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

Predrag Pilipovic, Adeline Samson, Susanne Ditlevsen

Submitted to the Annals of Statistics

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

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

[1]*Department of Mathematical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark,* [a]*predrag@math.ku.dk*

[2]*Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France,* [b]*adeline.leclercq-samson@univ-grenoble-alpes.fr;* [c]*susanne@math.ku.dk*

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

1. Introduction. Stochastic differential equations (SDEs) are popular models for physical, biological, and 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.

Keywords and phrases: Asymptotic normality, Consistency, L^p convergence, Splitting schemes, Stochastic differential equations, Stochastic Lorenz system.

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

Some pioneering papers on likelihood-based SDE estimators are Dacunha-Castelle and Florens-Zmirou (1986); Dohnal (1987); Florens-Zmirou (1989); Genon-Catalot and Jacod (1993); Kessler (1997). The first two only estimate the diffusion parameter. Florens-Zmirou (1989) used EM to estimate both parameters and derived asymptotic properties. Genon-Catalot and Jacod (1993) generalized to higher dimensions, 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 using a linear function. In the case of additive noise, this corresponds to an OU process with a known Gaussian transition density. Thus, the likelihood approximation is a product of Gaussian densities. Shoji (1998) proved that LL discretization is one-step consistent and L^p convergent with order 1.5. Shoji (2011), Jimenez, Mora and Selva (2017) extended the theory of LL for SDEs with multiplicative noise. Simulation studies show the superiority of the LL estimator compared to other estimators (Shoji and Ozaki, 1998; Hurn, Jeisman and Lindsay, 2007; Gloaguen, Etienne and Le Corff, 2018; Gu, Wu and Xue, 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, Wu and Xue (2020) proposed an efficient implementation that overcomes this. The main disadvantage of the LL method is its slow computational speed.

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

models. Hurn, Jeisman and Lindsay (2007) implemented HE up to third order in univariate cases, emphasizing the importance of symbolic computation tools like `Mathematica` or `Maple`. Their survey concluded that while LL is the best among discrete maximum likelihood estimators, HE is the preferred overall choice. They highlighted that the HE proposed by Aït-Sahalia (2002) has the best trade-off between speed and accuracy, proving more feasible than LL in most financial applications. ~~This finding aligns with the newer review study from López-Pérez, Febrero-Bande and González-Manteiga (2021).~~ Similar results are found in Jensen and Poulsen (2002); López-Pérez, Febrero-Bande and González-Manteiga (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, Beskos and Graham (2022) proposed sampling schemes for elliptic and hypoelliptic models that often result in conditionally non-Gaussian integrals, distinguishing their approach from prior works. As the transition density of their new scheme is typically complex, Iguchi, Beskos and Graham (2022) created a closed-form density expansion using Malliavin calculus. They recommended a transition density scheme that retained second-order precision through prudent truncation of the expansion. This closed-form expansion aligns with the works of Aït-Sahalia (2002, 2008) and Li (2013) on elliptic SDEs, although with a different approach. Iguchi, Beskos and Graham (2022) deliver asymptotic results with analytically available rates, beneficial for both elliptic and hypoelliptic models.

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

An essential aspect of any estimator is the practical execution in real-world applications. Although the previously mentioned research contributes significantly to the theoretical development and broadens our understanding of inference for SDEs, its practical implementations tend not to be user-friendly. Except for precomputed models, applications by 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

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

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 $J \to \infty$.

example, McLachlan and Quispel (2002); Blanes, Casas and Murua (2009)), but their extension to SDEs is straightforward. A few studies have investigated numerical properties (Bensoussan, Glowinski and Răşcanu, 1992; Ableidinger, Buckwar and Hinterleitner, 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, Vilmart and Zygalakis (2015) used LT splitting to investigate conditions for preserving the measure of the ergodic nonlinear Langevin equations. Recently, Bréhier, Cohen and Ulander (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, Tamborrino and Tubikanec (2020); Ditlevsen, Tamborrino and Tubikanec (2023) used splitting schemes for parametric inference in combination with Approximate Bayesian Computation, and Ditlevsen and Ditlevsen (2023) used it for prediction of a forthcoming collapse in the climate.

This paper presents five main contributions:

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

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

Notation. We use capital bold letters for random vectors, vector-valued functions, and matrices, while lowercase bold letters denote deterministic vectors. $\|\cdot\|$ denotes both the L^2 vector norm in $\mathbb{R}^d$ and the matrix norm induced by the L^2 norm, defined as the square root of the largest eigenvalue. Superscript (i) on a vector denotes the 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 $\partial^2_{ij} g(\mathbf{x})$ 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:

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

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, e.g., $\mathbb{E}$ implicitly refers to $\mathbb{E}_{\boldsymbol{\theta}}$.

