Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations based on Splitting Schemes

Predrag Pilipovic, Adeline Samson, Susanne Ditlevsen

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

A PREPRINT

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

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

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

ABSTRACT

Surprisingly, general estimators for nonlinear continuous time models based on stochastic differential equations are yet lacking. Most applications still use the Euler-Maruyama discretization, despite many proofs of its bias. More sophisticated methods, such as the Kessler, the Ozaki, or MCMC methods, lack a straightforward implementation and 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 also has an L^p convergence rate of order 1, which was already known for LT. We prove under the less restrictive one-sided Lipschitz assumption that the estimators are consistent and asymptotically normal. A numerical study on the 3-dimensional stochastic Lorenz chaotic system complements our theoretical findings. The simulation shows that the S estimator performs the best when measured on both precision and computational speed compared to the state-of-the-art.

Keywords Asymptotic normality · Consistency · Lp convergence · Splitting schemes · Stochastic differential equations · Stochastic Lorenz system

1 Introduction

Stochastic differential equations (SDEs) are popular models for physical, biological, and socio-economic processes. Some recent applications include tipping points in the climate (Ditlevsen and Ditlevsen, 2022), 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. This is particularly useful when the dynamics are not completely understood, and the unknown parts are described as random.

The following parametric form is often assumed for an SDE model with additive noise,

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

Our goal is 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 needed for likelihood-based estimators, and thus a closed-form solution to (1). However, the transition density is only available for a few SDEs including the Ornstein-Uhlenbeck (OU) process, which has a

linear drift function $\mathbf{F}$. Extensive literature exists on MCMC methods for the nonlinear case (Fuchs, 2013; Chopin and Papaspiliopoulos, 2020), however, these are often computationally intensive and do not always converge to the correct values for complex models. Thus, we need a valid approximation of the transition density to perform likelihood-based statistical inference.

The simplest discretization scheme is the Euler-Maruyama (EM) (Kloeden and Platen, 1992). Its main advantage is the easy-to-implement and intuitive Gaussian transition density. It is widely used in both frequentist and Bayesian approaches, across theoretical and applied studies. However, it has many disadvantages. First, the EM-based estimator suffers from a large bias when the discretization step is large (see Florens-Zmirou (1989) for a theoretical study, or Gloaguen et al. (2018), Gu et al. (2020) for applied studies). Second, Hutzenthaler et al. (2011) showed that it is not mean-square convergent when the drift function $\mathbf{F}$ of (1) grows super-linearly. Consequently, we should avoid EM for models with polynomial drifts. 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 Jacob (1993); Kessler (1997). The first two only estimate the diffusion parameter. Florens-Zmirou (1989) used EM to estimate both drift and diffusion parameters and derived asymptotic properties. Genon-Catalot and Jacob (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 (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 $\mathbf{F}$.

Specific setups have been studied, e.g., 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. Martingales estimating functions are explored in Bibby and Sørensen (1995); Forman and Sørensen (2008) for one-dimensional diffusions, however, these are difficult to extend to multidimensional SDEs. More recently, Ditlevsen and Samson (2019) used the 1.5 scheme to solve the problem of hypoellipticity when the diffusion matrix is not of full rank. However, it is complex to implement, especially for high-dimensional models.

A competitive likelihood-based method is based on local linearization (LL), first proposed by Ozaki (1985) and generalized by Shoji and Ozaki (1998). The drift between two consecutive observations is approximated by a linear function, which for additive noise corresponds to an OU process with 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. Simulation studies show the superiority of the LL estimator compared to others (Shoji and Ozaki, 1998; Gloaguen et al., 2018; Gu et al., 2020). Until recently, the implementation of the LL estimator was numerically ill-conditioned due to possible singularity of the Jacobian matrix of the drift function $\mathbf{F}$. Gu et al. (2020) proposed an efficient implementation that overcomes this. However, the main disadvantage of the LL method is its slow computational speed.

We propose to use the Lie-Trotter (LT) or the Strang (S) splitting schemes for statistical inference. These numerical approximations were first suggested for ordinary differential equations (ODEs) (Blanes et al., 2009; McLachlan and Quispel, 2002), but are straightforwardly extended to SDEs (Buckwar et al., 2022). Although splitting schemes are frequently used for ODEs, they were only recently introduced for SDEs. A few studies have investigated numerical properties (Alamo and Sanz-Serna, 2016; Ableidinger et al., 2017; Ableidinger and Buckwar, 2016; Buckwar et al., 2022), however, the statistical part is missing. To the best of our knowledge, only Buckwar et al. (2020) are using splitting schemes for statistical inference in combination with approximate Bayesian computation.

This paper contains five main contributions. First, we propose new, efficient, easy-to-implement, and computationally fast estimators for multidimensional nonlinear SDEs. Second, we prove convergence of the S-splitting scheme. Third, we prove consistency and asymptotic normality for the new estimators under the less restrictive assumption of one-sided Lipschitz. This requires original ideas for the proofs. Fourth, we show that the estimators work for a stochastic version of the chaotic Lorenz system. Only estimators for the deterministic system have previously been proposed. Fifth, we compare the new estimators to four likelihood-based estimators from literature in a simulation study, comparing accuracy, precision, and computational speed.

In Section 2 we introduce the SDE model 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. In Sections 4 and 5, we prove the estimator asymptotics under the less restrictive one-sided global Lipschitz assumption. In Section 6, we illustrate 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 to get faster and more reliable estimators. We compare the precision and speed of the EM, K, LL, LT, and S estimators. We show that the EM and LT estimators become biased before the others with the increase of discretization step h, and that the LL and S perform the best. However, S is much faster than LL, since the running time of LL increases with the sample size N.

Notation. We use capital bold letters for random vectors, vector-valued functions, and matrices. The L^2 norm is denoted $\|\cdot\|$. Superscript (i) on a vector denotes the i-th component, while on a matrix it denotes the i-th row. Double subscript ij on a matrix denotes the component in the i-th row and j-th column. If a matrix is a product of more matrices, square brackets with subscripts denote a component inside the matrix. The transpose is denoted by $\top$, operator $\mathrm{Tr}(\cdot)$ returns the trace of a matrix and $\det(\cdot)$ the determinant. Sometimes, we denote by $[a_i]_{i=1}^d$ a vector with coordinates a_i, and by $[b_{ij}]_{i,j=1}^d$ a matrix with coordinates b_{ij}, for $i, j = 1, \ldots, d$. We denote with $\partial_i g(\mathbf{x})$ the partial derivative of a generic function $g : \mathbb{R}^d \to \mathbb{R}$ with respect to $x^{(i)}$ and $\partial_{ij}^2 g(\mathbf{x})$ the second partial derivative. The nabla operator ∇ denotes the gradient vector of a function g, $\nabla g(\mathbf{x}) = [\partial_i g(\mathbf{x})]_{i=1}^d$. The differential operator D denotes the Jacobian matrix $D\mathbf{F}(\mathbf{x}) = [\partial_i F^{(j)}(\mathbf{x})]_{i,j=1}^d$, for a vector-valued function $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$. $\mathbf{H}$ denotes the Hessian matrix of a real-valued function g, $\mathbf{H}_g(\mathbf{x}) = [\partial_{ij} g(\mathbf{x})]_{i,j=1}^d$. We abuse the notation $\mathcal{O}(\cdot)$ and write $\mathcal{O}(h^p)$ in short for $\mathcal{O}(h^p(1 + \|\mathbf{x}\|)^C)$, where C is a positive constant and $\mathbf{x}$ is clear from the context. For a random vector $\mathbf{X}$ and positive constant C, we write $\mathcal{O}_{\mathbb{P}}(h^p) = \mathcal{O}_{\mathbb{P}}(h^p(1 + \|\mathbf{X}\|)^C)$. The Kronecker delta function is denoted by δ_i^j. For an open set A, the bar $\overline{A}$ indicates closure. We use $\overset{\boldsymbol{\theta}}{=}$ 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. $\mathbf{I}_d$ stands for d-dimensional identity matrix, while $\mathbf{0}_{d\times d}$ is a d-dimensional zero square matrix.

2 Problem setup

Let $\mathbf{X}_t$ be a d-dimensional stochastic process indexed by time t, defined as the solution to (1) on the time interval $[0, T]$. We rewrite the drift function $\mathbf{F}$ of (1) as follows

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

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \overline{\Theta}_\beta \times \overline{\Theta}_\Sigma = \overline{\Theta}$ be an unknown parameter with Θ_β and Θ_Σ being two open convex bounded subsets of $\mathbb{R}^r$ and $\mathbb{R}^{d\times d}$, respectively. The process $\mathbf{W}$ is a d-dimensional Wiener process; $\mathbf{F}, \mathbf{N} : \mathbb{R}^d \times \overline{\Theta}_\beta \to \mathbb{R}^d$; function $\mathbf{A}$ is defined on $\overline{\Theta}_\beta$ and takes values in $\mathbb{R}^{d\times d}$; and 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, when we write estimation of parameter $\boldsymbol{\Sigma}$, we mean estimation of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. The drift function $\mathbf{F}$ in (1) is split up into a linear part given by matrix $\mathbf{A}$ and a nonlinear part given by $\mathbf{N}$. This decomposition will be essential for the definition of the splitting schemes and the pseudo-likelihood that we later use for the estimation of $\boldsymbol{\theta}$.

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 $\mathbf{A}_0, \mathbf{N}_0(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$ instead of $\mathbf{A}(\boldsymbol{\beta}_0)$, $\mathbf{N}(\mathbf{x}; \boldsymbol{\beta}_0)$ and $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top$, when referring to the true parameters. We write $\mathbf{A}$, $\mathbf{N}(\mathbf{x})$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ for any parameter $\boldsymbol{\theta}$.

We assume additive noise, i.e. the diffusion matrix does not depend on the current state. This can be relaxed if the model is reducible, where we can apply the Lamperti transformation (see Aït-Sahalia (2008)).

2.1 Assumptions

The main assumption is that (2) has a unique strong solution $\mathbf{X}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\theta_0})$, which follows from the following first two assumptions (Buckwar et al., 2022). We need the last three assumptions for proving properties of the estimators.

(A1) Function $\mathbf{N}$ is twice continuously differentiable with respect to both $\mathbf{x}$ 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

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

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

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

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

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

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

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

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

Assume a sample $(\mathbf{X}_{t_k})_{k=0}^N \equiv \mathbf{X}_{0:t_N}$ from (2) at time steps $0 = t_0 < t_1 < \cdots < t_N = T$, which we, for notational simplicity, assume equidistant with step size $h = t_k - t_{k-1}$. We denote $\mathcal{F}_{t_k} \coloneqq \sigma\left(\mathbf{W}_s; s \leq t_k\right)$ the natural filtration of the paths.

2.2 Moments

Assumption (A1) ensures finiteness of the moments of the solution $\mathbf{X}$ (Buckwar et al., 2022), i.e.

$$\mathbb{E}[\sup_{t\in[0,T]} \|\mathbf{X}_t\|^p] < C\left(1 + \|\mathbf{x}_0\|^p\right), \qquad \forall p > 0. \tag{3}$$

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

$$L_{\theta_0} g\left(\mathbf{x}; \boldsymbol{\theta}\right) = \mathbf{F}\left(\mathbf{x}; \boldsymbol{\beta}_0\right)^\top \nabla g\left(\mathbf{x}; \boldsymbol{\theta}\right) + \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top \mathbf{H}_g(\mathbf{x}; \boldsymbol{\theta})\right). \tag{4}$$

The moments of SDE (1) are expanded using the following lemma (Lemma 1.10 in (Kessler et al., 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. Then*

$$\mathbb{E}_{\theta_0}[g(\mathbf{X}_{t_k}; \boldsymbol{\theta}) \mid \mathcal{F}_{t_{k-1}}] = \sum_{j=0}^{l} \frac{h^l}{j!} L_{\theta_0}^j g(\mathbf{X}_{t_{k-1}}; \boldsymbol{\theta}) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^{l+1}).$$

In the rest of this section, we suppress the parameter from the notation. We need terms up to order $\mathcal{O}(h^3)$. For $g(\mathbf{x}) = x^{(i)}$

$$\mathbb{E}[X_{t_k}^{(i)} \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}] = x^{(i)} + hF^{(i)}(\mathbf{x}) + \frac{h^2}{2}(\mathbf{F}(\mathbf{x})^\top \nabla F^{(i)}(\mathbf{x}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{F^{(i)}}(\mathbf{x}))) + \mathcal{O}(h^3). \tag{5}$$

2.3 Splitting Schemes

Consider the following splitting of (2)

$$\mathrm{d}\mathbf{X}_t^{[1]} = \mathbf{A}\mathbf{X}_t^{[1]}\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0^{[1]} = \mathbf{x}_0, \tag{6}$$

$$\mathrm{d}\mathbf{X}_t^{[2]} = \mathbf{N}(\mathbf{X}_t^{[2]})\,\mathrm{d}t, \qquad \mathbf{X}_0^{[2]} = \mathbf{x}_0. \tag{7}$$

Equation (6) is an OU process with explicit solution given by the following h-flow

$$\mathbf{X}_{t_k}^{[1]} = \Phi_h^{[1]}(\mathbf{X}_{t_{k-1}}^{[1]}) = e^{\mathbf{A}h}\mathbf{X}_{t_{k-1}}^{[1]} + \boldsymbol{\xi}_{h,k}, \tag{8}$$

where $\boldsymbol{\xi}_{h,k} \overset{i.i.d}{\sim} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_h)$ for $k = 1, \ldots, N$. Covariance matrix $\boldsymbol{\Omega}_h$ is given by (Vatiwutipong and Phewchean, 2019)

$$\boldsymbol{\Omega}_h = \int_0^h e^{\mathbf{A}(h-u)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top e^{\mathbf{A}^\top(h-u)}\,\mathrm{d}u = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}\left(\mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{A}^\top\right) + \mathcal{O}(h^3). \tag{9}$$

Assumptions (A1), (A2) ensure the existence and uniqueness of the solution of (7) (Thm 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

$$\mathbf{X}_{t_k}^{[2]} = \Phi_h^{[2]}(\mathbf{X}_{t_{k-1}}^{[2]}) = \boldsymbol{f}_h(\mathbf{X}_{t_{k-1}}^{[2]}; \boldsymbol{\beta}). \tag{10}$$

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

$$\boldsymbol{f}_0(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}, \qquad \boldsymbol{f}_{t+s}(\mathbf{x}; \boldsymbol{\beta}) = \boldsymbol{f}_t(\boldsymbol{f}_s(\mathbf{x}; \boldsymbol{\beta}); \boldsymbol{\beta}), \quad t, s \geq 0. \tag{11}$$

Remark Since only one-sided Lipschitz continuity is assumed, the solution to (7) 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. If needed, it can be approximated by a well-defined backward flow. In the globally Lipschitz case, this is not an issue.

We, therefore, introduce the following and last assumption.

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

We state a useful proposition for the nonlinear solution $\boldsymbol{f}_h$ (Section 1.8 in (Hairer et al., 1993)).

Proposition 2.2 *Let Assumptions (A1)-(A2) hold. When $h \to 0$, the h-flow of* (7) *is*

$$\boldsymbol{f}_h(\mathbf{x}) = \mathbf{x} + h\mathbf{N}(\mathbf{x}) + \frac{h^2}{2}\left(D\mathbf{N}(\mathbf{x})\right)\mathbf{N}(\mathbf{x}) + \mathcal{O}(h^3). \tag{12}$$

The two most common splitting approximations of the solution $\mathbf{X}$ are defined as follows.

Definition 2.3 *Let Assumptions (A1)-(A2) hold. The LT and S approximations of the solution of* (2) *are given by*

$$\mathbf{X}_{t_k}^{[\mathrm{LT}]} \coloneqq \Phi_h^{[\mathrm{LT}]}\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\right) = \left(\Phi_h^{[1]} \circ \Phi_h^{[2]}\right)\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\right) = e^{\mathbf{A}h}\boldsymbol{f}_h\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{LT}]}\right) + \boldsymbol{\xi}_{h,k}, \tag{13}$$

$$\mathbf{X}_{t_k}^{[\mathrm{S}]} \coloneqq \Phi_h^{[\mathrm{S}]}\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\right) = \left(\Phi_{h/2}^{[2]} \circ \Phi_h^{[1]} \circ \Phi_{h/2}^{[2]}\right)\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\right) = \boldsymbol{f}_{h/2}\left(e^{\mathbf{A}h}\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\right) + \boldsymbol{\xi}_{h,k}\right). \tag{14}$$

Remark The order of composition in the splitting schemes is not unique. Changing the order in the S-splitting leads to a sum of 2 independent random variables, one Gaussian and one non-Gaussian, whose likelihood is not trivial. Thus, we only use the splitting (14). The opposite order in the LT-splitting can be treated in the same way as the S-splitting.

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

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

Thus, it makes sense for LT and S to have the same order of L^p convergence. We prove this in Section 3.

2.4 Estimators

First, we introduce the two new estimators LT and S based on a sample $\mathbf{X}_{0:t_N}$. Then, we briefly recall the EM, K, and LL estimators that are compared in the simulation study.

2.4.1 Splitting estimators

The LT scheme (13) follows a Gaussian distribution. A pseudo negative log-likelihood of (2) is given as

$$\mathcal{L}^{[\mathrm{LT}]}(\mathbf{X}_{0:t_N}; \boldsymbol{\theta}) \overset{\theta}{=} \frac{N}{2}\log(\det \boldsymbol{\Omega}_h(\boldsymbol{\theta})) + \frac{1}{2}\sum_{k=1}^{N}(\mathbf{X}_{t_k} - e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}))^\top \boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1}(\mathbf{X}_{t_k} - e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta})). \tag{15}$$

The S-splitting (14) follows a nonlinear transformation of a Gaussian random variable $e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}) + \boldsymbol{\xi}_{h,k}$. We define

$$\mathbf{Z}_{t_k}(\boldsymbol{\beta}) \coloneqq \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta}) - \boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}\right), \qquad \boldsymbol{\mu}_h(\mathbf{x}; \boldsymbol{\beta}) \coloneqq e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_{h/2}(\mathbf{x}; \boldsymbol{\beta}), \tag{16}$$

and apply change of variables to obtain the pseudo-likelihood

$$\mathcal{L}^{[\mathrm{S}]}(\mathbf{X}_{0:t_N}; \boldsymbol{\theta}) \overset{\theta}{=} \frac{N}{2}\log\left(\det \boldsymbol{\Omega}_h(\boldsymbol{\theta})\right) + \frac{1}{2}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top \boldsymbol{\Omega}_h(\boldsymbol{\theta})^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}) - \sum_{k=1}^{N}\log\left|\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta})\right|. \tag{17}$$

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, it can often be solved numerically. When $\boldsymbol{f}_h^{-1}$ is well defined, we use the identity $-\log|\det D\boldsymbol{f}_h^{-1}(\mathbf{x};\boldsymbol{\beta})| = \log|\det D\boldsymbol{f}_h(\mathbf{x};\boldsymbol{\beta})|$ in (17) to increase the speed and numerical stability.

