Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations Based on Splitting Schemes. A preprint

Predrag Pilipovic, Adeline Samson, Susanne Ditlevsen

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

A PREPRINT

Predrag Pilipovic
Department of Mathematics
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 Mathematics
University of Copenhagen
2100 Copenhagen, Denmark
susanne@math.ku.dk

ABSTRACT

Surprisingly, general estimators for nonlinear continuous time models based on stochastic differential equations are yet lacking. Most applications still use the Euler-Maruyama discretization, despite many proofs of its bias. More sophisticated methods, such as Kessler's Gaussian approximation, Ozak's Local Linearization, Aït-Sahalia's Hermite expansions, or MCMC methods, lack a straightforward implementation, 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.

Keywords Asymptotic normality · Consistency · L^p convergence · Splitting schemes · Stochastic differential equations · Stochastic Lorenz system

1 Introduction

Stochastic differential equations (SDEs) are popular models for physical, biological, and socio-economic 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:

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

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 (Fuchs, 2013; Chopin and Papaspiliopoulos, 2020) 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.

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 et al. (2018), Gu et al. (2020) for applied studies). Second, Hutzenthaler et al. (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, non-equidistant 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 by 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 et al. (2017) extended the theory of LL for SDEs with multiplicative noise. Simulation studies show the superiority of the LL estimator compared to other estimators (Shoji and Ozaki, 1998; Hurn et al., 2007; Gloaguen et al., 2018; Gu et al., 2020). 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 et al. (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 et al. (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 non-linear models. Hurn et al. (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. This finding aligns with the newer review study from López-Pérez et al. (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 (2020, 2021) have extended the research of Uchida and Yoshida (2012). Gloter and Yoshida (2020) introduced a non-adaptive approach and offered similar analytic asymptotic results as Ditlevsen and Samson (2019) without imposing strict limitations on the model class. Iguchi et al. (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 et al. (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 et al. (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 non-specialists 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 for example, McLachlan and Quispel (2002); Blanes et al. (2009)), but their extension to SDEs is straightforward. A few studies have investigated numerical properties (Bensoussan et al., 1992; Ableidinger et al., 2017; Ableidinger and Buckwar, 2016; Buckwar et al., 2022). Barbu (1988) applied LT splitting on nonlinear optimal control problems, while Hopkins and Wong (1986) used it for nonlinear filtering. Bou-Rabee and Owhadi (2010); Abdulle et al. (2015) used LT splitting to investigate conditions for preserving the measure of the ergodic nonlinear Langevin equations. Recently, Bréhier et al. (2023) 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 Misawa (2001); Milstein and Tretyakov (2003); Leimkuhler and Matthews (2015); Alamo and Sanz-Serna (2016); Bréhier and Goudenège (2019). Regarding statistical applications, to the best of our knowledge, only Buckwar et al. (2020); Ditlevsen et al. (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 three 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

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 et al., 2017)	$h \to 0$, $Nh \to \infty$, $Nh^2 \to 0$ (Ozaki, 1992)	Slowest discrete ML approximations. (Hurn et al., 2007) Straightforward for any dimension.	Best among all discrete ML approximations. (Hurn et al., 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 et al., 2007)	For larger h, a higher order of J is needed. Better than LL in the univariate case. (Hurn et al., 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.

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.

[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 $J \to \infty$.

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, LT, and S estimators. We show that the EM and LT estimators become biased before the others with increasing discretization step h and that 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 i-th component, while on a matrix it denotes the i-th row. Double subscript ij on a matrix denotes the component in the i-th row and j-th 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:

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

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 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. In this case, the splitting proposed below will result in a Brownian motion (3) and a nonlinear ODE (4).

Remark We assume additive noise, meaning that the diffusion matrix does not depend on the current state. While this assumption is natural in some applications, it can be restrictive in others. The proposed methodology could potentially be extended to reducible diffusions by applying the Lamperti transform to obtain a unit diffusion coefficient, as demonstrated by Aït-Sahalia (2008). 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 (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 $\mathbf{x}$ and $\boldsymbol{\theta}$, i.e., $\mathbf{N} \in C^2$. Additionally, it is one-sided globally Lipschitz continuous with respect to $\mathbf{x}$ on $\mathbb{R}^d \times \overline{\Theta}_\beta$, i.e., there exists a constant $C > 0$ such that:

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

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

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

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 $\boldsymbol{\beta}$, i.e., 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 the ellipticity of model (1), which is not needed for the S estimator. On the contrary, 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 $0 = t_0 < t_1 < \cdots < t_N = T$, which we, for notational simplicity, assume equidistant with 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), i.e.,

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

Furthermore, we need the infinitesimal generator L of (1) defined on sufficiently smooth functions $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ given by:

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

The moments of SDE (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* $p \geq 2$*. Then,*

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

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

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

2.3 Splitting Schemes

Consider the following splitting of (2):

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

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

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

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

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:

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

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

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:

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

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

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

Remark 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 expending the nonlinear solution $\boldsymbol{f}_h$ (Section 1.8 in (Hairer et al., 1993)).

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

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

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:*

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

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

Remark 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 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 Overall trajectories of the S and LT splittings coincide up to the first $h/2$ and the last $h/2$ move of the flow $\Phi_{h/2}^{[2]}$. Indeed, when applied k times, the S splitting can be written as:

$$(\Phi_h^{[\mathrm{S}]})^k(\mathbf{x}_0) = (\Phi_{h/2}^{[2]} \circ (\Phi_h^{[\mathrm{LT}]})^k \circ \Phi_{-h/2}^{[2]})(\mathbf{x}_0).$$

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). The S splitting outputs the middle points of the smooth steps of the deterministic flow (8), while the 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 EM, K, and LL estimators, 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:

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

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

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

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

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

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 it 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:

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

2.4.2 Euler-Maruyama

The EM method uses first-order Taylor expansion of (1):

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

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]} \overset{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}, h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ for $k = 1,\ldots,N$ (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.

2.4.3 Kessler

The K estimator uses 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). 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:

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

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 et al., 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:

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

where $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}) = \int_0^h e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})(h-u)} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})^\top (h-u)} \,\mathrm{d}u$ and $\boldsymbol{\xi}_{h,k}^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}))$. Moreover,

$$\Phi_h^{[\mathrm{LL}]}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) + (h\mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) - \mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})))\mathbf{M}(\mathbf{x};\boldsymbol{\theta}),$$

$$\begin{aligned}
\mathbf{R}_{h,i}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) &= \int_0^h \exp(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})u) u^i \,\mathrm{d}u, \quad i = 0, 1, \\
\mathbf{M}(\mathbf{x};\boldsymbol{\theta}) &= \frac{1}{2}(\mathrm{Tr}\,\mathbf{H}_1(\mathbf{x};\boldsymbol{\theta}), \mathrm{Tr}\,\mathbf{H}_2(\mathbf{x};\boldsymbol{\theta}), ..., \mathrm{Tr}\,\mathbf{H}_d(\mathbf{x};\boldsymbol{\theta}))^\top, \\
\mathbf{H}_k(\mathbf{x};\boldsymbol{\theta}) &= \left[[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{ij} \frac{\partial^2 F^{(k)}}{\partial x^{(i)} \partial x^{(j)}}(\mathbf{x}) \right]_{i,j=1}^{d}.
\end{aligned}$$

Building on the approach by Gu et al. (2020), we can efficiently compute $\mathbf{R}_{h,i}$ and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ using the following procedure. To begin, let us define three block matrices:

$$\mathbf{P}_1(\mathbf{x}) = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) \end{bmatrix}, \mathbf{P}_2(\mathbf{x}) = \begin{bmatrix} -D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \mathbf{I}_d & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{bmatrix}, \mathbf{P}_3(\mathbf{x}) = \begin{bmatrix} D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\ \mathbf{0}_{d\times d} & -D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})^\top \end{bmatrix}. \tag{19}$$

Then, we compute the matrix exponential of matrices $h\mathbf{P}_1(\mathbf{x})$ and $h\mathbf{P}_2(\mathbf{x})$:

$$\exp(h\mathbf{P}_1(\mathbf{x})) = \begin{bmatrix} \star & \mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) \\ \mathbf{0}_{d\times d} & \star \end{bmatrix}, \qquad \exp(h\mathbf{P}_2(\mathbf{x})) = \begin{bmatrix} \star & \star & \mathbf{B}_{\mathbf{R}_{h,1}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) \\ \mathbf{0}_{d\times d} & \star & \star \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \star \end{bmatrix}.$$

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 $\mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) = \exp(hD\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))\mathbf{B}_{\mathbf{R}_{h,1}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$. The terms marked with $\star$ symbols can be disregarded. Finally, we obtain $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ from the matrix exponential:

$$\begin{aligned}
\exp(h\mathbf{P}_3(\mathbf{x})) &= \begin{bmatrix} \mathbf{B}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta}) & \mathbf{C}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta}) \\ \mathbf{0}_{d\times d} & \star \end{bmatrix}, \\
\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}) &= \mathbf{C}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta})\mathbf{B}_{\boldsymbol{\Omega}_{h,k}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta});\boldsymbol{\theta})^\top.
\end{aligned}$$

Thus, we have a Gaussian density $p^{[\mathrm{LL}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}}; \boldsymbol{\theta})$ and standard likelihood inference. Like in the case of K, the covariance matrix $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ 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 use Taylor expansions of the approximated drift and covariance matrix, so the influence of the sample size N on computational times is much stronger. For details on the derivations of the previous formulas, see Gu et al. (2020).