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

REMARK 2. *We assume additive noise, sometimes referred to as constant volatility, meaning that the diffusion matrix does not depend on the current state. ~~While this assumption is natural in some applications, it can be restrictive in others.~~ This assumption can be restrictive and even rejected by the data in some applications. The proposed methodology ~~could potentially be extended to~~ can be extended if the diffusion is reducible ~~diffusions~~ (Definition 1 in (Aït-Sahalia, 2008)) 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:

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

(A2) Function $\mathbf{N}$ grows at most polynomially in $\mathbf{x}$, uniformly in $\boldsymbol{\theta}$, i.e., there exist constants $C>0$ and $\chi\ge1$ such that:

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

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

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

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

$$\mathbb{E}[\sup_{t\in[0,T]}\|\mathbf{X}_t\|^{2p}]<C(1+\|\mathbf{x}_0\|^{2p}),\qquad\forall p\ge1.$$

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

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

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

LEMMA 2.1. *Let Assumptions (A1)-(A2) hold. Let* $\mathbf{X}$ *be a solution of* (1)*. Let* $g\in C^{(2l+2)}$ *be of polynomial growth and* $p\ge2$*. 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 $R(h^3,\mathbf{X}_{t_{k-1}})$. Applying $L_{\boldsymbol{\theta}}$ on $g(\mathbf{x})=x^{(i)}$, Lemma 2.1 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):

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

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

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

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

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:

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

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

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:

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

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

We, therefore, introduce the following and last assumption.

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

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

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

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

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:*

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

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

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

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

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

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

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)*. Strang splitting outputs the middle points of the smooth steps of the deterministic flow* (8)*, while LT splitting outputs the stochastic increments in the rough steps. We conjecture that this is one of the reasons why the S splitting has superior statistical properties.*

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

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

$$\mathcal{L}^{[\mathrm{LT}]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}) \overset{\theta}{=} N\log(\det \boldsymbol{\Omega}_h(\boldsymbol{\theta}))$$

$$+\sum_{k=1}^{N}(\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}))^\top \boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1}(\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta})). \tag{12}$$

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:

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

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

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

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 can often be solved numerically. When $\boldsymbol{f}_h^{-1}$ is well defined, we use the identity $-\log|\det D\boldsymbol{f}_h^{-1}(\mathbf{x};\boldsymbol{\beta})| = \log|\det D\boldsymbol{f}_h(\mathbf{x};\boldsymbol{\beta})|$ in (14) to increase the speed and numerical stability.

Finally, we define the estimators as:

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

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

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

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

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

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, Shoji and Ozaki, 1999). The LL method consists of the following steps:

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

The approximation becomes:

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

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

$$\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta}) := \int_0^h e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})(h-u)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\beta})^\top(h-u)}\,\mathrm{d}u,$$

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

$$\mathbf{R}_{h,i}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) := \int_0^h \exp(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})u)u^i\,\mathrm{d}u, \qquad i = 0,1,$$

$$\mathbf{M}(\mathbf{x};\boldsymbol{\theta}) := \frac{1}{2}(\mathrm{Tr}\,\mathbf{H}_1(\mathbf{x};\boldsymbol{\theta}),\dots,\mathrm{Tr}\,\mathbf{H}_d(\mathbf{x};\boldsymbol{\theta}))^\top, \qquad \mathbf{H}_k(\mathbf{x};\boldsymbol{\theta}) := \left[[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{ij}\frac{\partial^2 F^{(k)}}{\partial x^{(i)}\partial x^{(j)}}(\mathbf{x})\right]_{i,j=1}^d.$$

We can efficiently compute $\mathbf{R}_{h,i}$ and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ using formulas from (Van Loan, 1978), see (Gu, Wu and Xue, 2020). For more details, see Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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

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

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

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

2.5. *An example: the stochastic Lorenz system.* The Lorenz system is a 3D system introduced by Lorenz (1963) to model atmospheric convection. The model is originally deterministic exhibiting deterministic chaos, i.e., tiny differences in initial conditions lead to unpredictable and widely diverging trajectories. The Lorenz system evolves around two strange attractors, implying that trajectories remain within some bounded region, while points that start in close proximity may eventually separate by arbitrary distances as time progresses (Hilborn and Hilborn, 2000). We add noise to include unmodelled forces and randomness. The stochastic Lorenz system is given by:

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

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, Rivera and López-Caraballo

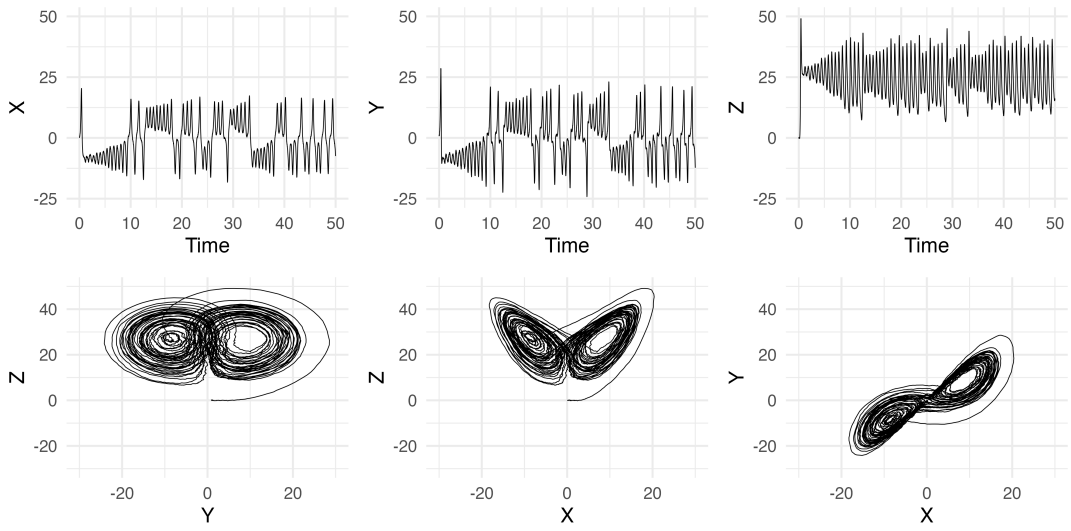

Fig 1: An example trajectory of the stochastic Lorenz system (20) starting at $(0, 1, 0)$ for $N = 10000$ and $h = 0.005$. The first row shows the evolution of the individual components X, Y, and Z. The second row shows the evolution of component pairs: (Y, Z), (X, Z) and (X, Y). Parameters are $p =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

