Parameter estimation for SDEs from observational data

Vincent Guan (18533281), Joseph Janssen (28738152)

April 24, 2024

1 Introduction

In applied stochastic analysis, we study stochastic differential equations (SDEs) of the form

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t. \tag{1}$$

To analyze its solutions and related properties, we usually assume that the drift and diffusion terms, $b(X_t, t)$ and $\sigma(X_t, t)$, are known. In practice, however, we often observe stochastic processes without knowing its corresponding SDE. In these cases, the objective is to uncover the data generating process by learning the functional form of the drift and diffusion as well as their parameters from observed time series data. Parameter estimation from observational data is a particular focus for research in causal inference, since it enables scientists to identify the causes of each variable, and to quantify these causal relationships. Furthermore, parameter estimation enables us to obtain the system's post-intervention distribution [5, 15]. This means that we can recover the new system dynamics after intervening on one of its components.

In this paper, we will overview how parameter estimation for various SDEs of the form (1) can be performed via quadratic variation analysis and maximum likelihood estimation [10]. The theory and methodology will be supplemented with empirical evaluations to convey how the effectiveness of parameter estimation crucially depends on the quality of the data, specifically the time granularity Δt and the observation period $[0, T]$. We will begin with parameter estimation for simple SDEs in one dimension, and generalize to higher dimensional and more complex functional models.

2 Notations

$$\begin{aligned}
T &= \text{length of observation period } [0, T] \\
N &= \text{number of discrete observations for the process } (X_t)_{0\le t\le T} \\
\Delta t &= \frac{T}{N}, \text{ time granularity of observed discrete data (assumed to be constant)} \\
\{X_i\}_{i=0}^{N-1} &= \text{discrete observations of the process } (X_t)_{0\le t\le T} \\
X_i &= X_{i\Delta t}, \ i\text{th observation of process} \\
\Delta X_i &= X_{i+1} - X_i \\
[X]_T &= \lim_{\Delta t_i \to 0} \sum_{t_i \le T} |\Delta X_i|^2, \text{quadratic variation of } (X_t)_{0\le t\le T} \\
[X, Y]_T &= \lim_{\Delta t_i \to 0} \sum_{t_i \le T} \Delta X_i \Delta Y_i, \text{quadratic covariation of } (X_t)_{0\le t\le T} \text{ and } (Y_t)_{0\le t\le T} \\
P_X &= \text{law of the process } (X_t)_{0\le t\le T}, \text{ also known as the observational distribution} \\
P_W &= \text{law of Brownian motion } (W_t)_{0\le t\le T}
\end{aligned}$$

3 Parameter estimation for one-dimensional OU processes

Estimating b and σ typically requires the analyst to assume some functional form of the SDE such that parameter estimation can be well defined. In this section, we overview the procedure from Pavliotis [10] for

estimating the drift $b(X_t, t) = -\alpha X_t$ and diffusion σ from a one-dimensional Ornstein-Uhlenbeck SDE:

$$dX_t = -\alpha X_t dt + \sigma dW_t. \tag{2}$$

3.1 Estimating diffusion via quadratic variation

To estimate σ from our observations of X_t given the known functional form (2), we note that

$$dX_t \cdot dX_t = \sigma^2 dt, \tag{3}$$

due to the stochastic calculus rules: $dt \cdot dt = dt \cdot dW_t = dW_t \cdot dt = 0$ and $dW_t \cdot dW_t = dt$. Let $[0, T]$ be the observation period of our process X_t. Then, we observe that the diffusion σ is related to the quadratic variation of X_t:

$$[X]_T = \lim_{\Delta t_i \to 0} \sum_{t_i \leq T} |\Delta X_i|^2 = \int_0^T \sigma^2 ds = \sigma^2 T, \tag{4}$$

where $\Delta X_i = X_{t_{i+1}} - X_i$. Thus, as long as we observe a high frequency of observations, such that $\Delta t \approx 0$, then we may approximate the squared diffusion σ^2 via

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{i=1}^{n} [\Delta X_i]^2 \approx \frac{[X]_T}{T} = \sigma^2 \tag{5}$$