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. It means that tiny differences in initial conditions lead to unpredictable and widely diverging trajectories. The Lorenz system evolves around two strange attractors. It means that the trajectories remain within some bounded region, while points that start in close proximity may eventually separate

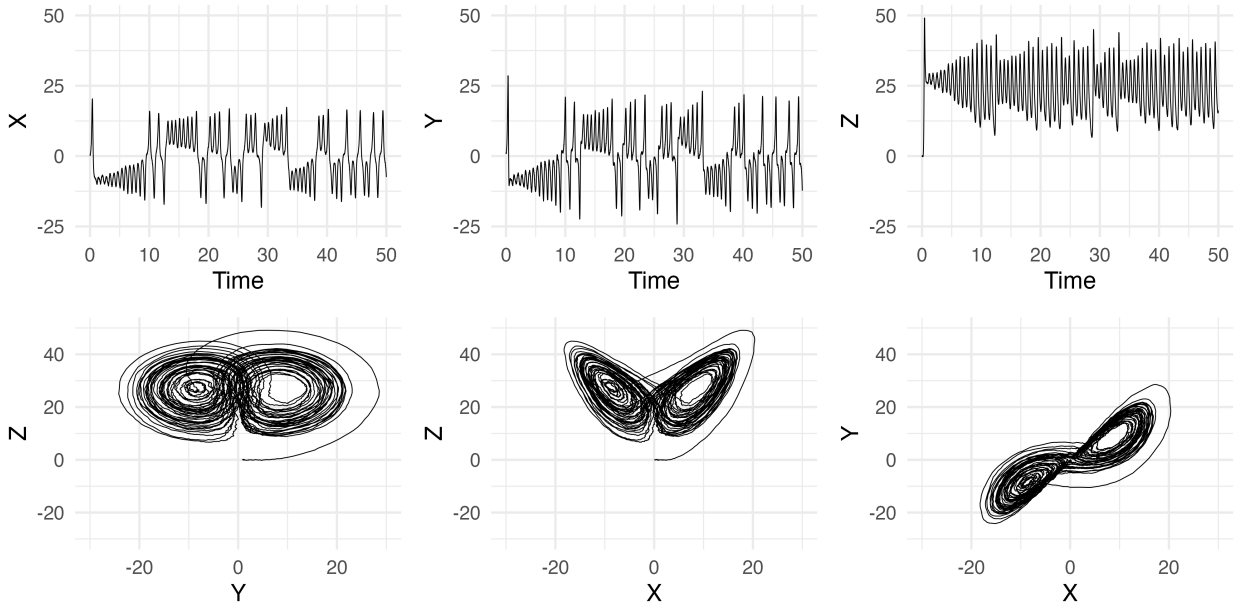

Figure 1: An example trajectory of the stochastic Lorenz system (20) starting at $(0, 1, 0)$ for $N = 10000$ and $h = 0.005$. The first row shows the evolution of the individual components X, Y, and Z. The second row shows the evolution of component pairs: (Y, Z), (X, Z) and (X, Y). Parameters are $p = 10$, $r = 28$, $c = 8/3$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\sigma_3^2 = 1.5$.

by arbitrary distances as time progresses. (Hilborn and Hilborn, 2000). We add noise to include unmodelled forces and randomness in the Lorenz system . 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 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 et al. (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 et al. (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 $N = 10000$ 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:

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

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

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

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, $\det D\boldsymbol{f}^{-1}_{\text{mix},h}(x,y,z) = 1$.

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 LT_{avg} and S_{avg}.The splitting is given by:

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

The nonlinear solution is:

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

where we define $\boldsymbol{f}_{\text{avg},h}(\mu_x, y, z) = (\mu_x, y + h\mu_x(r-1-\mu_z), z + h\mu_x^2 - c\mu_z)^\top$. Again, $\det D\boldsymbol{f}^{-1}_{\text{avg},h}(x,y,z) = 1$.

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, 1988) 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 $q_1 \geq q_2 + 1/2$:*

$$\|\mathbb{E}[\mathbf{X}_{t_k} - \widetilde{\Phi}_h(\mathbf{X}_{t_{k-1}}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]\| = R(h^{q_1}, \mathbf{x}),$$
$$(\mathbb{E}[\|\mathbf{X}_{t_k} - \widetilde{\Phi}_h(\mathbf{X}_{t_{k-1}})\|^{2p} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])^{\frac{1}{2p}} = R(h^{q_2}, \mathbf{x}),$$

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 $k = 1, \ldots, N$:*

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

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, the numerical method $\widetilde{\mathbf{X}}$ is L^p convergent, $p \geq 1$, of order $q_2 - 1/2$, i.e., for $k = 1, \ldots, N$, it holds:

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

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 et al., 2023) for scheme (10), as well as for the reversed order of composition. We demonstrate that the order of one-step prediction can not be improved unless the drift $\mathbf{F}$ is linear.

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

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

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) *Let Assumptions (A1) and (A2) hold, 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* $k = 1, \dots, N$*, it holds:*

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

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 Supplementary Material (Pilipovic et al., 2023).

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

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

Remark 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 Section 7.1.

Theorem 3.7 (L^p convergence of S splitting) *Let Assumptions (A1), (A2) and (A6) hold, 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* $k = 1, \dots, N$*, it holds:*

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

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:

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

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,

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

We used Proposition 2.2, together with the fact that $\mathbf{X}$ and $\mathbf{X}^{[\mathrm{S}]}$ have finite moments and $\boldsymbol{f}_{h/2}$ and $\boldsymbol{f}_{h/2}^{-1}$ 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 p-th moments of the solution of the SDE increments and also provides a moment bound of a polynomial map of the solution. The proof of this lemma, presented in Section 7.2, 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 for the drift function $\mathbf{F}$ in conjunction with the generalized Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020) stated in Supplementary Material (Pilipovic et al., 2023)) to establish the result.

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 *Let Assumptions (A1) and (A2) hold. 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:*

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

(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:*

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

Lemma 4.2 *Let Assumptions (A1), (A2) and (A3) hold, 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,*

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

uniformly in $\boldsymbol{\theta}$*.*

Lastly, we state the moment bounds needed for the estimator asymptotics. The proof is in Supplementary Material (Pilipovic et al., 2023).

Proposition 4.3 (Moment Bounds) *Let Assumptions (A1), (A2) and (A6) hold. 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) $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] = \mathbf{R}(h^3, \mathbf{X}_{t_{k-1}})$

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

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

5 Asymptotics

The estimators $\hat{\boldsymbol{\theta}}_N$ are defined in (15). However, we do not need the full objective functions (12) and (14) to prove consistency and asymptotic normality. It is enough to approximate the covariance matrix $\boldsymbol{\Omega}_h$ 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:

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

Similarly, we approximate the log-determinant as:

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

Using the same approximation we obtain:

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

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

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

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

Unlike other likelihood-based methods, such as Kessler (1997), Aït-Sahalia (2002, 2008), Choi (2013, 2015), Yang et al. (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) without requiring approximations. However, we leverage the approximations (22) and (23) for the mathematical analysis and 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 Section 7.3.

Theorem 5.1 *Let Assumptions (A1)-(A6) hold,* $\mathbf{X}$ *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 one of objective functions* (22) *or* (23)*. If* $h \to 0$ *and* $Nh \to \infty$*, then,*

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

5.2 Asymptotic normality

In this Section, we state the asymptotic normality of the estimator. First, we need some preliminaries. Let $\rho > 0$ and $\mathcal{B}_\rho(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \rho\}$ 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 one of the two objective functions (22) or (23). For $\hat{\boldsymbol{\theta}}_N \in \mathcal{B}_\rho(\boldsymbol{\theta}_0)$, the mean value theorem yields:

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

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

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

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

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

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

where:

$$[\mathbf{C}_\beta(\boldsymbol{\theta}_0)]_{i_1,i_2} \coloneqq \int (\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}),\ 1 \leq i_1, i_2 \leq r,$$
$$[\mathbf{C}_\varsigma(\boldsymbol{\theta}_0)]_{j_1,j_2} \coloneqq \frac{1}{2}\,\mathrm{Tr}((\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}),\ 1 \leq j_1, j_2 \leq s.$$

Now, we state the theorem for asymptotic normality, whose proof is in Section 7.4.

Theorem 5.2 *Let Assumptions (A1)-(A6) hold,* $\mathbf{X}$ *be the solution of* (1)*, and* $\widehat{\boldsymbol{\theta}}_N = (\widehat{\boldsymbol{\beta}}_N, \widehat{\boldsymbol{\varsigma}}_N)$ *be the estimator that minimizes one of the objective functions* (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,*

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

under $\mathbb{P}_{\boldsymbol{\theta}_0}$.

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 from the literature. 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:

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

We do not write the closed-form distributions for the K (17) and LL (18) estimators, 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 the one suggested by Gu et al. (2020) for calculating $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$. Namely, for the splitting schemes, we adapt $\mathbf{P}_3$ from (19) accordingly:

$$\mathbf{P}_3 = \begin{bmatrix} \mathbf{A} & \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\ \mathbf{0}_{d\times d} & -\mathbf{A}^\top \end{bmatrix}.$$

Therefore, computing $\exp(h\mathbf{P}_3)$ circumvents the need for evaluating the integral in $\boldsymbol{\Omega}_h$ (6), following the approach described in Section 2.4.4.

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 sub-sample 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 converged after approximately 80 iterations.

6.4 Comparing criteria

We compare seven estimators based on their precision and speed. For the precision, we compute the absolute relative error (ARE) for each component $\hat{\theta}_N^{(i)}$ of the estimator $\hat{\boldsymbol{\theta}}_N$:

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

For S and LL, we compare the distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ to investigate the precision 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 as a function of the discretization step h. For a clearer comparison, we use the log scale on the y axis. 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

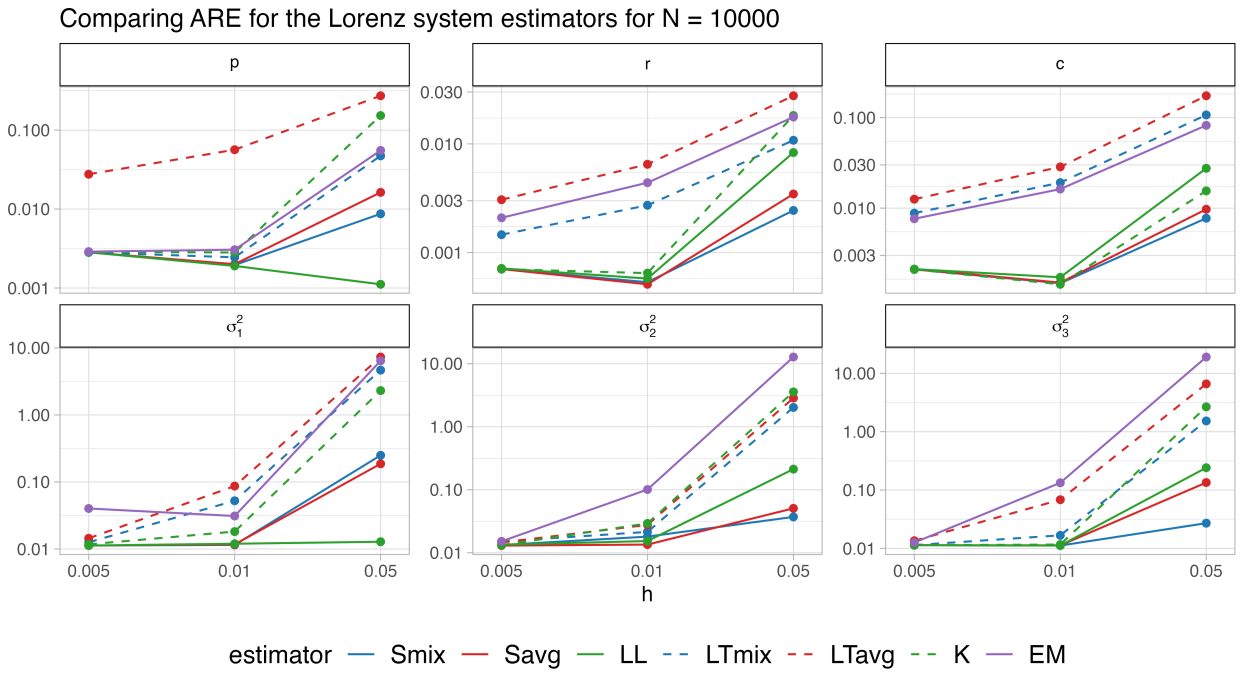

Figure 2: Comparing the absolute relative error (ARE) as a function of increasing h for seven different estimators in the stochastic Lorenz system. The estimators are obtained for sample sizes of $N = 10000$. The y-axis is on a log scale.

diffusion parameters, which is due to the poor approximation of $\mathbf{\Omega}_h^{\mathrm{EM}}$. K is less biased than EM except for p and r when $h = 0.05$. Note how some parameters are estimated better for larger h when N is fixed. This is due to a longer observation interval $T = Nh$, reflecting the $\sqrt{Nh}$ rate of convergence.