(2016) used sophisticated optimization algorithms to achieve better precision. Dubois et al. (2020) and Ann et al. (2022) used deep neural networks in combination with other machine learning algorithms. Ozaki, Jimenez and Haggan-Ozaki (2000) used Kalman filtering based on LL on the stochastic Lorenz system.

Figure 1 shows an example trajectory of the stochastic Lorenz system. The trajectory was generated by subsampling from an EM simulation, such that $N = 10000$ 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 $\mathbf{b}_{\mathrm{mix}} = (x^\star, y^\star, z^\star)$. The nonlinear solution is

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

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, $\det D\boldsymbol{f}_{\mathrm{mix},h}^{-1}(\cdot) = 1$.

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

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

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

The nonlinear solution is:

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

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

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

3. Order of one-step predictions and L^p convergence. In this Section, we investigate L^p convergence of the splitting schemes and the order of the one-step predictions. Theorem 2.1 in Tretyakov and Zhang (2013) extends Milstein's fundamental theorem on L^p convergence for global Lipschitz coefficients (Milstein, 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 $k=1,\dots,N$, 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 $k=1,\dots,N$:*

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

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 $\widetilde{\mathbf{X}}$ is L^p convergent, $p \geq 1$, of order $q_2 - 1/2$, i.e., for $k=1,\dots,N$, 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, Samson and Ditlevsen, 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). *Assume (A1)-(A2), let* $\mathbf{X}$ *be the solution to* (1) *and let* $\Phi_h^{[\mathrm{LT}]}$ *be the LT approximation* (10)*. Then, for* $k = 1, \dots, N$*, 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). *Assume (A1)-(A2), let* $\mathbf{X}^{[\mathrm{LT}]}$ *be the LT approximation defined in* (10)*, and let* $\mathbf{X}$ *be the solution of* (1)*. Then, there exists* $C \geq 1$ *such that for all* $p \geq 2$*, and* $k = 1, \dots, N$*, 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, Samson and Ditlevsen, 2023).

PROPOSITION 3.6. *Assume (A1)-(A2), let* $\mathbf{X}$ *be the solution to* (1)*, and let* $\Phi_h^{[\mathrm{S}]}$ *be the S splitting approximation* (11)*. Then, for* $k = 1, \dots, N$*, 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 7. *Even though LT and S have the same order of* L^p *convergence, the crucial difference is in the one-step prediction. The approximated transition density between two consecutive data points depends on the one-step approximation. Thus, the objective function based on pseudo-likelihood from the S splitting is more precise than the one from the LT.*

To prove L^p convergence of the S splitting scheme for (1) with one-sided Lipschitz drift, we follow the same procedure as in Buckwar et al. (2022). The proof of the following theorem is in Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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

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

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

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

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

Before presenting the central results for the estimator, we establish the groundwork with two essential lemmas that rely on the model assumptions. Lemma 4.1 (Lemma 6 in Kessler (1997)) deals with the p-th moments of the SDE increments and also provides a moment bound of a polynomial map of the solution. The proof of this lemma in Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023) differs from that in Kessler (1997) due to our relaxation of the global Lipschitz assumption of the drift $\mathbf{F}$. Instead, we use a one-sided Lipschitz condition in conjunction with the generalized Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020) to establish the result, see Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023)).

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

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

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

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

(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:*

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

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

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

uniformly in $\boldsymbol{\theta}$.

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

PROPOSITION 4.3 (Moment Bounds). *Assume (A1), (A2), (A6). Let* $\mathbf{X}$ *be the solution of* (1), *and* $\mathbf{Z}_{t_k}$ *as defined in* (13). *Let* $\mathbf{g}(\mathbf{x};\boldsymbol{\beta})$ *be a generic function with derivatives of polynomial growth, and* $\boldsymbol{\beta} \in \Theta_\beta$. *Then, for* $k = 1, \dots, N$, *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, the full objective functions (12) and (14) are not needed to prove consistency and asymptotic normality. It is enough to approximate $\boldsymbol{\Omega}_h$ up to the second order by $h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}(\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}^\top)$ (see equation (6)). Indeed, after applying Taylor series on the inverse of $\boldsymbol{\Omega}_h$, we get:

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

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:

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

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

$$\mathcal{L}_N^{[\mathrm{LT}]}(\boldsymbol{\theta}) \coloneqq N \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + Nh \operatorname{Tr} \mathbf{A}(\boldsymbol{\beta})$$

$$+ \frac{1}{h}\sum_{k=1}^{N} (\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} (\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta})) \tag{22}$$

$$- \sum_{k=1}^{N} (\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{A}(\boldsymbol{\beta}) (\mathbf{X}_{t_k} - \boldsymbol{\mu}_h(\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta});\boldsymbol{\beta}))$$

$$\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\theta}) \coloneqq N \log\det \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + Nh \operatorname{Tr} \mathbf{A}(\boldsymbol{\beta}) + \frac{1}{h}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta})$$

$$- \sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{A}(\boldsymbol{\beta}) \mathbf{Z}_{t_k}(\boldsymbol{\beta}) + h\sum_{k=1}^{N} \operatorname{Tr} D\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}). \tag{23}$$