Finally, we define the estimators as

$$\widehat{\boldsymbol{\theta}}_N^{[k]} \coloneqq \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{[k]}(\mathbf{X}_{0:t_N};\boldsymbol{\theta}), \qquad k \in \{\mathrm{LT},\mathrm{S}\}. \tag{18}$$

2.4.2 Euler-Maruyama

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

$$\mathbf{X}_{t_k}^{[\mathrm{EM}]} \coloneqq \mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]} + h\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{EM}]};\boldsymbol{\beta}) + \boldsymbol{\xi}_{h,k}^{[\mathrm{EM}]}, \tag{19}$$

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 and the pseudo-likelihood follows trivially.

2.4.3 Kessler

The K estimator uses Gaussian transition densities $p^{[\mathrm{K}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ with the true mean and covariance of the solution $\mathbf{X}$ (Kessler, 1997). When the moments are not known, they are approximated using the infinitesimal generator (Lemma 2.1). We implement the estimator based on the 2nd order approximation (K2). It is given as follows

$$\begin{aligned}\mathbf{X}_{t_k}^{[\mathrm{K2}]} \coloneqq \mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]} &+ h\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta}) \\ &+ \frac{h^2}{2}\Big(D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta})\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta}) + \frac{1}{2}[\mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{H}_{F^{(i)}}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta}))]_{i=1}^d\Big) + \boldsymbol{\xi}_{h,k}^{[\mathrm{K2}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]}),\end{aligned} \tag{20}$$

where $\boldsymbol{\xi}_{h,k}^{[\mathrm{K2}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]}) \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Omega}_{h,k}^{[\mathrm{K2}]}(\boldsymbol{\theta}))$, and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{K2}]}(\boldsymbol{\theta}) = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \frac{h^2}{2}(D\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta})\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top D^\top\mathbf{F}(\mathbf{X}_{t_{k-1}}^{[\mathrm{K2}]};\boldsymbol{\beta}))$. The covariance matrix is not constant which makes the algorithm slower with larger sample size.

2.4.4 Ozaki's local linearization

Ozaki's LL method approximates the drift of (1) between every two observations by a linear function (Jimenez et al., 1999). The LL method consists of the following steps:

(1) Perform LL of the drift $\mathbf{F}$ in each time interval $[t, t+h)$ by the Itô-Taylor series;

(2) Compute the analytic solution of the resulting linear SDE.

The approximation becomes

$$\mathbf{X}_{t_k}^{[\mathrm{LL}]} \coloneqq \mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]} + \Phi_h^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]};\boldsymbol{\theta}) + \boldsymbol{\xi}_{h,k}^{[\mathrm{LL}]}(\mathbf{X}_{t_{k-1}}^{[\mathrm{LL}]}), \tag{21}$$

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

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

Following Gu et al. (2020), a fast and efficient way to compute $\mathbf{R}_{h,i}$ and $\boldsymbol{\Omega}_{h,k}^{[\mathrm{LL}]}(\boldsymbol{\theta})$ is as follows. First, define three block matrices

$$\mathbf{P}_1(\mathbf{x}) = \begin{bmatrix}\mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})\end{bmatrix},\ \mathbf{P}_2(\mathbf{x}) = \begin{bmatrix}-D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \mathbf{I}_d & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d}\end{bmatrix},\ \mathbf{P}_3(\mathbf{x}) = \begin{bmatrix}D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}) & \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \\ \mathbf{0}_{d\times d} & -D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})^\top\end{bmatrix}. \tag{22}$$

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

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

From the first matrix, we obtain $\mathbf{R}_{h,0}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$. Then, we compute $\mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$ from the formula $\mathbf{R}_{h,1}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta})) = \exp(hD\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))\mathbf{B}_{\mathbf{R}_{h,1}}(D\mathbf{F}(\mathbf{x};\boldsymbol{\beta}))$. The terms at the $\star$ symbols are of no importance. Finally, $\boldsymbol{\Omega}^{[\mathrm{LL}]}_{h,k}(\boldsymbol{\theta})$ is obtained from the matrix exponential,

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

This provides a Gaussian density $p^{[\mathrm{LL}]}(\mathbf{X}_{t_k} \mid \mathbf{X}_{t_{k-1}};\boldsymbol{\theta})$ and standard likelihood inference. Like in K2, the covariance matrix $\boldsymbol{\Omega}^{[\mathrm{LL}]}_{h,k}(\boldsymbol{\theta})$ depends on the previous state of the process $\mathbf{X}^{[\mathrm{LL}]}_{t_{k-1}}$, 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 K2, LL does not use Taylor expansion of the approximated drift and covariance matrix, so the influence of sample size N is much stronger. For details on derivations of previous formulas, see Gu et al. (2020).

2.5 An Example: Stochastic Lorenz system

The Lorenz system is a 3D chaotic system introduced by Lorenz (1963) to model atmospheric convection. The model is originally deterministic exhibiting deterministic chaos. It means that tiny differences in initial conditions lead to unpredictable and widely diverging trajectories. The Lorenz system still exhibits some structure around two strange attractors, i.e., the trajectories remain within some bounded region, however, two nearby points at one time will be arbitrarily far apart at later times (Hilborn and Hilborn, 2000). To include unmodelled forces and randomness in the Lorenz system, we add noise. The stochastic Lorenz system is given by the equations

$$\begin{aligned} \mathrm{d}X_t &= p(Y_t - X_t)\,\mathrm{d}t + \sigma_1\,\mathrm{d}W^{(1)}_t, \\ \mathrm{d}Y_t &= (rX_t - Y_t - X_tZ_t)\,\mathrm{d}t + \sigma_2\,\mathrm{d}W^{(2)}_t, \\ \mathrm{d}Z_t &= (X_tY_t - cZ_t)\,\mathrm{d}t + \sigma_3\,\mathrm{d}W^{(3)}_t. \end{aligned} \tag{23}$$

Variables X_t, Y_t, and Z_t denote variables proportional to convective intensity, horizontal and vertical temperature differences, respectively. Parameters p, r, and c denote the Prandtl number, Rayleigh number, and a geometric factor, respectively (Tabor, 1989). Lorenz (1963) used parameters $p = 10$, $r = 28$ and $c = 8/3$, for system (23) to show chaotic behaviors.

The system does not fulfill the global Lipschitz condition because it is a second-order polynomial, nor does it fulfill the one-sided Lipschitz condition (Humphries and Stuart, 1994). However, it has a unique global solution and an invariant probability (Keller, 1996). Assumptions (A2) is fulfilled due to the polynomial structure. Thus, all assumptions (A2)-(A5), except (A1) hold. Even so, the solution exists and we show in Section 6 that our estimators still work.

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

In Figure 1 an example trajectory of the stochastic Lorenz system is illustrated. The trajectory was generated by subsampling from an EM simulation such that $N = 5000$ and $h = 0.01$, 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 chaotic properties.

In Section 5 we prove that the asymptotic properties do not depend on the splitting choice of the matrix $\mathbf{A}$ (and thus the function $\mathbf{N}$). However, before the asymptotics is reached the performance is influenced by the choice of splitting. Nonetheless, the choice of the optimal splitting strategy is beyond the scope of this paper and will be treated in a separate paper.

Based on preliminary simulations (results not shown), we choose the following splitting strategy

$$\mathbf{A} = \begin{bmatrix} -p/2 & p & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -c \end{bmatrix}, \qquad \mathbf{N}(x,y,z) = \begin{bmatrix} -px/2 \\ x(r-z) \\ xy \end{bmatrix}. \tag{24}$$

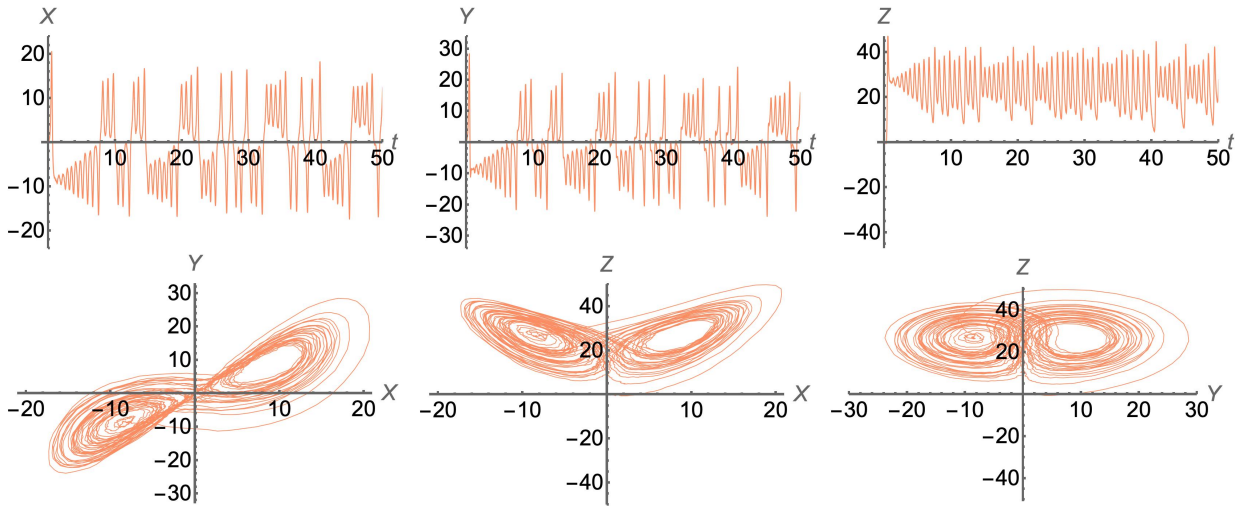

Figure 1: An example trajectory of the stochastic Lorenz system (23). The first row shows the evolution for the individual components X_t, Y_t, and Z_t. The second row shows the evolution of component pairs: (X_t, Y_t), (X_t, Z_t) and (Y_t, Z_t). Parameters are $p = 10$, $r = 28$, $c = 8/3$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\sigma_3^2 = 1.5$.

The corresponding nonlinear solution is

$$\boldsymbol{f}_h(x, y, z) = \begin{bmatrix} \exp(-ph/2)x \\ y\cos(2x(1-\exp(-ph/2))/p) - (z-r)\sin(2x(1-\exp(-ph/2))/p) \\ y\sin(2x(1-\exp(-ph/2))/p) + (z-r)\cos(2x(1-\exp(-ph/2))/p) + r \end{bmatrix}. \tag{25}$$

The solution $\boldsymbol{f}_h$ is a composition of a 3D rotation and translation of (y, z) around the scaled x-axis. The inverse always exists. Thus, Assumption (A6) holds. Moreover, $\det D\boldsymbol{f}_h^{-1}(x, y, z) = \exp(ph/2)$.

3 Order of one-step predictions and L^p convergence

In this Section, we investigate L^p convergence of the splitting schemes, as well as the order of the one-step predictions. First, we define L^p consistency of a one-step approximation (Definition 1 in Buckwar et al. (2022)).

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 with order $q_2 - 1/2$ if for $k = 1, \ldots, N$, and some $q_1 \geq q_2 + 1/2$*

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

Furthermore, we need the following definition (Definition 2 in Buckwar et al. (2022)).

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

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

The following theorem (Theorem 1 in Buckwar et al. (2022)) gives sufficient conditions for the L^p convergence of a numerical scheme in a one-sided Lipschitz framework. Since we work with the local Lipschitz case, we need to check that the moments of the numerical schemes are bounded.

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

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

(2) $\widetilde{\mathbf{X}}$ *has bounded moments,*

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

$$\left(\mathbb{E}\left[\left\|\mathbf{X}_{t_k} - \widetilde{\mathbf{X}}_{t_k}\right\|^p\right]\right)^{\frac{1}{p}} = \mathcal{O}(h^{q_2 - 1/2}).$$

3.1 Lie-Trotter splitting

We first show that the one-step LT approximation is of order $\mathcal{O}(h^2)$ in mean. The following proposition is proved in Supplementary Material for scheme (13) as well as for the reversed order of composition.

Proposition 3.4 (One step prediction of LT-splitting) *Let Assumptions (A1) and (A2) hold, let* $\mathbf{X}$ *be the solution to SDE* (1) *and let* $\Phi_h^{[\mathrm{LT}]}$ *be the LT approximation* (13)*. Then, for* $k = 1, \dots, N$

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

In fact, it can be shown that the convergence can not be improved upon, unless the drift $\mathbf{F}$ is linear. The L^p convergence of the LT-splitting scheme is provided in Theorem 2 in Buckwar et al. (2022), which we repeat here for convenience.

Theorem 3.5 (L^p convergence of the LT-splitting) *Let Assumptions (A1) and (A2) hold, let* $\mathbf{X}^{[\mathrm{LT}]}$ *be the LT approximation defined in* (13)*, and let* $\mathbf{X}$ *be the solution of* (1)*. Then, for all* $C \geq 1$ *and* $k = 1, \dots, N$*, it holds*

$$\left(\mathbb{E}\left[\left\|\mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{LT}]}\right\|^p\right]\right)^{\frac{1}{p}} = \mathcal{O}\left(h\left(1 + \|\mathbf{x}_0\|\right)^C\right). \tag{26}$$

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

3.2 Strang splitting

The following proposition states that the S-splitting (14) has higher order one-step predictions than the LT-splitting (13). The proof can be found in Supplementary Material.

Proposition 3.6 *Let Assumptions (A1) and (A2) hold, let* $\mathbf{X}$ *be the solution to* (1)*, and let* $\Phi_h^{[\mathrm{S}]}$ *be the S-splitting approximation* (14)*. Then, for* $k = 1, \dots, N$

$$\left\|\mathbb{E}\left[\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{\mathbf{t_{k-1}}}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right]\right\| = \mathcal{O}(h^3). \tag{27}$$

Remark Even though LT and S have the same order of L^p convergence, the crucial difference is in the one-step prediction. The approximated transition density between two consecutive data points depends on the one-step approximation, and the pseudo-likelihood from the S-splitting is therefore 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 given in Section 7.1.

Theorem 3.7 (L^p convergence of S-splitting) *Let Assumptions (A1), (A2) and (A6) hold, let* $\mathbf{X}^{[\mathrm{S}]}$ *be the S-splitting approximation defined in* (14)*, and let* $\mathbf{X}$ *be the solution of* (1)*. Then, for all* $C \geq 1$*, and* $i = 1, \dots, N$*, it holds*

$$\left(\mathbb{E}\left[\left\|\mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{S}]}\right\|^p\right]\right)^{\frac{1}{p}} = \mathcal{O}\left(h\left(1 + \|\mathbf{x}_0\|\right)^C\right). \tag{28}$$

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

Corollary 3.8 *Let all assumptions from Theorem 3.7 hold. Then, it holds that* $\mathbf{Z}_{t_k} = \boldsymbol{\xi}_{h,k} + \mathcal{O}_{\mathbb{P}}(h)$.

Proof From the definition of $\mathbf{Z}_{t_k}$ in (16), it is enough to prove that

$$\left(\mathbb{E}\left[\left\|\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - e^{\mathbf{A}h}\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_{k-1}}\right) - \boldsymbol{\xi}_{h,k}\right\|^p\right]\right)^{1/p} = \mathcal{O}\left(h\right).$$

From (14) we have that $\boldsymbol{\xi}_{h,k} = \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) - e^{\mathbf{A}h}\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]})$. Then

$$\begin{aligned}
&\left\| \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - e^{\mathbf{A}h}\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_{k-1}}\right) - \boldsymbol{\xi}_{h,k} \right\| \\
&\qquad \leq \left\| \boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}) - \boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k}^{[\mathrm{S}]}\right) \right\| + \left\| e^{\mathbf{A}h} \right\| \left\| \boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_{k-1}}\right) - \boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}\right) \right\| \\
&\qquad \leq C\left(\left\| \mathbf{X}_{t_k} - \mathbf{X}_{t_k}^{[\mathrm{S}]} \right\| + \left\| \mathbf{X}_{t_{k-1}} - \mathbf{X}_{t_{k-1}}^{[\mathrm{S}]} \right\| \right) + \mathcal{O}_{\mathbb{P}}(h).
\end{aligned}$$

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

4 Auxiliary properties

This paper is centered on proving properties of the S estimator. There are two reasons for this. First, in the literature, most numerical properties are proved only for LT-splitting because S-splitting is considered more complex. Here, we prove 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. Thus, standard tools, such as those from Kessler (1997) cannot be directly applied.

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-based estimator. The same reasoning can be used for the LT estimator.

Properties of the S estimator are proved based on the ergodicity of the solution of (1) as in Kessler (1997). However, without global Lipschitz drift, Lemma 6 in Kessler (1997) cannot be applied, since it uses the Lipschitz assumption together with the Grönwall's inequality. Instead, we use a generalization of Grönwall's inequality (Lemma 2.3 in Tian and Fan (2020)) stated in Supplementary Material. In Section 7.2, we prove the following extension of Lemma 6 in Kessler (1997).

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

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

$$\mathbb{E}_{\theta_0}\left[\left\| \mathbf{X}_t - \mathbf{X}_{t_{k-1}} \right\|^p \mid \mathcal{F}_{t_{k-1}} \right] \leq C_p (t - t_{k-1})^{p/2} \left(1 + \left\| \mathbf{X}_{t_{k-1}} \right\| \right)^{C_p}. \tag{29}$$

(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*

$$\mathbb{E}_{\theta_0}\left[\left| g\left(\mathbf{X}_t; \boldsymbol{\theta}\right) \right| \mid \mathcal{F}_{t_{k-1}} \right] \leq C_{t-t_{k-1}} \left(1 + \left\| \mathbf{X}_{t_{k-1}} \right\| \right)^{C} \tag{30}$$

Finally, we state a central ergodic property needed for the asymptotic behavior of the estimator. It is equivalent to Lemma 8 in Kessler (1997) or Lemma 2 in Sørensen and Uchida (2003). The proof for one-sided Lipschitz is the same as in Kessler (1997) when combined with the previous Lemma.

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

$$\frac{1}{N}\sum_{k=1}^{N} g\left(\mathbf{X}_{t_k}, \boldsymbol{\theta}\right) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} \int g\left(\mathbf{x}, \boldsymbol{\theta}\right) \mathrm{d}\nu_0(\mathbf{x}), \tag{31}$$

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

Lastly, we state moment bounds needed for the estimator asymptotics, the proof is in Supplementary Material.

Proposition 4.3 (Moment Bounds) *Let Assumptions (A1), (A2) and (A6) hold. Let* $\mathbf{X}$ *be the solution of* (1)*, and* $\boldsymbol{f}_h$*,* $\boldsymbol{\mu}_h$ *and* $\mathbf{Z}_{t_k}$ *as defined in* (10) *and* (16)*. Let* $\mathbf{g}$ *be a generic function with derivatives of polynomial growth and* $\boldsymbol{\beta} \in \Theta_\beta$*. Then, for* $k = 1, \ldots, N$ *we have the following moment bounds:*

(1) $\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x} \right] = \mathcal{O}(h^3)$*;*

(2) $\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{g}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_k} = \mathbf{x} \right] = \frac{h}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top \mathbf{g}\left(\mathbf{x}; \boldsymbol{\beta}\right) + D\mathbf{g}\left(\mathbf{x}; \boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top \right) + \mathcal{O}(h^2)$*;*

(3) $\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x} \right] = h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathcal{O}(h^2)$.