Since $\mathrm{S}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$, and LL perform the best with large time steps, we zoom in on their distributions in Figure 3. To make the figure clearer, 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. The three estimators perform similarly, especially for small h. 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 exhibited superior performance compared to LL, except for the parameters 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 second slowest estimator is $\mathrm{S}_{\mathrm{mix}}$, 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 the $\mathrm{S}_{\mathrm{mix}}$ and $\mathrm{LT}_{\mathrm{mix}}$ estimators. 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 S estimators are.

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

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

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.

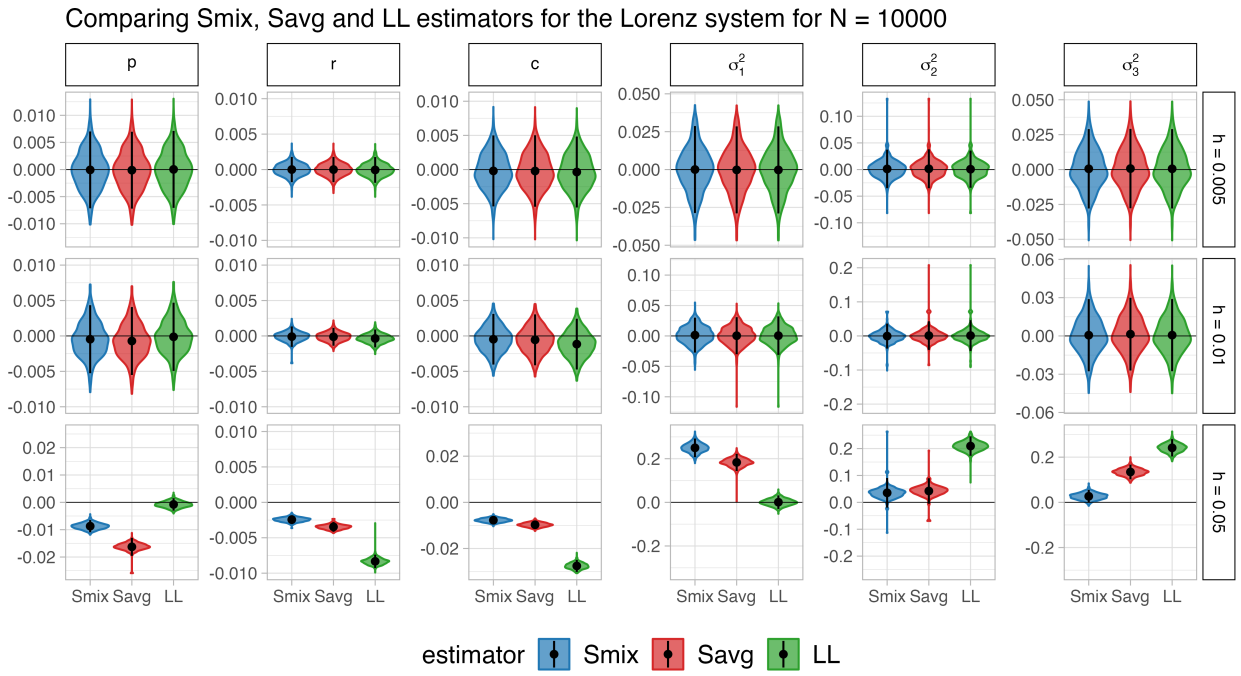

Figure 3: Comparison of the normalized distributions of $(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \oslash \boldsymbol{\theta}_0$ (where $\oslash$ is the element-wise division) in the Lorenz system for the $\mathrm{S}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$, and LL estimators for $N = 10000$. Each column represents one parameter, and each row represents one value of the discretization step h. A black dot with a vertical bar in each violin plot represents the mean and the standard deviation.

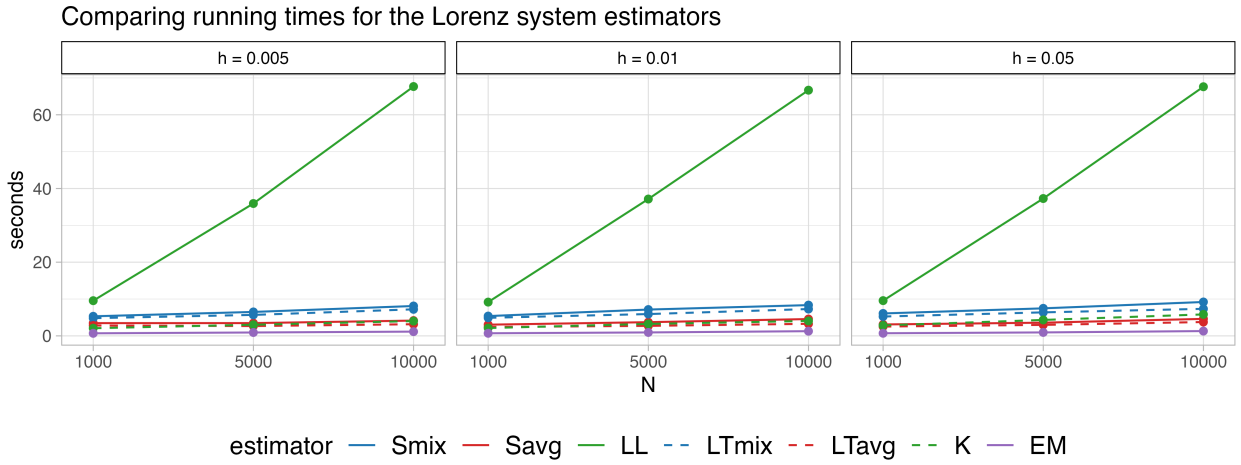

Figure 4: Running times as a function of N for different estimators of the Lorenz system. Each column shows one value of h. On the x-axis is the sample size N, and on the y-axis is the running time in seconds.

7 Proofs

7.1 Proof of L^p convergence of the splitting scheme

The proof of Proposition 3.6 is in Supplementary Material (Pilipovic et al., 2023). Here, we present the proof of L^p convergence stated in Theorem 3.7.

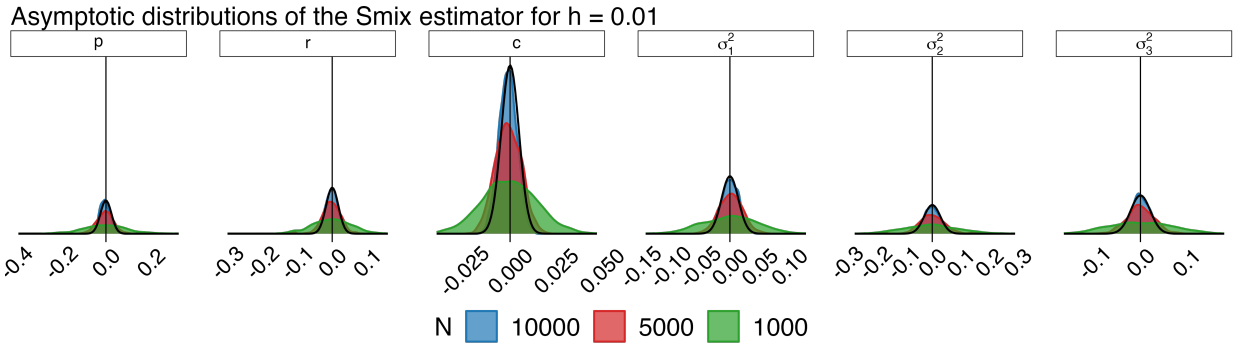

Figure 5: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ for the $\mathrm{S}_{\mathrm{mix}}$ estimator with theoretical asymptotic distributions (28) for each parameter (columns), for $h = 0.01$ and $N \in \{1000, 5000, 10000\}$ (colors). The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for $N = 10000$ and $h = 0.01$.

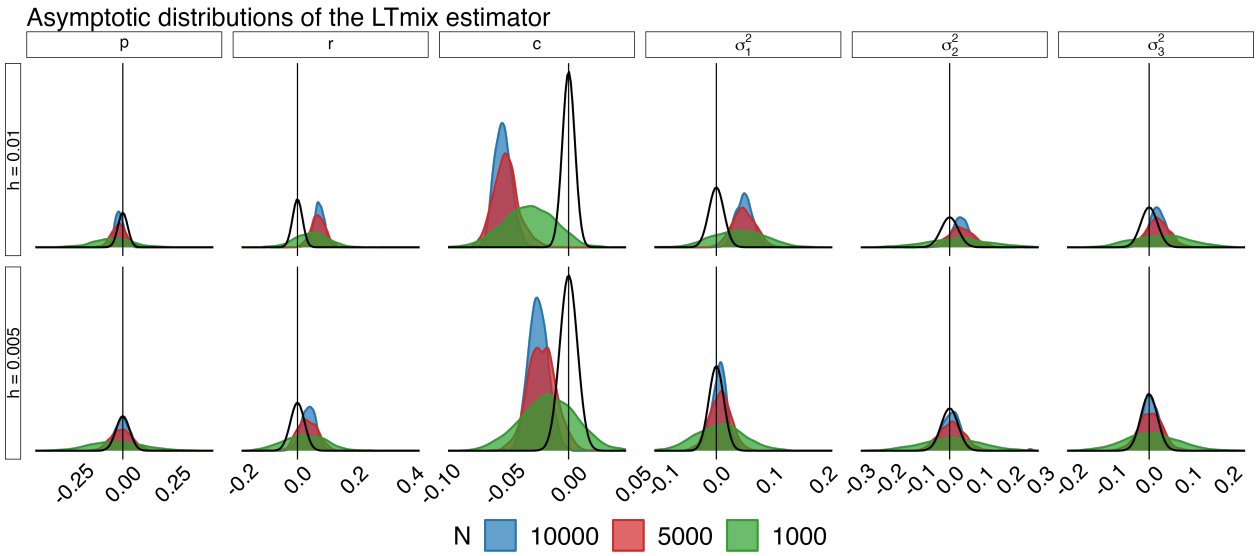

Figure 6: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ for the $\mathrm{LT}_{\mathrm{mix}}$ estimator with theoretical asymptotic distributions (28) for each parameter (columns), for $h \in \{0.005, 0.01\}$ (rows) and $N \in \{1000, 5000, 10000\}$ (colors). The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for $N = 10000$ and corresponding h.

