be a sequence of measurable real-valued functions defined on $\Omega \times \Theta$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and Θ is product of compact intervals of $\mathbb{R}$. We assume that $S_N(\cdot, \boldsymbol{\theta})$ converges to a constant C in probability for all $\boldsymbol{\theta} \in \Theta$; and that there exists an open neighbourhood of Θ on which $S_N(\omega, \cdot)$ is continuously differentiable for all $\omega \in \Omega$. Furthermore, we suppose that:*

$$\sup_{N \in \mathbb{N}} \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} |\nabla_{\boldsymbol{\theta}} S_N(\boldsymbol{\theta})|] < \infty.$$

Then, $S_N(\boldsymbol{\theta}) \xrightarrow[N \to \infty]{\mathbb{P}} C$ uniformly in $\boldsymbol{\theta}$.

LEMMA S2.4 (Lemma 3.1 in Yoshida (1990)). *Let $F \subset \mathbb{R}^d$ be a convex compact set, and let $\{\xi_N(\boldsymbol{\theta}); \boldsymbol{\theta} \in F\}$, be a family of real-valued random processes for $N \in \mathbb{N}$. If there exist constants $p \geq l > d$ and $C > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, it holds:*

(1) $\mathbb{E}[|\xi_N(\boldsymbol{\theta}_1) - \xi_N(\boldsymbol{\theta}_2)|^p] \leq C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^l$;
(2) $\mathbb{E}[|\xi_N(\boldsymbol{\theta})|^p] \leq C$;
(3) $\xi_N(\boldsymbol{\theta}) \xrightarrow[N \to \infty]{\mathbb{P}} 0$,

then $\sup_{\boldsymbol{\theta} \in F} |\xi_N(\boldsymbol{\theta})| \xrightarrow[N \to \infty]{\mathbb{P}} 0$.

THEOREM S2.5 (Rosenthal's inequality, Theorem 2.12 in Hall and Heyde (1980)). *Let $(X_k^N)_{N \in \mathbb{N}, 1 \leq k \leq N}$ be a triangular array with each row N adapted to a filtration $(\mathcal{G}_k^N)_{1 \leq k \leq N}$ and let:*

$$S_N = \sum_{k=1}^{N} X_k^N, \quad N \in \mathbb{N}$$

be a martingale array. Then, for all $p \in [2, \infty)$ there exist constants C_1, C_2 such that:

$$C_1(\mathbb{E}[(\sum_{k=1}^{N} \mathbb{E}[(X_k^N)^2 \mid \mathcal{G}_{k-1}^N])^{\frac{p}{2}}] + \sum_{k=1}^{N} \mathbb{E}[|X_k^N|^p]) \leq \mathbb{E}[|S_N|^p] \leq C_2(\mathbb{E}[(\sum_{k=1}^{N} \mathbb{E}[(X_k^N)^2 \mid \mathcal{G}_{k-1}^N])^{\frac{p}{2}}] + \sum_{k=1}^{N} \mathbb{E}[|X_k^N|^p]).$$

THEOREM S2.6 (Proposition 3.1. in Crimaldi and Pratelli (2005)). *Let $(\mathbf{X}_{N,k})_{N \in \mathbb{N}, 1 \leq k \leq N}$ be a triangular array of d-dimensional random vectors, such that, for each N, the finite sequence $(\mathbf{X}_{N,k})_{1 \leq k \leq N}$ is a martingale difference array with respect to a given filtration $(\mathcal{G}_k^N)_{1 \leq k \leq N}$ such that:*

$$\mathbf{S}_N = \sum_{k=1}^{N} \mathbf{X}_{N,k}, \ N \in \mathbf{N}.$$

If

(1) $\mathbb{E}[\sup_{1 \leq k \leq N} \|\mathbf{X}_{N,k}\|_1] \xrightarrow[N \to \infty]{} 0$;
(2) $\sum_{k=1}^{N} \mathbf{X}_{N,k} \mathbf{X}_{N,k}^\top \xrightarrow[N \to \infty]{\mathbb{P}} \mathbf{U}$, *for some non-random positive semi-definite matrix* $\mathbf{U}$,

then, $\mathbf{S}_N \xrightarrow[N \to \infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$.

REMARK S1. *Instead of using the second condition of Theorem S2.6, Lemma S2.4 yields that it is sufficient to prove that, for all $i, j = 1, ..., d$, it holds:*

$$\sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N] \xrightarrow[N \to \infty]{\mathbb{P}} U_{ij}, \qquad \sum_{k=1}^{N} \mathbb{E}[(X_{N,k}^{(i)} X_{N,k}^{(j)})^2 \mid \mathcal{G}_{k-1}^N] \xrightarrow[N \to \infty]{\mathbb{P}} 0.$$

REMARK S2. *For a martingale difference array the conditional expectations need to be zero almost surely, i.e:*

$$\mathbb{E}[\mathbf{X}_{N,k} \mid \mathcal{G}_{k-1}^N] = 0, \text{ a.s. for all } N \in \mathbb{N},\ 1 \le k \le N.$$

In our case, $(\mathbf{X}_{N,k})_{N\in\mathbb{N},1\le k\le N}$ *does not fulfil the previous condition. Hence, similar to the approach in Corollary 2.6 of McLeish (1974), we need the following two additional conditions on* $(\mathbf{X}_{N,k})_{N\in\mathbb{N},1\le k\le N}$*:*

$$\text{(S29)} \qquad \sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N] \xrightarrow[N\to\infty]{\mathbb{P}} 0, \qquad \sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N]\mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N] \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Indeed, martingale difference array $\mathbf{Y}_{N,k} = \mathbf{X}_{N,k} - \mathbb{E}[\mathbf{X}_{N,k} \mid \mathcal{G}_{k-1}^N]$ *satisfies conditions of the previous theorem. To prove that the first condition is satisfied, we write:*

$$\begin{aligned}\mathbb{E}[\sup_{1\le k\le N} \|\mathbf{Y}_{N,k}\|_1] &\le \mathbb{E}[\sup_{1\le k\le N} \|\mathbf{X}_{N,k}\|_1] + \mathbb{E}[\sup_{1\le k\le N} \mathbb{E}[\|\mathbf{X}_{N,k}\|_1 \mid \mathcal{G}_{k-1}^N]] \\ &\le \mathbb{E}[\sup_{1\le k\le N} \|\mathbf{X}_{N,k}\|_1] + \mathbb{E}[\sup_{1\le k\le N} \mathbb{E}[\sup_{1\le j\le N} \|\mathbf{X}_{N,j}\|_1 \mid \mathcal{G}_{k-1}^N]] \le 3\mathbb{E}[\sup_{1\le k\le N} \|\mathbf{X}_{N,k}\|_1] \xrightarrow[N\to\infty]{} 0.\end{aligned}$$

We used the Doob's inequality for the last submartingale. To demonstrate the second condition we fix i, j *to get:*

$$\begin{aligned}\sum_{k=1}^{N} Y_{N,k}^{(i)} Y_{N,k}^{(j)} &= \sum_{k=1}^{N} X_{N,k}^{(i)} X_{N,k}^{(j)} - \sum_{k=1}^{N} X_{N,k}^{(i)} \mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N] \\ &\quad - \sum_{k=1}^{N} X_{N,k}^{(j)} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N] + \sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N]\mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N].\end{aligned}$$

The first term goes to U_{ij}*, and the last term goes to zero. To prove that middle terms also vanish, we use the following inequalities:*

$$\begin{aligned}|\sum_{k=1}^{N} X_{N,k}^{(i)} \mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N]| &\le \sum_{k=1}^{N} |X_{N,k}^{(i)}||\mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N]| \\ &\le (\sum_{k=1}^{N} (X_{N,k}^{(i)})^2 \sum_{k=1}^{N} \mathbb{E}^2[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N])^{\frac{1}{2}} \xrightarrow[N\to\infty]{} 0.\end{aligned}$$

Theorem S2.6 yields that $\sum_{k=1}^{N} \mathbf{Y}_{N,k} \xrightarrow[N\to\infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$*, which together with (S29), gives* $\mathbf{S}_N \xrightarrow[N\to\infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$.

S3. Estimators. In this section, we treat the computation of integrals involving matrix exponentials, using formulas from (Van Loan, 1978) and apply it to the LL estimator, following (Gu, Wu and Xue, 2020). In the main paper, we extend this approach to calculate $\mathbf{\Omega}_h$ for the splitting schemes.

Additionally, we present the coefficients for the HE log-likelihood expansion up to order $J = 2$ for the Lorenz system, with our gratitude to the third reviewer for providing these formulas. The section concludes with a detailed analysis of the simulation results for the HE method.

S3.1. *Ozaki's local linearization.* Building on the approach by Gu, Wu and Xue (2020), we can efficiently compute $\mathbf{R}_{h,i}$ and $\mathbf{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ using the following procedure. To begin, define the three block matrices:

$$\text{(S30)} \quad \mathbf{P}_1(\mathbf{x}) = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) \end{bmatrix}, \mathbf{P}_2(\mathbf{x}) = \begin{bmatrix} -D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \mathbf{I}_d & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix}, \mathbf{P}_3(\mathbf{x}) = \begin{bmatrix} D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \mathbf{\Sigma}\mathbf{\Sigma}^\top \\ \mathbf{0}_{d\times d} & -D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})^\top \end{bmatrix}.$$

Then, we compute the matrix exponential of matrices $h\mathbf{P}_1(\mathbf{x})$ and $h\mathbf{P}_2(\mathbf{x})$:

$$\exp(h\mathbf{P}_1(\mathbf{x})) = \begin{bmatrix} \star & \mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) \\ \mathbf{0}_{d\times d} & \star \end{bmatrix}, \qquad \exp(h\mathbf{P}_2(\mathbf{x})) = \begin{bmatrix} \star & \star & \mathbf{B}_{\mathbf{R}_{h,1}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) \\ \mathbf{0}_{d\times d} & \star & \star \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \star \end{bmatrix}.$$

The terms marked with $\star$ symbols can be disregarded. Starting with the first matrix, we derive $\mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$. Then, we compute $\mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$ using the formula $\mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) = \exp(hD\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))\mathbf{B}_{\mathbf{R}_{h,1}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$. Finally, we obtain $\boldsymbol{\Omega}^{[\mathrm{LL}]}_{h,k}(\boldsymbol{\theta})$ from the matrix exponential:

$$\exp(h\mathbf{P}_3(\mathbf{x})) = \begin{bmatrix} \mathbf{B}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta})\, \mathbf{C}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta}) \\ \mathbf{0}_{d\times d} \qquad\qquad \star \end{bmatrix},$$
$$\boldsymbol{\Omega}^{[\mathrm{LL}]}_{h,k}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta})\mathbf{B}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta})^\top.$$