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

5.1. *Consistency.* Now, we state the consistency of $\hat{\boldsymbol{\beta}}_N$ and $\widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top$. The proof of Theorem 5.1 is in Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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

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

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

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

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

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

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

Now, we state the theorem for asymptotic normality, the proof is in Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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

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

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

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

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

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

We do not write the closed-form distributions for K (17), LL (18) and HE (19), but we use the corresponding formulas to implement the likelihoods. We implement the two splitting strategies proposed in Section 2.5, leading to four estimators: $\mathrm{LT}_{\mathrm{mix}}, \mathrm{LT}_{\mathrm{avg}}, \mathrm{S}_{\mathrm{mix}}$, and $\mathrm{S}_{\mathrm{avg}}$. To further speed up computation time, we use the same trick for calculating $\boldsymbol{\Omega}_h$ in (6) as for calculating $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$, see Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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 except HE converged after approximately 80 iterations. The HE estimator only converged with the smallest time step, $h = 0.005$, achieving convergence in $43\% - 72\%$ of cases across various sample sizes N. This probably occurs due to a polynomial approximation of the likelihood that can be unstable at the boundaries, especially for larger h. Incorporating higher-order approximations and adding constraints in the optimization step might improve performance. For further analysis, see the Supplementary Material (Pilipovic, Samson and Ditlevsen, 2023).

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

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

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

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

6.5. *Results.* In Figure 2, AREs are shown on log scale as a function of h. While most estimators work well for a step size no greater than 0.01, only LL, $\mathrm{S}_{\mathrm{mix}}$, and $\mathrm{S}_{\mathrm{avg}}$ perform well for $h = 0.05$. The $\mathrm{LT}_{\mathrm{avg}}$ is not competitive even for $h = 0.005$. The performance of $\mathrm{LT}_{\mathrm{mix}}$ varies, sometimes approaching the performance of K, while other times performing similarly to EM. Thus, $\mathrm{LT}_{\mathrm{mix}}$ is not a good choice for this specific model. The bias of EM starts to show for $h = 0.01$ escalating for $h = 0.05$. The largest bias appears in the diffusion parameters due to the poor approximation of $\boldsymbol{\Omega}_h^{\mathrm{EM}}$. K is less biased than EM except for p and r when $h = 0.05$. The HE estimator converged only for $h = 0.005$. The ARE is calculated from the 601 simulations out of a total of 1000 in which convergence was achieved. For these, the performance of HE in estimating drift parameters is comparable to the best estimators. However, the diffusion parameters are not well estimated, with the estimation of σ_3^2 being the least accurate. Drift parameters are generally estimated better for larger h for fixed N due to a longer observation interval $T = Nh$, reflecting the $\sqrt{Nh}$ rate of convergence.

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

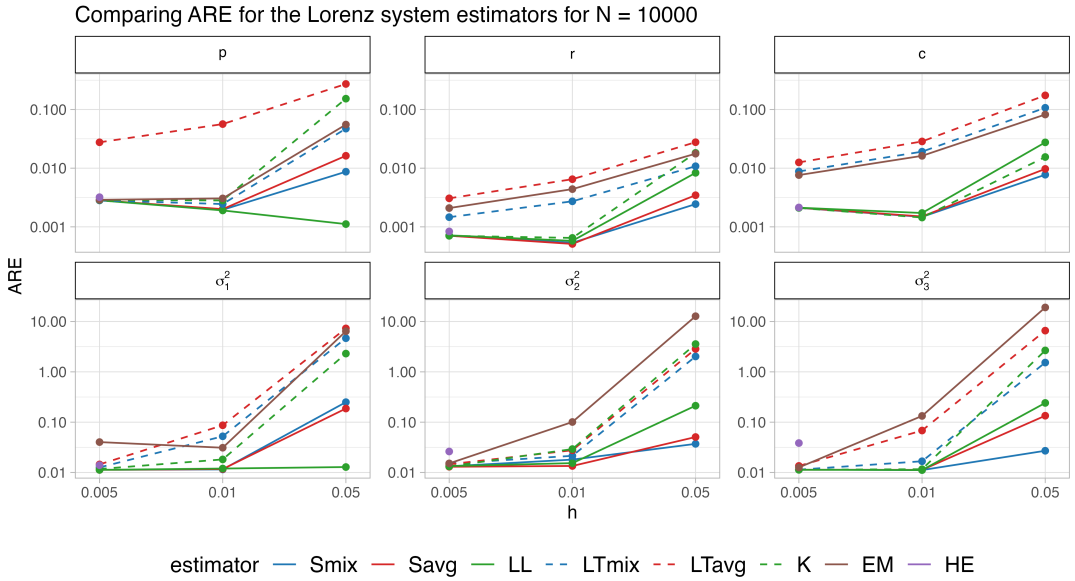

Fig 2: Comparing the absolute relative error (ARE) as a function of increasing discretization step h for eight estimators in the stochastic Lorenz system. The sample size is $N = 10000$. The y-axis is on log scale. The HE estimator (purple dot) converged only for $h = 0.005$, and only for 60% of the simulated data sets.