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 Supplementary Material (Pilipovic et al., 2023). 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:

$$\begin{aligned}
\boldsymbol{\xi}_{h,k} &= \int_{t_{k-1}}^{t_k} (\mathbf{I} + (t_k - s)\mathbf{A})\boldsymbol{\Sigma} \, \mathrm{d}\mathbf{W}_s + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}}) \\
&= \boldsymbol{\Sigma}(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}}) + \mathbf{A}\boldsymbol{\Sigma} \int_{t_{k-1}}^{t_k} (t_k - s) \, \mathrm{d}\mathbf{W}_s + \mathbf{R}(h^2, \mathbf{X}_{t_{k-1}}).
\end{aligned}$$

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:

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})\|^p \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]\\
&\leq C(\mathbb{E}[h^{p-1}\int_{t_{k-1}}^{t_k}\|\mathbf{F}(\mathbf{X}_s) - \mathbf{F}(\mathbf{X}_{t_{k-1}})\|^p\,\mathrm{d}s \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])\\
&\leq C(h^{p-1}\int_{t_{k-1}}^{t_k}\mathbb{E}[\|\mathbf{X}_s - \mathbf{X}_{t_{k-1}}\|^p\|\int_0^1 D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_s - u(\mathbf{X}_s - \mathbf{X}_{t_{k-1}}))\,\mathrm{d}u\|^p \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]\,\mathrm{d}s)\\
&\leq C\Bigg(h^{p-1}\int_{t_{k-1}}^{t_k}(\mathbb{E}[\|\mathbf{X}_s - \mathbf{X}_{t_{k-1}}\|^{2p} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])^{\frac{1}{2}}\\
&\qquad\qquad(\mathbb{E}[\|\int_0^1 D_{\mathbf{x}}\mathbf{F}(\mathbf{X}_s - u(\mathbf{X}_s - \mathbf{X}_{t_{k-1}}))\,\mathrm{d}u\|^{2p} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}])^{\frac{1}{2}}\,\mathrm{d}s\Bigg)\\
&\leq C(h^{p-1}\int_{t_{k-1}}^{t_k}h^{\frac{p}{2}}\,\mathrm{d}s) = R(h^{3p/2}, \mathbf{x}).
\end{aligned}$$

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 $\mathbf{R}_{t_k} \coloneqq e^{\mathbf{A}h}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) - \mathbf{X}_{t_k}^{[1]})$, 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.

7.2 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:

$$\begin{aligned}
&\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\|^p \leq 2^{p-1}(\|\int_{t_{k-1}}^t \mathbf{F}(\mathbf{X}_s;\boldsymbol{\theta})\,\mathrm{d}s\|^p + \|\boldsymbol{\Sigma}(\mathbf{W}_t - \mathbf{W}_{t_{k-1}})\|^p)\\
&\leq 2^{p-1}((\int_{t_{k-1}}^t C_1(1 + \|\mathbf{X}_s\|)^{C_1}\,\mathrm{d}s)^p + \|\boldsymbol{\Sigma}(\mathbf{W}_t - \mathbf{W}_{t_{k-1}})\|^p)\\
&\leq 2^{p-1}C_1^p(\int_{t_{k-1}}^t(1 + \|\mathbf{X}_s - \mathbf{X}_{t_{k-1}}\| + \|\mathbf{X}_{t_{k-1}}\|)^{C_1}\,\mathrm{d}s)^p + 2^{p-1}\|\boldsymbol{\Sigma}(\mathbf{W}_t - \mathbf{W}_{t_{k-1}})\|^p\\
&\leq 2^{C_1+2p-3}C_1^p(t - t_{k-1})^{p-1}(\int_{t_{k-1}}^t\|\mathbf{X}_s - \mathbf{X}_{t_{k-1}}\|^{pC_1}\,\mathrm{d}s + (t - t_{k-1})^p(1 + \|\mathbf{X}_{t_{k-1}}\|)^{pC_1})\\
&+ 2^{p-1}\|\boldsymbol{\Sigma}(\mathbf{W}_t - \mathbf{W}_{t_{k-1}})\|^p.
\end{aligned}$$

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 $\mathbb{E}[\|\boldsymbol{\Sigma}(\mathbf{W}_t - \mathbf{W}_{t_{k-1}})\|^p \mid \mathcal{F}_{t_{k-1}}] = (t - t_{t_{k-1}})^{p/2}C_2(p)$. Then, for $h < 1$, there exists a constant C_p that depends on p, such that:

$$C_p(t - t_{k-1})^{2p-1}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p} + C_p(t - t_{t_{k-1}})^{p/2} \leq C_p(t - t_{k-1})^{p/2}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p}.$$

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:

$$m(t) \leq C_p(t - t_{k-1})^{p/2}(1 + \|\mathbf{X}_{t_{k-1}}\|)^{C_p} + C_p\int_{t_{k-1}}^t m^{C_1}(s)\,\mathrm{d}s. \tag{29}$$

Now, we apply the generalized Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020), stated in Supplementary Material (Pilipovic et al., 2023)) on (29). Since we consider a super-linear growth, we can assume that there exist $C_1 > 1$ and $C_p > 0$, such that:

$$\begin{aligned} m(t) &\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p} + (\kappa^{1-C_1}(t) - (C_1-1)2^{C_1-1}C_p(t-t_{k-1}))^{\frac{1}{1-C_1}} \\ &\leq C_p(t-t_{k-1})^{p/2}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_p} + C\kappa(t), \end{aligned} \tag{30}$$

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 (30) 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 (30) 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:

$$\mathbb{E}[|g(\mathbf{X}_t;\boldsymbol{\theta})| \mid \mathcal{F}_{t_{k-1}}] \leq C_2(1+\|\mathbf{X}_{t_{k-1}}\|^{C_1} + C'_{t-t_{k-1}}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C_3}) \leq C_{t-t_{k-1}}(1+\|\mathbf{X}_{t_{k-1}}\|)^{C}.$$

That concludes the proof.

7.3 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. The proof of the Lemma is in Supplementary Material (Pilipovic et al., 2023).

Lemma 7.1 *Let Assumptions (A1)-(A6) hold, and* $\mathbf{X}$ *be the solution of* (1)*. Let* $\mathbf{g}, \mathbf{g}_1, \mathbf{g}_2 : \mathbb{R}^d \times \Theta \times \Theta \to \mathbb{R}$ *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. $\frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} \int \mathrm{Tr}(D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\,\mathrm{d}\nu_0(\mathbf{x})$;
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 $\boldsymbol{\theta}$.

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:

$$\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \to \log(\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)) + \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) =: G_1(\boldsymbol{\varsigma}, \boldsymbol{\varsigma}_0), \tag{31}$$

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:

$$\frac{1}{N}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) = \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + T_1 + T_2 + T_3 + 2(T_4 + T_5 + T_6) + R(h, \mathbf{x}_0). \tag{32}$$

The terms $T_1, \ldots, T_6$ 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 (32) is a constant. Properties 1, 2, 3, 5, and 7 from Lemma 7.1 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 (31) 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 $\boldsymbol{\theta}$:

$$\frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0,\boldsymbol{\varsigma})) \to G_2(\boldsymbol{\beta}_0,\boldsymbol{\varsigma}_0,\boldsymbol{\beta},\boldsymbol{\varsigma}), \tag{33}$$

where:

$$\begin{aligned} G_2(\boldsymbol{\beta}_0,\boldsymbol{\varsigma}_0,\boldsymbol{\beta},\boldsymbol{\varsigma}) \coloneqq &\int (\mathbf{F}_0(\mathbf{x}) - \mathbf{F}(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{F}_0(\mathbf{x}) - \mathbf{F}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}) \\ &+ \int \mathrm{Tr}(D(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} - \mathbf{I}))\,\mathrm{d}\nu_0(\mathbf{x}). \end{aligned}$$

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:

$$\begin{aligned} \frac{1}{Nh}(\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) - \mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta}_0,\boldsymbol{\varsigma})) &= \mathrm{Tr}(\mathbf{A}(\boldsymbol{\beta}) - \mathbf{A}(\boldsymbol{\beta}_0)) + \frac{1}{h}(T_2 + T_3 + 2(T_4 + T_5 + T_6)) \\ &+ \frac{1}{Nh}\sum_{k=1}^N (\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) - \mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{A}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta})) \\ &+ \frac{1}{N}\sum_{k=1}^N \mathrm{Tr}\, D(\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)) + R(h,\mathbf{x}_0). \end{aligned} \tag{34}$$

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 Jacod (1993) and Property 4 from Lemma 7.1. Lemma 4.2 yields:

$$\begin{aligned} \frac{1}{h}T_5 \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} &\frac{1}{4}\int (\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}) \\ &+ \frac{1}{2}\int (\mathbf{A}_0(\mathbf{x}-\mathbf{b}_0) - \mathbf{A}(\mathbf{x}-\mathbf{b}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}). \end{aligned}$$

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 7.1. 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 (33). 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 (32) 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.

7.4 Proof of asymptotic normality of the estimator

Proof of Theorem 5.2 According to Theorem 1 in Kessler (1997) or Theorem 1 in Sørensen and Uchida (2003), Lemmas 7.2 and 7.3 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 7.2 *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 7.3 *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 7.2 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 7.2 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 7.2 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 7.3 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 $\mathbf{E}_N \coloneqq \mathbf{D}_N$ 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 7.3, $\mathbf{s}_N \mathbb{1}_{\mathcal{E}_N} \to \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_0)^{-1})$ 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 7.2 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 7.1 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}}\\
&\qquad \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 \mathrm{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:

$$\frac{1}{N}\sum_{k=1}^N \partial_{\beta_{i_1}\beta_{i_2}} \mathrm{Tr}\, D\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \int \mathrm{Tr}(D\partial_{\beta_{i_1}\beta_{i_2}} \mathbf{N}_0(\mathbf{x}))\, \mathrm{d}\nu_0(\mathbf{x}).$$

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 $\partial_{\beta_{i_1}\beta_{i_2}} \mathrm{Tr}\, \mathbf{A}(\boldsymbol{\beta}_0)$ (Property 1 Lemma 7.1).

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:

$$\partial_{\varsigma_{j_1}\varsigma_{j_2}} \log(\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) = \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \partial_{\varsigma_{j_1}\varsigma_{j_2}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) - \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_1}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \partial_{\varsigma_{j_2}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top).$$

On the other hand, we have:

$$\begin{aligned}
&\partial_{\varsigma_{j_1}\varsigma_{j_2}} \frac{1}{Nh}\sum_{k=1}^N \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\\
&= -\frac{1}{Nh}\sum_{k=1}^N \mathrm{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_1}\varsigma_{j_2}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\\
&+ \frac{1}{Nh}\sum_{k=1}^N \mathrm{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_1}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_2}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\\
&+ \frac{1}{Nh}\sum_{k=1}^N \mathrm{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_2}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\partial_{\varsigma_{j_1}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}).
\end{aligned}$$

Then, from Property 1 of Lemma 7.1, we get:

$$\partial_{\varsigma_{j_1}\varsigma_{j_2}} \frac{1}{Nh} \sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\Big|_{\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}$$
$$\xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\boldsymbol{\theta}_0}} 2\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) - \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top).$$

Thus, $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0} \to \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$. 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 7.3 First, we compute the first derivatives. We start with:

$$\partial_{\beta_i}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) = -2\sum_{k=1}^{N}\mathrm{Tr}(D\boldsymbol{f}_{h/2,k}(\boldsymbol{\beta})D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}))$$
$$+\frac{2}{h}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\partial_{\beta_i}\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta})).$$

The first derivative with respect to $\boldsymbol{\varsigma}$ is:

$$\begin{aligned}\partial_{\varsigma_j}\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\varsigma}) &= N\partial_{\varsigma_j}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\\ &\quad+\frac{1}{h}\partial_{\varsigma_j}\sum_{k=1}^{N}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))\\ &= -\frac{1}{h}\sum_{k=1}^{N}\Bigg(\mathrm{Tr}\Big((\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))^\top\\ &\qquad(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\Big)+\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\Bigg).\end{aligned}$$

Define:

$$\begin{aligned}\eta_{N,k}^{(i)}(\boldsymbol{\theta}) &\coloneqq \frac{2}{\sqrt{Nh}}\mathrm{Tr}(D\boldsymbol{f}_{h/2,k}(\boldsymbol{\beta})D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta}))\\ &\quad-\frac{2}{\sqrt{Nhh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\beta_i}(\boldsymbol{f}_{h/2,k}^{-1}(\boldsymbol{\beta})-\boldsymbol{\mu}_{h,k-1}(\boldsymbol{\beta}))\end{aligned} \tag{35}$$

$$\begin{aligned}\zeta_{N,k}^{(j)}(\boldsymbol{\theta}) &\coloneqq \frac{1}{\sqrt{Nh}}\mathrm{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\\ &\quad-\frac{1}{\sqrt{N}}\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top),\end{aligned} \tag{36}$$

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 7.4. For more details, see Supplementary Material (Pilipovic et al., 2023).