S3.2. *Aït-Sahalia's Infinite Hermite Expansion.* Polynomial coefficients $C_Y^{(j)}(\boldsymbol{\gamma}(\mathbf{X}_{t_k}) \mid \boldsymbol{\gamma}(\mathbf{X}_{t_{k-1}}))$, for $j = -1, 0, 1, \ldots, J$ are calculated recursively according to Theorem 1 in (Aït-Sahalia, 2008). In the following, we present $C_Y^{(j)}$ for the Lorenz system up to order $J = 2$ (provided by the third reviewer):

$$C_Y^{(-1)}(\boldsymbol{\gamma}(x,y,z) \mid \boldsymbol{\gamma}(x_0,y_0,z_0)) = -\frac{1}{2}\left(\frac{(x-x_0)^2}{\sigma_1^2} + \frac{(y-y_0)^2}{\sigma_2^2} + \frac{(z-z_0)^2}{\sigma_3^2}\right);$$

$$C_Y^{(0)}(\boldsymbol{\gamma}(x,y,z) \mid \boldsymbol{\gamma}(x_0,y_0,z_0)) = \frac{1}{3}(x-x_0)(y-y_0)(z-z_0)\left(-\frac{1}{\sigma_2^2} + \frac{1}{\sigma_3^2}\right)$$
$$-\frac{1}{2}\left(\frac{p(x-x_0)^2}{\sigma_1^2} + \frac{(y-y_0)^2}{\sigma_2^2} + \frac{c(z-z_0)^2}{\sigma_3^2}\right)$$
$$+\frac{1}{2}x_0(y-y_0)(z-z_0)\left(-\frac{1}{\sigma_2^2} + \frac{1}{\sigma_3^2}\right) + \frac{1}{2}(x-x_0)(y-y_0)\left(\frac{p}{\sigma_1^2} + \frac{r-z_0}{\sigma_2^2}\right) + \frac{1}{2}(x-x_0)(z-z_0)\frac{y_0}{\sigma_3^2}$$
$$+(x-x_0)\frac{p(-x_0+y_0)}{\sigma_1^2} + (y-y_0)\frac{(rx_0 - y_0 - x_0z_0)}{\sigma_2^2} + (z-z_0)\frac{(x_0y_0 - cz_0)}{\sigma_3^2};$$

$$C_Y^{(1)}(\boldsymbol{\gamma}(x,y,z) \mid \boldsymbol{\gamma}(x_0,y_0,z_0)) = \frac{1}{24}(x-x_0)^2\left(\frac{p^2\sigma_2^2}{\sigma_1^4} - \frac{4p^2 + 2p(r-z_0)}{\sigma_1^2} - \frac{3(r-z_0)^2}{\sigma_2^2} - \frac{3y_0^2}{\sigma_3^2}\right)$$
$$+\frac{1}{24}(y-y_0)^2\left(\frac{\sigma_1^2(r-z_0)^2 + \sigma_3^2x_0^2}{\sigma_2^4} - \frac{3p^2}{\sigma_1^2} + \frac{2(x_0^2 - p(r-z_0) - 2)}{\sigma_2^2} - \frac{3x_0^2}{\sigma_3^2}\right)$$
$$+\frac{1}{24}(z-z_0)^2\left(\frac{\sigma_1^2y_0^2 + \sigma_2^2x_0^2}{\sigma_3^4} - \frac{3x_0^2}{\sigma_2^2} + \frac{2(x_0^2 - 2c^2)}{\sigma_3^2}\right)$$
$$+\frac{1}{12}(x-x_0)(y-y_0)\left(\frac{4p^2}{\sigma_1^2} + \frac{x_0y_0 + 4(r-z_0)}{\sigma_2^2} - \frac{7x_0y_0 - 4cz_0}{\sigma_3^2}\right)$$
$$+\frac{1}{12}(y-y_0)(z-z_0)\left(\frac{\sigma_1^2y_0(r-z_0)}{\sigma_2^2\sigma_3^2} - \frac{4x_0}{\sigma_2^2} - \frac{py_0 - 4cx_0}{\sigma_3^2}\right)$$
$$+\frac{1}{12}(x-x_0)(z-z_0)\left(\frac{px_0\sigma_2^2}{\sigma_1^2\sigma_3^2} + \frac{px_0}{\sigma_1^2} - \frac{4y_0 + 7x_0(r-z_0)}{\sigma_2^2} + \frac{4cy_0 - x_0(r-z_0)}{\sigma_3^2}\right)$$
$$+\frac{1}{2}(x-x_0)\left(\frac{p^2(-x_0+y_0)}{\sigma_1^2} + \frac{-x_0(r-z_0)^2 + y_0(r-z_0)}{\sigma_2^2} + \frac{y_0(-x_0y_0 + cz_0)}{\sigma_3^2}\right)$$
$$+\frac{1}{2}(y-y_0)\left(\frac{p^2(x_0-y_0)}{\sigma_1^2} + \frac{x_0(r-z_0) - y_0}{\sigma_2^2} + \frac{x_0(-x_0y_0 + cz_0)}{\sigma_3^2}\right)$$
$$+\frac{1}{2}(z-z_0)\left(\frac{x_0(-y_0 + x_0(r-z_0))}{\sigma_2^2} + \frac{c(x_0y_0 - cz_0)}{\sigma_3^2}\right)$$
$$+\frac{1}{2}\left(1 + p + c - \frac{p^2(x_0-y_0)^2}{\sigma_1^2} - \frac{(x_0y_0 - cz_0)^2}{\sigma_3^2} - \frac{(-x_0(r-z_0) + y_0)^2}{\sigma_2^2}\right);$$

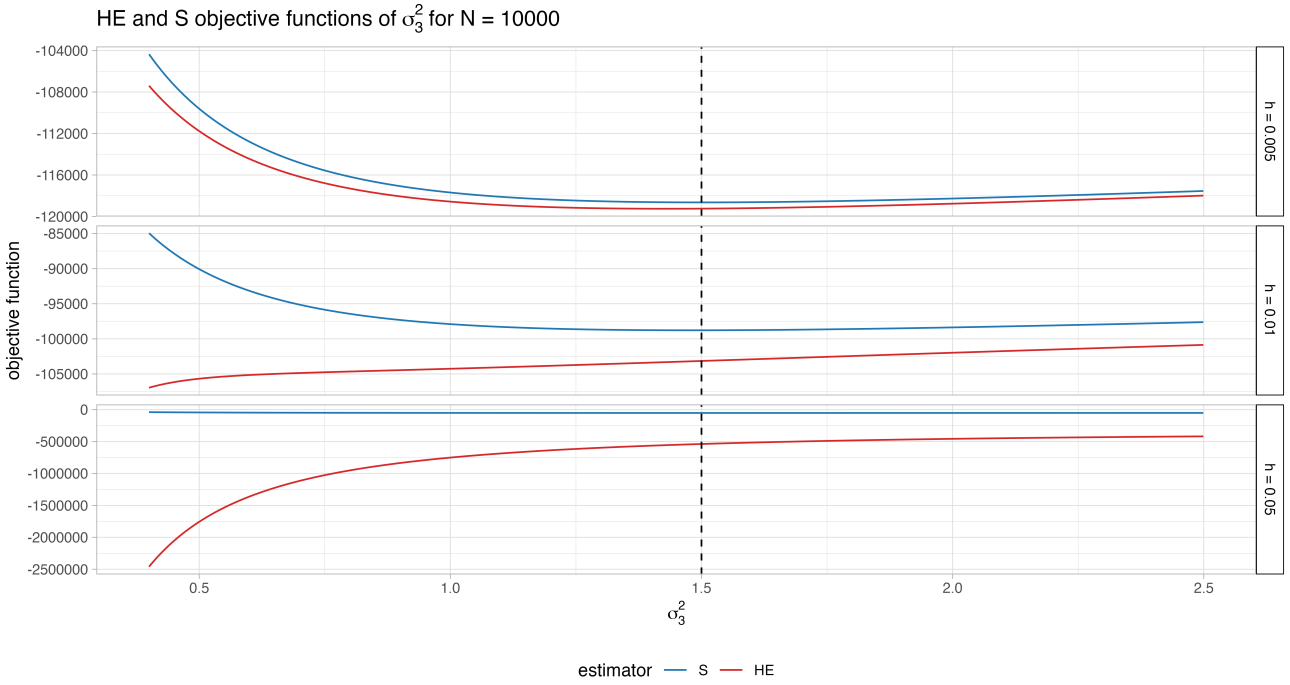

Fig 1: Comparing S (red) and HE (blue) objective functions of a data set generated from the Lorenz system where all parameters except σ_3^2 are fixed to the true values. The sample size is fixed to $N = 10000$. Each row represents one value of the discretization step h. The black vertical dashed line is the true value of σ_3^2.

$$C_Y^{(2)}(\boldsymbol{\gamma}(x,y,z) \mid \boldsymbol{\gamma}(x_0,y_0,z_0)) = -\frac{1}{12}\left(\frac{p^2\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2(r-z_0)^2 + \sigma_3^2 x_0^2}{\sigma_2^2} + \frac{\sigma_1^2 y_0^2 + \sigma_2^2 x_0^2}{\sigma_3^2}\right) - \frac{1}{6}(1 + p^2 + c^2 - x_0^2 + rp - pz_0).$$

The poor performance of the HE estimator (no convergence for larger discretization step sizes h, and only $\approx 43 - 72\%$ convergence for small h) in the simulation study can probably be attributed to the polynomial approximation of the likelihood function, which can become unstable, particularly for larger h, as illustrated in Figure 1. Additional coefficients $C_Y^{(j)}$ in the approximation might mitigate this problem.

Figure 1 shows the objective functions of HE and S for a fixed trajectory, h, and N, with all parameters fixed to their true values except for σ_3^2. Consequently, the objective functions are presented as functions of σ_3^2. The HE function tends towards $-\infty$ as σ_3^2 approaches zero. This is also the case for the smallest h, although it is not evident in the figure due to the x-scale used. However, in this case the objective function do possess a local minimum close to the true value. As a result, the global minimum of the HE objective function is always at $-\infty$. For sufficiently small h, this issue can be mitigated by imposing constraints on σ_3^2. However, as h increases, the local minimum vanishes. In contrast, the objective functions of other estimators like S tend towards $+\infty$, when σ_3^2 goes to zero, ensuring that the minimum around the true value of σ_3^2 is also the global minimum of their objective functions.

REFERENCES

AÏT-SAHALIA, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics* **36** 906 – 937. https://doi.org/10.1214/009053607000000622

BUCKWAR, E., SAMSON, A., TAMBORRINO, M. and TUBIKANEC, I. (2022). A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *Applied Numerical Mathematics* **179** 191-220. https://doi.org/10.1016/j.apnum.2022.04.018

CRIMALDI, I. and PRATELLI, L. (2005). Convergence results for multivariate martingales. *Stochastic Processes and their Applications* **115** 571-577. https://doi.org/10.1016/j.spa.2004.10.004

GENON-CATALOT, V. and JACOB, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'I.H.P. Probabilités et statistiques* **29** 119–151. MR1204521

GLOTER, A. (2006). Parameter Estimation for a Discretely Observed Integrated Diffusion Process. *Scandinavian Journal of Statistics* **33** 83–104.

GU, W., WU, H. and XUE, H. (2020). *Parameter Estimation for Multivariate Nonlinear Stochastic Differential Equation Models: A Comparison Study* In *Statistical Modeling for Biological Systems: In Memory of Andrei Yakovlev* 245–258. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-34675-1_13

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. *Probability and mathematical statistics*. Academic Press.

KESSLER, M. (1997). Estimation of an Ergodic Diffusion from Discrete Observations. *Scandinavian Journal of Statistics* **24** 211–229.

KUMAR, A. (1973). Expectation of Product of Quadratic Forms. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)* **35** 359–362.