Theorem 1. *(Proposition 5.2 in [10]) Let $\{X_{i\Delta t}\}_{i=0}^{N-1}$ be a sequence of equidistant observations of*

$$dX_t = b(X_t; \theta)dt + \sigma dW_t \tag{6}$$

with time step $\Delta t = \frac{T}{N}$. Assume that the drift $b(X; \theta)$ is bounded. Then, the estimator

$$\hat{\sigma}_N^2 = \frac{1}{T} \sum_{i=0}^{N-1} (\Delta X_i)^2$$

obeys

$$|\mathbb{E}\hat{\sigma}_N^2 - \sigma^2| \leq C \left(\Delta t + \Delta t^{1/2} \right) \tag{7}$$

for some $C > 0$. In other words, $\hat{\sigma}^2$ is an unbiased estimator.

Proof.

$$\Delta X_i = B_i + \sigma \Delta W_i$$

where $B_i = \int_{i\Delta t}^{(i+1)\Delta t} b(X_s; \theta) ds$ and $\Delta W_i \sim \mathcal{N}(0, \Delta t)$. Hence,

$$\hat{\sigma}_N^2 = \frac{1}{T} \sum_{i=0}^{N-1} \left(\sigma^2 (\Delta W_i)^2 + 2\sigma B_i \Delta W_i + B_i^2 \right)$$

Note that

$$\begin{aligned}
\mathbb{E}B_i^2 &= \mathbb{E} \left(\int_{i\Delta t}^{(i+1)\Delta t} b(X_s; \theta) ds \right)^2 && (8) \\
&\leq \mathbb{E} \left(\Delta t \int_{i\Delta t}^{(i+1)\Delta t} b(X_s; \theta)^2 ds \right) && \text{(Cauchy-Schwarz)} \quad (9) \\
&= \Delta t \int_{i\Delta t}^{(i+1)\Delta t} \mathbb{E} b(X_s; \theta)^2 ds && (10) \\
&\leq K(\Delta t)^2 && (11)
\end{aligned}$$

with $K = \|b(x;\theta)\|_\infty^2$, since b is bounded. Thus,

$$\begin{aligned}
\mathbb{E}\hat{\sigma}_N^2 &= \frac{\sigma^2}{T}\sum_{i=0}^{N-1}\mathbb{E}(\Delta W_i)^2 + 2\frac{\sigma}{T}\sum_{i=0}^{N-1}\mathbb{E}[B_i\Delta W_i] + \frac{1}{T}\sum_{i=0}^{N-1}\mathbb{E}[B_i^2] \\
&\leq \frac{\sigma^2 N(\Delta t)}{T} + 2\frac{\sigma}{T}\sum_{i=0}^{N-1}\sqrt{\mathbb{E}[B_i^2]\mathbb{E}[(\Delta W_i)^2]} + \frac{KN(\Delta t)^2}{T} && (\mathbb{E}(\Delta W_i)^2 = \Delta t, \text{ C-S, and (11)}) \\
&\leq \sigma^2 + 2\frac{\sigma}{T}\sum_{i=0}^{N-1}\sqrt{K}\sqrt{(\Delta t)^3} + K\Delta t && (\mathbb{E}(\Delta W_i)^2 = \Delta t \text{ and (11) and } N\Delta t = T) \\
&\leq \sigma^2 + 2\sqrt{K}\sigma\frac{N}{T}(\Delta t)^{3/2} + K\Delta t \\
&= \sigma^2 + 2\sqrt{K}\sigma(\Delta t)^{1/2} + K\Delta t
\end{aligned}$$

The desired bound then follows immediately

$$|\mathbb{E}\hat{\sigma}_N^2 - \sigma^2| \leq C((\Delta t)^{1/2} + \Delta t)$$

with $C = \max\{2\sqrt{K}\sigma, K\}$ depending on b, σ. □

Remark 1. *The error depends on the time step Δt, via the dominant term $(\Delta t)^{1/2}$. This indicates that the accuracy of the estimator depends crucially on the granularity of the observed discrete data.*

Remark 2. *We note that this estimate for σ^2 extends beyond OU processes, since the theorem applies for any SDE with bounded drift $b(X_t, \theta)$, and a constant diffusion σ. Furthermore, the stronger claim $\hat{\sigma}_N^2 \xrightarrow{a.s.} \sigma^2$ also holds, but is more difficult to prove [10].*