5 Asymptotics

The estimators $\hat{\boldsymbol{\theta}}_N$ are defined in (18). However, for the proofs, we do not need full pseudo-likelihoods. It is enough to approximate the covariance matrix $\boldsymbol{\Omega}_h$ by $h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$. Indeed, after applying Taylor series on the inverse of $\boldsymbol{\Omega}_h$ (9), we get $(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} + \mathcal{O}(1)$. Then, rewriting likelihoods (17) and (15) yields expressions (32) and (33) plus a term of order $\mathcal{O}_{\mathbb{P}_0}(h)$. Thus, we find estimators of $\boldsymbol{\theta}$ as the minimum of the following objective functions (this is only for the proofs)

$$\mathcal{L}_N^{[\mathrm{LT}]}(\boldsymbol{\theta}) \stackrel{\theta}{=} N \log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) + \frac{1}{h}\sum_{k=1}^{N}(\mathbf{X}_{t_k} - e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\mathbf{X}_{t_k} - e^{\mathbf{A}(\boldsymbol{\beta})h}\boldsymbol{f}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})) \tag{32}$$

$$\mathcal{L}_N^{[\mathrm{S}]}(\boldsymbol{\theta}) \stackrel{\theta}{=} N \log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top) + \frac{1}{h}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}) - 2\sum_{k=1}^{N}\log\left|\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta})\right|. \tag{33}$$

5.1 Consistency

Now, we state the consistency of $\hat{\boldsymbol{\beta}}_N$ and $\widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top$. The proof of the following theorem is in Section 7.3.

Theorem 5.1 *Let Assumptions (A1)-(A6) hold, let* $\mathbf{X}$ *be the solution of* (1)*, let* $\widehat{\boldsymbol{\theta}}_N = (\widehat{\boldsymbol{\beta}}_N, \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top)$ *be the estimator that minimizes one of objective functions* (32) *or* (33)*. If* $h \to 0$ *and* $Nh \to \infty$*, then*

$$\hat{\boldsymbol{\beta}}_N \xrightarrow{\mathbb{P}_{\theta_0}} \boldsymbol{\beta}_0, \qquad \widehat{\boldsymbol{\Sigma}\boldsymbol{\Sigma}}_N^\top \xrightarrow{\mathbb{P}_{\theta_0}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top. \tag{34}$$

5.2 Asymptotic normality

In this section, we state the asymptotic normality of the estimator. First, we need some preliminaries. Let $\rho > 0$ and $\mathcal{B}_\rho(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \rho\}$ be a ball around $\boldsymbol{\theta}_0$. Let $\mathcal{L}_N$ be one of the two pseudo negative log-likelihoods (32) or (33). For $\hat{\boldsymbol{\theta}}_N \in \mathcal{B}_\rho(\boldsymbol{\theta}_0)$, the mean value theorem yields

$$\left(\int_0^1 \mathbf{H}_{\mathcal{L}_N}\left(\boldsymbol{\theta}_0 + t\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right)\right)\mathrm{d}t\right)\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) = -\nabla\mathcal{L}_N(\boldsymbol{\theta}_0). \tag{35}$$

We half-vectorize $\boldsymbol{\Sigma}$ as $\boldsymbol{\varsigma} \coloneqq \mathrm{vech}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right) = \left(\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right]_{11}, \left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right]_{12}, \left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right]_{22}, ..., \left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right]_{1d}, ..., \left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right]_{dd}\right)$ 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 $s = d(d+1)/2$. In case of a diagonal matrix, instead of a half-vectorization, we use $\boldsymbol{\varsigma} \coloneqq \mathrm{diag}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)$. Define

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

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, (35) is equivalent to $\mathbf{D}_N\mathbf{s}_N = \mathbf{L}_N$. Let

$$\mathbf{C}(\boldsymbol{\theta}_0) \coloneqq \begin{bmatrix} \mathbf{C}_\beta(\boldsymbol{\theta}_0) & \mathbf{0}_{r\times s} \\ \mathbf{0}_{s\times r} & \mathbf{C}_\varsigma(\boldsymbol{\theta}_0) \end{bmatrix}, \tag{36}$$

$$[\mathbf{C}_\beta(\boldsymbol{\theta}_0)]_{i_1,i_2} \coloneqq \int \left(\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\left(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)\right)\mathrm{d}\nu_0(\mathbf{x}),\ 1 \le i_1, i_2 \le r, \tag{37}$$

$$[\mathbf{C}_\varsigma(\boldsymbol{\theta}_0)]_{j_1,j_2} \coloneqq \frac{1}{2}\,\mathrm{Tr}\left(\left(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\left(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\right),\ 1 \le j_1, j_2 \le s. \tag{38}$$

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

Theorem 5.2 *Let Assumptions (A1)-(A6) hold, let* $\mathbf{X}$ *be the solution of* (1)*, and let* $\widehat{\boldsymbol{\theta}}_N = (\widehat{\boldsymbol{\beta}}_N, \widehat{\boldsymbol{\varsigma}}_N)$ *be the estimator that minimizes one of objective functions* (32) *or* (33)*. If* $\boldsymbol{\theta}_0 \in \Theta$*,* $\mathbf{C}(\boldsymbol{\theta}_0)$ *is positive definite,* $h \to 0$*,* $Nh \to \infty$*, and* $Nh^2 \to 0$*, then*

$$\begin{bmatrix} \sqrt{Nh}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\right) \\ \sqrt{N}(\hat{\boldsymbol{\varsigma}}_N - \boldsymbol{\varsigma}_0) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{-1}(\boldsymbol{\theta}_0)\right), \tag{39}$$

under $\mathbb{P}_{\theta_0}$.

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

6 Simulation study

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

6.1 Estimators used in the study

The EM transition distribution (19) for the Lorenz system (23) is

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

We do not write the closed-form distributions for the K2 (20) and LL (21) estimators, but we use the corresponding formulas to implement the likelihoods.

The LT-splitting induces a Gaussian transition density with mean vector $\boldsymbol{\mu}^{[\mathrm{LT}]}(\mathbf{x}) = e^{\mathbf{A}h}\boldsymbol{f}_h(\mathbf{x})$. The S-splitting has density a nonlinear transformation of a Gaussian, where $\boldsymbol{\mu}^{[\mathrm{S}]}(\mathbf{x}) = e^{\mathbf{A}h}\boldsymbol{f}_{h/2}(\mathbf{x})$ enters. Both splitting schemes utilise $\boldsymbol{\Omega}_h$ from (9). This is the covariance matrix of the LT likelihood (15), and is used as an intermediate step of the S likelihood (17). More importantly, $\boldsymbol{\Omega}_h$ needs to be computed only once, unlike $\boldsymbol{\Omega}_h^{[\mathrm{LL}]}$ and $\boldsymbol{\Omega}_h^{[\mathrm{K2}]}$. To further speed up computation time, we use the trick suggested by Gu et al. (2020). For the splitting schemes, we adapt $\mathbf{P}_3$ from (22) accordingly

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

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 and decrease discretization errors. We perform $M = 1000$ Monte Carlo repetitions.

6.3 Optimization in `R`

To optimize the likelihoods we use `R` package `torch` (Falbel and Luraschi, 2022), which uses automatic differentiation (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), whereas AD is exact and fast and can be used in numerous applications, such as MLE or training neural networks.

We tried all optimizers available in the `torch` package and decided to use 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 old parameters as the stopping criteria. For starting values, we used a vector of the same value of 0.1. Additionally, we added a `nnf_softplus` function to the optimizer to ensure that all estimated parameters are positive. All estimators converged after approximately 80 iterations.

6.4 Comparing criteria

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

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

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

The running times from the beginning of the optimization step until the estimator is obtained are calculated with the `tictoc` package in R. To avoid the influence of outliers, we compute the median of running times over M repetitions.

6.5 Results

In Figure 2, AREs are shown as a function of the discretization step h. For clearer comparison, we use log-scale on the y axis. While most estimators work well for a step size no greater than 0.01, only LL and S perform well for $h = 0.05$. The LT estimator is not competitive even for $h = 0.005$ and is not the best choice for this model. The bias of EM starts to show for $h = 0.01$ with escalation for $h = 0.05$. The largest bias appears in the diffusion parameters, which is due to poor approximation of $\boldsymbol{\Omega}_h^{\mathrm{EM}}$. K2 is less biased than EM for the diffusion parameters, but performs worse than EM for the drift parameters. Note how some parameters are better determined for larger h, when N is fixed. This is due to a longer observation interval $T = Nh$.

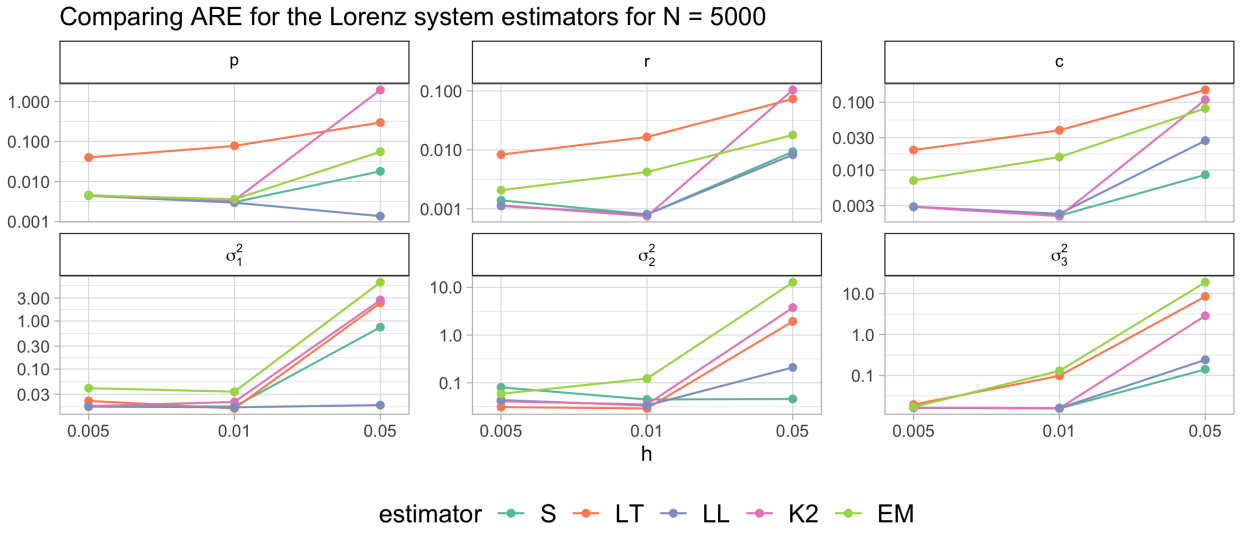

Figure 2: Comparing the ARE as a function of increasing h for 5 different estimators in the stochastic Lorenz system. The estimators are obtained for sample sizes of $N = 5000$. Each column represents one parameter. The y axis is on log-scale

Since S and LL perform the best with large time steps, we zoom in on their distributions in Figure 3. To make the figure clearer, we removed 76 outliers for σ_2^2 in the first two rows. This did not change the shape of the distributions, it only truncated the tails. The two estimators perform similarly, especially for small h. For $h = 0.05$, both estimators are a little biased. S is better for parameters c, σ_2^2 and σ_3^2, whereas LL is better for p, r and σ_1^2.

While LL and S perform similarly in terms of precision, Figure 4 shows the superiority of the S estimator over LL. While the speed of all estimators look linear in N, the slopes differ. Namely, the running time of the LL estimator increases approximately with the function $f(N) = N/250$. The second slowest estimator is K2, followed by the splitting schemes. EM speed is almost constant and does not depend on the sample size. While this is not a general rule but specific to the Lorenz system, it comes from the fact the LL estimator uses N covariance matrices. Additionally, running time does not depend on h. Thus, we recommend using the S estimator, especially for large N.

Figures 5 and 6 show that the theoretical results hold for the S and LT estimators. We compare how the distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ change with sample size N and h. With increasing N the variance decreases, whereas the mean does

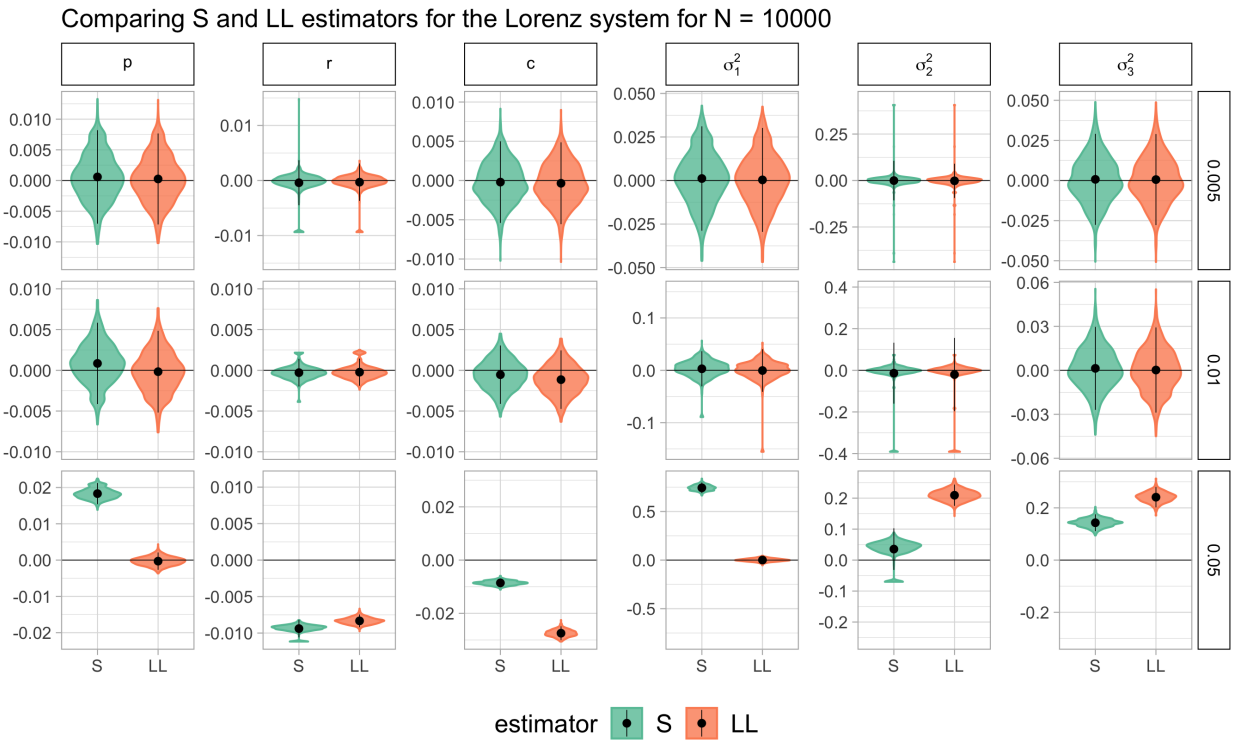

Figure 3: Comparison of the distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ in the Lorenz system for the S and LL estimators for $N = 10000$. Each column represents one parameter and each row represents one value of the discretization step h. A black dot with a vertical bar in each violin plot represents the mean and the standard deviation.

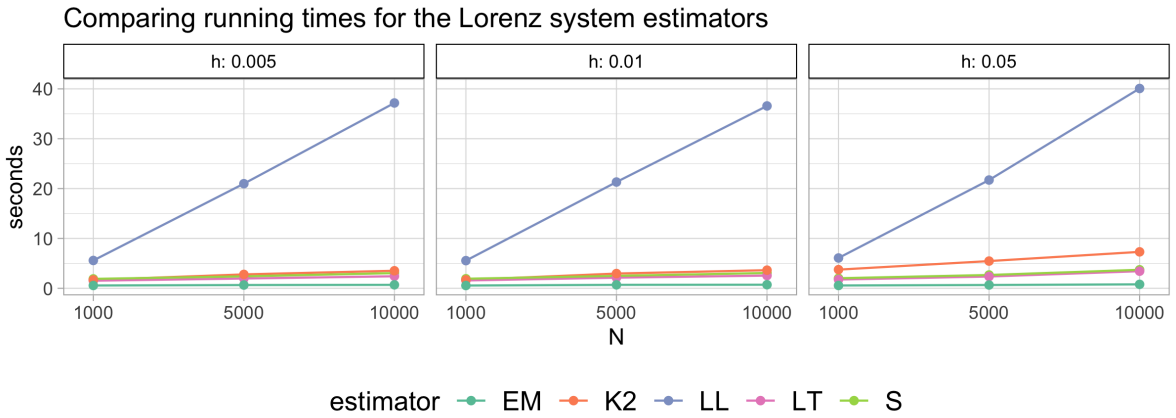

Figure 4: Running times as a function of N for different estimators of the Lorenz system. Each column shows one value of h. On the x-axis is the sample size N and on the y-axis is the running time in seconds.

not change. For that, we need to decrease h, as well. To obtain asymptotic results for LT we need small $h = 0.001$. However, S is unbiased up to $h = 0.01$. This shows that LT is not a good choice in practice, while S is.

The solid black lines in Figures 6 and 5 represent the theoretical asymptotic distributions for each parameter computed from (39). For the Lorenz system (23), the precision matrix (36) is given by

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

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

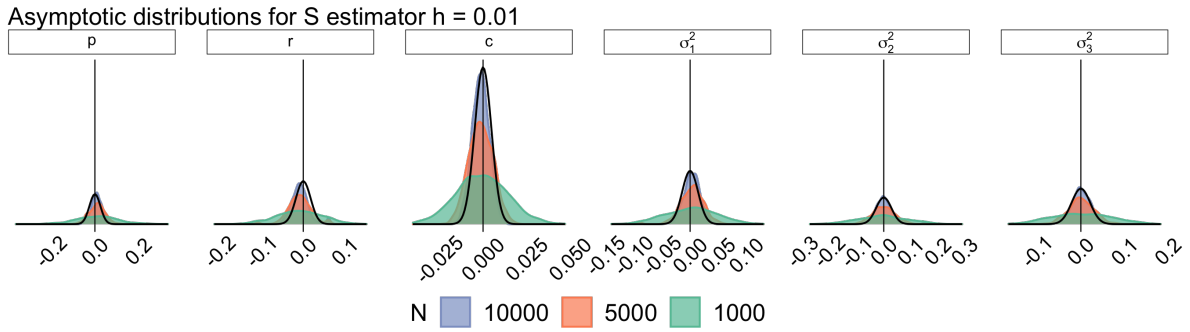

Figure 5: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ from the S estimator with theoretical asymptotic distributions (39) for each parameter (columns), for $h = 0.01$ and $N \in \{1000, 5000, 10000\}$ (colors). The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for $N = 10000$ and $h = 0.01$.