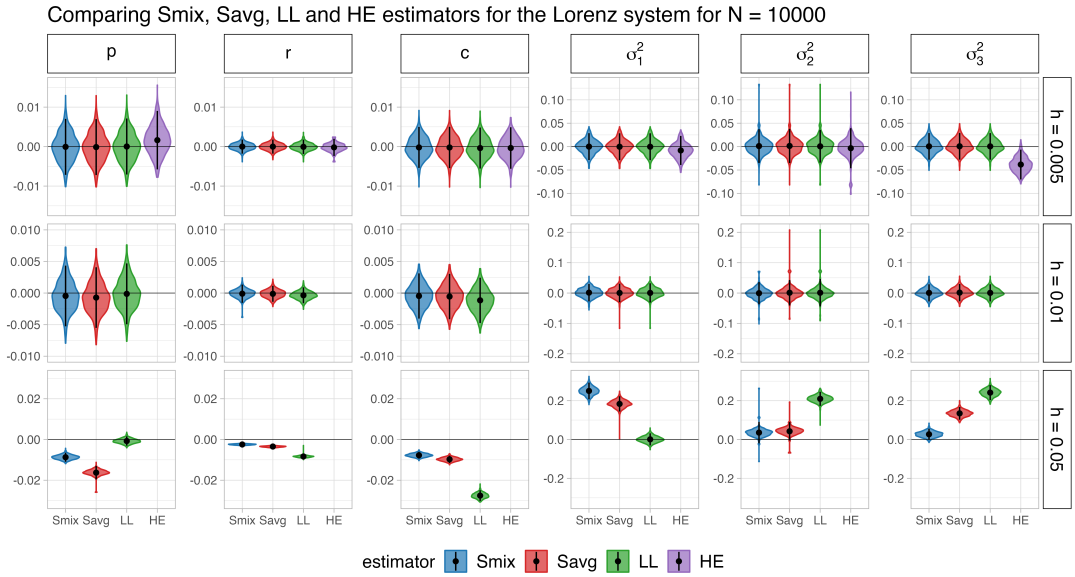

Fig 3: Comparing the normalized distributions of $(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \oslash \boldsymbol{\theta}_0$ (where $\oslash$ is the element-wise division) of the Lorenz system for the $\mathrm{S}_{\mathrm{mix}}$, $\mathrm{S}_{\mathrm{avg}}$, LL and HE estimators for $N = 10000$. Each column represents one parameter, and each row represents one value of the discretization step h. The black dot with a vertical bar in each violin plot represents the mean and the standard deviation. The HE estimator (purple) converged only for $h = 0.005$, and only for 60% of the simulated data sets.

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

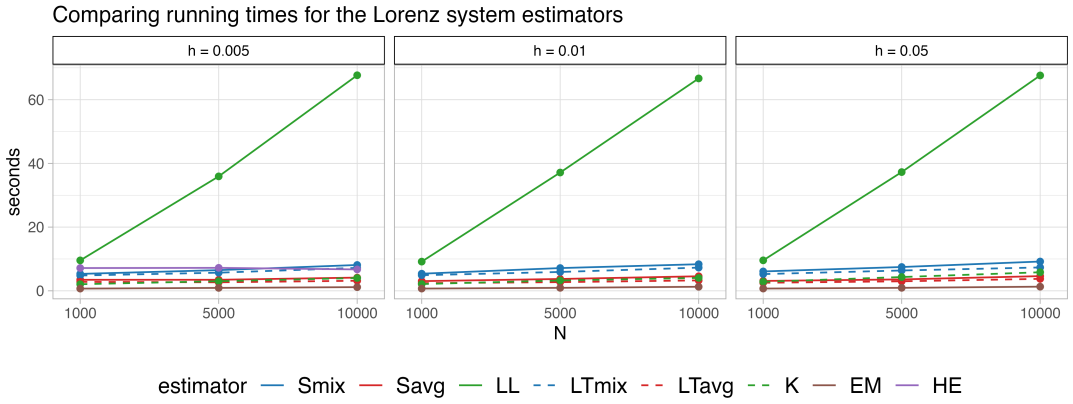

Fig 4: Running times as a function of N for different estimators of the Lorenz system. Each column shows one value of h. On the x-axis is the sample size N, and on the y-axis is the running time in seconds. The HE estimator (purple) achieved convergence only for $h = 0.005$, and only in $43\% - 72\%$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

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

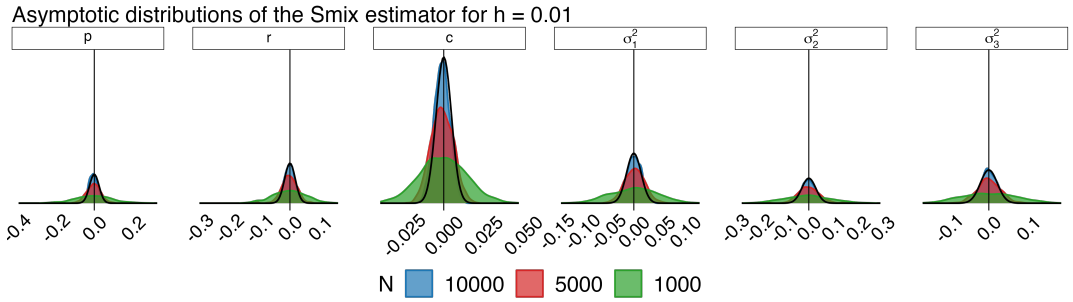

Fig 5: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ for the $\mathrm{S}_{\mathrm{mix}}$ estimator with theoretical asymptotic distributions (28) for each parameter (columns), for $h = 0.01$ and $N \in \{1000, 5000, 10000\}$ (colors). The black lines correspond to the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