Remark 3. *The estimate is done at the level of σ^2, which gives σ up to a sign. Indeed, it is impossible to estimate σ itself from observational data, since both SDEs would correspond to the same distribution. Since $\sigma dW_t \overset{d}{=} -\sigma dW_t \sim N(0, \sigma^2 dt)$, it follows that if*

$$\begin{aligned}
dX_t^{(1)} &= b(X_t, \theta)dt + \sigma dW_t \\
dX_t^{(2)} &= b(X_t, \theta)dt - \sigma dW_t,
\end{aligned}$$

then $X_t^{(1)} \overset{d}{=} X_t^{(2)}$ for each t.

3.2 Estimating drift via MLE

In [10], the estimation of the drift $b = -\alpha X_t$ is done separately via maximum likelihood estimation (MLE), following the estimation of σ^2 via the quadratic variation, $\hat{\sigma}^2 = \frac{[X]_T}{T}$.

Recall that given a set of independent observations $\{x_i\}_{i=1}^N$ sampled from a distribution X, MLE is performed by maximizing a likelihood function $\mathcal{L}(\{x_i\}_{i=1}^N|\theta)$ over a set of parameters θ in a parameter space Θ. These parameters determine a probability distribution P_θ, where we seek to find the true parameters $\hat{\theta} \in \Theta$ such that $P_X = P_{\hat{\theta}}$. Intuitively, $\mathcal{L}(\theta, \omega)$ is proportional to the probability density P_θ, evaluated on the observations ω. The formal definition of a likelihood function is given below.

Definition 1. *(Absolute continuity of measures) Let μ and ν be measures on a σ-algebra $\mathcal{F}$. Then, μ is absolutely continuous with respect to ν, denoted $\mu << \nu$, if $\mu(A) = 0 \implies \nu(A) = 0$ for all μ-null sets $A \in \mathcal{F}$.*

Definition 2. *(Likelihood function) [4] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model, where P_θ is a probability measure parameterized by θ. Let ν be any σ-finite measure such that $P_\theta << \nu \ \forall \theta \in \Theta$. Then, the likelihood function $\mathcal{L}(\theta, \omega)$ is proportional to the Radon-Nikodym derivative $\frac{dP_\theta}{d\nu}(\omega)$ for all $\theta \in \Theta$.*

Remark 4. *For a stochastic process X_t solving (1) with non-zero diffusion σ, the law of X_t, P_X, is absolutely continuous with respect to the Wiener measure, i.e. $P_X << P_W$. This result follows from Girsanov's theorem [10]. Hence, by Definition (2), the likelihoood function is given by the Radon-Nikodym derivative $\frac{dP_X}{dP_W}$.*

Theorem 2. *(Adapted from [10] and [1]) Let X_t be an Ornstein-Uhlenbeck process solving (2) with known diffusion σ and unknown drift $b(X_t, t) = -\alpha X_t$. Let $\theta = \alpha$. The likelihood function given observations $\{X_i\}_{i=0}^{N-1}$ is given by*

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) = \exp\left(-\frac{\alpha}{\sigma^2}\int_0^T X_t dX_t - \frac{\alpha^2}{2\sigma^2}\int_0^T X_t^2 dt\right) \tag{12}$$

and thus admits the discrete approximation

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) = \exp\left(-\frac{\alpha}{\sigma^2}\sum_{i=0}^{N-1} X_i \Delta X_i - \frac{\alpha^2}{2\sigma^2}\sum_{i=0}^{N-1} X_i^2(\Delta t)\right) \tag{13}$$

Proof. As noted in remark 4, P_X is absolutely continuous with respect to P_W. The likelihood function is therefore defined by $\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) = \frac{dP_X}{dP_W}$. We will first consider the discretized laws P_X^N and P_W^N for $\{X_i\}_{i=0}^{N-1}$ and $\{W_i\}_{i=0}^{N-1}$ respectively via Euler-Maruyama. To do so, we compute the densities

$$p_X^N = \prod_{i=0}^{N-1} p(X_{i+1}|X_i) \qquad \text{(Markov property)} \tag{14}$$