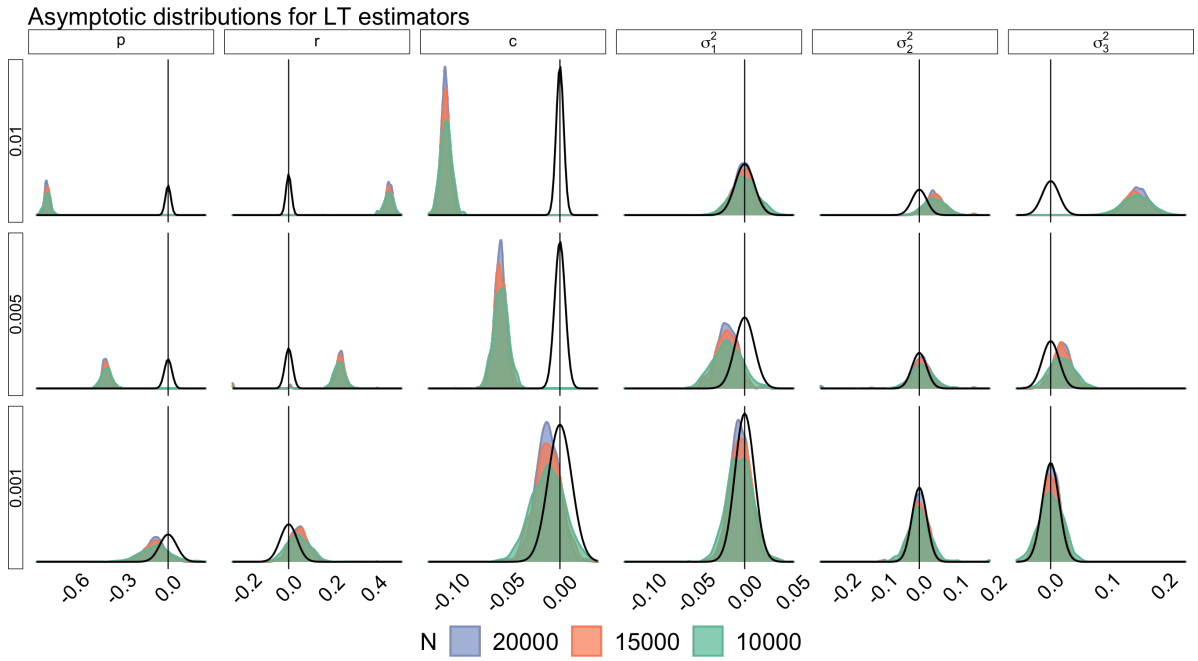

Figure 6: Comparing distributions of $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ from the LT estimator with theoretical asymptotic distributions (39) for each parameter (columns), for $h \in \{0.001, 0.005, 0.01\}$ (rows) and $N \in \{10000, 15000, 20000\}$ (colors). The black lines correspond to the theoretical asymptotic distributions computed from data and true parameters for $N = 20000$ and corresponding h.

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

7 Proofs

7.1 Proof of Lp convergence

Proof of Theorem 3.7 We use Theorem 3.3. To prove condition (1), we use (27) and need to prove the following

$$\left(\mathbb{E}\left[\left\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}\left(\mathbf{X}_{t_{k-1}}\right)\right\|^p \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right]\right)^{\frac{1}{p}} = \mathcal{O}(h^{q_2}),$$

where $q_2 = 3/2$. We start with $\|\mathbf{X}_{t_k} - \Phi_h^{[\mathrm{S}]}(\mathbf{X}_{t_{k-1}})\|^p = \|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\mathbf{F}(\mathbf{X}_{t_{k-1}}) - \boldsymbol{\xi}_{h,k} + \mathcal{O}(h^{3/2})\|^p$. For more details on the expansion of $\Phi_h^{[\mathrm{S}]}$, see Supplementary Material. Approximate $\boldsymbol{\xi}_{h,k} = e^{\mathbf{A}h}\int_{t_{k-1}}^{t_k} e^{-\mathbf{A}s}\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_s$ by

$$\boldsymbol{\xi}_{h,k} = (\mathbf{I} + h\mathbf{A})\int_{t_{k-1}}^{t_k}(\mathbf{I} + s\mathbf{A})\,\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_s + \mathcal{O}_{\mathbb{P}}(h^2) = \boldsymbol{\Sigma}\left(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}}\right) + \mathbf{A}\boldsymbol{\Sigma}\int_{t_{k-1}}^{t_k}\mathbf{W}_s\,\mathrm{d}s + \mathcal{O}_{\mathbb{P}}(h^2).$$

We have that $\int_{t_{k-1}}^{t_k}\mathbf{W}_s\,\mathrm{d}s \sim \mathcal{N}(\mathbf{0}, \frac{h^3}{3}\mathbf{I})$, thus $\boldsymbol{\xi}_{h,k} = \boldsymbol{\Sigma}(\mathbf{W}_{t_k} - \mathbf{W}_{t_{k-1}}) + \mathcal{O}_{\mathbb{P}}(h^{3/2})$. The Hölder inequality then yields

$$\|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}} - h\mathbf{F}(\mathbf{X}_{t_{k-1}}) - \boldsymbol{\xi}_{h,k} + \mathcal{O}(h^{3/2})\|^p \le h^{p-1}\int_{t_{k-1}}^{t_k}\left\|\left(\mathbf{F}\left(\mathbf{X}_s\right) - \mathbf{F}\left(\mathbf{X}_{t_{k-1}}\right)\right)\right\|^p\mathrm{d}s.$$

Use Assumption (A2) together with some standard inequalities and the mean value theorem to get

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

In the last line, we used Lemma 4.1. This proves condition (1) of Theorem 3.3.

Now, we prove condition (2). Use (8) and (14) to write $\mathbf{X}_{t_k}^{[\mathrm{S}]} = \boldsymbol{f}_{h/2}(e^{\mathbf{A}h}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_{k-1}}^{[\mathrm{S}]}) - \mathbf{X}_{t_{k-1}}^{[1]}) + \mathbf{X}_{t_k}^{[1]})$. Define $\mathbf{R}_{t_k} := e^{\mathbf{A}h}(\boldsymbol{f}_{h/2}(\mathbf{X}_{t_k}^{[\mathrm{S}]}) - \mathbf{X}_{t_k}^{[1]})$, and use the associativity (11) to get $\mathbf{R}_{t_k} = e^{\mathbf{A}h}(\boldsymbol{f}_h(\mathbf{R}_{t_{k-1}} + \mathbf{X}_{t_k}^{[1]}) - \mathbf{X}_{t_k}^{[1]})$. The proof of the boundness of the moments of $\mathbf{R}_{t_k}$ is the same as in Lemma 2 in Buckwar et al. (2022). Finally, we have $\mathbf{X}_{t_k}^{[\mathrm{S}]} = \boldsymbol{f}_{h/2}^{-1}(e^{-\mathbf{A}h}\mathbf{R}_{t_k} + \mathbf{X}_{t_k}^{[1]})$. Since $\boldsymbol{f}_{h/2}^{-1}$ grows polynomially and $\mathbf{X}_{t_k}^{[1]}$ has finite moments, $\mathbf{X}_{t_k}^{[\mathrm{S}]}$ must have finite moments too. This concludes the proof.

7.2 Proof of Lemma 4.1

Proof of Lemma 4.1 We first prove (1). We have the following series of inequalities

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

In the second inequality, we used the polynomial growth (A2) of $\mathbf{F}$. Furthermore, for some constant C_2 that depends on p we have $\mathbb{E}\left[\left\|\boldsymbol{\Sigma}\left(\mathbf{W}_t - \mathbf{W}_{t_{k-1}}\right)\right\|^p \mid \mathcal{F}_{t_{k-1}}\right] = (t - t_{t_{k-1}})^{p/2} C_2(p)$. Then, for $h < 1$

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

where constants C_p refer to different constants that depend on p. The last inequality holds because for $t - t_{k-1} < 1$ the power $p/2$ is dominating. Denote $m(t) = \mathbb{E}_{\theta_0}\left[\left\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\right\|^p \mid \mathcal{F}_{t_{k-1}}\right]$, then

$$m(t) \leq C_p\left(t - t_{k-1}\right)^{p/2}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_p} + C_p\int_{t_{k-1}}^{t} m^{C_1}(s)\,\mathrm{d}s. \tag{40}$$

Now, apply Lemma 2.3 from Tian and Fan (2020) on (40). Since we consider super-linear growth, we can assume that $C_1 > 1$, so

$$\begin{aligned} m(t) &\leq C_p\left(t - t_{k-1}\right)^{p/2}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_p} + \left(\kappa^{1-C_1}(t) - (C_1 - 1)2^{C_1-1}C_p\left(t - t_{k-1}\right)\right)^{\frac{1}{1-C_1}} \\ &\leq C_p\left(t - t_{k-1}\right)^{p/2}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_p} + C\kappa(t), \end{aligned} \tag{41}$$

where $\kappa(t) = C_p\left(t - t_{k-1}\right)^{C_1 p/2+1}\left(1 + \|\mathbf{X}_{t_{k-1}}\|\right)^{C_p}$. The bound C in inequality (41) makes sense, because the term

$$\left(1 - (C_1 - 1)2^{C_1-1}C_p\left(t - t_{k-1}\right)\kappa^{\frac{1}{1-C_1}}(t)\right)^{\frac{1}{1-C_1}}$$

is positive by Lemma 2.3 from Tian and Fan (2020). Additionally, it is at most 1 for $t = t_{k-1}$. In constant C in (41) there are terms that depend on $t - t_{k-1}$. However, these terms will not change the dominating term of $\kappa(t)$ (since $h < 1$). Finally, the terms in $\kappa(t)$ are dominated by the power of $p/2$, thus for large enough constant C_p we have $m(t) \leq C_p\left(t - t_{k-1}\right)^{p/2}\left(1 + \|\mathbf{X}_{t_{k-1}}\|\right)^{C_p}$.

To prove (2), use that g is of polynomial growth

$$\begin{aligned} \mathbb{E}_{\theta_0}\left[\left|g\left(\mathbf{X}_t;\boldsymbol{\theta}\right)\right| \mid \mathcal{F}_{t_{k-1}}\right] &\leq C_1\mathbb{E}_{\theta_0}\left[\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\| + \left\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\right\|\right)^{C_1} \mid \mathcal{F}_{t_{k-1}}\right] \\ &\leq C_2\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|^{C_1} + \mathbb{E}_{\theta_0}\left[\left\|\mathbf{X}_t - \mathbf{X}_{t_{k-1}}\right\|^{C_1} \mid \mathcal{F}_{t_{k-1}}\right]\right). \end{aligned}$$

Now, apply the first part of the lemma to get

$$\mathbb{E}_{\theta_0}\left[\left|g\left(\mathbf{X}_t;\boldsymbol{\theta}\right)\right| \mid \mathcal{F}_{t_{k-1}}\right] \leq C_2\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|^{C_1} + C'_{t-t_{k-1}}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_3}\right) \leq C_{t-t_{k-1}}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C}.$$

That concludes the proof.

7.3 Proof of consistency

The following lemma is central to proving consistency and asymptotic normality. The proof is in Supplementary Material.

Lemma 7.1 *Let Assumptions (A1)-(A6) hold and* $\mathbf{X}$ *be the solution of* (1)*. Let* $\mathbf{g}, \mathbf{g}_1, \mathbf{g}_2 : \mathbb{R}^d \times \Theta \times \Theta \to \mathbb{R}$ *be differentiable functions with respect to* $\mathbf{x}$ *and* $\boldsymbol{\theta}$ *with derivatives of polynomial growth in* $\mathbf{x}$*, uniformly in* $\boldsymbol{\theta}$*. If* $h \to 0$ *and* $Nh \to \infty$*, then,*

1. $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} \mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)$;
2. $\frac{h}{N}\sum_{k=1}^{N}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} 0$;
3. $\frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} 0$;
4. $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} 0$;

5. $\frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}}0$;

6. $\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}}\int\mathrm{Tr}\left(D\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right)\mathrm{d}\nu_0(\mathbf{x})$;

7. $\frac{h}{N}\sum_{k=1}^{N}\mathbf{g}_1\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}_2\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}}0$,

uniformly in $\boldsymbol{\theta}$.

Rewrite the objective function (33) as follows

$$\frac{1}{N}\mathcal{L}_N\left(\boldsymbol{\beta},\boldsymbol{\varsigma}\right)=\log\left(\det\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)\right)-\frac{2}{N}\sum_{k=1}^{N}\log\left|\det D\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)\right|+T_1+T_2+T_3+2\left(T_4+T_5+T_6\right). \quad (42)$$

where

$$\begin{aligned}
T_1 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0),\\
T_2 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\left(\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)-\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0\right)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)-\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0\right)\right),\\
T_3 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\left(\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0\right)-\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}\right)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0\right)-\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}\right)\right),\\
T_4 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0\right)-\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}\right)\right),\\
T_5 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\left(\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)-\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0\right)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0\right)-\boldsymbol{\mu}_h\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}\right)\right),\\
T_6 &\coloneqq \frac{1}{Nh}\sum_{k=1}^{N}\left(\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)-\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0).
\end{aligned}$$

Use Proposition 2.2 and Taylor expansion of function $\boldsymbol{\mu}_h$ to approximate T_i, for $1\leq i\leq 6$.

Proof of Theorem 5.1 The proof follows Kessler (1997). First, we prove that

$$\frac{1}{N}\mathcal{L}_N\left(\boldsymbol{\beta},\boldsymbol{\varsigma}\right)\to\log\left(\det\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)\right)+\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)=:G_1\left(\boldsymbol{\varsigma},\boldsymbol{\varsigma}_0\right), \quad (43)$$

in $\mathbb{P}_{\theta_0}$, for $Nh\to\infty$, $h\to 0$, uniformly in $\boldsymbol{\theta}$. The first term of (42) is constant. The second term converges to 0. This follows from derivations (47) below. Properties 1, 2, 3, 5, and 7 from Lemma 7.1 give the following limits $T_1\to\mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$ and for $l=2,3,...,6$, $T_l\to 0$, uniformly in $\boldsymbol{\theta}$. The convergence in probability is equivalent to the existence of a subsequence converging almost surely. Thus, the convergence in (43) is almost sure for a subsequence $(\widehat{\boldsymbol{\beta}}_{N_l},\widehat{\boldsymbol{\varsigma}}_{N_l})$, which implies

$$\widehat{\boldsymbol{\varsigma}}_{N_l}\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}-a.s.}\boldsymbol{\varsigma}_0. \quad (44)$$

The compactness of $\overline{\Theta}$ implies that $(\widehat{\boldsymbol{\beta}}_{N_l},\widehat{\boldsymbol{\varsigma}}_{N_l})$ converges to a limit $(\boldsymbol{\beta}_\infty,\boldsymbol{\varsigma}_\infty)$ almost surely. By continuity of mapping $\boldsymbol{\varsigma}\mapsto G_1(\boldsymbol{\varsigma},\boldsymbol{\varsigma}_0)$ we have $\frac{1}{N_l}\mathcal{L}_{N_l}(\hat{\boldsymbol{\beta}}_{N_l},\widehat{\boldsymbol{\varsigma}}_{N_l})\to G_1(\boldsymbol{\varsigma}_\infty^\top,\boldsymbol{\varsigma}_0)$, in $\mathbb{P}_{\theta_0}$, for $Nh\to\infty$, $h\to 0$, uniformly in $\boldsymbol{\theta}$. By the

definition of the estimator, $G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) \leq G_1(\boldsymbol{\varsigma}_0, \boldsymbol{\varsigma}_0)$. We also have

$$\begin{aligned} G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) \geq G_1(\boldsymbol{\varsigma}_0, \boldsymbol{\varsigma}_0) &\Leftrightarrow \log\left(\det\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top\right)\right) + \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right) \geq \log\left(\det\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\right) + \operatorname{Tr}(\mathbf{I}_d) \\ &\Leftrightarrow \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right) - \log\left(\det\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\right) \geq d \\ &\Leftrightarrow \sum_{i=1}^d \lambda_i - \log\prod_{i=1}^d \lambda_i \geq \sum_{i=1}^d 1 \Leftrightarrow \sum_{i=1}^d (\lambda_i - 1 - \log\lambda_i) \geq 0, \end{aligned}$$

where λ_i represent the eigenvalues of $\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$, which is a positive semi-definite matrix. The last inequality follows since for any positive x, $\log x \leq x - 1$. Thus, $G_1(\boldsymbol{\varsigma}_\infty, \boldsymbol{\varsigma}_0) = G_1(\boldsymbol{\varsigma}_0.\boldsymbol{\varsigma}_0)$. Then, all the eigenvalues λ_i must be equal to 1, hence $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$. We proved that a convergent subsequence of $\widehat{\boldsymbol{\varsigma}}_N$ tends to $\boldsymbol{\varsigma}_0$ almost surely, from there, consistency of the diffusion estimator follows.

For consistency of $\widehat{\boldsymbol{\beta}}_N$, it is sufficient to show in $\mathbb{P}_{\theta_0}$, for $Nh \to \infty$, $h \to 0$, uniformly with respect to $\boldsymbol{\theta}$, it holds

$$\frac{1}{Nh}\left(\mathcal{L}_N(\boldsymbol{\beta}, \boldsymbol{\varsigma}) - \mathcal{L}_N(\boldsymbol{\beta}_0, \boldsymbol{\varsigma})\right) \to G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}), \tag{45}$$

where

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

Rewrite

$$\frac{1}{Nh}\left(\mathcal{L}_N(\boldsymbol{\beta}, \boldsymbol{\varsigma}) - \mathcal{L}_N(\boldsymbol{\beta}_0, \boldsymbol{\varsigma})\right) = \frac{2}{Nh}\sum_{k=1}^N \log\left|\frac{\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)}{\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta})}\right| + \frac{1}{h}\left(T_2 + T_3 + 2\left(T_4 + T_5 + T_6\right)\right).$$

Lemma 4.2 gives the uniform convergence of $\frac{1}{h}T_2$ with respect to $\boldsymbol{\theta}$

$$\begin{aligned} \frac{1}{h}T_2 &= \frac{1}{4N}\sum_{k=1}^N \left(\mathbf{N}_0(\mathbf{X}_{t_k}) - \mathbf{N}(\mathbf{X}_{t_k})\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\mathbf{N}_0(\mathbf{X}_{t_k}) - \mathbf{N}(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h) \\ &\to \frac{1}{4}\int \left(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x})\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x})\right) \mathrm{d}\nu_0(\mathbf{x}). \end{aligned}$$

We compute the limit for $\frac{1}{h}T_3$ analogously. To prove $\frac{1}{h}T_4 \to 0$, we use Lemma 9 in Genon-Catalot and Jacob (1993) and Property 4 from Lemma 7.1. Lemma 4.2 yields

$$\begin{aligned} \frac{1}{h}T_5 \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} &\frac{1}{4}\int \left(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x})\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x})\right) \mathrm{d}\nu_0(\mathbf{x}) \\ &+ \frac{1}{2}\int \left(\mathbf{A}_0\mathbf{x} - \mathbf{A}\mathbf{x}\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x})\right) \mathrm{d}\nu_0(\mathbf{x}). \end{aligned}$$

Finally, Property 6 of Lemma 7.1 gives $\frac{1}{h}T_6 \to \frac{1}{2}\int \operatorname{Tr}(D(\mathbf{N}_0(\mathbf{x}) - \mathbf{N}(\mathbf{x}))^\top \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}) \mathrm{d}\nu_0(\mathbf{x})$ uniformly in $\boldsymbol{\theta}$.