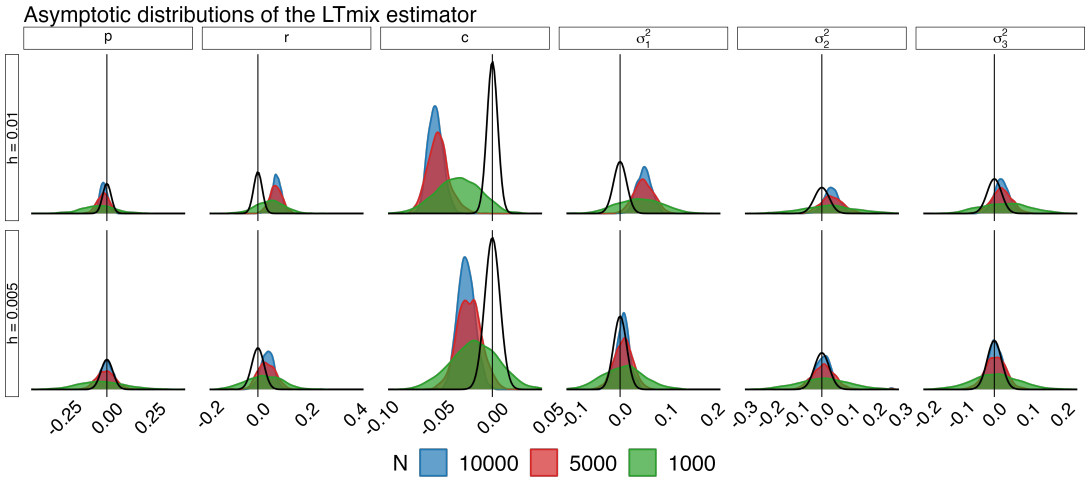

Fig 6: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ for the $\mathrm{LT}_{\mathrm{mix}}$ estimator with theoretical asymptotic distributions (28) for each parameter (columns), for $h \in \{0.005, 0.01\}$ (rows) and $N \in \{1000, 5000, 10000\}$ (colors). The black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

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

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

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

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

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

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

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

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

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

Acknowledgments. PP is also affiliated with the Bielefeld Graduate School of Economics and Management at the University of Bielefeld in Germany. We would like to thank three anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved paper. We are thankful to the third reviewer for providing the HE method implementation for the Lorenz system.

Funding. The European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)"; and Novo Nordisk Foundation NNF20OC0062958. This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

SUPPLEMENTARY MATERIAL

Supplement to "Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations Based on Splitting Schemes"
This supplement provides proofs of all the propositions, lemmas, and theorems from the paper that are not proved in the main text.

REFERENCES

ABDULLE, A., VILMART, G. and ZYGALAKIS, K. C. (2015). Long Time Accuracy of Lie–Trotter Splitting Methods for Langevin Dynamics. *SIAM Journal on Numerical Analysis* **53** 1-16. https://doi.org/10.1137/140962644

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

ABLEIDINGER, M., BUCKWAR, E. and HINTERLEITNER, H. (2017). A Stochastic Version of the Jansen and Rit Neural Mass Model: Analysis and Numerics. *J. Math. Neurosci.* **7**. https://doi.org/10.1186/s13408-017-0046-4

AÏT-SAHALIA, Y. (2002). Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach. *Econometrica* **70** 223-262.

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

ALAMO, A. and SANZ-SERNA, J. M. (2016). A Technique for Studying Strong and Weak Local Errors of Splitting Stochastic Integrators. *SIAM J. Num. Anal.* **54** 3239-3257. https://doi.org/10.1137/16M1058765

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

ANN, N., PEBRIANTI, D., ABAS, M. and BAYUAJI, L. (2022). *Parameter Estimation of Lorenz Attractor: A Combined Deep Neural Network and K-Means Clustering Approach* In *Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering* **730** 321-331. Springer, Singapore.

ARNST, M., LOUPPE, G., VAN HULLE, R., GILLET, L., BUREAU, F. and DENOËL, V. (2022). A hybrid stochastic model and its Bayesian identification for infectious disease screening in a university campus with application to massive COVID-19 screening at the University of Liège. *Math. Biosci.* **347** 108805.

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

BAYDIN, A. G., PEARLMUTTER, B. A., RADUL, A. A. and SISKIND, J. M. (2017). Automatic Differentiation in Machine Learning: A Survey. *J. Mach. Learn. Res.* **18** 5595–5637.

BENSOUSSAN, A., GLOWINSKI, R. and RĂŞCANU, A. (1992). Approximation of Some Stochastic Differential Equations by the Splitting Up Method. *Applied Mathematics and Optimization* **25** 81-106.

BIBBY, B. M. and SØRENSEN, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1** 17-39.

BLANES, S., CASAS, F. and MURUA, A. (2009). Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Mat. Apl.* **45**.

BOU-RABEE, N. and OWHADI, H. (2010). Long-Run Accuracy of Variational Integrators in the Stochastic Context. *SIAM Journal on Numerical Analysis* **48** 278-297. https://doi.org/10.1137/090758842

BRÉHIER, C.-E., COHEN, D. and ULANDER, J. (2023). Analysis of a Positivity-preserving Splitting Scheme for Some Nonlinear Stochastic Heat Equations. *arXiv:2302.08858*.