$$= \frac{1}{(2\pi\sigma^2\Delta t)^{N/2}} \exp\left(-\sum_{i=0}^{N-1}\frac{1}{2\sigma^2\Delta t}(X_{i+1} - X_i - b_i\Delta t)^2\right) \qquad (X_{i+1}|X_i \sim \mathcal{N}(X_i + b_i\Delta t, \sigma^2\Delta t)) \tag{15}$$

$$= \frac{1}{(2\pi\sigma^2\Delta t)^{N/2}} \exp\left(-\frac{1}{2\sigma^2\Delta t}\sum_{i=0}^{N-1}(\Delta X_i)^2 - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1} b_i^2\Delta t - \frac{1}{\sigma^2}\sum_{i=0}^{N-1} b_i\Delta X_i\right) \tag{16}$$

and

$$p_W^N = \prod_{i=0}^{N-1} p(W_{i+1}|W_i) \qquad \text{(Markov property)}$$

$$= \frac{1}{(2\pi\Delta t)^{N/2}} \exp\left(-\frac{1}{2\Delta t}\sum_{i=0}^{N-1}(\Delta W_i)^2\right) \qquad (W_{i+1}|W_i \sim \mathcal{N}(0, \Delta t))$$

Hence, the Radon-Nikodym derivative can be approximated by $\frac{dP_X^N}{dP_W^N} = \frac{p_X^N (dX_t)^N}{p_W^N (dW_t)^N}$, where we note the change of variable factor $(\frac{dX_t}{dW_t})^N = \sigma^N$ arising from (2).

$$\frac{dP_X^N}{dP_W^N} = \sigma^N \frac{1}{\sigma^N} \exp\left(-\frac{1}{2\sigma^2\Delta t}\sum_{i=0}^{N-1}(\Delta X_i)^2 + \frac{1}{2\Delta t}\sum_{i=0}^{N-1}(\Delta W_i)^2 - \frac{1}{\sigma^2}\sum_{i=0}^{N-1} b_i\Delta X_i - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1} b_i^2(\Delta t)\right) \tag{17}$$

$$= \exp\left(-\frac{1}{2\sigma^2\Delta t}\sum_{i=0}^{N-1}\left((\Delta X_i)^2 - \sigma^2(\Delta W_i)^2\right) - \frac{1}{\sigma^2}\sum_{i=0}^{N-1} b_i\Delta X_i - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1} b_i^2(\Delta t)\right) \tag{18}$$

$$= \exp\left(-\frac{1}{\sigma^2}\sum_{i=0}^{N-1} b_i\Delta X_i - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1} b_i^2(\Delta t)\right) \qquad ((\Delta X_i)^2 \approx \sigma^2(\Delta W_i)^2) \tag{19}$$

The discrete approximation (13) of the likelihood function for a 1D OU process is obtained by substituting $b_i = -\alpha X_i$. Informally taking the limit as $N \to \infty$ yields the likelihood function (12).

□

Corollary 1. *Given $\theta = \alpha$, with σ independently estimated via (5), we obtain a closed form for alpha:*

$$\hat{\alpha}_T = -\frac{\int_0^T X_t dX_t}{\int_0^T X_t^2 dt} \approx -\frac{\sum_{i=0}^{N-1} X_i\Delta X_i}{\sum_{i=0}^{N-1} X_i^2\Delta t} \tag{20}$$

Proof. To solve for $\hat{\alpha}$, we can equivalently maximize the log-likelihood function

$$\ell(\{X_i\}_{i=0}^{N-1}|\theta) = -\frac{\alpha}{\sigma^2}\int_0^T X_t dX_t - \frac{\alpha^2}{2\sigma^2}\int_0^T X_t^2 dt.$$

Setting $\frac{d}{d\alpha}\ell = 0$ yields

$$\frac{\alpha}{\sigma^2}\int_0^T X_t^2 dt = -\frac{1}{\sigma^2}\int_0^T X_t dt$$
$$\implies \hat{\alpha}_T = -\frac{\int_0^T X_t dX_t}{\int_0^T X_t^2 dt}$$