Use properties of the Jacobian and Taylor expansion of a determinant and Lemma 4.2 to get

$$\begin{aligned} &\frac{2}{Nh}\sum_{k=1}^N \log\left|\frac{\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)}{\det D\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k}; \boldsymbol{\beta})}\right| = \frac{2}{Nh}\sum_{k=1}^N \log\left|\det\left(\left(\mathbf{I} + \frac{h}{2}D\mathbf{N}(\mathbf{X}_{t_k})\right)\left(\mathbf{I} - \frac{h}{2}D\mathbf{N}_0(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\right)\right| \\ &= \frac{2}{Nh}\sum_{k=1}^N \log\left|\det\left(\mathbf{I} + \frac{h}{2}D\left(\mathbf{N}(\mathbf{X}_{t_k}) - \mathbf{N}_0(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\right)\right| \qquad (46) \\ &= \frac{2}{Nh}\sum_{k=1}^N \log\left|1 + \frac{h}{2}\operatorname{Tr} D\left(\mathbf{N}(\mathbf{X}_{t_k}) - \mathbf{N}_0(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\right| \\ &= \frac{1}{N}\sum_{k=1}^N \operatorname{Tr} D\left(\mathbf{N}(\mathbf{X}_{t_k}) - \mathbf{N}_0(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h) \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} \int \operatorname{Tr} D\left(\mathbf{N}(\mathbf{x}) - \mathbf{N}_0(\mathbf{x})\right) \mathrm{d}\nu_0(\mathbf{x}), \qquad (47) \end{aligned}$$

uniform in $\boldsymbol{\theta}$. This proves (45). Then, there exists a subsequence N_l such that $(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l})$ converges to a limit $(\boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty)$ almost surely. By continuity of mapping $(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \mapsto G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma})$, for $N_l h \to \infty$, $h \to 0$, we have the following convergence in $\mathbb{P}_{\theta_0}$

$$\frac{1}{N_l h}\left(\mathcal{L}_{N_l}\left(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l}\right) - \mathcal{L}_{N_l}\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\varsigma}}_{N_l}\right)\right) \to G_2\left(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty\right).$$

Then, $G_2(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}_0, \boldsymbol{\beta}_\infty, \boldsymbol{\varsigma}_\infty) = \int (\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_0) - \mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_\infty))^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1} (\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_0) - \mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_\infty)) \, \mathrm{d}\nu_0(\mathbf{x}) \geq 0$, since $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\infty^\top = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top$. Moreover, $\mathcal{L}_{N_l}(\widehat{\boldsymbol{\beta}}_{N_l}, \widehat{\boldsymbol{\varsigma}}_{N_l}) - \mathcal{L}_{N_l}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\varsigma}}_{N_l}) \leq 0$, by definition of the estimator. Thus, the identifiability assumption (A5) concludes the proof for the S estimator.

To prove the same statement for the LT estimator, the representation of the objective function (42) has to be adapted. In the LT case, this representation is simpler. There is no extra logarithmic term of a Jacobian, and there are only 3 instead of 6 auxiliary T terms. This is due to the Gaussian transition density in the LT approximation.

7.4 Proofs of asymptotic normality

Proof of Theorem 5.2 According to Sørensen and Uchida (2003), it is enough to prove the following two lemmas.

Lemma 7.2 *If $h \to 0$, $Nh \to \infty$ and $\varepsilon_N \to 0$, then*

$$\mathbf{C}_N\left(\boldsymbol{\theta}_0\right) \xrightarrow{\mathbb{P}_{\theta_0}} 2\mathbf{C}\left(\boldsymbol{\theta}_0\right), \qquad \sup_{\|\boldsymbol{\theta}\| \leq \varepsilon_N} \left\|\mathbf{C}_N\left(\boldsymbol{\theta}_0 + \boldsymbol{\theta}\right) - \mathbf{C}_N\left(\boldsymbol{\theta}_0\right)\right\| \xrightarrow{\mathbb{P}_{\theta_0}} 0.$$

Lemma 7.3 *If $h \to 0$, $Nh \to \infty$ and $Nh^2 \to 0$, then under $\mathbb{P}_{\theta_0}$, we have*

$$\mathbf{L}_N \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\mathbf{C}\left(\boldsymbol{\theta}_0\right)\right).$$

Proof of Lemma 7.2 To prove the first part of the lemma, use equation (42) and compute corresponding derivatives to obtain $\mathbf{C}_N$. We start with

$$\begin{aligned}\frac{1}{Nh}\partial_{\beta_{i_1}\beta_{i_2}}\mathcal{L}_N\left(\boldsymbol{\beta}, \boldsymbol{\varsigma}\right) = -2\frac{1}{Nh}\sum_{k=1}^{N}\partial_{\beta_{i_1}\beta_{i_2}}\log\left|\det D\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}\right)\right| \\ + \partial_{\beta_{i_1}\beta_{i_2}}\frac{1}{h}\Big(T_2\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}\right) + T_3\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}\right) + 2\left(T_4\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}\right) + T_5\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}\right) + T_6\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}\right)\right)\Big).\end{aligned}$$

Use results from Lemma 7.1 and Theorem 5.1 to get the following limits

$$\begin{aligned}\partial_{\beta_{i_1}\beta_{i_2}}\frac{1}{h}T_2\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &\xrightarrow{\mathbb{P}_{\theta_0}} \frac{1}{2}\int\left(\partial_{\beta_{i_1}}\mathbf{N}_0\left(\mathbf{x}\right)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\partial_{\beta_{i_2}}\mathbf{N}_0\left(\mathbf{x}\right)\mathrm{d}\nu_0(\mathbf{x}), \\ \partial_{\beta_{i_1}\beta_{i_2}}\frac{1}{h}T_3\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &\xrightarrow{\mathbb{P}_{\theta_0}} \frac{1}{2}\int\left(\partial_{\beta_{i_1}}\mathbf{N}_0\left(\mathbf{x}\right) + 2\partial_{\beta_{i_1}}\mathbf{A}_0\mathbf{x}\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\left(\partial_{\beta_{i_2}}\mathbf{N}_0\left(\mathbf{x}\right) + 2\partial_{\beta_{i_2}}\mathbf{A}_0\mathbf{x}\right)\mathrm{d}\nu_0(\mathbf{x}), \\ \partial_{\beta_{i_1}\beta_{i_2}}\frac{1}{h}T_5\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &\xrightarrow{\mathbb{P}_{\theta_0}} \frac{1}{2}\int\left(\partial_{\beta_{i_1}}\mathbf{F}_0\left(\mathbf{x}\right)\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\partial_{\beta_{i_2}}\mathbf{N}_0\left(\mathbf{x}\right)\mathrm{d}\nu_0(\mathbf{x}) \\ &\quad + \frac{1}{2}\int\left(\partial_{\beta_{i_2}}\mathbf{A}_0\mathbf{x}\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\partial_{\beta_{i_1}}\mathbf{N}_0\left(\mathbf{x}\right)\mathrm{d}\nu_0(\mathbf{x}), \\ \partial_{\beta_{i_1}\beta_{i_2}}\frac{1}{h}T_6\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\varsigma}_0\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} &\xrightarrow{\mathbb{P}_{\theta_0}} -\frac{1}{2}\int\mathrm{Tr}(D\partial_{\beta_{i_1}\beta_{i_2}}\mathbf{N}_0(\mathbf{x}))\,\mathrm{d}\nu_0(\mathbf{x}),\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. Since $\frac{1}{h}T_4 \to 0$, the partial derivatives go to zero too. Like in (47), for $Nh \to \infty$, $h \to 0$, we have

$$-2\frac{1}{Nh}\sum_{k=1}^{N}\partial_{\beta_{i_1}\beta_{i_2}}\log\left|\det D\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}\right)\right|\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{\mathbb{P}_{\theta_0}} \int\mathrm{Tr}\left(D\partial_{\beta_{i_1}\beta_{i_2}}\mathbf{N}_0\left(\mathbf{x}\right)\right)\mathrm{d}\nu_0(\mathbf{x}).$$

Thus, $\frac{1}{Nh}\partial_{\beta_{i_1}\beta_{i_2}}\mathcal{L}_N(\boldsymbol{\beta}, \boldsymbol{\varsigma}_0)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \to 2\int(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_0))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{x}; \boldsymbol{\beta}_0)\,\mathrm{d}\nu_0(\mathbf{x})$, in $\mathbb{P}_{\theta_0}$ for $Nh \to \infty$, $h \to 0$.

Now, we prove $\frac{1}{\sqrt{Nh}}\partial_{\beta\varsigma}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0} \to 0$, in $\mathbb{P}_{\theta_0}$ for $Nh\to\infty$, $h\to 0$. The following term is at most of the order $\mathcal{O}(h)$

$$\partial_{\beta_i} T_l(\boldsymbol{\beta},\boldsymbol{\varsigma}) = C_h \sum_{k=1}^{N} \left(\mathbf{g}\left(\boldsymbol{\beta}_0;\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}\right) - \mathbf{g}\left(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}\right)\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}_1\left(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}\right),$$

for $l=2,3,...,6$, where C_h is a constant depending on h, and $\mathbf{g},\mathbf{g}_1$ are generic functions. Then, term $\partial_{\beta\varsigma}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})$ still contains $\mathbf{g}(\boldsymbol{\beta}_0;\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}}) - \mathbf{g}(\boldsymbol{\beta};\mathbf{X}_{t_k},\mathbf{X}_{t_{k-1}})$ which is 0 for $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. Thus, $\frac{1}{\sqrt{Nh}}\partial_{\beta\varsigma}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0} = 0$. Finally, we compute $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})$. As before, it holds $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}} T_l(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0} \to 0$, for $l=2,3,...,6$. So, we need to compute the following second derivatives $\partial_{\varsigma_{j_1}\varsigma_{j_2}}\log(\det\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)$ and $\partial_{\varsigma_{j_1}\varsigma_{j_2}}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$. The first one yields

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

On the other hand, we have

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

Then, from Property 1 of Lemma 7.1, we get

$$\begin{aligned}
&\partial_{\varsigma_{j_1}\varsigma_{j_2}}\frac{1}{Nh}\sum_{k=1}^{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0}\\
&\xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}} 2\,\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\left(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right) - \mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right).
\end{aligned}$$

Thus, $\frac{1}{N}\partial_{\varsigma_{j_1}\varsigma_{j_2}}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\boldsymbol{\varsigma}=\boldsymbol{\varsigma}_0} \to \mathrm{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$. Since all the limits used in this proof are uniform in $\boldsymbol{\theta}$, the first part of the lemma is proved. The second part is trivial, due to the fact that all limits are continuous in $\boldsymbol{\theta}$.

Proof of Lemma 7.3 First, we compute the first derivatives. We start with

$$\begin{aligned}
\partial_{\beta_i}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma}) = \frac{2}{h}\sum_{k=1}^{N}&\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\partial_{\beta_i}\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \partial_{\beta_i}\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})\right)\\
&- 2\sum_{k=1}^{N}\mathrm{Tr}\left(D\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)\right).
\end{aligned}$$

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

$$\begin{aligned}
&\partial_{\varsigma_j}\mathcal{L}_N(\boldsymbol{\beta},\boldsymbol{\varsigma})\\
&= \frac{1}{h}\partial_{\varsigma_j}\sum_{k=1}^{N}\left((\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}))^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})) + \partial_{\varsigma_j}\log\det(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\right)\\
&= -\frac{1}{h}\sum_{k=1}^{N}\Bigg(\mathrm{Tr}\left(\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})\right)\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})\right)^\top\right.\\
&\qquad\left.\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right) + \mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)\Bigg).
\end{aligned}$$

Introduce

$$\eta_{N,k}^{(i)}(\boldsymbol{\theta}) \coloneqq \frac{2}{\sqrt{Nh}}\left(\operatorname{Tr}\left(D\boldsymbol{f}_{h/2}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right) D_{\mathbf{x}}\partial_{\beta_i}\boldsymbol{f}_{h/2}^{-1}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)\right)\right.$$
$$\left. -\frac{1}{h}\mathbf{Z}_{t_k}(\boldsymbol{\beta})^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)^{-1}\partial_{\beta_i}\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta})-\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta})\right)\right)$$
$$\zeta_{N,k}^{(j)}(\boldsymbol{\theta}) \coloneqq \frac{1}{\sqrt{N}}\left(\frac{1}{h}\operatorname{Tr}\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta})\mathbf{Z}_{t_k}(\boldsymbol{\beta})^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)^{-1}\left(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)^{-1}\right)-\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)^{-1}\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\right)\right),$$

and rewrite $\mathbf{L}_N$ as $\mathbf{L}_N = \sum_{k=1}^{N}[\eta_{N,k}^{(1)}(\boldsymbol{\theta}_0),\ldots,\eta_{N,k}^{(r)}(\boldsymbol{\theta}_0),\zeta_{N,k}^{(1)}(\boldsymbol{\theta}_0),\ldots,\zeta_{N,k}^{(s)}(\boldsymbol{\theta}_0)]^{\top}$. Now, according to Proposition 3.1 from Crimaldi and Pratelli (2005) (for more details see Supplementary Material), it is sufficient to prove

$$\mathbb{E}_{\theta_0}\left[\sup_{1\le k\le n}\left|\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0)\right|\right]\xrightarrow[n\to\infty]{} 0, \qquad \mathbb{E}_{\theta_0}\left[\sup_{1\le k\le n}\left|\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0)\right|\right]\xrightarrow[n\to\infty]{} 0, \tag{48}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \qquad \sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{49}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{50}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_1)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{51}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{52}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_1)}(\boldsymbol{\theta}_0)\,\eta_{n,k}^{(i_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 4\left[\mathbf{C}_{\beta}(\boldsymbol{\theta}_0)\right]_{i_1 i_2}, \tag{53}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_1)}(\boldsymbol{\theta}_0)\,\zeta_{n,k}^{(j_2)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 4\left[\mathbf{C}_{\varsigma}(\boldsymbol{\theta}_0)\right]_{j_1 j_2}, \tag{54}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0)\,\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{55}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\left(\eta_{n,k}^{(i_1)}(\boldsymbol{\theta}_0)\,\eta_{n,k}^{(i_2)}(\boldsymbol{\theta}_0)\right)^2\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{56}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\left(\zeta_{n,k}^{(j_1)}(\boldsymbol{\theta}_0)\,\zeta_{n,k}^{(j_2)}(\boldsymbol{\theta}_0)\right)^2\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0, \tag{57}$$

$$\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\left(\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0)\,\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0)^2\right)\mid\mathbf{X}_{t_{k-1}}\right]\xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0,, \tag{58}$$

for all $i, i_1, i_2 = 1, 2, ..., r$ and $j, j_1, j_2 = 1, 2, ..., s$. The proof of the previous limits is quite technical and is shown in Supplementary Material.

8 Conclusion

We proposed two new estimators for nonlinear multivariate SDEs. These estimators are based on splitting schemes, a numerical approximation that preserves all important properties of the model. It was known that the LT-splitting scheme has L^p convergence rate of order 1. We proved that the same holds for the S-splitting. This is 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 $\Phi_{h/2}^{[2]}$. However, S-splitting is more precise in one-step predictions. This is crucial because the transition densities between two consecutive data points entering the likelihood are therefore better approximated for the S estimator. Since S one-step prediction has one order more compared to LT, the obtained estimator is less biased.

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

In the simulation study of the stochastic Lorenz system, we showed the superiority of the S estimators. We compared five estimators based on different discretization schemes. Estimators based on Ozaki's LL and the S-splitting schemes performed the best in terms of precision. However, the running time of LL is greatly dependent on the sample size N, which is not the case for the S estimator. This property makes the S estimator a more appropriate choice for large sample sizes. We showed that the LT estimator does not perform well in real examples, despite its asymptotic properties. The EM and K estimators perform well for small h, but for $h = 0.01$ the bias is large, especially for the diffusion parameters in the EM case.

While proposed estimators cover a wide range of models, we still have some restrictive assumptions such as additive noise. However, this can be relaxed if the Lamperti transformation can be applied. Moreover, we assume equidistant observations. This can easily be relaxed due to the continuous-time formulation. Finally, we assumed that the diffusion parameter $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is invertible. Sometimes, models with degenerate noise naturally arise in applications, for example, second-order differential equations. If the covariance matrices have components with different time scales, they are called hypoelliptic models. These will be thoroughly investigated in another paper, where the proofs are more involved.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)"; Novo Nordisk Foundation NNF20OC0062958; and Independent Research Fund Denmark | Natural Sciences 9040-00215B.

References

Peter Ditlevsen and Susanne Ditlevsen. Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Preprint at Research Square*, pages 1–15, 2022. URL https://www.researchsquare.com/article/rs-2034845/v1.

M. Arnst, G. Louppe, R. Van Hulle, L. Gillet, F. Bureau, and V. Denoël. A hybrid stochastic model and its Bayesian identification for infectious disease screening in a university campus with application to massive COVID-19 screening at the University of Liège. *Mathematical Biosciences*, 347:108805, 2022. ISSN 0025-5564. doi:https://doi.org/10.1016/j.mbs.2022.108805. URL https://www.sciencedirect.com/science/article/pii/S0025556422000219.

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

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

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

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

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

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

P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1992. ISBN 9783540540625. doi:10.1007/978-3-662-12616-5. URL https://books.google.dk/books?id=BCvtssom1CMC.

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

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

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

Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2130):1563–1576, 2011. doi:10.1098/rspa.2010.0348. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2010.0348.

Evelyn Buckwar, Adeline Samson, Massimiliano Tamborrino, and Irene Tubikanec. A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *Applied Numerical Mathematics*, 179:191–220, 2022. ISSN 0168-9274. doi:https://doi.org/10.1016/j.apnum.2022.04.018. URL https://www.sciencedirect.com/science/article/pii/S0168927422001118.

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

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

Valentine Genon-Catalot and Jean Jacob. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'I.H.P. Probabilités et statistiques*, 29(1):119–151, 1993. URL http://www.numdam.org/item/AIHPB_1993__29_1_119_0/.

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

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

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

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

Masayuki Uchida and Nakahiro Yoshida. Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications*, 122(8):2885–2924, 2012. ISSN 0304-4149. doi:https://doi.org/10.1016/j.spa.2012.04.001. URL https://www.sciencedirect.com/science/article/pii/S0304414912000622.

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

Julie Lyng Forman and Michael Sørensen. The Pearson diffusions: A class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, 35(3):438–465, 2008.

Susanne Ditlevsen and Adeline Samson. Hypoelliptic diffusions: filtering and inference from complete and partial observations, 2019. ISSN 1369-7412.

Tohru Ozaki. Statistical Identification of Storage Models with Application to Stochastic Hydrology. *Journal of The American Water Resources Association*, 21:663–675, 1985.

Isao Shoji and Tohru Ozaki. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4):733–752, 1998. doi:10.1080/07362999808809559. URL https://doi.org/10.1080/07362999808809559.

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

Sergio Blanes, Fernando Casas, and Ander Murua. Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Mat. Apl.*, 45, 01 2009.