BRÉHIER, C.-E. and GOUDENÈGE, L. (2019). Analysis of some splitting schemes for the stochastic Allen-Cahn equation. *Disc. Cont. Dyn. Sys. - B* **24** 4169-4190. https://doi.org/10.3934/dcdsb.2019077

BUCKWAR, E., TAMBORRINO, M. and TUBIKANEC, I. (2020). Spectral density-based and measure-preserving ABC for partially observed diffusion processes. An illustration on Hamiltonian SDEs. *Stat. Comput.* **30**.

BUCKWAR, E., SAMSON, A., TAMBORRINO, M. and TUBIKANEC, I. (2022). A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *App. Num. Math.* **179** 191-220.

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

CHOI, S. (2013). Closed-form likelihood expansions for multivariate time-inhomogeneous diffusions. *Journal of Econometrics* **174** 45-65. https://doi.org/10.1016/j.jeconom.2011.12

CHOI, S. (2015). Explicit form of approximate transition probability density functions of diffusion processes. *Journal of Econometrics* **187** 57-73. https://doi.org/10.1016/j.jeconom.2015.02

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

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

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

DITLEVSEN, P. and DITLEVSEN, S. (2023). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications* **14** 4254.

DITLEVSEN, S. and SAMSON, A. (2019). Hypoelliptic diffusions: filtering and inference from complete and partial observations. *J. R. Stat. Soc. B-Statistical Methodology* **81** 361-384. https://doi.org/10.1111/rssb.12307

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

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

DOHNAL, G. (1987). On Estimating the Diffusion Coefficient. *Journal of Applied Probability* **24** 105–114.

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

FALBEL, D. and LURASCHI, J. (2022). torch: Tensors and Neural Networks with 'GPU' Acceleration https://torch.mlverse.org/docs, https://github.com/mlverse/torch.

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

FORMAN, J. L. and SØRENSEN, M. (2008). The Pearson diffusions: A class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics* **35** 438–465.

FUCHS, C. (2013). *Inference for Diffusion Processes with Applications in Life Sciences*. Springer Berlin, Heidelberg.

GENON-CATALOT, V. and JACOD, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'I.H.P. Probabilités et statistiques* **29** 119–151. MR1204521

GLOAGUEN, P., ETIENNE, M.-P. and LE CORFF, S. (2018). Stochastic differential equation based on a multimodal potential to model movement data in ecology. *J. R. Stat. Soc. C Applied Statistics* **67**.

GLOTER, A. (2006). Parameter Estimation for a Discretely Observed Integrated Diffusion Process. *Scandinavian Journal of Statistics* **33** 83–104.

GLOTER, A. and YOSHIDA, N. (2020). Adaptive and non-adaptive estimation for degenerate diffusion processes.

GLOTER, A. and YOSHIDA, N. (2021). Adaptive estimation for degenerate diffusion processes. *Elec. J. Stat.* **15** 1424 – 1472.

GOBET, E. (2002). LAN property for ergodic diffusions with discrete observations. *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38** 711-737. https://doi.org/10.1016/S0246-0203(02)01107-X

GU, W., WU, H. and XUE, H. (2020). *Parameter Estimation for Multivariate Nonlinear Stochastic Differential Equation Models: A Comparison Study* In *Statistical Modeling for Biological Systems: In Memory of Andrei Yakovlev* 245–258. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-34675-1_13

HAIRER, E., NØRSETT, S. P. and WANNER, G. (1993). *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg.

HILBORN, R. C. and HILBORN, A. L. C. P. P. R. (2000). *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*. *Oxford scholarship online: Physics module*. Oxford University Press.

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

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

HUMPHRIES, A. R. and STUART, A. M. (2002). *Deterministic and random dynamical systems: theory and numerics* In *Modern Methods in Scientific Computing and Applications* 211–254. Springer Netherlands, Dordrecht.

HURN, A. S., JEISMAN, J. I. and LINDSAY, K. A. (2007). Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations. *J. Finan. Econ.* **5** 390-455.

HUTZENTHALER, M., JENTZEN, A. and KLOEDEN, P. E. (2011). Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proc. Roy. Soc. A: Mathematical, Physical and Engineering Sciences* **467** 1563-1576.

IGUCHI, Y., BESKOS, A. and GRAHAM, M. M. (2022). Parameter Estimation with Increased Precision for Elliptic and Hypo-elliptic Diffusions. *ArXiv* 2211.16384.

JENSEN, B. and POULSEN, R. (2002). Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques. *Journal of Derivatives* **9** 18–32.

JIMENEZ, J. C., MORA, C. and SELVA, M. (2017). A weak Local Linearization scheme for stochastic differential equations with multiplicative noise. *J. Comput. Appl. Math.* **313** 202-217.

JIMENEZ, J., SHOJI, I. and OZAKI, T. (1999). Simulation of Stochastic Differential Equations Through the Local Linearization Method. A Comparative Study. *J. Stat. Phys.* **94** 587-602.

KAREEM, A. M. and AL-AZZAWI, S. N. (2021). A Stochastic Differential Equations Model for Internal COVID-19 Dynamics. *J. Phys.: Conference Series* **1818** 012121. https://doi.org/10.1088/1742-6596/1818/1/012121

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

KESSLER, M. (1997). Estimation of an Ergodic Diffusion from Discrete Observations. *Scand. J. Stat.* **24** 211–229.