□

Remark 5. *We note that the provided derivation of the likelihood function via the Euler-Maruyama discretization is informal. Indeed, $(\Delta X_i)^2 = \sigma^2(\Delta W_i)^2$ holds for infinitesimally small Δt , due to the cancellations from the stochastic laws $(\Delta t)^2 = \Delta t \Delta W_i = 0$. However, distributing the factor $\frac{1}{2\sigma^2\Delta t}$ in line (18) would muddle this argument, as the cancellations would no longer be apparent. Nevertheless, the likelihood function* (12) *and corresponding estimates are frequently presented in the literature [11, 1, 10, 8], as the Radon-Nikodym derivative of P_{X_t} with respect to the Wiener measure is well-established via Girsanov's Theorem. Indeed, given*

$$dX_t = f(\theta, X_t)dt + dW_t$$

it has been shown [13, 1] that

$$\frac{dP_X}{dP_W} = \exp\left(\int_0^T f(\theta, X_t)dX_t - \frac{1}{2}\int_0^T f^2(\theta, X_t)dt\right), \tag{21}$$

which may also heuristically be seen from (19)

Remark 6. *(Parameters in the likelihood function) Pavliotis [10] assumes that $\theta = \alpha$, following the estimation of σ via* (5)*. This yields closed form estimates $\hat{\alpha}, \hat{\sigma}$. However, we may also set $\theta = (\alpha, \sigma)$ and optimize the likelihood function* (12) *over both parameters, as done in [15]. So far, we have found the latter approach to outperform the former, particularly when Δt is large.*

Remark 7. *(Alternative likelihood formulations) There seem to be two schools of thought in terms of defining the likelihood function as the Radon-Nikodym derivative with respect to Brownian motion (Equation* (19)*) versus the the discretized density of X_t (Equation 16). [10, 1] use the former while [7, 15] use the latter. We performed some simple experiments comparing the two approaches (not shown), and found that they often give similar results, with the discretized density of X_t usually having a slight advantage. We have yet to encounter a comparative theoretical or empirical analysis in the literature with respect to these two approaches.*

Remark 8. *(Asymptotic optimality) [8, 1] The estimate $\hat{\alpha}_T$ is strongly consistent. Moreover, it has been shown that if X_t is stationary (which is true for* (2) *for $\alpha > 0$), then*

$$\lim_{T\to\infty} \frac{\hat{\alpha}_T - \alpha}{\sqrt{T}} \sim \mathcal{N}(0, \frac{1}{2\alpha}) \tag{22}$$

In contrast to the estimate for the diffusion $\hat{\sigma}$, which converges to the true value as $\Delta t \to 0$, $\hat{\alpha}$ converges to α as $T \to \infty$. The condition $\Delta t = \frac{T}{N} \to 0$ is satisfied when we fix the observation period $[0, T]$ and let $N \to \infty$. The condition $T \to \infty$ is satisfied when we fix the time granularity Δt and let $N \to \infty$. We explore these relationships in depth in the following section via empirical evaluation.

3.3 Experiments for 1D OU processes

Pavliotis [10] considered an experiment which estimated the drift from a single trajectory of Equation 2. We recreate this experiment by simulating a single trajectory of an OU process with $\alpha = 2$ and $\sigma = 1$, then we estimate $\hat{\alpha}$ while fixing Δt and increasing T. A plot of the randomly generated trajectory is given on the left

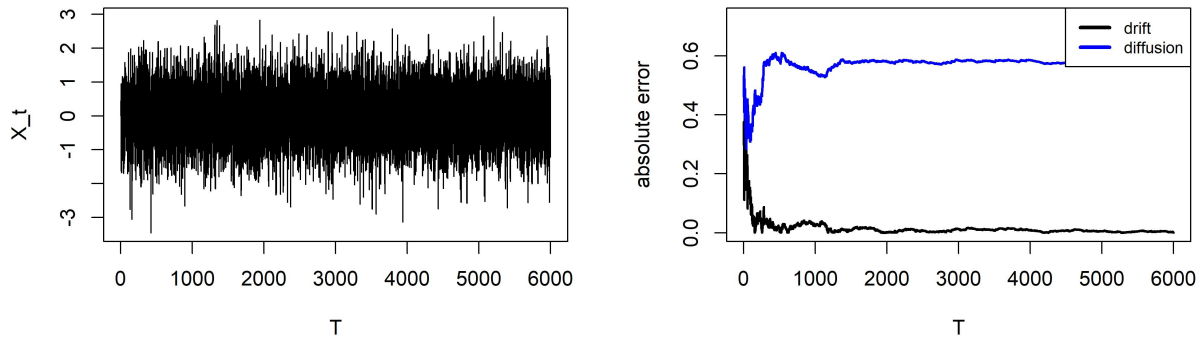

Figure 1: Recreation of the Pavliotis experiment, with the left side showing a realization of a 1D OU process and the right side showing the estimation error of drift and diffusion as we increase T and keep dt constant.