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

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

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

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

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

Yacine Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906 – 937, 2008. doi:10.1214/009053607000000622. URL https://doi.org/10.1214/009053607000000622.

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

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

A. R. Humphries and A. M. Stuart. *Deterministic and random dynamical systems: theory and numerics*, pages 211–254. Springer Netherlands, Dordrecht, 2002. ISBN 978-94-010-0510-4. URL https://doi.org/10.1007/978-94-010-0510-4_6.

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

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

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

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

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

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

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

Li Zhuang, Longpeng Cao, Yong Wu, Yonghong Zhong, Lili Zhangzhong, Wengang Zheng, and Long Wang. Parameter Estimation of Lorenz Chaotic System Based on a Hybrid Jaya-Powell Algorithm. *IEEE Access*, 8:20514–20522, 2020. doi:10.1109/ACCESS.2020.2968106.

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

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

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

T. Ozaki, J. C. Jimenez, and V. Haggan-Ozaki. The Role of the Likelihood Function in the Estimation of Chaos Models. *Journal of Time Series Analysis*, 21(4):363–387, 2000. doi:https://doi.org/10.1111/1467-9892.00189. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00189.

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

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

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

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

Atılım Günes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic Differentiation in Machine Learning: A Survey. *J. Mach. Learn. Res.*, 18(1):5595–5637, jan 2017. ISSN 1532-4435.

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

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

Nakahiro Yoshida. Asymptotic behavior of M-estimator and related random field for diffusion process. *Annals of the Institute of Statistical Mathematics*, 42(2):221–251, June 1990. doi:10.1007/BF00050834. URL https://ideas.repec.org/a/spr/aistmt/v42y1990i2p221-251.html.

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

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

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

S1 Supplementary Material

This section gives proofs of all the propositions, lemmas, and theorems. If not differently stated, we assume the parameters are the true ones $\boldsymbol{\theta}_0$ and the expectations are taken under the probability measure $\mathbb{P}_{\theta_0}$.

S1.1 Proof for the Lie-Trotter splitting

Proof of Proposition 3.4 First, we prove the proposition for splitting (13). Taylor-expansion of $\mathbb{E}[\Phi_h^{[\mathrm{LT}]}(\mathbf{x})] = e^{\mathbf{A}h}\boldsymbol{f}_h(\mathbf{x})$ as a function of h around $h = 0$ using (12) yields

$$e^{\mathbf{A}h}\boldsymbol{f}_h(\mathbf{x}) = \mathbf{x} + h(\mathbf{A}\mathbf{x} + \mathbf{N}(\mathbf{x})) + \frac{h^2}{2}\left(\mathbf{A}^2\mathbf{x} + 2\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\,\mathbf{N}(\mathbf{x})\right) + \mathcal{O}(h^3). \tag{S1}$$

The coefficient of h equals $\mathbf{F}(\mathbf{x})$, which coincides with the coefficient of h of the theoretical moment of the solution of (1) given in (5). In (5), $\boldsymbol{\Sigma}$ appears in the coefficient of h^2, however, it does not appear in (S1). Thus, to obtain the order of convergence $\mathcal{O}(h^3)$, we need the following unrealistic assumption.

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

Comparing (S1) and (5) under Assumption (A*) we need $\mathbf{A}\mathbf{N}(\mathbf{x}) = (D\mathbf{N}(\mathbf{x}))\,\mathbf{A}\mathbf{x}$ to obtain equality of the coefficient of h^2. This only holds for all $\mathbf{x} \in \mathbb{R}^d$ when $\mathbf{N}$ is linear. Thus order $\mathcal{O}(h^3)$ one-step convergence is only possible for LT if SDE (1) is linear.

Now, define the reversed Lie-Trotter splitting

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

We need to compute $\mathbb{E}[\boldsymbol{f}_h(e^{\mathbf{A}h}\mathbf{X}_{t_{k-1}} + \boldsymbol{\xi}_{h,k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}]$, which is equivalent to computing $\mathbb{E}[\boldsymbol{f}_h(\mathbf{X}_{t_k}^{[1]}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}] = \mathbb{E}[\boldsymbol{f}_h(e^{\mathbf{A}h}\mathbf{X}_{t_{k-1}}^{[1]} + \boldsymbol{\xi}_{h,k}) \mid \mathbf{X}_{t_{k-1}}^{[1]} = \mathbf{x}]$. The infinitesimal generator $L_{[1]}$ for SDE (6) is defined on the class of sufficiently smooth functions $g : \mathbb{R}^d \to \mathbb{R}$ by $L_{[1]}g(\mathbf{x}) = (\mathbf{A}\mathbf{x})^\top \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_g(\mathbf{x})\right)$. This yields

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

We apply (S2) on $g(\mathbf{x}) = f_h^{(i)}(\mathbf{x})$. To compute $L_{[1]}f_h^{(i)}(\mathbf{x})$ and $L_{[1]}^2 f_h^{(i)}(\mathbf{x})$, we use the Taylor expansion of $\boldsymbol{f}_h(\mathbf{x})$ around $h = 0$, given in (12). The partial derivatives are $\partial_j f_h^{(i)}(\mathbf{x}) = \delta_j^i + h\partial_j N^{(i)}(\mathbf{x}) + \mathcal{O}(h^2)$ and $\partial_{jk}^2 f_h^{(i)}(\mathbf{x}) = h\partial_{jk}^2 N^{(i)}(\mathbf{x}) + \mathcal{O}(h^2)$. We only need to calculate $L_{[1]}f_h^{(i)}(\mathbf{x})$ up to order $\mathcal{O}(h)$ because it is multiplied by h; $L_{[1]}f_h^{(i)}(\mathbf{x}) = (\mathbf{A}\mathbf{x})^{(i)} + h\,(\mathbf{A}\mathbf{x})^\top \nabla N^{(i)}(\mathbf{x}) + \frac{h}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})\right) + \mathcal{O}(h^2)$. Likewise, $L_{[1]}^2 f_h^{(i)}(\mathbf{x}) = (\mathbf{A}\mathbf{x})^\top \nabla(\mathbf{A}\mathbf{x})^{(i)} + \mathcal{O}(h) = \mathbf{A}^{(i)}\mathbf{A}\mathbf{x} + \mathcal{O}(h)$. Thus

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

Using that $F^{(i)}(\mathbf{x}) = (\mathbf{A}\mathbf{x})^{(i)} + N^{(i)}(\mathbf{x})$, $\frac{\partial F^{(i)}(\mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A}^{(i)})^\top + \nabla N^{(i)}(\mathbf{x})$ and $\mathbf{H}_{F^{(i)}}(\mathbf{x}) = \mathbf{H}_{N^{(i)}}(\mathbf{x})$, the expectation of the true process (5) rewrites as

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

In conclusion, the last equation coincides with equation (S3) only up to order $\mathcal{O}(h)$. To obtain order $\mathcal{O}(h^2)$, $(\mathbf{N}(\mathbf{x}))^\top \nabla F^{(i)}(\mathbf{x}) - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top \mathbf{H}_{N^{(i)}}(\mathbf{x})\right) = (\mathbf{F}(\mathbf{x}))^\top \nabla N^{(i)}(\mathbf{x})$, for all $i = 1, \dots, d$ should hold.

S1.2 Proof for the Strang Splitting

Proof of Proposition 3.6 Let $\mathbf{Q}_h(\mathbf{x}) \coloneqq \frac{h}{2}(2\mathbf{A}\mathbf{x} + \mathbf{N}(\mathbf{x})) + \frac{h^2}{8}(4\mathbf{A}^2\mathbf{x} + 4\mathbf{A}\mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{N}(\mathbf{x}))$. Then, use Proposition 2.2 to rewrite

$$\begin{aligned}\boldsymbol{f}_{h/2}\left(e^{\mathbf{A}h}\boldsymbol{f}_{h/2}(\mathbf{X}) + \boldsymbol{\xi}_h\right) &= \boldsymbol{f}_{h/2}\left(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \mathcal{O}(h^3)\right) = \mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h + \frac{h}{2}\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \\ &\quad + \frac{h^2}{8}(D\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h))\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) + \mathcal{O}_{\mathbb{P}}(h^3). \end{aligned} \tag{S4}$$

Then,

$$\begin{aligned}\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) &= \mathbf{N}(\mathbf{X}) + (D\mathbf{N}(\mathbf{X}))(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \\ &\quad + \frac{1}{2}\Big[(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})(\mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h)\Big]_{i=1}^d + \mathcal{O}_{\mathbb{P}}(h^2) \\ &= \mathbf{N}(\mathbf{X}) + (D\mathbf{N}(\mathbf{X}))\mathbf{Q}_h(\mathbf{X}) + (D\mathbf{N}(\mathbf{X}))\boldsymbol{\xi}_h + \frac{1}{2}\Big[\mathbf{Q}_h(\mathbf{X})^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\mathbf{Q}_h(\mathbf{X})\Big]_{i=1}^d \\ &\quad + \frac{1}{2}\Big[\boldsymbol{\xi}_h^\top(\mathbf{H}_{N^{(i)}}(\mathbf{X}))\boldsymbol{\xi}_h\Big]_{i=1}^d + \mathcal{O}_{\mathbb{P}}(h^2). \end{aligned} \tag{S5}$$

The term $\Big[\mathbf{Q}_h(\mathbf{X})^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\mathbf{Q}_h(\mathbf{X})\Big]_{i=1}^d$ is $\mathcal{O}(h^2)$. Terms with only one $\boldsymbol{\xi}_h$ have zero mean. Thus,

$$\mathbb{E}\left[\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}\right] = \mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x}) + \frac{1}{2}\Big[\mathbb{E}\left[\boldsymbol{\xi}_h^\top \mathbf{H}_{N^{(i)}}(\mathbf{X})\boldsymbol{\xi}_h \mid \mathbf{X} = \mathbf{x}\right]\Big]_{i=1}^d + \mathcal{O}(h^2). \tag{S6}$$

Lastly, we compute

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

We use the approximation of the variance of the random vector $\boldsymbol{\xi}_h$ from equation (9) to get $\mathbb{E}[\mathbf{N}(\mathbf{X} + \mathbf{Q}_h(\mathbf{X}) + \boldsymbol{\xi}_h) \mid \mathbf{X} = \mathbf{x}] = \mathbf{N}(\mathbf{x}) + (D\mathbf{N}(\mathbf{x}))\mathbf{Q}_h(\mathbf{x}) + \frac{h}{2}[\sum_{j,k=1}^d [\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top]_{jk}\partial_{jk}^2 F^{(i)}(\mathbf{x})]_{i=1}^d + \mathcal{O}(h^2)$. Taking the expectation of (S4) and using the previous equation conclude the proof.

S1.3 Proofs for Moment Bounds

Before proving moment bounds, we need some auxiliary properties of the infinitesimal generator L.

Lemma S1.1 *Let L be the infinitesimal generator given by* (4) *of SDE* (1)*. For sufficiently smooth functions $\alpha, \beta : \mathbb{R}^d \to \mathbb{R}$*

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

Proof We use the operator L and the product rule to get

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

This concludes the proof.

We add one more auxiliary lemma regarding the mean function $\boldsymbol{\mu}_h$.

Lemma S1.2 *For the mean function $\boldsymbol{\mu}_h$ we have the following three identities*

1. $\boldsymbol{\mu}_h(\mathbf{x}) = \boldsymbol{f}_{h/2}(\mathbf{x}) + h\mathbf{A}\mathbf{x} + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathcal{O}(h^3)$
2. $\boldsymbol{\mu}_h(\mathbf{x}) = \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathcal{O}(h^3)$.
3. $\boldsymbol{\mu}_h(\mathbf{x}) = \mathbf{x} + h\mathbf{A}\mathbf{x} + \frac{h}{2}\mathbf{N}(\mathbf{x}) + \mathcal{O}(h^2)$.

Proof We prove only the first two identities. The last one is a direct consequence. We use definition of $\boldsymbol{\mu}_h$, Taylor expansion, and expansion of $\boldsymbol{f}_{h/2}$ to obtain $\boldsymbol{\mu}_h(\mathbf{x}) = \left(\mathbf{I} + h\mathbf{A} + \frac{h^2}{2}\mathbf{A}^2\right)\boldsymbol{f}_{h/2}(\mathbf{x}) + \mathcal{O}(h^3) = \boldsymbol{f}_{h/2}(\mathbf{x}) + h\mathbf{A}\mathbf{x} + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathcal{O}(h^3)$, which concludes the first part.

For the second part, formula (12) gives $\boldsymbol{f}_{h/2}(\mathbf{x}) - \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) = h\mathbf{N}(\mathbf{x}) + \mathcal{O}(h^3)$. Then we get $\boldsymbol{\mu}_h(\mathbf{x}) = \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + h\mathbf{F}(\mathbf{x}) + \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathcal{O}(h^3)$. This concludes the proof.

Proof of Proposition 4.3 Proof of (1). Lemma S1.2 yields

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] &= \mathbb{E}\left[\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] - \boldsymbol{\mu}_h(\mathbf{x}) \\ &= \mathbb{E}\left[\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] - \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) - h\mathbf{F}(\mathbf{x}) \\ &\quad - \frac{h^2}{2}\mathbf{A}\mathbf{F}(\mathbf{x}) + \mathcal{O}(h^3).\end{aligned}$$

Now, use the infinitesimal generator L to find the expectation in the last line. Here, we abuse the notation and apply the generator L directly to a vector-valued function, instead of applying it on each coordinate. We have

$$\mathbb{E}\left[\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] = \boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + hL\boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + \frac{h^2}{2}L^2\boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) + \mathcal{O}(h^3).$$

Use that $\boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) = \boldsymbol{f}_{-h/2}(\mathbf{x})$ and expansion (12) to get

$$\begin{aligned}L\boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) &= L\mathbf{x} - \frac{h}{2}L\mathbf{N}(\mathbf{x}) + \mathcal{O}(h^2) = \mathbf{F}(\mathbf{x}) - \frac{h}{2}L\mathbf{N}(\mathbf{x}) + \mathcal{O}(h^2), \\ L^2\boldsymbol{f}^{-1}_{h/2}(\mathbf{x}) &= L\mathbf{A}\mathbf{x} + L\mathbf{N}(\mathbf{x}) + \mathcal{O}(h) = \mathbf{A}\mathbf{F}(\mathbf{x}) + L\mathbf{N}(\mathbf{x}) + \mathcal{O}(h).\end{aligned}$$

Now, it is clear that $\mathbb{E}\left[\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k}) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] = \mathcal{O}(h^3)$.

Proof of (2). Here, parameters are specified. Start with expansions of $\boldsymbol{f}_h^{-1}$ and $\boldsymbol{\mu}_h$

$$\begin{aligned}&\mathbb{E}_{\theta_0}\left[\left(\boldsymbol{f}^{-1}_{h/2}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\right)\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] \\ &= \mathbb{E}_{\theta_0}\left[\mathbf{X}_{t_k}\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] - \frac{h}{2}\mathbb{E}_{\theta_0}\left[\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] \\ &- \mathbf{x}\mathbb{E}_{\theta_0}\left[\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] - \frac{h}{2}\left(2\mathbf{A}^0\mathbf{x} + \mathbf{N}_0\left(\mathbf{x}\right)\right)\mathbb{E}_{\theta_0}\left[\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] + \mathcal{O}(h^2) \\ &= \mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + hL_{\theta_0}\left(\mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top\right) - \frac{h}{2}\mathbf{N}_0\left(\mathbf{x}\right)\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top \\ &- \mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top - h\mathbf{x}L_{\theta_0}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top - h\mathbf{A}^0\mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top - \frac{h}{2}\mathbf{N}_0\left(\mathbf{x}\right)\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \mathcal{O}(h^2) \\ &= hL_{\theta_0}\left(\mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top\right) - h\mathbf{x}L_{\theta_0}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top - h\mathbf{F}\left(\mathbf{x};\boldsymbol{\beta}_0\right)\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \mathcal{O}(h^2).\end{aligned}$$

Now, we use Lemma S1.1 and equation (4) to get

$$\begin{aligned}L_{\theta_0}\left(\mathbf{x}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top\right) &= \mathbf{x}L_{\theta_0}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \left(L_{\theta_0}\mathbf{x}\right)\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \frac{1}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right) + D\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right) \\ &= \mathbf{x}L_{\theta_0}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \mathbf{F}(\mathbf{x};\boldsymbol{\beta}_0)\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)^\top + \frac{1}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right) + D\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right),\end{aligned}$$

which concludes the proof of (2).

Proof of (3). We use the previous property. Introduce $\mathbf{g}\left(\mathbf{X}_{t_k};\boldsymbol{\beta}_0\right)=\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)$, then,

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)-\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\right)\left(\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)-\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\right)^{\top}\mid \mathbf{X}_{t_{k-1}}=\mathbf{x}\right]\\
&=\frac{h}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}D^{\top}\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}_0\right)+D\mathbf{g}\left(\mathbf{x};\boldsymbol{\beta}_0\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\\
&-\mathbb{E}_{\theta_0}\left[\boldsymbol{f}_{h/2}^{-1}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0)-\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\mid \mathbf{X}_{t_{k-1}}=\mathbf{x}\right]\boldsymbol{\mu}_h(\mathbf{x};\boldsymbol{\beta}_0)^{\top}+\mathcal{O}(h^2).
\end{aligned}$$

Using the first property of this proposition together with $D\mathbf{g}(\mathbf{x};\boldsymbol{\beta}_0)=\mathbf{I}+\mathcal{O}(h)$, the result follows.

S1.4 Auxiliary properties

Here we list some helpful properties needed to prove the consistency and asymptotic normality of the estimator. First, we state Lemma 2.3 from Tian and Fan (2020) needed for the proof of 4.1. This lemma generalizes Grönwall's inequality.

Lemma S1.3 *Let $p>1$ and $b>0$ be constants, and let $a:(0,+\infty)\to(0,+\infty)$ be a continuous function. If*

$$u(t)\leq a(t)+b\int_0^t u^p(s)\,\mathrm{d}s,$$

then $u(t)\leq a(t)+(\kappa^{1-p}(t)-(p-1)2^{p-1}bt)^{\frac{1}{1-p}}$ and $\kappa^{1-p}(t)>(p-1)2^{p-1}bt$, where

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

Lemma 9 in Genon-Catalot and Jacob (1993) provides conditions for the converging of a sum of a triangular array.

Lemma S1.4 *Let $\left(X_k^N\right)_{N\in\mathbb{N},1\leq k\leq N}$ be a triangular array with each row N adapted to a filtration $\left(\mathcal{G}_k^N\right)_{1\leq k\leq N}$, and let U be a random variable. If*

$$\sum_{k=1}^N\mathbb{E}\left[X_k^N\mid\mathcal{G}_{k-1}^N\right]\xrightarrow[n\to\infty]{\mathbb{P}}U,\qquad\qquad\sum_{k=1}^N\mathbb{E}\left[\left(X_k^N\right)^2\mid\mathcal{G}_{k-1}^N\right]\xrightarrow[n\to\infty]{\mathbb{P}}0,$$

then $\sum_{k=1}^N X_k^N\xrightarrow[n\to\infty]{\mathbb{P}}U$.