MCLEISH, D. L. (1974). Dependent Central Limit Theorems and Invariance Principles. *The Annals of Probability* **2** 620–628.

SØRENSEN, M. and UCHIDA, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9 (6)** 1051 – 1069.

TIAN, Y. and FAN, M. (2020). Nonlinear integral inequality with power and its application in delay integro-differential equations. *Advances in Difference Equations* **2020**. https://doi.org/10.1186/s13662-020-02596-y

VAN LOAN, C. (1978). Computing Integrals Involving the Matrix Exponential. *IEEE Trans. Aut. Cont.* **23** 395-404.

YOSHIDA, N. (1990). Asymptotic behavior of M-estimator and related random field for diffusion process. *Annals of the Institute of Statistical Mathematics* **42** 221-251. https://doi.org/10.1007/BF00050834

B Appendix to Parameter Estimation in Nonlinear Multivariate Second-order Stochastic Differential Equations with Additive Noise

S1 Supplementary Material

In this section, we provide proofs for all the nontrivial lemmas, propositions, and theorems presented in Sections 3 and 6. The majority of these proofs, especially those in Section 6, heavily rely on Itô or Taylor expansions in h around $\mathbf{Y}_{t_{k-1}}$. Additionally, we frequently employ Fubini's theorem as a useful tool. Our initial focus is on the results from Section 6, as they constitute technical auxiliary properties essential for understanding the main results outlined in Section 3.

S1.1 Proofs of results from Section 6

Proof of Lemma 6.2 To prove (i), calculate

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1} &= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{I}+\frac{h}{2}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})^{-1})+\mathbf{R}(h,\mathbf{y}_0)\\ &= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{I}-\frac{h}{2}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})+\mathbf{R}(h,\mathbf{y}_0)\\ &= \frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}-\frac{1}{2}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})+\mathbf{R}(h,\mathbf{y}_0).\end{aligned}$$

Proof of (ii):

$$\begin{aligned}\boldsymbol{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\theta})\boldsymbol{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1} &= (\frac{h^2}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\frac{h^3}{6}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top))\frac{1}{h}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\\ &\quad-\frac{h^2}{4}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}))+\mathbf{R}(h^3,\mathbf{y}_0)\\ &= \frac{h}{2}\mathbf{I}-\frac{h^2}{12}(\mathbf{A}_\mathbf{v}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})+\mathbf{R}(h^3,\mathbf{y}_0).\end{aligned}$$

To prove (iii), use the previous result to obtain:

$$\begin{aligned}&\boldsymbol{\Omega}_h^{[\mathrm{SR}]}(\boldsymbol{\theta})\boldsymbol{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta})^{-1}\boldsymbol{\Omega}_h^{[\mathrm{RS}]}(\boldsymbol{\theta})\\ &= (\frac{h}{2}\mathbf{I}-\frac{h^2}{12}(\mathbf{A}_\mathbf{v}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}))(\frac{h^2}{2}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\frac{h^3}{6}(2\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top))+\mathbf{R}(h^5,\mathbf{y}_0)\\ &= \frac{h^3}{4}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\frac{h^4}{8}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top)+\mathbf{R}(h^5,\mathbf{y}_0).\end{aligned}$$

Proof of (iv) follows from (iii) and Lemma 6.1. To prove (v), approximate the log-determinant as:

$$\begin{aligned}\log\det\boldsymbol{\Omega}_h^{[\mathrm{RR}]}(\boldsymbol{\theta}) &= \log\det\left(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\frac{h^2}{2}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top)\right)+R(h^2,\mathbf{y}_0)\\ &= d\log h+\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\log\det\left(\mathbf{I}+\frac{h}{2}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\right)+R(h^2,\mathbf{y}_0)\\ &= d\log h+\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+\frac{h}{2}\operatorname{Tr}(\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})+R(h^2,\mathbf{y}_0)\\ &= d\log h+\log\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top+h\operatorname{Tr}\mathbf{A}_\mathbf{v}(\boldsymbol{\beta})+R(h^2,\mathbf{y}_0).\end{aligned}\tag{S1}$$

To prove (vi), repeat the previous reasoning on (iv). The proof of (vii) follows from $\det\widetilde{\boldsymbol{\Omega}}_h=\det\boldsymbol{\Omega}_h^{[\mathrm{RR}]}\det\boldsymbol{\Omega}_h^{[\mathrm{S|R}]}$ and properties (v) and (vi).

Proof of Lemma 6.4 Using the same approximation as in the previous proof of (v), we obtain:

$$\begin{aligned}2\log|\det D\boldsymbol{f}_{h/2}(\mathbf{y};\boldsymbol{\beta})| &= 2\log|\det(\mathbf{I}+\frac{h}{2}D\mathbf{N}(\mathbf{y};\boldsymbol{\beta}))|+R(h^2,\mathbf{y})\\ &= 2\log|1+\frac{h}{2}\operatorname{Tr}D\mathbf{N}(\mathbf{y};\boldsymbol{\beta})|+R(h^2,\mathbf{y})\\ &= h\operatorname{Tr}D\mathbf{N}(\mathbf{y};\boldsymbol{\beta})+R(h^2,\mathbf{y})=h\operatorname{Tr}D_\mathbf{v}\mathbf{N}(\mathbf{y};\boldsymbol{\beta})+R(h^2,\mathbf{y}).\end{aligned}\tag{S2}$$

In complete observation, put $\mathbf{Y}_{t_k}$ instead of $\mathbf{y}$ and use Itô's lemma on $\mathbf{N}(\mathbf{Y}_{t_k})$ as in (S6). In partial observation, put $(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_{k+1}})$ instead of $\mathbf{y}$ and approximate $\mathbf{N}(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_{k+1}})$ as in (S7).

Proof of Lemma 6.5 We use definition (6) and approximation (55), and plug them in (17) to obtain:

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_h(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y})) &= e^{\widetilde{\mathbf{A}}h}(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y}) - \widetilde{\mathbf{b}}) + \widetilde{\mathbf{b}} \\
&= \left(\mathbf{I}_{2d} + h\widetilde{\mathbf{A}} + \frac{h^2}{2}\widetilde{\mathbf{A}}^2 + \mathbf{R}(h^3, \mathbf{y})\right)(\widetilde{\boldsymbol{f}}_{h/2}(\mathbf{y}) - \widetilde{\mathbf{b}}) + \widetilde{\mathbf{b}} \\
&= \begin{bmatrix} \mathbf{I}_d + \frac{h^2}{2}\mathbf{A}_{\mathbf{x}} + \mathbf{R}(h^3, \mathbf{y}) & h\mathbf{I}_d + \frac{h^2}{2}\mathbf{A}_{\mathbf{v}} + \mathbf{R}(h^3, \mathbf{y}) \\ h\mathbf{A}_{\mathbf{x}} + \mathbf{R}(h^2, \mathbf{y}) & \mathbf{I}_d + h\mathbf{A}_{\mathbf{v}} + \mathbf{R}(h^2, \mathbf{y}) \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{b} \\ \mathbf{v} + \frac{h}{2}\mathbf{N}(\mathbf{y}) + \mathbf{R}(h^2, \mathbf{y}) \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{x} + h\mathbf{v} + \frac{h^2}{2}\mathbf{F}(\mathbf{y}) + \mathbf{R}(h^3, \mathbf{y}) \\ \mathbf{v} + h(\mathbf{F}(\mathbf{y}) - \frac{1}{2}\mathbf{N}(\mathbf{y})) + \mathbf{R}(h^2, \mathbf{y}) \end{bmatrix}.
\end{aligned}$$

This concludes the proof.

To prove Proposition 6.6, we need the following lemma that provides expansion of $\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k}$.

Lemma S1.1 *For process* $\Delta_h \mathbf{X}_{t_{k+1}}$ (33) *it holds:*

$$\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k} = \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}), \tag{S3}$$

$$\Delta_h \mathbf{X}_{t_k} - \mathbf{V}_{t_{k-1}} = \sqrt{h}\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \frac{h}{2}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\zeta}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}). \tag{S4}$$

Proof of Lemma S1.1 Proof of (S3). Equation (2) in integral form and (33) yield:

$$\begin{aligned}
\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k} &= \frac{1}{h}\int_{t_k}^{t_{k+1}} \mathbf{V}_t \,\mathrm{d}t - \frac{1}{h}\int_{t_{k-1}}^{t_k} \mathbf{V}_t \,\mathrm{d}t = \frac{1}{h}\int_{t_k}^{t_{k+1}} (\mathbf{V}_t - \mathbf{V}_{t_k}) \,\mathrm{d}t + \frac{1}{h}\int_{t_{k-1}}^{t_k} (\mathbf{V}_{t_k} - \mathbf{V}_t) \,\mathrm{d}t \\
&= \frac{1}{h}\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathbf{F}_0(\mathbf{Y}_s) \,\mathrm{d}s \,\mathrm{d}t + \frac{1}{h}\boldsymbol{\Sigma}_0 \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathrm{d}\mathbf{W}_s \,\mathrm{d}t \\
&+ \frac{1}{h}\int_{t_{k-1}}^{t_k} \int_{t}^{t_k} \mathbf{F}_0(\mathbf{Y}_s) \,\mathrm{d}s \,\mathrm{d}t + \frac{1}{h}\boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k} \int_{t}^{t_k} \mathrm{d}\mathbf{W}_s \,\mathrm{d}t.
\end{aligned}$$

Apply Fubini's theorem on double integrals to obtain:

$$\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k} = \frac{1}{h}\int_{t_k}^{t_{k+1}} (t_{k+1} - t)\mathbf{F}_0(\mathbf{Y}_t) \,\mathrm{d}t + \frac{1}{h}\int_{t_{k-1}}^{t_k} (t - t_{k-1})\mathbf{F}_0(\mathbf{Y}_t) \,\mathrm{d}t + \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1}. \tag{S5}$$

Applying Itô's lemma on $\mathbf{F}_0(\mathbf{Y}_t)$ yields the following approximation:

$$\mathbf{F}_0(\mathbf{Y}_t) = \mathbf{F}_0(\mathbf{Y}_{t_k}) + D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0 \int_{t_k}^{t} \mathrm{d}\mathbf{W}_s + \mathbf{R}(h, \mathbf{Y}_{t_k}). \tag{S6}$$

Plugging (S6) into (S5) gives:

$$\begin{aligned}
\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k} &= h\mathbf{F}_0(\mathbf{Y}_{t_k}) + \frac{1}{h} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0 \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} (t_{k+1} - t) \,\mathrm{d}\mathbf{W}_s \,\mathrm{d}t \\
&- \frac{1}{h} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k} \int_{t}^{t_k} (t - t_{k-1}) \,\mathrm{d}\mathbf{W}_s \,\mathrm{d}t + \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Once again, apply Fubini's theorem on the double integrals to get:

$$\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} (t_{k+1} - t) \,\mathrm{d}\mathbf{W}_s \,\mathrm{d}t = \frac{1}{2}h^{5/2}\boldsymbol{\zeta}'_k, \qquad \int_{t_{k-1}}^{t_k} \int_{t}^{t_k} (t - t_{k-1}) \,\mathrm{d}\mathbf{W}_s \,\mathrm{d}t = \frac{1}{2}h^{5/2}\boldsymbol{\zeta}_{k-1}.$$

So far, we have

$$\Delta_h \mathbf{X}_{t_{k+1}} - \Delta_h \mathbf{X}_{t_k} = \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h\mathbf{F}_0(\mathbf{Y}_{t_k}) + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0(\boldsymbol{\zeta}'_k - \boldsymbol{\zeta}_{k-1}) + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).$$

To conclude the proof, use Itô's lemma to get $\mathbf{F}_0(\mathbf{Y}_{t_k}) = \mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \sqrt{h} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} + \mathbf{R}(h, \mathbf{Y}_{t_{k-1}})$.