Lemma 7.4 *Let $\eta_{N,k}^{(i)}(\boldsymbol{\theta})$ and $\zeta_{N,k}^{(j)}(\boldsymbol{\theta})$ be defined as in* (35) *and* (36)*, 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)]

1. $\mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{1\le k\le N}|\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)|] \longrightarrow 0$, *and* $\mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{1\le k\le N}|\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)|] \longrightarrow 0$;
2. $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$, *and* $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)\mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
3. $\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\mid \mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;

4. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
5. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
6. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 4[\mathbf{C}_\beta(\boldsymbol{\theta}_0)]_{i_1 i_2}$;
7. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 4[\mathbf{C}_\varsigma(\boldsymbol{\theta}_0)]_{j_1 j_2}$;
8. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
9. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[(\eta_{N,k}^{(i_1)}(\boldsymbol{\theta}_0)\eta_{N,k}^{(i_2)}(\boldsymbol{\theta}_0))^2 \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
10. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[(\zeta_{N,k}^{(j_1)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j_2)}(\boldsymbol{\theta}_0))^2 \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$;
11. $\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[(\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0)\zeta_{N,k}^{(j)}(\boldsymbol{\theta}_0)^2) \mid \mathbf{X}_{t_{k-1}}] \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0$.

The proof of the previous Lemma is technical and is shown in Supplementary Material (Pilipovic et al., 2023).

8 Conclusion

We proposed two new estimators for nonlinear multivariate SDEs. These estimators 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 discretized 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 conducted with the stochastic Lorenz system, we the superior performance of the S estimators. We compared seven 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 property makes the S estimator a more appropriate choice for large sample sizes. The LT, EM, and K estimators perform well for small h, but for larger h the bias increases, especially for the diffusion parameters in the EM case.

While the proposed estimators are versatile, they come with certain limitations. These include assumptions like the presence of 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. We will thoroughly investigate these cases in another paper with more involved proofs.

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)"; Novo Nordisk Foundation NNF20OC0062958; and Independent Research Fund Denmark | Natural Sciences 9040-00215B.

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. doi:10.1137/140962644. URL `https:`

//doi.org/10.1137/140962644.

M. Ableidinger and E. Buckwar. Splitting Integrators for the Stochastic Landau–Lifshitz Equation. *SIAM Journal on Scientific Computing*, 38(3):A1788–A1806, 2016. doi:10.1137/15M103529X. URL https://doi.org/10.1137/15M103529X.

M. Ableidinger, E. Buckwar, and H. Hinterleitner. A Stochastic Version of the Jansen and Rit Neural Mass Model: Analysis and Numerics. *The Journal of Mathematical Neuroscience*, 7, 08 2017. doi:10.1186/s13408-017-0046-4.

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.

A. Alamo and J. M. Sanz-Serna. A Technique for Studying Strong and Weak Local Errors of Splitting Stochastic Integrators. *SIAM Journal on Numerical Analysis*, 54(6):3239–3257, 2016. doi:10.1137/16M1058765. URL https://doi.org/10.1137/16M1058765.

L. A. Alyushina. Euler Polygonal Lines for Itô Equations with Monotone Coefficients. *Theory of Probability & Its Applications*, 32(2):340–345, 1988. doi:10.1137/1132046. URL https://doi.org/10.1137/1132046.

N. Ann, D. Pebrianti, M. Abas, and L. Bayuaji. *Parameter Estimation of Lorenz Attractor: A Combined Deep Neural Network and K-Means Clustering Approach*, volume 730, pages 321–331. Springer, Singapore, 01 2022. ISBN 978-981-33-4596-6. doi:10.1007/978-981-33-4597-3_30.

M. Arnst, G. Louppe, R. Van Hulle, L. Gillet, F. Bureau, and V. Denoël. 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. *Mathematical Biosciences*, 347:108805, 2022. ISSN 0025-5564. doi:https://doi.org/10.1016/j.mbs.2022.108805. URL https://www.sciencedirect.com/science/article/pii/S0025556422000219.

V. Barbu. A Product Formula Approach to Nonlinear Optimal Control Problems. *SIAM Journal on Control and Optimization*, 26(3):497–520, 1988. doi:10.1137/0326030. URL https://doi.org/10.1137/0326030.

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic Differentiation in Machine Learning: A Survey. *J. Mach. Learn. Res.*, 18(1):5595–5637, jan 2017. ISSN 1532-4435.

A. Bensoussan, R. Glowinski, and A. Răşcanu. Approximation of Some Stochastic Differential Equations by the Splitting Up Method. *Applied Mathematics and Optimization*, 25:81–106, 1992.

B. Bibby and M. Sørensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1 (1/2):17–39, 1995.

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.

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, 2010. doi:10.1137/090758842. URL https://doi.org/10.1137/090758842.

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. ISSN 1531-3492. doi:10.3934/dcdsb.2019077. URL /article/id/e1b6175e-c608-4553-a1fd-da01425e6811.

C.-E. Bréhier, D. Cohen, and J. Ulander. Analysis of a Positivity-preserving Splitting Scheme for Some Nonlinear Stochastic Heat Equations, 2023.

E. Buckwar, M. Tamborrino, and I. Tubikanec. Spectral density-based and measure-preserving ABC for partially observed diffusion processes. An illustration on Hamiltonian SDEs. *Statistics and Computing*, 30, 05 2020. doi:10.1007/s11222-019-09909-6.

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. ISSN 0168-9274. doi:https://doi.org/10.1016/j.apnum.2022.04.018. URL https://www.sciencedirect.com/science/article/pii/S0168927422001118.

J. Chang and S. X. Chen. On the approximate maximum likelihood estimation for diffusion processes. *The Annals of Statistics*, 39(6):2820 – 2851, 2011. doi:10.1214/11-AOS922. URL https://doi.org/10.1214/11-AOS922.

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.

N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in statistics. Springer Cham, 2020. doi:https://doi.org/10.1007/978-3-030-47845-2.

I. Crimaldi and L. Pratelli. Convergence results for multivariate martingales. *Stochastic Processes and their Applications*, 115(4):571–577, 2005. ISSN 0304-4149. doi:https://doi.org/10.1016/j.spa.2004.10.004. URL https://www.sciencedirect.com/science/article/pii/S030441490400167X.

D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986. doi:10.1080/17442508608833428. URL https://doi.org/10.1080/17442508608833428.

S. Dipple, A. Choudhary, J. Flamino, B. Szymanski, and G. Korniss. Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities. *Applied Network Science*, 5, 03 2020. doi:10.1007/s41109-020-00259-1.

P. Ditlevsen and S. Ditlevsen. Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications*, 14:4254, 2023. URL https://doi.org/10.1038/s41467-023-39810-w.

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. ISSN 1369-7412. doi:10.1111/rssb.12307.

S. Ditlevsen and M. Sørensen. Inference for observations of integrated diffusion processes. *Scandinavian Journal of Statistics*, 31(3):417–429, 2004. ISSN 0303-6898. doi:10.1111/j.1467-9469.2004.02_023.x.

S. Ditlevsen, M. Tamborrino, and I. Tubikanec. Network inference in a stochastic multi-population neural mass model via approximate bayesian computation. *ArXiv*, page 2306.15787, 2023.

G. Dohnal. On Estimating the Diffusion Coefficient. *Journal of Applied Probability*, 24(1):105–114, 1987. ISSN 00219002. URL http://www.jstor.org/stable/3214063.

P. Dubois, T. Gomez, L. Planckaert, and L. Perret. Data-driven predictions of the Lorenz system. *Physica D Nonlinear Phenomena*, 408:132495, 6 2020. doi:10.1016/j.physd.2020.132495.

D. Falbel and J. Luraschi. *torch: Tensors and Neural Networks with 'GPU' Acceleration*, 2022. https://torch.mlverse.org/docs, https://github.com/mlverse/torch.

D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989. doi:10.1080/02331888908802205. URL https://doi.org/10.1080/02331888908802205.

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.

C. Fuchs. *Inference for Diffusion Processes with Applications in Life Sciences*. Springer Berlin, Heidelberg, 01 2013. ISBN 978-3-642-25968-5 (Print) 978-3-642-25969-2 (Online). doi:10.1007/978-3-642-25969-2.

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. URL http://www.numdam.org/item/AIHPB_1993__29_1_119_0/.

P. Gloaguen, M.-P. Etienne, and S. Le Corff. Stochastic differential equation based on a multimodal potential to model movement data in ecology. *Journal of the Royal Statistical Society: Series C Applied Statistics*, 67(3), Apr 2018. URL https://hal.archives-ouvertes.fr/hal-01207001.

A. Gloter. Parameter Estimation for a Discretely Observed Integrated Diffusion Process. *Scandinavian Journal of Statistics*, 33(1):83–104, 2006. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616910.

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. doi:10.1214/20-EJS1777. URL https://doi.org/10.1214/20-EJS1777.

E. Gobet. LAN property for ergodic diffusions with discrete observations. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(5):711–737, 2002. ISSN 0246-0203. doi:https://doi.org/10.1016/S0246-0203(02)01107-X. URL https://www.sciencedirect.com/science/article/pii/S024602030201107X.

W. Gu, H. Wu, and H. Xue. *Parameter Estimation for Multivariate Nonlinear Stochastic Differential Equation Models: A Comparison Study*, pages 245–258. Springer International Publishing, Cham, 2020. ISBN 978-3-030-34675-1. doi:10.1007/978-3-030-34675-1_13. URL https://doi.org/10.1007/978-3-030-34675-1_13.

E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg, 1993. ISBN 0387566708.

P. Hall and C. Heyde. *Martingale Limit Theory and Its Application*. Probability and mathematical statistics. Academic Press, 1980. ISBN 9781483240244. URL https://books.google.dk/books?id=wdLajgEACAAJ.