The two following lemmas give sufficient conditions for uniform convergence. The first one is Proposition A1 in Gloter (2006), and the second one is Lemma 3.1 from Yoshida (1990).

Lemma S1.5 *Let $S_N(\omega,\boldsymbol{\theta})$ be a sequence of measurable real-valued functions defined on $\Omega\times\Theta$, where $(\Omega,\mathcal{F},\mathbb{P})$ is a probability space, and Θ is product of compact intervals of $\mathbb{R}$. We assume that $S_N(\cdot,\boldsymbol{\theta})$ converges to a constant C in probability for all $\boldsymbol{\theta}\in\Theta$; and that there exists an open neighbourhood of Θ on which $S_N(\omega,\cdot)$ is continuously differentiable for all $\omega\in\Omega$. Furthermore, we suppose that*

$$\sup_{N\in\mathbb{N}}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|\nabla_{\boldsymbol{\theta}}S_N(\boldsymbol{\theta})\right|\right]<\infty.$$

Then, $S_N(\boldsymbol{\theta})\xrightarrow[n\to\infty]{\mathbb{P}}C$ uniformly in $\boldsymbol{\theta}$.

Sometimes, we can not use the previous lemma and the following lemma will be used instead.

Lemma S1.6 *Let $F\subset\mathbb{R}^d$ be a convex compact set, and let $\{\xi_N(\boldsymbol{\theta});\boldsymbol{\theta}\in F\}$, be a family of real-valued random processes for $N\in\mathbb{N}$. If there exist constants $p\geq l>d$ and $C>0$ such that for all $\boldsymbol{\theta},\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ it holds*

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

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

(3) $\xi_N\left(\boldsymbol{\theta}\right)\xrightarrow[n\to\infty]{\mathbb{P}}0$,

then $\sup_{\boldsymbol{\theta}\in F}|\xi_N(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

Another useful tool for checking if the conditions of the previous lemma hold, is Rosenthal's inequality for martingales (Theorem 2.12 in Hall and Heyde (1980)).

Theorem S1.7 (Rosenthal's inequality) *Let* $\left(X_k^N\right)_{N\in\mathbb{N},1\le k\le N}$ *be a triangular array with each row* N *adapted to a filtration* $\left(\mathcal{G}_k^N\right)_{1\le k\le N}$ *and let*

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

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

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

A special case of multivariate martingale triangular arrays central limit theorem (Proposition 3.1 from Crimaldi and Pratelli (2005)) that we present here will be useful to prove asymptotic normality.

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

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

If

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

(2) $\sum_{k=1}^{N}\mathbf{X}_{N,k}\mathbf{X}_{N,k}^{\top}\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{U}$*, for some non-random positive semi-definite matrix* $\mathbf{U}$ *,*

then $\mathbf{S}_N\xrightarrow[n\to\infty]{d}\mathcal{N}_d(\mathbf{0},\mathbf{U})$.

Remark To prove the second condition, we use Lemma S1.6. It is sufficient to prove that, for all $i,j=1,...,d$,

$$\sum_{k=1}^{N}\mathbb{E}\left[X_{N,k}^{(i)}X_{N,k}^{(j)}\mid\mathcal{G}_{k-1}^N\right]\xrightarrow[n\to\infty]{\mathbb{P}}U_{ij},\qquad\qquad\sum_{k=1}^{N}\mathbb{E}\left[\left(X_{N,k}^{(i)}X_{N,k}^{(j)}\right)^2\mid\mathcal{G}_{k-1}^N\right]\xrightarrow[n\to\infty]{\mathbb{P}}0.$$

Remark For a martingale difference array we need conditional expectations to be zero almost surely, i.e.

$$\mathbb{E}\left[\mathbf{X}_{N,k}\mid\mathcal{G}_{k-1}^N\right]=0,\ \text{a.s. for all } N\in\mathbb{N},\ 1\le k\le N.$$

In our case, $\left(\mathbf{X}_{N,k}\right)_{N\in\mathbb{N},1\le k\le N}$ is not a martingale difference array. So, in the same manner as in Corollary 2.6 in McLeish (1974) we need to assume two additional conditions on $\left(\mathbf{X}_{N,k}\right)_{N\in\mathbb{N},1\le k\le N}$

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

Indeed, we have a martingale difference array $\mathbf{Y}_{N,k} = \mathbf{X}_{N,k} - \mathbb{E}\left[\mathbf{X}_{N,k} \mid \mathcal{G}_{k-1}^N\right]$ that satisfies conditions of the previous theorem. To prove the first condition we write

$$\begin{aligned}\mathbb{E}\left[\sup_{1\le k\le N}\|\mathbf{Y}_{N,k}\|_1\right] &\le \mathbb{E}\left[\sup_{1\le k\le N}\|\mathbf{X}_{N,k}\|_1\right] + \mathbb{E}\left[\sup_{1\le k\le N}\mathbb{E}\left[\|\mathbf{X}_{N,k}\|_1 \mid \mathcal{G}_{k-1}^N\right]\right]\\ &\le \mathbb{E}\left[\sup_{1\le k\le N}\|\mathbf{X}_{N,k}\|_1\right] + \mathbb{E}\left[\sup_{1\le k\le N}\mathbb{E}\left[\sup_{1\le j\le N}\|\mathbf{X}_{N,j}\|_1 \mid \mathcal{G}_{k-1}^N\right]\right]\\ &\le 3\mathbb{E}\left[\sup_{1\le k\le N}\|\mathbf{X}_{N,k}\|_1\right] \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

We used Doob's inequality for the last submartingale. To prove the second condition we fix i, j to get

$$\begin{aligned}\sum_{k=1}^N Y_{N,k}^{(i)}Y_{N,k}^{(j)} &= \sum_{k=1}^N X_{N,k}^{(i)}X_{N,k}^{(j)} - \sum_{k=1}^N X_{N,k}^{(i)}\mathbb{E}\left[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N\right]\\ &\quad - \sum_{k=1}^N X_{N,k}^{(j)}\mathbb{E}\left[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N\right] + \sum_{k=1}^N \mathbb{E}\left[X_{N,k}^{(i)} \mid \mathcal{G}_{k-1}^N\right]\mathbb{E}\left[X_{N,k}^{(j)} \mid \mathcal{G}_{k-1}^N\right].\end{aligned}$$

The first term goes to U_{ij}, and the last term goes to zero. To prove that middle terms also go to zero we use the following inequalities

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

The previous theorem yields $\sum_{k=1}^N \mathbf{Y}_{N,k} \xrightarrow[n\to\infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$. Together with (S8), we get $\mathbf{S}_N \xrightarrow[n\to\infty]{d} \mathcal{N}_d(\mathbf{0}, \mathbf{U})$.

S1.5 Proof of Lemma 7.1

Proof of Lemma 7.1 To prove point-wise convergence, we mostly use Lemma S1.4. To prove uniform convergence we use interchangeably Lemma S1.5 and Lemma S1.6

Proof of 1. We denote $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}) \coloneqq \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$. We have

$$\begin{aligned}\sum_{k=1}^N \mathbb{E}_{\theta_0}\left[Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}\right) \mid \mathbf{X}_{t_{k-1}}\right] &= \frac{1}{Nh}\sum_{k=1}^N \mathbb{E}_{\theta_0}\left[\operatorname{Tr}\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right) \mid \mathbf{X}_{t_{k-1}}\right]\\ &= \frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}\right]\right)\\ &= \frac{1}{Nh}\sum_{k=1}^N \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\right) \xrightarrow[\substack{Nh\to\infty\\ h\to 0}]{\mathbb{P}_{\theta_0}} \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right).\end{aligned}$$

To use Lemma S1.4, we need to prove that covariance of $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\varsigma})$ goes to zero. Recall that if $\mathbf{A}$ is a Gaussian random vector $\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Pi})$, then $\mathbb{E}[(\mathbf{A}^T\mathbf{B}\mathbf{A})^2] = 2\operatorname{Tr}((\mathbf{B}\boldsymbol{\Pi})^2) + (\operatorname{Tr}(\mathbf{B}\boldsymbol{\Pi}))^2$. We use Corollary 3.8 to write $\mathbf{Z}_{t_k}$ in terms of $\boldsymbol{\xi}_{h,k}$

$$\begin{aligned}&\sum_{k=1}^N \mathbb{E}_{\theta_0}\left[Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\varsigma}\right)^2 \mid \mathbf{X}_{t_{k-1}}\right]\\ &= \frac{1}{N^2h^2}\sum_{k=1}^N\left(\mathbb{E}_{\theta_0}\left[\left(\boldsymbol{\xi}_{h,k}^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{\xi}_{h,k}\right)^2 \mid \mathbf{X}_{t_{k-1}}\right] + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^{3/2})\right)\\ &= \frac{1}{N^2h^2}\sum_{k=1}^N\left(2\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}h\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top\right)^2 + \left(\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}h\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\top\right)\right)^2\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(\sqrt{h}/N) \xrightarrow{\mathbb{P}_{\theta_0}} 0,\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. By Lemma S1.4 $\frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0) \xrightarrow{\mathbb{P}_{\theta_0}} \operatorname{Tr}((\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)$, for $Nh \to \infty$, $h \to 0$. We still need to prove that the limits hold uniformly in $\boldsymbol{\theta}$. For that recall Frobenius inner product of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$: $\langle \mathbf{A}, \mathbf{B}\rangle_F \coloneqq \operatorname{Tr}(\mathbf{A}^\top\mathbf{B})$. Then, Hölder inequality applied on the Frobenius norm gives the following bound of a trace of a product of matrices $|\operatorname{Tr}(\mathbf{A}^\top\mathbf{B})| \le |\operatorname{Tr}(\mathbf{A})|\|\mathbf{B}\|$. Here, $\|\cdot\|$ denotes the operator 2-norm defined as $\|\mathbf{A}\| \coloneqq \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}$, where $\lambda_{\max}$ is the largest eigenvalue. Also, we use the operator norm inequality $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$. Now, it is enough to prove the uniformity in $\boldsymbol{\varsigma}$, because $\boldsymbol{\beta}$ is fixed to the true value. To prove the uniformity use Lemma S1.5. Denote $\boldsymbol{\varsigma} = (\varsigma_1, \varsigma_2, \ldots, \varsigma_s) \in \Theta_{\varsigma_1} \times \Theta_{\varsigma_2} \times \cdots \times \Theta_{\varsigma_s} = \Theta_\varsigma$. It is enough to show that for all $j = 1, \ldots, s$

$$\sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\sup_{\varsigma_j\in\Theta_{\varsigma_j}} \left|\partial_{\varsigma_j} \frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right|\right] < \infty. \tag{S9}$$

Use the well-known rule of matrix differentiation $\partial_{\mathbf{X}}(\mathbf{a}^\top\mathbf{X}^{-1}\mathbf{a}) = -\mathbf{X}^{-1}\mathbf{a}\mathbf{a}^\top\mathbf{X}^{-1}$, where $\mathbf{a}$ is a vector and $\mathbf{X}$ is a symmetric matrix. Then, we get

$$\partial_{x^{(i)}} \operatorname{Tr}\left(\mathbf{a}^\top\mathbf{C}^{-1}(\mathbf{x})\mathbf{a}\right) = -\operatorname{Tr}\left(\mathbf{C}^{-1}(\mathbf{x})\mathbf{a}\mathbf{a}^\top\mathbf{C}^{-1}(\mathbf{x})\partial_{x^{(i)}}\mathbf{C}(\mathbf{x})\right) = -\operatorname{Tr}\left(\mathbf{a}\mathbf{a}^\top\mathbf{C}^{-1}(\mathbf{x})\left(\partial_{x^{(i)}}\mathbf{C}(\mathbf{x})\right)\mathbf{C}^{-1}(\mathbf{x})\right).$$

For ease of notation, we omit writing $\boldsymbol{\beta}_0$. Then, we have

$$\begin{aligned}
&\sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\sup_{\varsigma_j\in\Theta_{\varsigma_j}} \left|\partial_{\varsigma_j} \frac{1}{Nh}\sum_{k=1}^{N} \mathbf{Z}_{t_k}^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{Z}_{t_k}\right|\right] \le \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \left|\partial_{\varsigma_j} \operatorname{Tr}\left(\mathbf{Z}_{t_k}^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{Z}_{t_k}\right)\right|\right] \\
&\le \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}\left(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\right) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \left\|\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right\|\right] \\
&\le \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}\left(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\right) \sup_{\varsigma_j\in\Theta_{\varsigma_j}} \left\|\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right\|^2 \left\|\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right\|\right] \le C \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}\left(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\right)\right] \\
&= C \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}\left(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top\right) \mid \mathbf{X}_{t_{k-1}}\right]\right] = C \sup_{N\in\mathbb{N}} \mathbb{E}_{\theta_0}\left[\frac{1}{Nh}\sum_{k=1}^{N} \operatorname{Tr}\left(h\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\right)\right] < \infty.
\end{aligned}$$

We used the trace bound in the second inequality, the operator norm inequality in the third inequality, and (A4) in the last one.

Proof of 2. Lemma 4.2 yields

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right) \xrightarrow{\mathbb{P}_{\theta_0}} \int \mathbf{g}\left(\mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right) \mathrm{d}\nu_0(\mathbf{x}),$$

uniformly in $\boldsymbol{\theta}$, for $Nh \to \infty$, $h \to 0$. Property 2 follows from the boundedness of $\mathbf{g}$.

Proof of 3. Introduce $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$. We start with

$$\begin{aligned}
\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}\right] &= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\operatorname{Tr}\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\right) \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{1}{N}\sum_{k=1}^{N} \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right) \mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}\right]\right) \\
&= \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^3) \xrightarrow{\mathbb{P}_{\theta_0}} 0,
\end{aligned}$$

for $h \to 0$. Now, repeat the same derivation for the second moment of Y_k^N

$$\begin{aligned}
&\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right) \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \mid \mathbf{X}_{t_{k-1}}\right] \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right) \\
&= \mathcal{O}_{\mathbb{P}_{\theta_0}}(h/N) \xrightarrow{\mathbb{P}_{\theta_0}} 0,
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. Lemma S1.4 yields $\frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}) \xrightarrow{\mathbb{P}_{\theta_0}} 0$, for $Nh \to \infty$, $h \to 0$. We still need to prove that the limits hold uniformly in $\boldsymbol{\theta}$. We use Lemma S1.6. It is enough to prove that there exist constants $p \geq l > r + s$ and $C > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ it holds

$$\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N} Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)\right|^p\right] \leq C, \tag{S10}$$

$$\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N}\left(Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2\right)\right)\right|^p\right] \leq C\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|^l. \tag{S11}$$

We start with the first one. Note that

$$\left\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right\|^p \leq \left\|\mathbf{X}_{t_k} - \mathbf{X}_{t_{k-1}}\right\|^p + C_1 h^p\left(1 + \left\|\mathbf{X}_{t_k}\right\|\right)^{C_1} + C_2 h^p\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_2}, \tag{S12}$$

for some constants $C_1, C_2 > 0$, due to the shape of $\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)$ and assumptions of $\mathbf{N}$ and $h < 1$. Then, from Lemma 4.1

$$\mathbb{E}_{\theta_0}\left[\left\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right\|^p \mid \mathbf{X}_{t_{k-1}}\right] \leq C h^{p/2}\left(1 + \left\|\mathbf{X}_{t_{k-1}}\right\|\right)^C. \tag{S13}$$

Use the inequality of norms to get

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N} Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)\right|^p\right] &\leq N^{p-1}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\left|Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)\right|^p\right] &\text{(S14)}\\
&= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\mathbb{E}_{\theta_0}\left[\left|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1} \mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\right|^p \mid \mathbf{X}_{t_{k-1}}\right]\right] \\
&\leq \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\mathbb{E}_{\theta_0}\left[\left\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right\|^p \mid \mathbf{X}_{t_{k-1}}\right] \left\|\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right\|^p \left\|\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\right\|^p\right] \\
&\leq \frac{1}{N}\cdot n \cdot C. &\text{(S15)}
\end{aligned}$$

In the last line we used (S13) together with both statements of Lemma 4.1. Now, we can prove the other condition. We start in the same manner

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N}\left(Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2\right)\right)\right|^p\right] &\text{(S16)}\\
&\leq \frac{2^{p-1}}{N}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\left|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1}\left(\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0\right) - \mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0\right)\right)\right|^p\right] &\text{(S17)}\\
&+ \frac{2^{p-1}}{N}\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[\left|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1} - \left(\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top\right)^{-1}\right)\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0\right)\right|^p\right]. &\text{(S18)}
\end{aligned}$$

We start with (S17). Use the mean value theorem and the norm inequalities, like before, to get

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[\left|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1}\left(\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0\right)-\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0\right)\right)\right|^p\right]\\
&\leq\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[\mathbb{E}_{\theta_0}\left[\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\|^p\mid\mathbf{X}_{t_{k-1}}\right]\left\|\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1}\right\|^p\left\|\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_1,\boldsymbol{\beta}_0\right)-\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2,\boldsymbol{\beta}_0\right)\right\|^p\right]\\
&\leq\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[C_p\left(1+\left\|\mathbf{X}_{t_{k-1}}\right\|\right)^{C_p}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|^p\left\|\int_0^1 D_\beta\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_2+t\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right),\boldsymbol{\beta}_0\right)\mathrm{d}t\right\|^p\right] \qquad \text{(S19)}\\
&\leq C\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|^p. \qquad \text{(S20)}
\end{aligned}$$

To prove (S18), introduce the following multivariate matrix-valued function $\mathbf{G}(\boldsymbol{\varsigma}) \coloneqq (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}$. Repeating the same series of inequalities as before, we get $\|\mathbf{G}(\boldsymbol{\varsigma}_1)-\mathbf{G}(\boldsymbol{\varsigma}_2)\|$. Use the inequality between operator 2-norm and Frobenius norm, and the definition of the Frobenius norm to get

$$\|\mathbf{G}(\boldsymbol{\varsigma}_1)-\mathbf{G}(\boldsymbol{\varsigma}_2)\|\leq\left(\sum_{i,j=1}^{d}\|G_{ij}(\boldsymbol{\varsigma}_1)-G_{ij}(\boldsymbol{\varsigma}_2)\|^2\right)^{\frac{1}{2}}.$$

Now, apply the mean value theorem on each G_{ij} to get

$$\|\mathbf{G}(\boldsymbol{\varsigma}_1)-\mathbf{G}(\boldsymbol{\varsigma}_2)\|\leq\left(\sum_{i,j=1}^{d}\|\boldsymbol{\varsigma}_1-\boldsymbol{\varsigma}_2\|^2\left\|\int_0^t\nabla_{\boldsymbol{\varsigma}}G_{ij}\left(\boldsymbol{\varsigma}_2+t(\boldsymbol{\varsigma}_1-\boldsymbol{\sigma_2})\right)\mathrm{d}t\right\|^2\right)^{\frac{1}{2}}\leq C\|\boldsymbol{\varsigma}_1-\boldsymbol{\varsigma}_2\|,$$

due to Assumption (A4). The rest of the proof is the same. Now, we get

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N}\left(Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_1\right)-Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}_2\right)\right)\right|^p\right]&\leq C\left(\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|^p+\|\boldsymbol{\varsigma}_1-\boldsymbol{\varsigma}_2\|^p\right)\\
&\leq C\left(\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|^2+\|\boldsymbol{\varsigma}_1-\boldsymbol{\varsigma}_2\|^2\right)^{p/2}=C\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^p,
\end{aligned}$$

for $p\geq 2$. This concludes the proof.