Proof of (S4). As before, start with (33) and use (S6) to get:

$$\begin{aligned}\Delta_h \mathbf{X}_{t_k} - \mathbf{V}_{t_{k-1}} &= \frac{1}{h}\int_{t_{k-1}}^{t_k} (t_k - t)\mathbf{F}_0(\mathbf{Y}_t)\,\mathrm{d}t + \frac{1}{h}\boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k} (t_k - t)\,\mathrm{d}\mathbf{W}_t \\ &= \sqrt{h}\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \frac{h}{2}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \frac{1}{h}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k}\int_{t_{k-1}}^{t} (t_k - t)\,\mathrm{d}\mathbf{W}_s\,\mathrm{d}t + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

This concludes the proof.

Proof of Proposition 6.6 The expansion of $\mathbf{Z}^{[\mathrm{S}]}_{k,k-1}$ follows directly from Lemma 6.5 and S1.1. Indeed, it holds

$$\begin{aligned}\mathbf{Z}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) &= \mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\mathbf{V}_{t_{k-1}} - \frac{h^2}{2}\mathbf{F}(\mathbf{Y}_{t_{k-1}}) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}) \\ &= h(\Delta_h \mathbf{X}_{t_k} - \mathbf{V}_{t_{k-1}}) - \frac{h^2}{2}\mathbf{F}(\mathbf{Y}_{t_{k-1}}) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

To expand $\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}$, we use definition (27) and approximations (58) and (60), as follows:

$$\begin{aligned}\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) &= \mathbf{V}_{t_k} - \mathbf{V}_{t_{k-1}} - h\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \frac{h}{2}(\mathbf{N}(\mathbf{Y}_{t_k}) - \mathbf{N}(\mathbf{Y}_{t_{k-1}})) + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}) \\ &= \boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k} \mathrm{d}\mathbf{W}_t + \int_{t_{k-1}}^{t_k} \mathbf{F}_0(\mathbf{Y}_t)\,\mathrm{d}t - h\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \frac{h}{2}(\mathbf{N}(\mathbf{Y}_{t_k}) - \mathbf{N}(\mathbf{Y}_{t_{k-1}})) + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}) \\ &= \sqrt{h}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h}{2}(\mathbf{N}(\mathbf{Y}_{t_k}) - \mathbf{N}(\mathbf{Y}_{t_{k-1}})) \\ &+ D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \int_{t_{k-1}}^{t_k}\int_{t_{k-1}}^{t} \mathrm{d}\mathbf{W}_s\,\mathrm{d}t + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

In the last line, we used Itô's lemma on $\mathbf{F}_0(\mathbf{Y}_t)$ as in (S6). Again, apply Itô's lemma on $\mathbf{N}(\mathbf{Y}_{t_k})$ to get:

$$\begin{aligned}\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) &= \sqrt{h}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2}D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} \\ &+ h^{3/2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

The expansion of $\mathbf{Z}^{[\mathrm{S}]}_{k,k-1}$ follows from definition (36) and plugging $(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k})$ in approximation (60):

$$\begin{aligned}\overline{\mathbf{Z}}^{[\mathrm{S}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\Delta_h\mathbf{X}_{t_k} - \frac{h^2}{2}\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k}) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}) \\ &= -\frac{h^2}{2}\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k}) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

Use Taylor's formula on $\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k})$ to get

$$\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k}) = \mathbf{F}(\mathbf{Y}_{t_{k-1}}) + D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})(\Delta_h\mathbf{X}_{t_k} - \mathbf{V}_{t_{k-1}}) + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}). \tag{S7}$$

Now, the rest follows from Lemma S1.1.

Finally, to expand $\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}$, start with definition (37) and approximations (58), and (60):

$$\begin{aligned}\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \Delta_h\mathbf{X}_{t_{k+1}} - \Delta_h\mathbf{X}_{t_k} - h\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k}) \\ &- \frac{h}{2}(\mathbf{N}(\mathbf{X}_{t_k}, \Delta_h\mathbf{X}_{t_{k+1}}) - \mathbf{N}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k})) + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

Lemma S1.1 yields:

$$\begin{aligned}\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \sqrt{h}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k})) \\ &- \frac{h}{2}(\mathbf{N}(\mathbf{X}_{t_k}, \Delta_h\mathbf{X}_{t_{k+1}}) - \mathbf{N}(\mathbf{X}_{t_{k-1}}, \Delta_h\mathbf{X}_{t_k})) + \frac{h^{3/2}}{2}D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0\mathbf{Q}_{k,k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).\end{aligned}$$

Apply Taylor's formula on $\mathbf{F}(\mathbf{X}_{t_{k-1}}, \Delta_h \mathbf{X}_{t_k})$, $\mathbf{N}(\mathbf{X}_{t_k}, \Delta_h \mathbf{X}_{t_{k+1}})$, and $\mathbf{N}(\mathbf{X}_{t_{k-1}}, \Delta_h \mathbf{X}_{t_k})$, to get:

$$\begin{aligned}
\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h}{2}(\mathbf{N}(\mathbf{Y}_{t_k}) - \mathbf{N}(\mathbf{Y}_{t_{k-1}})) \\
&\quad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} - h^{3/2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_k})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_k \\
&\quad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}}))\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Finally, applying Itô's lemma on $\mathbf{N}(\mathbf{Y}_{t_k})$ yields

$$\begin{aligned}
\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} \\
&\quad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} - h^{3/2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_k \\
&\quad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}}))\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}) \\
&= \sqrt{h}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} \\
&\quad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} - h^{3/2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \mathbf{R}(h^2, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

This concludes the proof.

Proof of Lemma 6.7 Combining Proposition 6.6 and property (ii) of Lemma 6.2 yields:

$$\begin{aligned}
\mathbf{Z}^{[\mathrm{S|R}]}_{k,k-1}(\boldsymbol{\beta}) &= \mathbf{Z}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) - \boldsymbol{\Omega}^{[\mathrm{SR}]}_h(\boldsymbol{\theta})\boldsymbol{\Omega}^{[\mathrm{RR}]}_h(\boldsymbol{\theta})^{-1}\mathbf{Z}^{[\mathrm{R}]}_{k,k-1}(\boldsymbol{\beta}) \\
&= h^{3/2}\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \frac{h^2}{2}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) + \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\zeta}'_{k-1} \\
&\quad - \left(\frac{h}{2}\mathbf{I} - \frac{h^2}{12}(\mathbf{A}_{\mathbf{v}} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\right)\Big(h^{1/2}\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&\qquad\qquad - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} + h^{3/2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1})\Big) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}) \\
&= h^{3/2}\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \frac{h^2}{2}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) + \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\zeta}'_{k-1} \\
&\quad - h^{3/2}\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} - \frac{h^2}{2}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) + \frac{h^{5/2}}{4} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} \\
&\quad - \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} + \frac{h^{5/2}}{12}(\mathbf{A}_{\mathbf{v}} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1} + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Additionally,

$$\begin{aligned}
\overline{\mathbf{Z}}^{[\mathrm{S|R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) &= \overline{\mathbf{Z}}^{[\mathrm{S}]}_{k,k-1}(\boldsymbol{\beta}) - \boldsymbol{\Omega}^{[\mathrm{SR}]}_h(\boldsymbol{\theta})\boldsymbol{\Omega}^{[\mathrm{RR}]}_h(\boldsymbol{\theta})^{-1}\overline{\mathbf{Z}}^{[\mathrm{R}]}_{k+1,k,k-1}(\boldsymbol{\beta}) \\
&= -\frac{h^2}{2}\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} - \left(\frac{h}{2}\mathbf{I} - \frac{h^2}{12}(\mathbf{A}_{\mathbf{v}} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\right) \\
&\quad \cdot \Bigg(h^{1/2}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} + h(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) - \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} \\
&\qquad + \frac{h^{3/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} - h^{3/2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1}\Bigg) + \mathbf{R}(h^3, \mathbf{Y}_{t_{k-1}}) \\
&= -\frac{h^2}{2}\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1} - h^{3/2}\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} - \frac{h^2}{2}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&\quad + \frac{h^{5/2}}{4} D_{\mathbf{v}}\mathbf{N}(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1} - \frac{h^{5/2}}{4} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \mathbf{Q}_{k,k-1} + \frac{h^{5/2}}{2} D_{\mathbf{v}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})\boldsymbol{\Sigma}_0 \boldsymbol{\xi}'_{k-1}
\end{aligned}$$

$$+\frac{h^{5/2}}{12}(\mathbf{A}_{\mathbf{v}}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}_{\mathbf{v}}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}+\mathbf{R}(h^3,\mathbf{Y}_{t_{k-1}}).$$

S1.2 Proofs from Section 3

Before we start the proofs, we state the following ergodic property, which is proved in Kessler [1997] in case of complete observations, and in Samson and Thieullen [2012] for both complete and partial observation.

Lemma S1.2 (Proposition 4 in Samson and Thieullen [2012]) *Let Assumptions (A1), (A2) and (A3) hold, and let* $\mathbf{Y}$ *be the solution to* (3)*. Let* $g:\mathbb{R}^{2d}\times\Theta\to\mathbb{R}$ *be a differentiable function with respect to* $\mathbf{y}$ *and* $\boldsymbol{\theta}$ *with derivatives of polynomial growth in* $\mathbf{y}$*, uniformly in* $\boldsymbol{\theta}$*. If* $h\to 0$ *and* $Nh\to\infty$*, then,*

$$\frac{1}{N-1}\sum_{k=1}^{N} g\left(\mathbf{Y}_{t_k};\boldsymbol{\theta}\right)\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}}\int g\left(\mathbf{y};\boldsymbol{\theta}\right)\mathrm{d}\nu_0(\mathbf{y}), \tag{S8}$$

$$\frac{1}{N-2}\sum_{k=1}^{N-1} g\left(\mathbf{X}_{t_k},\Delta_h\mathbf{X}_{t_k};\boldsymbol{\theta}\right)\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}}\int g\left(\mathbf{y};\boldsymbol{\theta}\right)\mathrm{d}\nu_0(\mathbf{y}), \tag{S9}$$

uniformly in $\boldsymbol{\theta}$*.*

S1.3 Proof of consistency

Proof of Theorem 3.1 The proof of the consistency of the estimators follows a similar path as in Theorem 5.1 of [Pilipovic et al., 2024]. With $\boldsymbol{\sigma}:=\mathrm{vech}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)=([\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{11},[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{12},[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{22},...,[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{1d},...,[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{dd})$, we half-vectorize $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ to avoid working with tensors when computing derivatives with respect to $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. Since $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is a symmetric $d\times d$ matrix, $\boldsymbol{\sigma}$ is of dimension $s=d(d+1)/2$. For a diagonal matrix, instead of a half-vectorization, we use $\boldsymbol{\sigma}:=\mathrm{diag}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ and $s=d$ in that case.