R. Hilborn and A. Hilborn. *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*. Oxford scholarship online: Physics module. Oxford University Press, 2000. ISBN 9780198507239. URL https://books.google.lu/books?id=fwybfh-nIyEC.

W. E. J. Hopkins and W. S. Wong. Lie-Trotter Product Formulas for Nonlinear Filtering. *Stochastics*, 17(4):313–337, 1986. doi:10.1080/17442508608833395. URL https://doi.org/10.1080/17442508608833395.

A. R. Humphries and A. M. Stuart. Runge–Kutta Methods for Dissipative and Gradient Dynamical Systems. *SIAM Journal on Numerical Analysis*, 31(5):1452–1485, 1994. doi:10.1137/0731075. URL https://doi.org/10.1137/0731075.

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. S. Hurn, J. I. Jeisman, and K. A. Lindsay. Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations. *Journal of Financial Econometrics*, 5(3):390–455, 06 2007. ISSN 1479-8409. doi:10.1093/jjfinec/nbm009. URL https://doi.org/10.1093/jjfinec/nbm009.

M. Hutzenthaler, A. Jentzen, and P. E. 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(2130):1563–1576, 2011. doi:10.1098/rspa.2010.0348. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2010.0348.

Y. Iguchi, A. Beskos, and M. M. Graham. Parameter Estimation with Increased Precision for Elliptic and Hypo-elliptic Diffusions. *ArXiv*, page 2211.16384, 2022.

J. Jimenez, I. Shoji, and T. Ozaki. Simulation of Stochastic Differential Equations Through the Local Linearization Method. A Comparative Study. *Journal of Statistical Physics*, 94:587–602, 02 1999. doi:10.1023/A:1004504506041.

J. Jimenez, C. Mora, and M. Selva. A weak Local Linearization scheme for stochastic differential equations with multiplicative noise. *Journal of Computational and Applied Mathematics*, 313:202–217, 2017. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2016.09.013. URL https://www.sciencedirect.com/science/article/pii/S0377042716304332.

A. M. Kareem and S. N. Al-Azzawi. A Stochastic Differential Equations Model for Internal COVID-19 Dynamics. *Journal of Physics: Conference Series*, 1818(1):012121, mar 2021. doi:10.1088/1742-6596/1818/1/012121. URL https://doi.org/10.1088/1742-6596/1818/1/012121.

H. Keller. *Attractors and Bifurcations of the Stochastic Lorenz System*. Technical Report. Institut für Dynamische syteme, Universität Bremen, 1996.

M. Kessler. Estimation of an Ergodic Diffusion from Discrete Observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616449.

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.

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. doi:10.1137/1135082. URL https://doi.org/10.1137/1135082.

A. Kumar. Expectation of Product of Quadratic Forms. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 35(3):359–362, 1973. ISSN 05815738. URL http://www.jstor.org/stable/25051849.

J. A. Lazzús, M. Rivera, and C. H. López-Caraballo. Parameter estimation of Lorenz chaotic system using a hybrid swarm intelligence algorithm. *Physics Letters A*, 380(11):1164–1171, 2016. ISSN 0375-9601. doi:https://doi.org/10.1016/j.physleta.2016.01.040. URL https://www.sciencedirect.com/science/article/pii/S0375960116000839.

B. Leimkuhler and C. Matthews. Molecular dynamics. *Interdisciplinary applied mathematics*, 39:443, 2015.

C. Li. Maximum-likelihood estimation for diffusion processes via closed-form density expansions. *The Annals of Statistics*, 41(3):1350 – 1380, 2013. doi:10.1214/13-AOS1118. URL https://doi.org/10.1214/13-AOS1118.

E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences*, 20(2):130 – 141, 1963. doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

A. López-Pérez, M. Febrero-Bande, and W. González-Manteiga. Parametric Estimation of Diffusion Processes: A Review and Comparative Study. *Mathematics*, 9(8), 2021. ISSN 2227-7390. doi:10.3390/math9080859. URL https://www.mdpi.com/2227-7390/9/8/859.

X. Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

R. I. McLachlan and G. R. W. Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002. URL www.scopus.com. Cited By :491.

D. L. McLeish. Dependent Central Limit Theorems and Invariance Principles. *The Annals of Probability*, 2(4):620–628, 1974. ISSN 00911798. URL http://www.jstor.org/stable/2959412.

T. Michelot, P. Gloaguen, P. Blackwell, and M.-P. Etienne. The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods in Ecology and Evolution*, 10, 08 2019. doi:10.1111/2041-210x.13275.

T. Michelot, R. Glennie, C. Harris, and L. Thomas. Varying-Coefficient Stochastic Differential Equations with Applications in Ecology. *Journal of Agricultural, Biological and Environmental Statistics*, 26, 03 2021. doi:10.1007/s13253-021-00450-6.

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.

G. N. Milstein and M. V. Tretyakov. Quasi-symplectic methods for Langevin-type equations. *IMA Journal of Numerical Analysis*, 23(4):593–626, 10 2003. ISSN 0272-4979. doi:10.1093/imanum/23.4.593. URL https://doi.org/10.1093/imanum/23.4.593.

T. Misawa. A lie algebraic approach to numerical integration of stochastic differential equations. *SIAM Journal on Scientific Computing*, 23(3):866–890, 2001.

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. A Bridge between Nonlinear Time Series Models and Nonlinear Stochastic Dynamical Systems: A Local Linearization Approach. *Statistica Sinica*, 2(1):113–135, 1992. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24304122.

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. doi:https://doi.org/10.1111/1467-9892.00189. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00189.

U. Picchini and S. Ditlevsen. Practical estimation of high dimensional stochastic differential mixed-effects models. *Computational Statistics & Data Analysis*, 55(3):1426–1444, 2011. ISSN 0167-9473. doi:https://doi.org/10.1016/j.csda.2010.10.003. URL https://www.sciencedirect.com/science/article/pii/S0167947310003774.

P. Pilipovic, A. Samson, and S. Ditlevsen. Supplement to "Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations Based on Splitting Schemes". 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

M. Riedmiller and H. Braun. RPROP - A Fast Adaptive Learning Algorithm. Technical report, Proc. of ISCIS VII), Universitat, 1992.

I. Shoji. Approximation of Continuous Time Stochastic Processes by a Local Linearization Method. *Mathematics of Computation*, 67(221):287–298, 1998. ISSN 00255718, 10886842. URL http://www.jstor.org/stable/2584984.

I. Shoji. A note on convergence rate of a linearization method for the discretization of stochastic differential equations. *Communications in Nonlinear Science and Numerical Simulation*, 16(7):2667–2671, 2011. ISSN 1007-5704. doi:https://doi.org/10.1016/j.cnsns.2010.09.008. URL https://www.sciencedirect.com/science/article/pii/S1007570410004806.

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. doi:10.1080/07362999808809559. URL https://doi.org/10.1080/07362999808809559.

M. Sørensen and M. Uchida. Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli*, 9 (6):1051 – 1069, 2003. ISSN 1350-7265.

M. Sørensen. *Estimating functions for diffusion-type processes*, chapter 1, pages 1–97. Chapman and Hall/CRC, 05 2012. ISBN 978-1-4398-4940-8. doi:10.1201/b12126-2.

M. Tabor. *Chaos and Integrability in Nonlinear Dynamics: An Introduction*. Wiley, 1989. ISBN 9780471827283. URL https://books.google.de/books?id=TkvvAAAAMAAJ.

Y. Tian and M. Fan. Nonlinear integral inequality with power and its application in delay integro-differential equations. *Advances in Difference Equations*, 2020, 03 2020. doi:10.1186/s13662-020-02596-y.

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. doi:10.1137/120902318. 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. ISSN 0304-4149. doi:https://doi.org/10.1016/j.spa.2012.04.001. URL https://www.sciencedirect.com/science/article/pii/S0304414912000622.

P. Vatiwutipong and N. Phewchean. Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations*, 2019:1–7, 2019.

N. Yang, N. Chen, and X. Wan. A new delta expansion for multivariate diffusions via the Itô-Taylor expansion. *Journal of Econometrics*, 209(2):256–288, 2019. doi:10.1016/j.jeconom.2019.01. URL https://ideas.repec.org/a/eee/econom/v209y2019i2p256-288.html.

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. doi:10.1007/BF00050834. URL https://ideas.repec.org/a/spr/aistmt/v42y1990i2p221-251.html.

L. Zhuang, L. Cao, Y. Wu, Y. Zhong, L. Zhangzhong, W. Zheng, and L. Wang. Parameter Estimation of Lorenz Chaotic System Based on a Hybrid Jaya-Powell Algorithm. *IEEE Access*, 8:20514–20522, 2020. doi:10.1109/ACCESS.2020.2968106.

S1 Supplementary Material

This section provides proofs for all propositions, lemmas, and theorems. References to equations and sections that do not begin with "S" refer to the main paper. 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 $\mathbb{E}_{\boldsymbol{\theta}}$.

In Section S1.1, we provide the proof for the Lie-Trotter splitting, while Section S1.2 contains the proofs for the Strang splitting. Additionally, the proofs of moment bounds are detailed in Section S1.3. These three sections rely on manipulating the Itô-Taylor expansion. The foundational properties necessary for subsequent proofs are outlined in Section S1.4. These properties encompass Grönwall's and Rosenthal's inequalities, as well as Central Limit Theorems for a sum of triangular arrays. The proof for Lemma 7.1 can be found in Section S1.5, while the section addressing asymptotic normality is presented in Section S1.6.

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 Lie-Trotter 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:

$$\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}((\mathbf{x})) = \mathbf{x} + h(\mathbf{A}(\mathbf{x}-\mathbf{b}) + \mathbf{N}(\mathbf{x})) + \frac{h^2}{2}(\mathbf{A}^2(\mathbf{x}-\mathbf{b}) + 2\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x})) + \mathbf{R}(h^3, \mathbf{x}). \tag{S1}$$

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.