KLOEDEN, P. E. and PLATEN, E. (1992). *Numerical Solution of Stochastic Differential Equations*. *Stochastic Modelling and Applied Probability*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-12616-5

KRYLOV, N. V. (1991). A Simple Proof of the Existence of a Solution of Itô's Equation with Monotone Coefficients. *Theory of Probability & Its Applications* **35** 583-587. https://doi.org/10.1137/1135082

LAZZÚS, J. A., RIVERA, M. and LÓPEZ-CARABALLO, C. H. (2016). Parameter estimation of Lorenz chaotic system using a hybrid swarm intelligence algorithm. *Physics Letters A* **380** 1164-1171.

LEIMKUHLER, B. and MATTHEWS, C. (2015). Molecular dynamics. *Interdisc. Appl. Math.* **39** 443.

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

LORENZ, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences* **20** 130 - 141.

LÓPEZ-PÉREZ, A., FEBRERO-BANDE, M. and GONZÁLEZ-MANTEIGA, W. (2021). Parametric Estimation of Diffusion Processes: A Review and Comparative Study. *Mathematics* **9**. https://doi.org/10.3390/math9080859

MAO, X. (2007). *Stochastic differential equations and applications*. Elsevier.

MCLACHLAN, R. I. and QUISPEL, G. R. W. (2002). Splitting methods. *Acta Numerica* **11** 341-434.

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

MICHELOT, T., GLENNIE, R., HARRIS, C. and THOMAS, L. (2021). Varying-Coefficient Stochastic Differential Equations with Applications in Ecology. *J. Agri., Biol. Environ. Stat.* **26**.

MILSTEIN, G. N. (1988). A Theorem on the Order of Convergence of Mean-Square Approximations of Solutions of Systems of Stochastic Differential Equations. *Theo. Prob. Appl.* **32** 738-741.

MILSTEIN, G. N. and TRETYAKOV, M. V. (2003). Quasi-symplectic methods for Langevin-type equations. *IMA Journal of Numerical Analysis* **23** 593-626. https://doi.org/10.1093/imanum/23.4.593

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

OZAKI, T. (1985). Statistical Identification of Storage Models with Application to Stochastic Hydrology. *Journal of The American Water Resources Association* **21** 663-675.

OZAKI, T. (1992). A Bridge between Nonlinear Time Series Models and Nonlinear Stochastic Dynamical Systems: A Local Linearization Approach. *Statistica Sinica* **2** 113–135.

OZAKI, T., JIMENEZ, J. C. and HAGGAN-OZAKI, V. (2000). The Role of the Likelihood Function in the Estimation of Chaos Models. *Journal of Time Series Analysis* **21** 363-387. https://doi.org/10.1111/1467-9892.00189

PICCHINI, U. and DITLEVSEN, S. (2011). Practical estimation of high dimensional stochastic differential mixed-effects models. *Comput. Stat. Data Anal.* **55** 1426-1444.

PILIPOVIC, P., SAMSON, A. and DITLEVSEN, S. (2023). Supplement to "Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations Based on Splitting Schemes".

RIEDMILLER, M. and BRAUN, H. (1992). RPROP - A Fast Adaptive Learning Algorithm Technical Report, Proc. of ISCIS VII), Universitat.

SHOJI, I. (1998). Approximation of Continuous Time Stochastic Processes by a Local Linearization Method. *Mathematics of Computation* **67** 287–298.

SHOJI, I. (2011). A note on convergence rate of a linearization method for the discretization of stochastic differential equations. *Comm. Nonlinear Sci. Num. Simul.* **16** 2667-2671.

SHOJI, I. and OZAKI, T. (1998). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications* **16** 733-752. https://doi.org/10.1080/07362999808809559

SØRENSEN, M. and UCHIDA, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9 (6)** 1051 – 1069.

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

TABOR, M. (1989). *Chaos and Integrability in Nonlinear Dynamics: An Introduction*. Wiley.

R CORE TEAM (2022). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

TIAN, Y. and FAN, M. (2020). Nonlinear integral inequality with power and its application in delay integro-differential equations. *Advances in Difference Equations* **2020**. https://doi.org/10.1186/s13662-020-02596-y

TRETYAKOV, M. V. and ZHANG, Z. (2013). A Fundamental Mean-Square Convergence Theorem for SDEs with Locally Lipschitz Coefficients and Its Applications. *SIAM Journal on Numerical Analysis* **51** 3135-3162.

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

VAN LOAN, C. (1978). Computing Integrals Involving the Matrix Exponential. *IEEE Trans. Aut. Cont.* **23** 395-404.

VATIWUTIPONG, P. and PHEWCHEAN, N. (2019). Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations* **2019** 1-7.

YANG, N., CHEN, N. and WAN, X. (2019). A new delta expansion for multivariate diffusions via the Itô-Taylor expansion. *Journal of Econometrics* **209** 256-288. https://doi.org/10.1016/j.jeconom.2019.01

ZHUANG, L., CAO, L., WU, Y., ZHONG, Y., ZHANGZHONG, L., ZHENG, W. and WANG, L. (2020). Parameter Estimation of Lorenz Chaotic System Based on a Hybrid Jaya-Powell Algorithm. *IEEE Access* **8** 20514-20522.