We start by finding the limits in $\mathbb{P}_{\boldsymbol{\theta}_0}$ of

$$\frac{1}{N-1}\mathcal{L}_N^{[\mathrm{C}\cdot]}(\boldsymbol{\beta},\boldsymbol{\sigma})\quad\text{and}\quad\frac{1}{N-2}\mathcal{L}_N^{[\mathrm{P}\cdot]}(\boldsymbol{\beta},\boldsymbol{\sigma}), \tag{S10}$$

for $Nh\to\infty$, $h\to 0$, uniformly in $\boldsymbol{\theta}$. We apply Lemma 9 in Genon-Catalot and Jacod [1993] to prove the convergence and use Proposition A1 in Gloter [2006] to prove the uniform convergence. For more detailed derivations, see proofs in Pilipovic et al. [2024]. Taking the expectations of (69)- (74), we conclude that:

$$\begin{aligned}
\frac{1}{N-1}\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),\\
\frac{1}{N-1}\mathcal{L}_N^{[\mathrm{CS|R}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),\\
\frac{1}{N-1}\mathcal{L}_N^{[\mathrm{CF}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to 2\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+2\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),\\
\frac{1}{N-2}\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to\frac{2}{3}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\frac{2}{3}\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),\\
\frac{1}{N-2}\mathcal{L}_N^{[\mathrm{PS|R}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to 2\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+2\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),\\
\frac{1}{N-2}\mathcal{L}_N^{[\mathrm{PF}]}(\boldsymbol{\beta},\boldsymbol{\sigma})&\to\frac{8}{3}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)+\frac{8}{3}\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top),
\end{aligned}$$

in $\mathbb{P}_{\boldsymbol{\theta}_0}$, for $Nh\to\infty$, $h\to 0$, uniformly in $\boldsymbol{\theta}$. From here, the rest of the proof of consistency for $\widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^{\top[\mathrm{C}\cdot]}$ and $\widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^{\top[\mathrm{P}\cdot]}$ is the same as in [Pilipovic et al., 2024]. The coefficients in front of $\log\det$ terms in the partial observation setup correspond to the correcting factors in definitions of objective functions (39)-(41). They are needed to match coefficients in front of Tr terms that come from the forward difference's under- or over-estimation of the noise effects.

To prove the consistency of the drift estimators $\widehat{\boldsymbol{\beta}}_N^{[\mathrm{CR}]}$ and $\widehat{\boldsymbol{\beta}}_N^{[\mathrm{PR}]}$, we start by finding the limits in $\mathbb{P}_{\boldsymbol{\theta}_0}$ of

$$\frac{1}{(N-1)h}(\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta},\boldsymbol{\sigma})-\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma}))\quad\text{and}\quad\frac{1}{(N-2)h}(\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\beta},\boldsymbol{\sigma})-\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})), \tag{S11}$$

for $Nh \to \infty$, $h \to 0$, uniformly in $\boldsymbol{\theta}$. Starting with expressions (69) and (72) we get

$$\begin{aligned}
\frac{1}{(N-1)h}(\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) &= \frac{2}{(N-1)\sqrt{h}}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&+ \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&- \frac{1}{N-1}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}(\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}_0(\mathbf{Y}_{t_{k-1}}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} \\
&+ \frac{1}{N-1}\sum_{k=1}^{N}\operatorname{Tr} D_{\mathbf{v}}(\mathbf{F}(\mathbf{Y}_{t_k}) - \mathbf{F}_0(\mathbf{Y}_{t_k})), \\
\frac{1}{(N-2)h}(\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) &= \frac{2}{(N-2)\sqrt{h}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}(\mathbf{Y}_{t_{k-1}})) \\
&- \frac{1}{N-2}\sum_{k=1}^{N-1}(\mathbf{U}_{k,k-1} + 2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}(\mathbf{F}(\mathbf{Y}_{t_{k-1}}) - \mathbf{F}_0(\mathbf{Y}_{t_{k-1}}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1}\operatorname{Tr} D_{\mathbf{v}}(\mathbf{F}(\mathbf{Y}_{t_k}) - \mathbf{F}_0(\mathbf{Y}_{t_k})).
\end{aligned}$$

To prove the convergence in probability of the previous two sequences, we use Lemma S1.2 and Lemma 9 in Genon-Catalot and Jacod [1993]. To apply Lemma 9 from Genon-Catalot and Jacod [1993], we need to show that the sum of expectations converges to a certain value, while the sum of covariances converges to zero. Here, we only show the former. Moreover, standard tools like Proposition A1 in Gloter [2006] or Lemma 3.1 in Yoshida [1990] can be used to prove uniform convergence. Thus, we just look at the expectation to find the limits of these sequences. We use the known covariances (66) and (68) to get:

$$\begin{aligned}
\frac{1}{Nh}(\mathcal{L}_N^{[\cdot\mathrm{R}]}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \mathcal{L}_N^{[\cdot\mathrm{R}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} &\int(\mathbf{F}_0(\mathbf{y}) - \mathbf{F}(\mathbf{y}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{y}) - \mathbf{F}(\mathbf{y}))\,\mathrm{d}\nu_0(\mathbf{y}) \\
&+ \int\operatorname{Tr}(D_{\mathbf{v}}\left(\mathbf{F}_0\left(\mathbf{y}\right) - \mathbf{F}\left(\mathbf{y}\right)\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} - \mathbf{I}\right))\,\mathrm{d}\nu_0(\mathbf{y}). \qquad \text{(S12)}
\end{aligned}$$

Thus, the consistency of the drift estimator in the partial case coincides with the complete case when using rough objective functions. This is because the right-hand side of (S12) is non-negative when $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$, and the left-hand side is non-positive, following the definition of the likelihood. The remainder of the proof is analogous to that in Pilipovic et al. [2024], and is therefore not repeated here.

Here, we also illustrate why the objective functions based on the conditional likelihood of smooth given rough coordinates do not provide identifiable drift estimators. Starting with the complete observations objective function (70) and using that $\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\eta}_{k-1} - 2\boldsymbol{\xi}'_{k-1})\boldsymbol{\eta}_{k-1}^\top|\mathcal{F}_{t_{k-1}}] = \mathbf{0}$, we conclude::

$$\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{CS|R}]}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \mathcal{L}_N^{[\mathrm{CS|R}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 0. \qquad \text{(S13)}$$

In the partial observation case, we need to add the term $4\log|\det D_{\mathbf{v}}\boldsymbol{f}_{h/2}|$ in (31). Due to this correction, we obtain the consistency of the drift estimator from the following derivations and the fact that the diffusion estimator converges faster:

$$\begin{aligned}
\frac{1}{(N-2)h}(\mathcal{L}_N^{[\mathrm{PS|R}]}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \mathcal{L}_N^{[\mathrm{PS|R}]}(\boldsymbol{\beta}_0,\boldsymbol{\sigma})) &= \frac{2}{N-2}\sum_{k=1}^{N-1}\operatorname{Tr} D_{\mathbf{v}}(\mathbf{N}(\mathbf{Y}_{t_k}) - \mathbf{N}_0(\mathbf{Y}_{t_k})) \\
&+ \frac{3}{N-2}\sum_{k=1}^{N-1}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}(\mathbf{N}_0(\mathbf{Y}_{t_{k-1}}) - \mathbf{N}(\mathbf{Y}_{t_{k-1}}))^\top)
\end{aligned}$$

$$\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 2\int \mathrm{Tr}(D_{\mathbf{v}}\left(\mathbf{N}_0\left(\mathbf{y}\right)-\mathbf{N}\left(\mathbf{y}\right)\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}-\mathbf{I}\right))\,\mathrm{d}\nu_0(\mathbf{y}).$$

Finally, the consistency of the estimators based on the full objective functions follows from the previous proofs, (71), and (74). That concludes the proof of consistency.

Proof of Lemma 3.3 We start by proving the first part of the lemma, for both complete and partial cases using the rough objective functions (69) and (72). First, we find their second derivatives with respect to $\boldsymbol{\beta}$:

$$\begin{aligned}
\frac{1}{(N-1)h}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathcal{L}_N^{[\mathrm{CR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right) &= \frac{2}{(N-1)\sqrt{h}}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&+\frac{1}{N-1}\sum_{k=1}^{N}\mathrm{Tr}\,D_{\mathbf{v}}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta})+\frac{2}{N-1}\sum_{k=1}^{N}\partial_{\beta^{(i_1)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&-\frac{2}{N-1}\sum_{k=1}^{N}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))\\
&-\frac{1}{N-1}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1},\\
\frac{1}{(N-2)h}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathcal{L}_N^{[\mathrm{PR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right) &= \frac{2}{(N-2)\sqrt{h}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&+\frac{1}{N-2}\sum_{k=1}^{N-1}\mathrm{Tr}\,D_{\mathbf{v}}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta})+\frac{2}{N-2}\sum_{k=1}^{N-1}\partial_{\beta^{(i_1)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&-\frac{2}{N-2}\sum_{k=1}^{N-1}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))\\
&-\frac{1}{N-2}\sum_{k=1}^{N-1}(\mathbf{U}_{k,k-1}+2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}.
\end{aligned}$$

As in the proof of consistency, it holds:

$$\frac{1}{Nh}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathcal{L}_N^{[\cdot\mathrm{R}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 2\int\partial_{\beta^{(i_1)}}\mathbf{F}_0(\mathbf{y})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i_2)}}\mathbf{F}_0(\mathbf{y})\,\mathrm{d}\nu_0(\mathbf{y}).$$

Now, we investigate the limit of $\frac{1}{(N-1)\sqrt{h}}\partial^2_{\beta^{(i_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\theta})$ and $\partial^2_{\beta^{(i_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\theta})$:

$$\begin{aligned}
\frac{1}{(N-1)\sqrt{h}}\partial^2_{\beta^{(i_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right) &= -\frac{2}{N-1}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top\partial_{\sigma^{(j_2)}}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i_1)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&+\frac{1}{N-1}\sum_{k=1}^{N}R(\sqrt{h},\mathbf{Y}_{t_{k-1}})\\
\frac{1}{(N-2)\sqrt{h}}\partial^2_{\beta^{(i_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right) &= -\frac{2}{N-2}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top\partial_{\sigma^{(j_2)}}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i_1)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})\\
&+\frac{1}{N-2}\sum_{k=1}^{N-1}R(\sqrt{h},\mathbf{Y}_{t_{k-1}})
\end{aligned}$$

Both previous sequences converge to zero due to Lemma 9 in Genon-Catalot and Jacod [1993]. Next, we look at the limits of $\frac{1}{N-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CR}]}(\boldsymbol{\theta})$ and $\frac{1}{N-2}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PR}]}(\boldsymbol{\theta})$:

$$\frac{1}{N-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CR}]}\left(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}\right)=\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$$

$$
\begin{aligned}
&+ \frac{1}{N-1}\sum_{k=1}^{N} \partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) + \frac{1}{N-1}\sum_{k=1}^{N} R(\sqrt{h}, \mathbf{Y}_{t_{k-1}}) \\
&= \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) - \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) \\
&- \frac{1}{N-1}\sum_{k=1}^{N} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-1}\sum_{k=1}^{N} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-1}\sum_{k=1}^{N} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-1}\sum_{k=1}^{N} R(\sqrt{h}, \mathbf{Y}_{t_{k-1}}),
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{N-2}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) &= \frac{2}{3}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1} \partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) + \frac{1}{N-2}\sum_{k=1}^{N-1} R(\sqrt{h}, \mathbf{Y}_{t_{k-1}}) \\
&= \frac{2}{3}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) - \frac{2}{3}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) \\
&- \frac{1}{N-2}\sum_{k=1}^{N-1} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1} \operatorname{Tr}(\boldsymbol{\Sigma}_0 \mathbf{U}_{k,k-1}\mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \\
&+ \frac{1}{N-2}\sum_{k=1}^{N-1} R(\sqrt{h}, \mathbf{Y}_{t_{k-1}}).
\end{aligned}
$$