Proof of 4. Introduce $Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})\coloneqq\frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)$. Derivations from the previous point yield

$$\begin{aligned}
&\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mid\mathbf{X}_{t_{k-1}}\right]\\
&=\frac{1}{Nh}\sum_{k=1}^{N}\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\mid\mathbf{X}_{t_{k-1}}\right]\right)=\mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2)\xrightarrow{\mathbb{P}_{\theta_0}}0,
\end{aligned}$$

for $h\to 0$. Now, $\sum_{k=1}^{N}\mathbb{E}_{\theta_0}[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2\mid\mathbf{X}_{t_{k-1}}]=\mathcal{O}_{\mathbb{P}_{\theta_0}}(\frac{1}{Nh})\xrightarrow{\mathbb{P}_{\theta_0}}0$, for $Nh\to\infty$, because

$$\begin{aligned}
&\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0,\boldsymbol{\theta})^2\mid\mathbf{X}_{t_{k-1}}\right] \qquad \text{(S21)}\\
&=\frac{1}{N^2h^2}\sum_{k=1}^{N}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\mid\mathbf{X}_{t_{k-1}}\right]\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0,\boldsymbol{\beta}\right).
\end{aligned}$$

To prove the uniform convergence, use Lemma S1.6 and Rosenthal's inequality (Theorem 2.12 in Hall and Heyde (1980), also stated as Theorem S1.7 in this paper) to get

$$\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N}Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)\right|^p\right]\leq C\left(\mathbb{E}\left[\left(\sum_{k=1}^{N}\mathbb{E}[Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)^2\mid\mathbf{X}_{t_{k-1}}]\right)^{p/2}\right]+\sum_{k=1}^{N}\mathbb{E}\left[\left|Y_k^N\left(\boldsymbol{\beta}_0,\boldsymbol{\theta}\right)\right|^p\right]\right).$$

The first term is bounded because of (S21). From (S15), $Nh \to \infty$ and $h \to 0$ and $p > 2$ we have

$$\sum_{k=1}^{N} \mathbb{E}\left[\left|Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}\right)\right|^p\right] \leq \frac{1}{(Nh)^p} \cdot Nh^{p/2} \cdot C = \frac{1}{(Nh)^{p-1}} \cdot h^{p/2-1} \cdot C \leq C.$$

To finish the proof of uniform convergence, we again use Rosenthal's inequality to get

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\left|\sum_{k=1}^{N}\left(Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2\right)\right)\right|^p\right] \\
&\leq C\mathbb{E}\left[\left(\sum_{k=1}^{N}\mathbb{E}[\left(Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2\right)\right)^2 \mid \mathbf{X}_{t_{k-1}}]\right)^{p/2}\right] + C\sum_{k=1}^{N}\mathbb{E}\left[\left|\left(Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2\right)\right)\right|^p\right].
\end{aligned}$$

For the second term use derivations from (S20) to get

$$\sum_{k=1}^{N}\mathbb{E}\left[\left|\left(Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2\right)\right)\right|^p\right] \leq \frac{1}{(Nh)^p} \cdot Nh^{p/2} \cdot C \cdot \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|^p \leq C\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|^p,$$

for the same reasons as above. Again, from (S20) we have

$$\begin{aligned}
&\mathbb{E}[\left(Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_1\right) - Y_k^N\left(\boldsymbol{\beta}_0, \boldsymbol{\theta}_2\right)\right)^2 \mid \mathbf{X}_{t_{k-1}}] \\
&\leq 2\mathbb{E}_{\theta_0}\left[\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1}\left(\mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_0\right) - \mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_2, \boldsymbol{\beta}_0\right)\right)\right)^2 \mid \mathbf{X}_{t_{k-1}}\right] \\
&+ 2\mathbb{E}_{\theta_0}\left[\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^\top\right)^{-1} - \left(\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^\top\right)^{-1}\right)\mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_2, \boldsymbol{\beta}_0\right)\right)^2 \mid \mathbf{X}_{t_{k-1}}\right].
\end{aligned}$$

Finally, (S21) and derivations (S20) conclude the proof.

Proof of 5. We introduce $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{N}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$. Since $\mathbb{E}[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})^\top \mid \mathbf{X}_{t_{k-1}}] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(h)$ holds by Proposition 4.3, then $\sum_{k=1}^{N}\mathbb{E}_{\theta_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}] \to 0$ in $\mathbb{P}_{\theta_0}$, for $Nh \to \infty$, $h \to 0$. To prove convergence of the sum of second moments $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})$ of we start with

$$\begin{aligned}
&\frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}\left[\mathrm{Tr}\left(\left(\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbf{g}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\right)^2\right) \mid \mathbf{X}_{t_{k-1}}\right] \\
&\leq \frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}\left[\left\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right\|^2\left\|\mathbf{g}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\right\|^2 \mid \mathbf{X}_{t_{k-1}}\right]\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right)\left\|\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right\| \\
&\leq \frac{C}{N^2}\sum_{k=1}^{N}\left(\mathbb{E}\left[\left\|\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\right\|^4 \mid \mathbf{X}_{t_{k-1}}\right]\mathbb{E}\left[\left\|\mathbf{g}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\right\|^4 \mid \mathbf{X}_{t_{k-1}}\right]\right)^{\frac{1}{2}} = \mathcal{O}_{\mathbb{P}_{\theta_0}}(h/N) \xrightarrow{\mathbb{P}_{\theta_0}} 0,
\end{aligned} \tag{S22}$$

for $Nh \to \infty$, $h \to 0$. Previously, we used the bound on the trace through Hölder inequality, Cauchy-Schwartz inequality, and Lemma 4.1 with (S13). To prove uniform convergence, we prove (S10) by repeating the same steps as in proof of (S15). Similarly, to prove (S11) we repeat the same steps as in (S20) with help of the previous series of inequalities.

Proof of 6. We introduce $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{1}{Nh}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$. Proposition 4.3 yields

$$\begin{aligned}
&\sum_{k=1}^{N}\mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}\right] = \frac{1}{Nh}\sum_{k=1}^{N}\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)\mathbf{g}\left(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)^\top \mid \mathbf{X}_{t_{k-1}}\right]\right) \\
&= \frac{1}{2N}\sum_{k=1}^{N}\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top D^\top\mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right) + D\mathbf{g}\left(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h) \\
&\xrightarrow{\mathbb{P}_{\theta_0}} \int \mathrm{Tr}\left(D\mathbf{g}\left(\mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\beta}\right)\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\right)\mathrm{d}\nu_0(\mathbf{x}),
\end{aligned}$$

for $Nh \to \infty$, $h \to 0$. On the other hand, $\sum_{k=1}^{N} \mathbb{E}_{\theta_0}[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(\frac{1}{Nh}) \to 0$, in $\mathbb{P}_{\theta_0}$, for $Nh \to \infty$, $h \to 0$, which follows from (S22). To prove the uniform convergence we repeat the same reasoning as in the previous two proofs.

Proof of 7. First, we use the fact that $\mathbb{E}\left[\boldsymbol{g}(\mathbf{X}_{t_k}; \beta_0, \beta) \mid \mathbf{X}_{t_{k-1}} = \mathbf{x}\right] = \boldsymbol{g}(\mathbf{x}; \beta_0, \beta) + \mathcal{O}(h)$, for a generic function $\boldsymbol{g}$. Then, for $Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \coloneqq \frac{h}{N}\mathbf{g}_1(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0, \boldsymbol{\beta})^\top (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)^{-1}\mathbf{g}_2(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0, \boldsymbol{\beta})$ it follows

$$\sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta}) \mid \mathbf{X}_{t_{k-1}}\right] \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} 0, \qquad \sum_{k=1}^{N} \mathbb{E}_{\theta_0}\left[Y_k^N(\boldsymbol{\beta}_0, \boldsymbol{\theta})^2 \mid \mathbf{X}_{t_{k-1}}\right] \xrightarrow[\substack{Nh\to\infty \\ h\to 0}]{\mathbb{P}_{\theta_0}} 0.$$

Again, the proofs of (S10) and (S11) are the same as in the third property, with the distinction of rewriting

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

S1.6 Proofs of Asymptotic Normality

In this section, we make a distinction between the true parameter $\boldsymbol{\theta}_0$ and any parameter $\boldsymbol{\theta}$.

Proof of Lemma 7.3 First, we transform $\eta_k^{(i)}$

$$\begin{aligned}
\eta_{N,k}^{(i)}(\boldsymbol{\theta}_0) &= \frac{2}{\sqrt{Nh}} \operatorname{Tr}\left(\left(\mathbf{I} + \frac{h}{2} D\mathbf{N}_0(\mathbf{X}_{t_k})\right)\left(-\frac{h}{2} D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\right)\right) \\
&\quad - \frac{2}{h\sqrt{Nh}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \left(-\frac{h}{2}\partial_{\beta_i}\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0) + \frac{h^2}{8}\partial_{\beta_i}\left(D\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)\right)\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)\right) \\
&\quad + \frac{2}{h\sqrt{Nh}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\beta_i}\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h^3}{N}}\right) \\
&= -\sqrt{\frac{h}{N}} \operatorname{Tr}\left(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\right) + \frac{1}{\sqrt{Nh}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\beta_i}\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0) \\
&\quad - \frac{1}{4}\sqrt{\frac{h}{N}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\beta_i}\left(D\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0)\right)\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0) \\
&\quad + \frac{2}{h\sqrt{Nh}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\beta_i}\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h^3}{N}}\right).
\end{aligned} \tag{S23}$$

Proof of (48). All functions in $\eta_{n,k}^{(i)}$ are bounded and the largest term is of order $\mathcal{O}_{\mathbb{P}_{\theta_0}}(1/\sqrt{nh})$ because $\partial_{\beta_i}\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}; \boldsymbol{\beta}_0)$ is of order $\mathcal{O}_{\mathbb{P}_{\theta_0}}(h)$. All the other terms converge to zero. Moreover, terms with coefficients $\frac{1}{\sqrt{nh}}$ are of the shape $\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top)^{-1}\mathbf{g}$, where $\mathbf{g}$ is a vector-valued function of either $\mathbf{X}_{t_{k-1}}$ or $\mathbf{X}_{t_k}$. Either way, the expectation is at least of order $\mathcal{O}_{\mathbb{P}_{\theta_0}}(h)$. Thus the largest order is $\mathcal{O}_{\mathbb{P}_{\theta_0}}(\sqrt{h/n})$ that converges to zero. On the other hand, the leading term in $\zeta_{n,k}^{(j)}$ is of order $\mathcal{O}_{\mathbb{P}_{\theta_0}}(1/\sqrt{nh^2})$. However, the expectation of the coefficient in front of it is of order $\mathcal{O}(h)$, thus (48) follows.

To prove limits (49) - (52) compute expectations of $\eta_{n,k}^{(i)}$ and $\zeta_{n,k}^{(i)}$. From (S23) follows that $\mathbb{E}_{\theta_0}[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(\sqrt{h^3/n})$, since

$$\mathbb{E}_{\theta_0}\left[\frac{1}{\sqrt{Nh}} \mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\beta_i}\mathbf{N}(\mathbf{X}_{t_k}; \boldsymbol{\beta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \sqrt{\frac{h}{N}} \operatorname{Tr}\left(D_{\mathbf{x}}\partial_{\beta_i}\mathbf{N}_0(\mathbf{X}_{t_k})\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h^3}{N}}\right),$$

which comes from Proposition 4.3. Similarly, from

$$\mathbb{E}_{\theta_0}\left[\operatorname{Tr}\left(\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^\top \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \left(\partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1}\right) \mid \mathbf{X}_{t_{k-1}}\right] = h \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right)^{-1} \partial_{\varsigma_j}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^\top\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2).$$

we have $\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(i)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\frac{h}{\sqrt{n}}\right)$. Then,

$$\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{nh^3}\right) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0,$$

$$\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{nh^2}\right) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0,$$

$$\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_1)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] \mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_2)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^3) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0,$$

$$\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_1)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] \mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_2)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(h^2) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0,$$

$$\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] \mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j)}(\boldsymbol{\theta}_0) \mid \mathbf{X}_{t_{k-1}}\right] = \mathcal{O}_{\mathbb{P}_{\theta_0}}(\sqrt{h^5}) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 0.$$

Now, we prove limit (53). When multiplying $\eta_{n,k}^{(i_1)}$ and $\eta_{n,k}^{(i_2)}$ the only terms that do not converge to zero are the ones with $1/\sqrt{nh}$ in front. Here, we take a closer look at these

$$\begin{aligned}
\eta_{n,k}^{(i)}(\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{nh}}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_i}\mathbf{N}(\mathbf{X}_{t_k}) + \frac{2}{h\sqrt{nh}}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_i}\boldsymbol{\mu}_h(\mathbf{X}_{t_{k-1}}) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h}{n}}\right) \\
&= \frac{1}{\sqrt{nh}}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_i}\mathbf{N}(\mathbf{X}_{t_k}) + \frac{1}{\sqrt{nh}}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_i}\left(\mathbf{N}(\mathbf{X}_{t_{k-1}}) + 2\mathbf{A}\mathbf{X}_{t_{k-1}}\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h}{n}}\right) \\
&= \frac{2}{\sqrt{nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_i}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0) \\
&+ \frac{1}{\sqrt{nh}}\mathbf{Z}_{t_k}(\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\psi}_i(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\sqrt{\frac{h}{n}}\right),
\end{aligned}$$

where $\boldsymbol{\psi}_i(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0) = \partial_{\beta_i}(\mathbf{N}(\mathbf{X}_{t_k};\boldsymbol{\beta}_0) - \mathbf{N}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0))$. Then,

$$\begin{aligned}
\eta_{n,k}^{(i_1)}(\boldsymbol{\theta}_0)\,\eta_{n,k}^{(i_2)}(\boldsymbol{\theta}_0) &= \frac{4}{nh}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{2}{nh}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\psi}_{i_1}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{2}{nh}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\boldsymbol{\psi}_{i_2}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \\
&+ \frac{1}{nh}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\psi}_{i_1}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\boldsymbol{\psi}_{i_2}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\frac{1}{n}\right).
\end{aligned}$$

In the previous equation, there are terms of order $\mathcal{O}_{\mathbb{P}_{\theta_0}}(\frac{1}{n})$, but when taking the expectation, an additional h appears, which makes the sum converge to zero. We need to prove that sum of expectations of all the terms except first converge to zero, in the previous equation. We only prove for the second row, the rest are analogous. Due to definition of $\boldsymbol{\psi}_i$, it is clear that $\mathbb{E}_0[\|\boldsymbol{\psi}_i(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\|^p \mid \mathbf{X}_{t_{k-1}}] = \mathcal{O}_{\mathbb{P}_0}(h)$, for all $p \geq 1$. Then,

$$\begin{aligned}
&\frac{1}{nh}\left|\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\psi}_{i_1}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}\right]\right| \\
&\leq \frac{1}{nh}\left|\mathrm{Tr}\left(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\right)\right| \|\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\|\mathbb{E}_{\theta_0}\left[\|\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^{\top}\|\|\boldsymbol{\psi}_{i_1}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\| \mid \mathbf{X}_{t_{k-1}}\right] \\
&\leq \frac{C}{nh}\left(\mathbb{E}_{\theta_0}\left[\|\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^{\top}\|^2 \mid \mathbf{X}_{t_{k-1}}\right]\mathbb{E}_{\theta_0}\left[\|\boldsymbol{\psi}_{i_1}(\mathbf{X}_{t_{k-1}},\mathbf{X}_{t_k};\boldsymbol{\beta}_0)\|^2 \mid \mathbf{X}_{t_{k-1}}\right]\right)^{\frac{1}{2}} \\
&= \frac{C}{nh}\left(\mathcal{O}_{\mathbb{P}_0}(h^2)\mathcal{O}_{\mathbb{P}_0}(h)\right)^{\frac{1}{2}} = \mathcal{O}_{\mathbb{P}_0}\left(\frac{\sqrt{h}}{n}\right).
\end{aligned}$$

We used property (S13) in the last line. Finally, we use Lemma 4.2 to get

$$\begin{aligned}
&\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\eta_{n,k}^{(i_1)}\left(\boldsymbol{\theta}_0\right)\eta_{n,k}^{(i_2)}\left(\boldsymbol{\theta}_0\right) \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{4}{nh}\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}\right] + \mathcal{O}_{\mathbb{P}_0}\left(\frac{\sqrt{h}}{n}\right) \\
&= \frac{4}{n}\sum_{k=1}^{n} \operatorname{Tr}\left(\partial_{\beta_{i_2}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\beta_{i_1}}\mathbf{F}(\mathbf{X}_{t_{k-1}};\boldsymbol{\beta}_0)\right) + \mathcal{O}_{\mathbb{P}_0}\left(\frac{\sqrt{h}}{n}\right) \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}} 4\left[\mathbf{C}_{\boldsymbol{\beta}}\left(\boldsymbol{\theta}_0\right)\right]_{i_1 i_2}.
\end{aligned}$$

To prove (54) we use Corollary 3.8. We start with

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\zeta_{n,k}^{(j_1)}\left(\boldsymbol{\theta}_0\right)\zeta_{n,k}^{(j_2)}\left(\boldsymbol{\theta}_0\right) \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{1}{h^2 n}\mathbb{E}_{\theta_0}\left[\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k}\mathbf{Z}_{t_k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\mathbf{Z}_{t_k} \mid \mathbf{X}_{t_{k-1}}\right] \\
&- \frac{1}{n}\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right) \\
&= \frac{1}{h^2 n}\mathbb{E}_{\theta_0}\left[\boldsymbol{\xi}_{h,k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}_{h,k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}\right] \\
&- \frac{1}{n}\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right) + \mathcal{O}_{\mathbb{P}_{\theta_0}}\left(\frac{\sqrt{h}}{n}\right).
\end{aligned}$$

Now, we use the expectation of a product of two quadratic forms of normally distributed random vectors (see for example Section 2 in Kumar (1973)) to get

$$\begin{aligned}
&\frac{1}{h^2 n}\mathbb{E}_{\theta_0}\left[\boldsymbol{\xi}_{h,k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_1}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\xi}_{h,k}\boldsymbol{\xi}_{h,k}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\left(\partial_{\varsigma_{j_2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\boldsymbol{\xi}_{h,k} \mid \mathbf{X}_{t_{k-1}}\right] \\
&= \frac{2}{n}\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}}{\partial\varsigma_{j_1}}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}}{\partial\varsigma_{j_2}}\right) + \frac{1}{n}\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}}{\partial\varsigma_{j_1}}\right)\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}\right)^{-1}\frac{\partial\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{\top}}{\partial\varsigma_{j_2}}\right).
\end{aligned}$$

This proves (54). We omit the proofs of (55)-(58) since they follow the same pattern. Namely, we find the leading term and make sure it goes to zero. For the expectations of squares, we can apply the same trick with a product of two quadratic forms.