(SA) $\sum_{i=1}^d \sum_{j=1}^d [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{ij} \partial_{ij}^2 F^{(i)}(\mathbf{x}) = 0, \quad \text{for all } k = 1, \dots, d.$

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 splitting does not affect the one-step convergence order. To demonstrate this, we define the reversed Lie-Trotter splitting:

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

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:

$$\mathbb{E}[g(\mathbf{X}_{t_k}^{[1]}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}] = g(\mathbf{x}) + hL_{[1]}g(\mathbf{x}) + \frac{h^2}{2}L_{[1]}^2 g(\mathbf{x}) + R(h^3, \mathbf{x}). \tag{S2}$$

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 $\partial_{jk}^2 f_h^{(i)}(\mathbf{x}) = h\partial_{jk}^2 N^{(i)}(\mathbf{x}) + R(h^2, \mathbf{x})$. 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 $L_{[1]}^2 f_h^{(i)}(\mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla(\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + R(h, \mathbf{x}) = (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \mathbf{A}^{(i)} + R(h, \mathbf{x})$. Thus,

$$\begin{aligned}
&\mathbb{E}[f_h^{(i)}(\mathbf{X}_{t_{k-1}}^{[1]}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}] = x^{(i)} + hN^{(i)}(\mathbf{x}) + \frac{h^2}{2}(\mathbf{N}(\mathbf{x}))^\top \nabla N^{(i)}(\mathbf{x}) \\
&+ h(\mathbf{A}(\mathbf{x}-\mathbf{b}))^{(i)} + h^2(\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla N^{(i)}(\mathbf{x}) + \frac{h^2}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})) + \frac{h^2}{2}(\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \mathbf{A}^{(i)} + R(h^3, \mathbf{x}) \\
&= x^{(i)} + hF^{(i)}(\mathbf{x}) + \frac{h^2}{2}((\mathbf{F}(\mathbf{x}))^\top (\nabla N^{(i)}(\mathbf{x})) + (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla F^{(i)}(\mathbf{x}) + \operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x}))) + R(h^3, \mathbf{x}).
\end{aligned} \tag{S3}$$

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:

$$\begin{aligned}\mathbb{E}[X_{t_k}^{(i)} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] &= x^{(i)} + hF^{(i)}(\mathbf{x}) \\ &\quad + \frac{h^2}{2}((\mathbf{N}(\mathbf{x}))^\top \nabla F^{(i)}(\mathbf{x}) + (\mathbf{A}(\mathbf{x}-\mathbf{b}))^\top \nabla F^{(i)}(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x}))) + R(h^3, \mathbf{x}).\end{aligned}$$

The final equation coincides with equation (S3) only up to order $R(h, \mathbf{x})$. Despite the reversed LT splitting having 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 $i = 1, \ldots, d$. Given Assumption (SA), the condition for achieving a higher one-step convergence order remains equivalent to the case of the original LT splitting.

S1.2 Proof for the Strang Splitting

We continue employing the Taylor expansion to establish the numerical properties of the Strang splitting 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 expends $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x}))$ 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. $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3, \mathbf{x})$.
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 $\boldsymbol{f}_{h/2}(\mathbf{x}) - \boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) = h\mathbf{N}(\mathbf{x}) + \mathbf{R}(h^3, \mathbf{x})$. This leads to: $\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{x})) = \boldsymbol{f}_{h/2}^{-1}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathbf{R}(h^3, \mathbf{x})$.

Proof of Proposition 3.6 We begin by introducing a new function of $\mathbf{x}$, arising from the third property of Lemma S1.1:

$$\mathbf{Q}_h(\mathbf{x}) \coloneqq \frac{h}{2}(2\mathbf{A}(\mathbf{x}-\mathbf{b}) + \mathbf{N}(\mathbf{x})) + \frac{h^2}{8}(4\mathbf{A}^2(\mathbf{x}-\mathbf{b}) + 4\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x})).$$

Then, for a generic random vector $\mathbf{X}$ we use Proposition 2.2 and Lemma S1.1 to write:

$$\begin{aligned}\boldsymbol{f}_{h/2}(\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X})) + \boldsymbol{\xi}_h) &= \boldsymbol{f}_{h/2}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \mathbf{R}(h^3, \mathbf{X})) \\ &= \mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \frac{h}{2}\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \\ &\quad + \frac{h^2}{8}(D\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h))\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) + \mathbf{R}(h^3, \mathbf{X}). \qquad \text{(S4)}\end{aligned}$$

Consequently, we expand:

$$\begin{aligned}\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) &= \mathbf{N}(\mathbf{X}) + (D\mathbf{N}(\mathbf{X}))(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \\ &\quad + \frac{1}{2}[(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)]_{i=1}^d + \mathbf{R}(h^2, \mathbf{X}). \qquad \text{(S5)}\end{aligned}$$

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,

$$\mathbb{E}[\mathbf{N}(\mathbf{X}+\mathbf{Q}_h(\mathbf{X})+\boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}] = \mathbf{N}(\mathbf{x})+(D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x})+\frac{1}{2}[\mathbb{E}[\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h \mid \mathbf{X} = \mathbf{x}]]_{i=1}^d+\mathbf{R}(h^2, \mathbf{x}). \quad \text{(S6)}$$

Lastly, we compute:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h \mid \mathbf{X}=\mathbf{x}] &= \mathbb{E}[\mathrm{tr}(\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h) \mid \mathbf{X}=\mathbf{x}] = \mathrm{tr}(\mathbf{H}_{N^{(i)}}(\mathbf{X})\mathbb{E}[\boldsymbol{\xi}_h\boldsymbol{\xi}_h^\top]) \\ &= \sum_{j,k=1}^{d} \partial^2_{jk} N^{(i)}(\mathbf{x})[\mathrm{var}(\boldsymbol{\xi}_h)]_{jk} = \sum_{j,k=1}^{d} \partial^2_{jk} F^{(i)}(\mathbf{x})[\boldsymbol{\Omega}_h]_{jk}.\end{aligned}$$

We use the approximation of the variance of the random vector $\boldsymbol{\xi}_h$ to get $\mathbb{E}[\mathbf{N}(\mathbf{X}+\mathbf{Q}_h(\mathbf{X})+\boldsymbol{\xi}_h) \mid \mathbf{X}=\mathbf{x}] = \mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x}) + \frac{h}{2}[\sum_{j,k=1}^{d}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{jk}\partial^2_{jk}F^{(i)}(\mathbf{x})]_{i=1}^{d} + \mathbf{R}(h^2,\mathbf{x})$. Taking the expectation of (S4) and incorporating the previous equation completes the proof.

S1.3 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:*

$$L(\alpha(\mathbf{x})\beta(\mathbf{x})) = \alpha(\mathbf{x})L\beta(\mathbf{x}) + \beta(\mathbf{x})L\alpha(\mathbf{x}) + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x}))).$$

Proof We use the generator L and the product rule to get:

$$\begin{aligned}L(\alpha(\mathbf{x})\beta(\mathbf{x})) &= \mathbf{F}(\mathbf{x})^\top\alpha(\mathbf{x})\nabla\beta(\mathbf{x}) + \mathbf{F}(\mathbf{x})^\top\beta(\mathbf{x})\nabla\alpha(\mathbf{x}) + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\alpha(\mathbf{x})\mathbf{H}_\beta(\mathbf{x}) + \beta(\mathbf{x})\mathbf{H}_\alpha(\mathbf{x})))\\ &\quad + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x})))\\ &= \alpha(\mathbf{x})L\beta(\mathbf{x}) + \beta(\mathbf{x})L\alpha(\mathbf{x}) + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top(\nabla\alpha(\mathbf{x})\nabla^\top\beta(\mathbf{x}) + \nabla\beta(\mathbf{x})\nabla^\top\alpha(\mathbf{x}))).\end{aligned}$$

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
$D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0) = \mathbf{I} + \mathbf{R}(h,\mathbf{x})$.

S1.4 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 S1.3, 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 Jacod (1993) provides conditions for the convergence of a sum of a triangular array and is recalled as Lemma S1.4.

Lemmas S1.5 and S1.6 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 S1.5 might not suffice, warranting the use of Lemma S1.6. Theorem S1.7 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 S1.8 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 S1.3 (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

$$\kappa(t) \coloneqq 2^{p-1}b\int_0^t a^p(s)\,\mathrm{d}s. \tag{S7}$$

Lemma S1.4 (Lemma 9 in Genon-Catalot and Jacod (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 S1.5 (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 S1.6 (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 $p\ge l>d$ and $C>0$ such that for all $\boldsymbol{\theta},\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, it holds:*

(1) $\mathbb{E}[|\xi_N(\boldsymbol{\theta}_1)-\xi_N(\boldsymbol{\theta}_2)|^p]\le C\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^l$;

(2) $\mathbb{E}[|\xi_N(\boldsymbol{\theta})|^p]\le C$;

(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 S1.7 (Rosenthal's inequality, Theorem 2.12 in Hall and Heyde (1980)) *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:*

$$S_N=\sum_{k=1}^{N}X_k^N,\ \ N\in\mathbb{N}$$

be a martingale array. Then, for all $p\in[2,\infty)$ there exist constants C_1,C_2 such that:

$$C_1(\mathbb{E}[(\sum_{k=1}^{N}\mathbb{E}[(X_k^N)^2\mid\mathcal{G}_{k-1}^N])^{\frac{p}{2}}]+\sum_{k=1}^{N}\mathbb{E}[|X_k^N|^p])\le\mathbb{E}[|S_N|^p]\le C_2(\mathbb{E}[(\sum_{k=1}^{N}\mathbb{E}[(X_k^N)^2\mid\mathcal{G}_{k-1}^N])^{\frac{p}{2}}]+\sum_{k=1}^{N}\mathbb{E}[|X_k^N|^p]).$$

Theorem S1.8 (Proposition 3.1. in Crimaldi and Pratelli (2005)) *Let $(\mathbf{X}_{N,k})_{N\in\mathbb{N},1\le k\le N}$ be a triangular array of d-dimensional random vectors, such that, for each N, the finite sequence $(\mathbf{X}_{N,k})_{1\le k\le N}$ is a martingale difference array with respect to a given filtration $(\mathcal{G}_k^N)_{1\le k\le N}$ such that:*

$$\mathbf{S}_N=\sum_{k=1}^{N}\mathbf{X}_{N,k},\ N\in\mathbf{N}.$$

If

(1) $\mathbb{E}[\sup_{1\le k\le N}\|\mathbf{X}_{N,k}\|_1] \xrightarrow[N\to\infty]{} 0$;

(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 Instead of using the second condition of Theorem S1.8, Lemma S1.6 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 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 $(\mathbf{X}_{N,k})_{N\in\mathbb{N},1\le k\le N}$:

$$\sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^{N}] \xrightarrow[N\to\infty]{\mathbb{P}} 0, \qquad \sum_{k=1}^{N} \mathbb{E}[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^{N}]\mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^{N}] \xrightarrow[N\to\infty]{\mathbb{P}} 0. \tag{S8}$$

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:

$$|\sum_{k=1}^{N} X_{N,k}^{(i)} \mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^{N}]| \le \sum_{k=1}^{N} |X_{N,k}^{(i)}||\mathbb{E}[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^{N}]| \le (\sum_{k=1}^{N} (X_{N,k}^{(i)})^2 \sum_{k=1}^{N} \mathbb{E}^2[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^{N}])^{\frac{1}{2}} \xrightarrow[N\to\infty]{} 0.$$

Theorem S1.8 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 (S8), gives $\mathbf{S}_N \xrightarrow[N\to\infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$.