Using the second moments of $\boldsymbol{\eta}_{k-1}$ and $\mathbf{U}_{k,k-1}$ with additional calculations, we can conclude that:

$$
\begin{aligned}
\frac{1}{N-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top), \\
\frac{1}{N-2}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{2}{3}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top).
\end{aligned}
$$

To extend the previous results on the objective functions (70) and (73), we start by acknowledging that

$$
\frac{1}{Nh}\partial^2_{\beta^{(i_1)}\beta^{(i_2)}}\mathcal{L}_N^{[\cdot\mathrm{S}|\mathrm{R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.
$$

The reasons behind this are the same as in the proof of consistency. The same can be said for the limit of $\frac{1}{N\sqrt{h}}\partial_{\boldsymbol{\beta}\boldsymbol{\sigma}}\mathcal{L}_N^{[\cdot\mathrm{S}|\mathrm{R}]}(\boldsymbol{\theta})$. Finally, repeating the same derivations as before, we get:

$$
\begin{aligned}
\frac{1}{N-1}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{CS}|\mathrm{R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top), \\
\frac{1}{N-2}\partial^2_{\sigma^{(j_1)}\sigma^{(j_2)}}\mathcal{L}_N^{[\mathrm{PS}|\mathrm{R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 2\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j_1)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j_2)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top).
\end{aligned}
$$

This concludes the first part of the lemma. The second part follows from the fact that all limits are continuous in $\boldsymbol{\theta}$.

To prove Lemma 3.4, we state another useful property that provides a general formula to calculate moments of a product of two quadratic forms with Gaussian vectors.

Lemma S1.3 *Let $(\boldsymbol{\alpha}_k)_{k=1}^N$, $(\boldsymbol{\beta}_k)_{k=1}^N$, be two sequences of independent $\mathcal{F}_{t_{k+1}}$-measurable Gaussian random variables with mean zero. If $\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}]$ is diagonal, and $\mathbf{A}$ and $\mathbf{B}$ are two symmetric positive definite matrices, then:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}^\top\mathbf{A}\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top\mathbf{B}\boldsymbol{\beta}_{k-1} \mid \mathcal{F}_{t_{k-1}}] &= 2\operatorname{Tr}(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}]\mathbf{A}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}]\mathbf{B}) \\ &\quad + \operatorname{Tr}(\mathbf{A}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}\boldsymbol{\alpha}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}])\operatorname{Tr}(\mathbf{B}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\beta}_{k-1}\boldsymbol{\beta}_{k-1}^\top \mid \mathcal{F}_{t_{k-1}}]).\end{aligned}$$

Proof We start with

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}^\top\mathbf{A}\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top\mathbf{B}\boldsymbol{\beta}_{k-1} \mid \mathcal{F}_{t_{k-1}}] &= \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\alpha_{k-1}^{(j)}\beta_{k-1}^{(l)}\beta_{k-1}^{(m)} \mid \mathcal{F}_{t_{k-1}}] \\ &= \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\operatorname{cov}(\alpha_{k-1}^{(i)}\alpha_{k-1}^{(j)}, \beta_{k-1}^{(l)}\beta_{k-1}^{(m)}) \qquad \text{(S14)} \\ &\quad + \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\alpha_{k-1}^{(j)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\beta_{k-1}^{(l)}\beta_{k-1}^{(m)} \mid \mathcal{F}_{t_{k-1}}].\end{aligned}$$

We compute (S14) using the formula for the covariance of products of centered Gaussian random variables [Bohrnstedt and Goldberger, 1969]. Then we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\alpha}_{k-1}^\top\mathbf{A}\boldsymbol{\alpha}_{k-1}\boldsymbol{\beta}_{k-1}^\top\mathbf{B}\boldsymbol{\beta}_{k-1} \mid \mathcal{F}_{t_{k-1}}] &= \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\beta_{k-1}^{(l)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(j)}\beta_{k-1}^{(m)} \mid \mathcal{F}_{t_{k-1}}] \\ &\quad + \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\beta_{k-1}^{(m)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(j)}\beta_{k-1}^{(l)} \mid \mathcal{F}_{t_{k-1}}] \\ &\quad + \sum_{i,j,l,m=1}^d A_{ij}B_{lm}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\alpha_{k-1}^{(j)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\beta_{k-1}^{(l)}\beta_{k-1}^{(m)} \mid \mathcal{F}_{t_{k-1}}] \\ &= 2\sum_{i,j=1}^d A_{ij}B_{ij}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\beta_{k-1}^{(i)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(j)}\beta_{k-1}^{(j)} \mid \mathcal{F}_{t_{k-1}}] \\ &\quad + \sum_{i,j=1}^d A_{ii}B_{jj}\mathbb{E}_{\boldsymbol{\theta}_0}[\alpha_{k-1}^{(i)}\alpha_{k-1}^{(i)} \mid \mathcal{F}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\beta_{k-1}^{(j)}\beta_{k-1}^{(j)} \mid \mathcal{F}_{t_{k-1}}].\end{aligned}$$

That concludes the proof.

Applying the previous Lemma to our setup, corollary S1.4 follows immediately.

Corollary S1.4 *Let $(\boldsymbol{\eta}_k)_{k=1}^N$, $(\boldsymbol{\xi}_k)_{k=1}^N$, $(\boldsymbol{\xi}'_k)_{k=1}^N$ be random sequences as defined in* (61) *and* (62)*. Let $\mathbf{B}_{j_1}$ and $\mathbf{B}_{j_2}$ be two symmetric positive definite matrices. Then, it holds:*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\eta}_{k-1}^\top\mathbf{B}_{j_1}\boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top\mathbf{B}_{j_2}\boldsymbol{\eta}_{k-1} \mid \mathcal{F}_{t_{k-1}}] = 2\operatorname{Tr}(\mathbf{B}_{j_1}\mathbf{B}_{j_2}) + \operatorname{Tr}\mathbf{B}_{j_1}\operatorname{Tr}\mathbf{B}_{j_2}, \qquad \text{(S15)}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-1}^\top\mathbf{B}_{j_1}\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}_{k-1}^\top\mathbf{B}_{j_2}\boldsymbol{\xi}_{k-1} \mid \mathcal{F}_{t_{k-1}}] = \frac{2}{9}\operatorname{Tr}(\mathbf{B}_{j_1}\mathbf{B}_{j_2}) + \frac{1}{9}\operatorname{Tr}\mathbf{B}_{j_1}\operatorname{Tr}\mathbf{B}_{j_2}, \qquad \text{(S16)}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-1}'^\top\mathbf{B}_{j_1}\boldsymbol{\xi}'_{k-1}\boldsymbol{\xi}_{k-1}'^\top\mathbf{B}_{j_2}\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] = \frac{2}{9}\operatorname{Tr}(\mathbf{B}_{j_1}\mathbf{B}_{j_2}) + \frac{1}{9}\operatorname{Tr}\mathbf{B}_{j_1}\operatorname{Tr}\mathbf{B}_{j_2}, \qquad \text{(S17)}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-1}^\top\mathbf{B}_{j_1}\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}_{k-1}'^\top\mathbf{B}_{j_2}\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] = \frac{1}{18}\operatorname{Tr}(\mathbf{B}_{j_1}\mathbf{B}_{j_2}) + \frac{1}{9}\operatorname{Tr}\mathbf{B}_{j_1}\operatorname{Tr}\mathbf{B}_{j_2}. \qquad \text{(S18)}$$

Proof of Lemma 3.4 To prove the lemma, we need to compute $\boldsymbol{\lambda}_N^{[\mathrm{obj}]}$. The main part of the proof focuses only on the rough estimators, while at the end of the proof we discuss how the same ideas can be adapted for other estimators. Thus, we start with $-\frac{1}{\sqrt{Nh}}\partial_{\beta^{(i)}}\mathcal{L}_N^{[\cdot\mathrm{R}]}$:

$$-\frac{1}{\sqrt{(N-1)h}}\partial_{\beta^{(i)}}\mathcal{L}_N^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = \frac{2}{\sqrt{N-1}}\sum_{k=1}^N \boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})$$

$$+ 2\sqrt{\frac{h}{N-1}}\sum_{k=1}^{N}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))$$
$$+ \sqrt{\frac{h}{N-1}}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} - \sqrt{\frac{h}{N-1}}\sum_{k=1}^{N}\operatorname{Tr} D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta}),$$

$$-\frac{1}{\sqrt{(N-2)h}}\partial_{\beta^{(i)}}\mathcal{L}_N^{[\mathrm{PR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = \frac{2}{\sqrt{N-2}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})$$
$$+ 2\sqrt{\frac{h}{N-2}}\sum_{k=1}^{N-1}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))$$
$$+ \sqrt{\frac{h}{N-2}}\sum_{k=1}^{N-1}(\mathbf{U}_{k,k-1}+2\boldsymbol{\xi}_{k-1})^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}$$
$$- \sqrt{\frac{h}{N-2}}\sum_{k=1}^{N-1}\operatorname{Tr} D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}(\mathbf{Y}_{t_k};\boldsymbol{\beta}).$$

Similarly, for $-\frac{1}{\sqrt{N}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\cdot\mathrm{R}]}$, we get:

$$-\frac{1}{\sqrt{(N-1)}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\mathrm{CR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = -\frac{1}{\sqrt{N-1}}\sum_{k=1}^{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$$
$$+ \frac{1}{\sqrt{N-1}}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1}$$
$$+ 2\sqrt{\frac{h}{N-1}}\sum_{k=1}^{N}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))$$
$$+ \sum_{k=1}^{N}R(\frac{h}{\sqrt{N}},\mathbf{Y}_{t_{k-1}}),$$

$$-\frac{1}{\sqrt{(N-2)}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\mathrm{PR}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = -\frac{2}{3\sqrt{N-2}}\sum_{k=1}^{N-1}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$$
$$+ \frac{1}{\sqrt{N-2}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}$$
$$+ 2\sqrt{\frac{h}{N-2}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)-\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}))$$
$$+ \sum_{k=1}^{N-1}R(\frac{h}{\sqrt{N}},\mathbf{Y}_{t_{k-1}}).$$

To prove the convergence in distribution of $\boldsymbol{\lambda}_N^{[\cdot\mathrm{R}]}$, we introduce the following triangular arrays that arise from the previous calculations:

$$\boldsymbol{\phi}_{N,k-1}^{[\mathrm{CR}](i)}(\boldsymbol{\theta}_0) \coloneqq \frac{2}{\sqrt{N-1}}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) \tag{S19}$$
$$+ \sqrt{\frac{h}{N-1}}(\operatorname{Tr}(\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1}\boldsymbol{\eta}_{k-1}^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}) - \operatorname{Tr} D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})),$$
$$\boldsymbol{\phi}_{N,k-1}^{[\mathrm{PR}](i)}(\boldsymbol{\theta}_0) \coloneqq \frac{2}{\sqrt{N-2}}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) \tag{S20}$$
$$+ \sqrt{\frac{h}{N-2}}(\operatorname{Tr}(\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}(\mathbf{U}_{k,k-1}+2\boldsymbol{\xi}_{k-1}')^\top\boldsymbol{\Sigma}_0^\top D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}) - \operatorname{Tr} D_{\mathbf{v}}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_k})),$$

$$\boldsymbol{\rho}_{N,k-1}^{[\mathrm{CR}](j)}(\boldsymbol{\theta}_0) \coloneqq \frac{1}{\sqrt{N-1}}(\boldsymbol{\eta}_{k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\eta}_{k-1} - \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)), \quad \text{(S21)}$$

$$\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PR}](j)}(\boldsymbol{\theta}_0) \coloneqq \frac{1}{\sqrt{N-2}}(\mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1} - \frac{2}{3}\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)). \quad \text{(S22)}$$

Then, $\boldsymbol{\lambda}_N^{[\cdot\mathrm{R}]}$ rewrites as:

$$\boldsymbol{\lambda}_N^{[\cdot\mathrm{R}]} = \sum_{k=1}^{N} \begin{bmatrix} \boldsymbol{\phi}_{N,k-1}^{[\cdot\mathrm{R}](1)}(\boldsymbol{\theta}_0) \\ \vdots \\ \boldsymbol{\phi}_{N,k-1}^{[\cdot\mathrm{R}](r)}(\boldsymbol{\theta}_0) \\ \boldsymbol{\rho}_{N,k-1}^{[\cdot\mathrm{R}](1)}(\boldsymbol{\theta}_0) \\ \vdots \\ \boldsymbol{\rho}_{N,k-1}^{[\cdot\mathrm{R}](s)}(\boldsymbol{\theta}_0) \end{bmatrix} + \frac{1}{N}\sum_{k=1}^{N} R(\sqrt{Nh^2}, \mathbf{Y}_{t_{k-1}}). \quad \text{(S23)}$$

Thus, to establish estimators' asymptotic normality, we need an extra convergence condition $Nh^2 \to 0$. This is common in literature, and it is necessary for most estimators.

To finish the proof, we apply the central limit theorem for martingale difference arrays (Proposition 3.1 in Crimaldi and Pratelli [2005]). However, we can not apply the same reasoning in complete and partial observation cases.

First, we notice that in both complete and partial cases, $\boldsymbol{\phi}_{N,k-1}^{[\cdot\mathrm{R}](i)}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\rho}_{N,k-1}^{[\cdot\mathrm{R}](j)}(\boldsymbol{\theta}_0)$ are centered conditionally to $\mathcal{F}_{t_{k-1}}$. Moreover, in the complete case, $\boldsymbol{\phi}_{N,k-1}^{[\mathrm{CR}](i)}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\rho}_{N,k-1}^{[\mathrm{CR}](j)}(\boldsymbol{\theta}_0)$ are adapted to the filtration $\mathcal{F}_{t_k}$. Thus, the proof follows directly by applying Proposition 3.1 in Crimaldi and Pratelli [2005]. This proposition assumes a martingale difference array centered conditionally to $\mathcal{F}_{t_{k-1}}$ and $\mathcal{F}_{t_k}$-measurable.

In the partial observation case, $\mathbf{U}_{k,k-1}$ is $\mathcal{F}_{t_{k+1}}$-measurable as it depends on random variables $\boldsymbol{\xi}_{k-1}$ and $\boldsymbol{\xi}_k'$. Consequently, to apply Proposition 3.1 in Crimaldi and Pratelli [2005], it is not enough for $\boldsymbol{\phi}_{N,k-1}^{[\mathrm{PR}](i)}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PR}](j)}(\boldsymbol{\theta}_0)$ to be centered conditionally to $\mathcal{F}_{t_{k-1}}$, they also need to be centered conditionally to $\mathcal{F}_{t_k}$. The previous condition, however, does not hold. Thus, we use the idea of reordering the sum in $\boldsymbol{\lambda}_N^{[\mathrm{PR}]}$ (S23) to obtain the $\mathcal{F}_{t_k}$-measurable and centered conditionally on $\mathcal{F}_{t_{k-1}}$, as proposed by Gloter [2000, 2006] and later used by Samson and Thieullen [2012].

First, use Lemma 9 from Genon-Catalot and Jacod [1993] to notice that:

$$\sum_{k=1}^{N-1} \boldsymbol{\phi}_{N,k-1}^{[\mathrm{PR}](i)}(\boldsymbol{\theta}_0) = \frac{2}{\sqrt{N-2}}\sum_{k=1}^{N-1} \mathbf{U}_{k,k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1).$$

Then, reorder the sum of $\boldsymbol{\phi}_{N,k-1}^{[\mathrm{PR}](i)}(\boldsymbol{\theta}_0)$ as follows:

$$\begin{aligned}
\sum_{k=1}^{N-1} \boldsymbol{\phi}_{N,k-1}^{[\mathrm{PR}](i)}(\boldsymbol{\theta}_0) &= \frac{2}{\sqrt{N-2}}\left(\boldsymbol{\xi}_0^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_0}) + \boldsymbol{\xi}_{N-1}'^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{N-2}})\right) \\
&\quad + \frac{2}{\sqrt{N-2}}\sum_{k=2}^{N-1}\left(\boldsymbol{\xi}_{k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \boldsymbol{\xi}_{k-1}'^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-2}})\right) + o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1) \\
&= \frac{2}{\sqrt{N-2}}\sum_{k=2}^{N-1}\left(\boldsymbol{\xi}_{k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \boldsymbol{\xi}_{k-1}'^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-2}})\right) + o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1).
\end{aligned}$$

Now, the triangular arrays under the sum are centered conditionally on $\mathcal{F}_{t_{k-1}}$ and $\mathcal{F}_{t_k}$ measurable. Thus, define:

$$\boldsymbol{\phi}_{N,k-1}^{\star[\mathrm{PR}](i)}(\boldsymbol{\theta}_0) \coloneqq \frac{2}{\sqrt{N-2}}\left(\boldsymbol{\xi}_{k-1}^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}) + \boldsymbol{\xi}_{k-1}'^\top \boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta^{(i)}}\mathbf{F}_0(\mathbf{Y}_{t_{k-2}})\right).$$

To apply Proposition 3.1 from Crimaldi and Pratelli [2005], we need the following limits in probability:

$$\sum_{k=2}^{N-1} \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}_{N,k-1}^{\star[\mathrm{PR}](i_1)}(\boldsymbol{\theta}_0)\boldsymbol{\phi}_{N,k-1}^{\star[\mathrm{PR}](i_2)}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 4[\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)]_{i_1 i_2},$$

$$\sum_{k=2}^{N-1} \mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\phi}^{\star[\mathrm{PR}](i_1)}_{N,k-1}(\boldsymbol{\theta}_0)\boldsymbol{\phi}^{\star[\mathrm{PR}](i_2)}_{N,k-1})^2(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

The first limit follows from properties (66) and (67). The second limit follows due to an additional order of $1/N$.

When looking at $\boldsymbol{\rho}^{[\cdot\mathrm{R}](j)}_{N,k-1}$, we repeat the same reasoning. For notational simplicity, start with defining:

$$\mathbf{B}_j(\boldsymbol{\theta}_0) \coloneqq \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\Sigma}_0.$$

It follows immediately that $\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0)) = \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$. Again, reorder the sum of $\boldsymbol{\rho}^{[\mathrm{PR}](j)}_{N,k-1}(\boldsymbol{\theta}_0)$ as follows:

$$\begin{aligned}
\sum_{k=1}^{N-1} \boldsymbol{\rho}^{[\mathrm{PR}](j)}_{N,k-1}(\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{N-2}} \sum_{k=1}^{N-1} (\mathbf{U}^\top_{k,k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\mathbf{U}_{k,k-1} - \frac{2}{3}\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0))) \\
&= \frac{1}{\sqrt{N-2}} \sum_{k=2}^{N-1} \left(\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1} + 2\boldsymbol{\xi}^\top_{k-2}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} + \boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} - \frac{2}{3}\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0))\right) \\
&+ \frac{1}{\sqrt{N-2}} \left(\boldsymbol{\xi}^\top_0\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}_0 + 2\boldsymbol{\xi}^\top_{N-2}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{N-1} + \boldsymbol{\xi}'^\top_{N-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{N-1} - \frac{2}{3}\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0))\right).
\end{aligned}$$

Since the last term in the previous equation is $o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1)$, we focus only on:

$$\boldsymbol{\rho}^{\star[\mathrm{PR}](j)}_{N,k-1}(\boldsymbol{\theta}_0) \coloneqq \frac{1}{\sqrt{N-2}} \left(\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1} + 2\boldsymbol{\xi}^\top_{k-2}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} + \boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} - \frac{2}{3}\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0))\right).$$

Notice that $\boldsymbol{\rho}^{\star[\mathrm{PR}](j)}_{N,k-1}(\boldsymbol{\theta}_0)$ is $\mathcal{F}_{t_k}$ measurable and centered conditionally on $\mathcal{F}_{t_{k-1}}$. Again, to apply Proposition 3.1 from Crimaldi and Pratelli [2005], we need the following limits in probability:

$$\begin{aligned}
\sum_{k=2}^{N-1} \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\rho}^{\star[\mathrm{PR}](j_1)}_{N,k-1}(\boldsymbol{\theta}_0)\boldsymbol{\rho}^{\star[\mathrm{PR}](j_2)}_{N,k-1}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} [\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)]_{j_1 j_2}, \\
\sum_{k=2}^{N-1} \mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\rho}^{\star[\mathrm{PR}](j_1)}_{N,k-1}(\boldsymbol{\theta}_0)\boldsymbol{\rho}^{\star[\mathrm{PR}](j_2)}_{N,k-1})^2(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] &\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.
\end{aligned}$$

Once again, the second limit follows trivially. To prove the first limit, start by noticing that:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1} \mid \mathcal{F}_{t_{k-1}}] = \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_j(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] = \frac{1}{3}\mathrm{Tr}(\mathbf{B}_j(\boldsymbol{\theta}_0)).$$

Then, we multiply the expectation with $N-2$ for notational simplicity and compute:

$$\begin{aligned}
&(N-2)\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\rho}^{\star[\mathrm{PR}](j_1)}_{N,k-1}(\boldsymbol{\theta}_0)\boldsymbol{\rho}^{\star[\mathrm{PR}](j_2)}_{N,k-1}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] \\
&= \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1} \mid \mathcal{F}_{t_{k-1}}] + 4\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{k-2}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1}\boldsymbol{\xi}^\top_{k-2}\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] \\
&+ \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1}\boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] + \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1}\boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1} \mid \mathcal{F}_{t_{k-1}}] \\
&+ \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}'^\top_{k-1}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\boldsymbol{\xi}'_{k-1}\boldsymbol{\xi}^\top_{k-1}\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-1} \mid \mathcal{F}_{t_{k-1}}] - \frac{4}{9}\mathrm{Tr}(\mathbf{B}_{j_1}(\boldsymbol{\theta}_0))\,\mathrm{Tr}(\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)).
\end{aligned} \tag{S24}$$

Applying Corollary S1.4 on (S24) yields:

$$\begin{aligned}
\sum_{k=2}^{N-1} \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\rho}^{\star[\mathrm{PR}](j_1)}_{N,k-1}(\boldsymbol{\theta}_0)\boldsymbol{\rho}^{\star[\mathrm{PR}](j_2)}_{N,k-1}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] &= \frac{5}{9}\mathrm{Tr}(\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)) \\
&+ \frac{4}{3}\frac{1}{N-2}\sum_{k=2}^{N-1} \boldsymbol{\xi}^\top_{k-2}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-2}.
\end{aligned}$$

Once again, applying Proposition 3.1 from Crimaldi and Pratelli [2005] yields:

$$\frac{4}{3}\frac{1}{N-2}\sum_{k=2}^{N-1} \boldsymbol{\xi}^\top_{k-2}\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-2} \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{4}{9}\mathrm{Tr}(\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)),$$

since

$$\frac{1}{N-2}\sum_{k=2}^{N-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}_{k-2}^\top \mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-2} \mid \mathcal{F}_{t_{k-2}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \frac{1}{3}\operatorname{Tr}(\mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)),$$
$$\frac{1}{(N-2)^2}\sum_{k=2}^{N-1}\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\xi}_{k-2}^\top \mathbf{B}_{j_1}(\boldsymbol{\theta}_0)\mathbf{B}_{j_2}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{k-2})^2 \mid \mathcal{F}_{t_{k-2}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

This concludes the convergence in distribution of $\boldsymbol{\lambda}_N^{[\mathrm{PR}]}$.