S1.5 Proof of Lemma 7.1

Lemma 7.1 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 7.1 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 S1.4. On the other hand, for proving uniform convergence, we utilize both Lemma S1.5 and Lemma S1.6. Throughout the proof of Lemma 7.1, 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 7.1 Proof of 1. As previously discussed, we introduce a martingale array that corresponds to the limit outlined in point 1. We then utilize Lemma S1.4 to facilitate our analysis. We denote $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}) \coloneqq \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$. 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 S1.4, 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 S1.4 $\frac{1}{Nh}\sum_{k=1}^N \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$, for $Nh \to \infty$, $h \to 0$. To establish the uniformity of the limits with respect to $\boldsymbol{\varsigma}$, we turn to Lemma S1.5 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:

$$\begin{aligned}&\sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{\varsigma_j\in\Theta_{\varsigma_j}} |\partial_{\varsigma_j}\frac{1}{Nh}\sum_{k=1}^N \mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}|] \le \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \sup_{\varsigma_j\in\Theta_{\varsigma_j}} |\partial_{\varsigma_j}\operatorname{Tr}(\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k})|] \\ &\le \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|] \\ &\le \sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|^2\|\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\|] \le C\sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top)] \\ &= C\sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top) \mid \mathbf{X}_{t_{k-1}}]] = C\sup_{N\in\mathbb{N}} \mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathbf{R}(h^2,\mathbf{X}_{t_{k-1}}))] < \infty.\end{aligned}$$

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 $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$, 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:

$$\begin{aligned}\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}] &= \frac{1}{N}\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[\mathrm{Tr}(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})) \mid \mathbf{X}_{t_{k-1}}] \\ &= \frac{1}{N}\sum_{k=1}^N \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}]) \\ &= \frac{1}{N}\sum_{k=1}^N R(h^3, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,\end{aligned}$$

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 S1.4 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 S1.6. 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:

$$\begin{aligned}
&\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}] \qquad \text{(S18)}\\
&= \frac{1}{N^2 h^2} \sum_{k=1}^{N} \mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}] (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})\\
&= \frac{1}{N} \sum_{k=1}^{N} R(\frac{1}{Nh}, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0,
\end{aligned}$$

for $Nh \to \infty$. To prove uniform convergence, we use Lemma S1.6 along with Rosenthal's inequality from Theorem S1.7, 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:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[|\sum_{k=1}^{N} (Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2))|^p]\\
&\leq C\mathbb{E}[(\sum_{k=1}^{N} \mathbb{E}[(Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2))^2 \mid \mathbf{X}_{t_{k-1}}])^{p/2}] + C\sum_{k=1}^{N} \mathbb{E}[|(Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1) - Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2))|^p]. \qquad \text{(S19)}
\end{aligned}$$

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:

$$\begin{aligned}
&\frac{1}{N^2} \sum_{k=1}^{N} \mathbb{E}[\mathrm{Tr}((\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta}))^2) \mid \mathbf{X}_{t_{k-1}}]\\
&\leq \frac{1}{N^2} \sum_{k=1}^{N} \mathbb{E}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^2 \|\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})\|^2 \mid \mathbf{X}_{t_{k-1}}] \,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1})\|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\|\\
&\leq \frac{C}{N^2} \sum_{k=1}^{N} (\mathbb{E}[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^4 \mid \mathbf{X}_{t_{k-1}}]\mathbb{E}[\|\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})\|^4 \mid \mathbf{X}_{t_{k-1}}])^{\frac{1}{2}} = \frac{1}{N} \sum_{k=1}^{N} R(h/N, \mathbf{X}_{t_{k-1}}) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}} 0, \qquad \text{(S20)}
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. To prove the uniform convergence, we use Lemma S1.6. 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 $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta})$ 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:

$$\begin{aligned}
&\mathbf{g}_1(\boldsymbol{\beta}_1)^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_1) - \mathbf{g}_1(\boldsymbol{\beta}_2)^\top(\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_2)\\
&\quad= (\mathbf{g}_1(\boldsymbol{\beta}_1) - \mathbf{g}_1(\boldsymbol{\beta}_2))^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}\mathbf{g}_2(\boldsymbol{\beta}_1) + \mathbf{g}_1(\boldsymbol{\beta}_2)^\top(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1}(\mathbf{g}_2(\boldsymbol{\beta}_1) - \mathbf{g}_2(\boldsymbol{\beta}_2))\\
&\quad+ \mathbf{g}_1(\boldsymbol{\beta}_2)^\top((\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top)^{-1} - (\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top)^{-1})\mathbf{g}_2(\boldsymbol{\beta}_2).
\end{aligned}$$

S1.6 Proof of the Asymptotic Normality

In this section, we distinguish between the true parameter $\boldsymbol{\theta}_0$ and a generic parameter $\boldsymbol{\theta}$. To complete the proof of asymptotic normality, we need to prove Lemma 7.4. The proof of this lemma 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.

Proof of Lemma 7.4 We begin by expanding $\eta_k^{(i)}$ to differentiate between terms that vanish and those that do not in the limits:

$$\begin{aligned}
\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0) &= \frac{2}{\sqrt{Nh}}\mathrm{Tr}((\mathbf{I} + \frac{h}{2}D\mathbf{N}_0(\mathbf{X}_{t_k}))(-\frac{h}{2}D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})))\\
&\quad- \frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(-\frac{h}{2}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}) + \frac{h^2}{8}\partial_{\beta_i}(D\mathbf{N}_0(\mathbf{X}_{t_k}))\mathbf{N}_0(\mathbf{X}_{t_k}))\\
&\quad+ \frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0) + R(\sqrt{h^3/N}, \mathbf{X}_{t_{k-1}})\\
&= -\sqrt{\frac{h}{N}}\,\mathrm{Tr}(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})) + \frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\\
&\quad- \frac{1}{4}\sqrt{\frac{h}{N}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}(D\mathbf{N}_0(\mathbf{X}_{t_k}))\mathbf{N}_0(\mathbf{X}_{t_k})\\
&\quad+ \frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0) + R(\sqrt{h^3/N}, \mathbf{X}_{t_{k-1}}).
\end{aligned} \tag{S21}$$

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 (S21), 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 $\sup_{1\le k\le N}|\zeta^{(j)}_{N,k}(\boldsymbol{\theta}_0)|$. The leading term in $\zeta^{(j)}_{N,k}(\boldsymbol{\theta}_0)$, 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 $\eta^{(i)}_{N,k}$ and $\zeta^{(i)}_{N,k}$. By analyzing (S21), we can deduce that $\mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}})$, since Proposition 4.3 gives:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\mid\mathbf{X}_{t_{k-1}}]=\sqrt{\frac{h}{N}}\,\mathrm{Tr}(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k}))+R(\sqrt{h^3/N},\mathbf{X}_{t_{k-1}}),$$

Similarly, from:

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\mathrm{Tr}(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1})\mid\mathbf{X}_{t_{k-1}}]=h\,\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)+R(h^2,\mathbf{X}_{t_{k-1}})$$

we conclude that $\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(i)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]=R(h/\sqrt{N},\mathbf{X}_{t_{k-1}})$. Then, combining the previous, we get:

$$\begin{aligned}
\sum_{k=1}^N\mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]&=R(\sqrt{Nh^3},\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,\\
\sum_{k=1}^N\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]&=R(\sqrt{Nh^2},\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,\\
\sum_{k=1}^N\mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]&=R(h^3,\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,\\
\sum_{k=1}^N\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j_1)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j_2)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]&=R(h^2,\mathbf{X}_{t_{k-1}})\xrightarrow{\mathbb{P}_{\boldsymbol{\theta}_0}}0,\\
\sum_{k=1}^N\mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j)}_{N,k}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}]&=R(h^{5/2},\mathbf{X}_{t_{k-1}})\xrightarrow[N\to\infty]{\mathbb{P}_{\boldsymbol{\theta}_0}}0.
\end{aligned}$$

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 $\eta^{(i_2)}_{N,k}$:

$$\begin{aligned}
\eta^{(i)}_{N,k}(\boldsymbol{\theta}_0)&=\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\\
&\quad+\frac{2}{h\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0);\boldsymbol{\beta}_0)+R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}})\\
&=\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})+\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}(\mathbf{N}_0(\mathbf{X}_{t_{k-1}})\\
&\quad+2\mathbf{A}_0(\mathbf{X}_{t_{k-1}}-\mathbf{b}_0))+R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}})\\
&=\frac{2}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_i}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})+\frac{1}{\sqrt{Nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^i_{k,k-1}(\boldsymbol{\beta}_0)+R(\sqrt{\frac{h}{N}},\mathbf{X}_{t_{k-1}}),
\end{aligned}$$

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 $1/Nh$:

$$\begin{aligned}
\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0) = &\frac{4}{Nh}\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k}\\
&+\frac{2}{Nh}\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k}\\
&+\frac{2}{Nh}\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\boldsymbol{\psi}^{i_2}_{k,k-1}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k}\\
&+\frac{1}{Nh}\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\boldsymbol{\psi}^{i_2}_{k,k-1}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} + R(1/N, \mathbf{X}_{t_{k-1}}).
\end{aligned}$$

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:

$$\begin{aligned}
&\frac{1}{Nh}|\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}]|\\
&\leq \frac{1}{Nh}|\operatorname{Tr}(\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1})|\|(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\|\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}\mathbf{Z}^\top_{t_k}\|\|\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\| \mid \mathbf{X}_{t_{k-1}}]\\
&\leq \frac{C}{Nh}(\mathbb{E}_{\boldsymbol{\theta}_0}[\|\mathbf{Z}_{t_k}\mathbf{Z}^\top_{t_k}\|^2 \mid \mathbf{X}_{t_{k-1}}]\mathbb{E}_{\boldsymbol{\theta}_0}[\|\boldsymbol{\psi}^{i_1}_{k,k-1}(\boldsymbol{\beta}_0)\|^2 \mid \mathbf{X}_{t_{k-1}}])^{\frac{1}{2}}\\
&= \frac{1}{Nh}(R(h^2, \mathbf{X}_{t_{k-1}})R(h, \mathbf{X}_{t_{k-1}}))^{\frac{1}{2}} = R(\sqrt{h}/N, \mathbf{X}_{t_{k-1}}).
\end{aligned}$$

Finally, we use Lemma 4.2 to get:

$$\begin{aligned}
&\sum_{k=1}^N \mathbb{E}_{\boldsymbol{\theta}_0}[\eta^{(i_1)}_{N,k}(\boldsymbol{\theta}_0)\eta^{(i_2)}_{N,k}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}]\\
&= \frac{4}{Nh}\sum_{k=1}^N(\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})\partial_{\beta_{i_2}}\mathbf{F}_0(\mathbf{X}_{t_{k-1}})^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}] + R(h^{3/2}, \mathbf{X}_{t_{k-1}}))\\
&= \frac{4}{N}\sum_{k=1}^N(\operatorname{Tr}(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)) + R(\sqrt{h}, \mathbf{X}_{t_{k-1}})) \xrightarrow[N\to\infty]{\mathbb{P}_{\boldsymbol{\theta}_0}} 4[\mathbf{C}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)]_{i_1 i_2}.
\end{aligned}$$

To prove (vii) we use Corollary 3.8:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[\zeta^{(j_1)}_{N,k}(\boldsymbol{\theta}_0)\zeta^{(j_2)}_{N,k}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}]\\
&= \frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k}\mathbf{Z}^\top_{t_k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}]\\
&- \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\\
&= \frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{h,k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}^\top_{h,k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}]\\
&- \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top) + R(\sqrt{h}/N, \mathbf{X}_{t_{k-1}}).
\end{aligned}$$

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:

$$\begin{aligned}
&\frac{1}{h^2N}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\xi}^\top_{h,k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}^\top_{h,k}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}]\\
&= \frac{2}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_1}}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_2}}) + \frac{1}{N}\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_1}})\operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top}{\partial\varsigma_{j_2}}).
\end{aligned}$$

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.