To find the asymptotic distributions of $\boldsymbol{\lambda}_N^{[\cdot\mathrm{S}|\mathrm{R}]}$, the main issue is the fact that $-\frac{1}{\sqrt{Nh}}\partial_{\beta^{(i)}}\mathcal{L}_N^{[\cdot\mathrm{S}|\mathrm{R}]} \to 0$ in probability. The proof of this follows the same ideas as in the proof of consistency. Thus, we focus only on $-\frac{1}{\sqrt{N}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\cdot\mathrm{S}|\mathrm{R}]}$. This is then used together with equations (71) and (74) to obtain the asymptotic distributions of $\boldsymbol{\lambda}_N^{[\cdot\mathrm{F}]}$. Thus, we start with $-\frac{1}{\sqrt{N}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\cdot\mathrm{S}|\mathrm{R}]}$:

$$\begin{aligned}
&-\frac{1}{\sqrt{(N-1)}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\mathrm{CS}|\mathrm{R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = -\frac{1}{\sqrt{N-1}}\sum_{k=1}^{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\\
&+\frac{3}{\sqrt{N-1}}\sum_{k=1}^{N}(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})+\sum_{k=1}^{N}R(\frac{h}{\sqrt{N}},\mathbf{Y}_{t_{k-1}}),\\
&-\frac{1}{\sqrt{(N-2)}}\partial_{\sigma^{(j)}}\mathcal{L}_N^{[\mathrm{PS}|\mathrm{R}]}(\mathbf{Y}_{0:t_N};\boldsymbol{\theta}) = -\frac{2}{\sqrt{N-2}}\sum_{k=1}^{N-1}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\\
&+\frac{3}{\sqrt{N-2}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}\\
&-6\sqrt{\frac{h}{N-2}}\sum_{k=1}^{N-1}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{F}(\mathbf{Y}_{t_{k-1}};\boldsymbol{\beta}_0)+\sum_{k=1}^{N-1}R(\frac{h}{\sqrt{N}},\mathbf{Y}_{t_{k-1}}).
\end{aligned}$$

Once again, we define:

$$\begin{aligned}
\boldsymbol{\rho}_{N,k-1}^{[\mathrm{CS}|\mathrm{R}](j)}(\boldsymbol{\theta}_0) \coloneqq \frac{1}{\sqrt{N-1}}\Bigg(&3(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\Sigma}_0(\boldsymbol{\eta}_{k-1}-2\boldsymbol{\xi}'_{k-1})\\
&-\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\Bigg), \qquad \text{(S25)}\\
\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j)}(\boldsymbol{\theta}_0) \coloneqq \frac{3}{\sqrt{N-2}}&(\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\Sigma}_0\mathbf{U}_{k,k-1}-\frac{2}{3}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top))\\
&-6\sqrt{\frac{h}{N-2}}\mathbf{U}_{k,k-1}^\top\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\sigma^{(j)}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{F}_0(\mathbf{Y}_{t_{k-1}}). \qquad \text{(S26)}
\end{aligned}$$

We skip the proof of the complete case, but it can be shown analogously that:

$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j_1)}(\boldsymbol{\theta}_0)\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j_2)}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} [\mathbf{C}_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_0)]_{j_1j_2},$$
$$\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j_1)}(\boldsymbol{\theta}_0)\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j_2)})^2(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0.$$

Focusing on the partial case, we first notice:

$$\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PS}|\mathrm{R}](j)}(\boldsymbol{\theta}_0) = 3\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PR}](j)}(\boldsymbol{\theta}_0) + o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1).$$

Thus, the same derivations from before hold. Moreover,

$$\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PF}](j)}(\boldsymbol{\theta}_0) = 4\boldsymbol{\rho}_{N,k-1}^{[\mathrm{PR}](j)}(\boldsymbol{\theta}_0) + o_{\mathbb{P}_{\boldsymbol{\theta}_0}}(1),$$

which concludes the proof.

Bibliography

Y. Aït-Sahalia. Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach. *Econometrica*, 70(1):223–262, January 2002. URL https://ideas.repec.org/a/ecm/emetrp/v70y2002i1p223-262.html.

Y. Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906 – 937, 2008. doi: 10.1214/009053607000000622. URL https://doi.org/10.1214/009053607000000622.

P. Bader, S. Blanes, and F. Casas. Computing the matrix exponential with an optimized taylor polynomial approximation. *Mathematics*, 7(12):1174, 2019. doi: 10.3390/math7121174. URL https://doi.org/10.3390/math7121174.

S. Blanes, F. Casas, and A. Murua. Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Mat. Apl.*, 45, 01 2009.

E. Buckwar, A. Samson, M. Tamborrino, and I. Tubikanec. A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *App. Num. Math.*, 179:191–220, 2022. ISSN 0168-9274. URL https://www.sciencedirect.com/science/article/pii/S0168927422001118.

F. Carbonell, J. Jímenez, and L. Pedroso. Computing multiple integrals involving matrix exponentials. *Journal of Computational and Applied Mathematics*, 213(1):300–305, 2008. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2007.01.007. URL https://www.sciencedirect.com/science/article/pii/S0377042707000283.

S. Choi. Closed-form likelihood expansions for multivariate time-inhomogeneous diffusions. *Journal of Econometrics*, 174(2):45–65, 2013. doi: 10.1016/j.jeconom.2011.12. URL https://ideas.repec.org/a/eee/econom/v174y2013i2p45-65.html.

S. Choi. Explicit form of approximate transition probability density functions of diffusion processes. *Journal of Econometrics*, 187(1):57–73, 2015. doi: 10.1016/j.jeconom.2015.02. URL https://ideas.repec.org/a/eee/econom/v187y2015i1p57-73.html.

S. Ditlevsen and A. Samson. Hypoelliptic diffusions: filtering and inference from complete and partial observations. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 81(2):361–384, 2019.

D. Falbel and J. Luraschi. *torch: Tensors and Neural Networks with 'GPU' Acceleration*, 2024. URL https://torch.mlverse.org/docs. R package version 0.13.0, https://github.com/mlverse/torch.

A. Gloter and N. Yoshida. Adaptive and non-adaptive estimation for degenerate diffusion processes, 2020.

A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, USA, second edition, 2008. ISBN 0898716594.

A. Hurn, K. Lindsay, and A. McClelland. A quasi-maximum likelihood method for estimating the parameters of multivariate diffusions. *Journal of Econometrics*, 172(1):106–126, 2013. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2012.09.002. URL https://www.sciencedirect.com/science/article/pii/S0304407612002187.

M. Hutzenthaler, A. Jentzen, and P. Kloeden. Strong and weak divergence in finite time of euler's method for stochastic differential equations with non-globally lipschitz continuous coefficients. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 467, 12 2010. doi: 10.1098/rspa.2010.0348.

Y. Iguchi and A. Beskos. Parameter inference for hypo-elliptic diffusions under a weak design condition, 2023.

J. C. Jimenez and R. J. Biscay. Approximation of continuous time stochastic processes by the local linearization method revisited. *Stochastic Analysis and Applications*, 20 (1):105–121, 2002. doi: 10.1081/SAP-120002423. URL https://doi.org/10.1081/SAP-120002423.

M. Kessler. Estimation of an Ergodic Diffusion from Discrete Observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997.

P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1992. ISBN 9783540540625. doi: 10.1007/978-3-662-12616-5. URL https://books.google.dk/books?id=BCvtssom1CMC.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21, 2016. doi: 10.18637/jss.v070.i05.

A. Melnykova. Parametric inference for hypoelliptic ergodic diffusions with full observations, 2020.

G. N. Milstein. A theorem on the order of convergence of mean-square approximations of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 32(4):738–741, 1988. doi: 10.1137/1132113. URL https://doi.org/10.1137/1132113.

J. K. Møller and H. Madsen. From state dependent diffusion to constant diffusion in stochastic differential equations by the lamperti transform. IMM-Technical Report 2010-16, Technical University of Denmark, DTU Informatics, Building 321, 2010.

D. Nualart. *The Malliavin Calculus and Related Topics*. Probability and Its Applications. Springer Berlin Heidelberg, 2006. ISBN 9783540283294.

T. Ozaki. Statistical Identification of Storage Models with Application to Stochastic Hydrology. *Journal of The American Water Resources Association*, 21:663–675, 1985.

T. Ozaki, J. C. Jimenez, and V. Haggan-Ozaki. The Role of the Likelihood Function in the Estimation of Chaos Models. *Journal of Time Series Analysis*, 21(4):363–387, 2000.

P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics*, 52(2):842 – 867, 2024a. doi: 10.1214/24-AOS2371. URL https://doi.org/10.1214/24-AOS2371.

P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting for parametric inference in second-order stochastic differential equations, 2024b.

M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. IEEE, 1993.

I. Shoji. Approximation of Continuous Time Stochastic Processes by a Local Linearization Method. *Mathematics of Computation*, 67(221):287–298, 1998.

I. Shoji and T. Ozaki. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4):733–752, 1998.

Stan Development Team. RStan: the R interface to Stan, 2024. URL https://mc-stan.org/. R package version 2.32.6.

M. V. Tretyakov and Z. Zhang. A Fundamental Mean-Square Convergence Theorem for SDEs with Locally Lipschitz Coefficients and Its Applications. *SIAM Journal on Numerical Analysis*, 51(6):3135–3162, 2013. URL https://doi.org/10.1137/120902318.

M. Uchida and N. Yoshida. Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications*, 122(8):2885–2924, 2012.

C. Van Loan. Computing Integrals Involving the Matrix Exponential. *IEEE Trans. Aut. Cont.*, 23(3):395–404, 1978.

Wolfram Research, Inc. Mathematica, Version 13.1. URL https://www.wolfram.com/mathematica. Champaign, IL, 2022.