side of Figure 1, while the errors in estimating drift and diffusion as a function of T are given on the right. Consistent with Remark (8), we observe that for 1D processes, drift estimation improves as $T \to \infty$, while diffusion error stays constant with increasing T.

Considering the same process, we extended the experiment of [10] to include two more scenarios. In the first scenario, we hold T constant and increase N exponentially. The results of this are shown on the left side of Figure 2. We observe that diffusion estimation gets exponentially better while drift error is fairly constant as N is increased. In the second scenario, we hold the number of observations N constant while dt, and therefore T, are decreased. The right side of Figure 2 shows that diffusion estimation gets better linearly while drift becomes exponentially worse as dt (and T) is reduced.

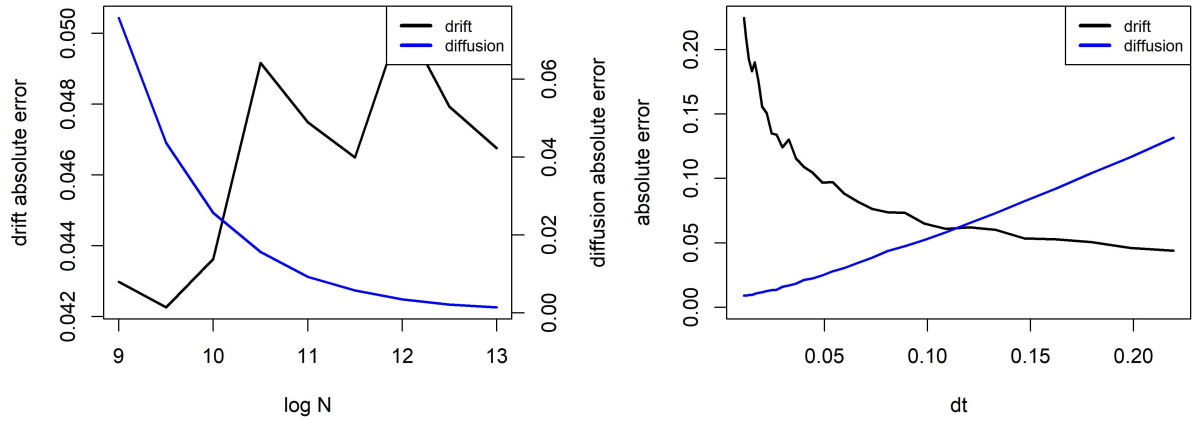

Figure 2: Extensions of the Pavliotis experiment, with the left side showing drift and diffusion error as T is held constant and N is increased. The right side shows the drift and diffusion error as dt is decreased and N is held constant.

These experiments confirm the theoretical findings discussed above. In Theorem 1, we saw that the error in estimating σ^2 via quadratic variation primarily comes from Δt. This is confirmed in all experiments since in Figure 1 we keep dt constant and the diffusion error keeps constant. Further, in Figure 2 we see that in both cases where we decrease dt, diffusion error also decreases. As noted in Remark 8, the strength of the drift estimate depends on T. Indeed, we observe that if dt is constant and T is increased or if N is held constant and dt increases, drift estimates improve (Figure 1 and right side of Figure 2). Meanwhile, if T is constant and N increases, drift error is fairly constant (Figure 2).

4 Other 1D SDEs

4.1 Experiments for 1D process with negative α

To ensure stationarity, Ornstein-Uhlenbeck processes require that X_t follows Equation 2 and that $\alpha > 0$. If we eliminate the assumption on the positivity of α and instead make it small but negative, we see very different behaviour of a single trajectory as it blows up quite quickly (Figure 3). We also observe that nonstationary drift can lead to larger diffusion error. This is consistent with the error bound from Theorem 1, where K depends on the supremum of b over $[0, T]$.

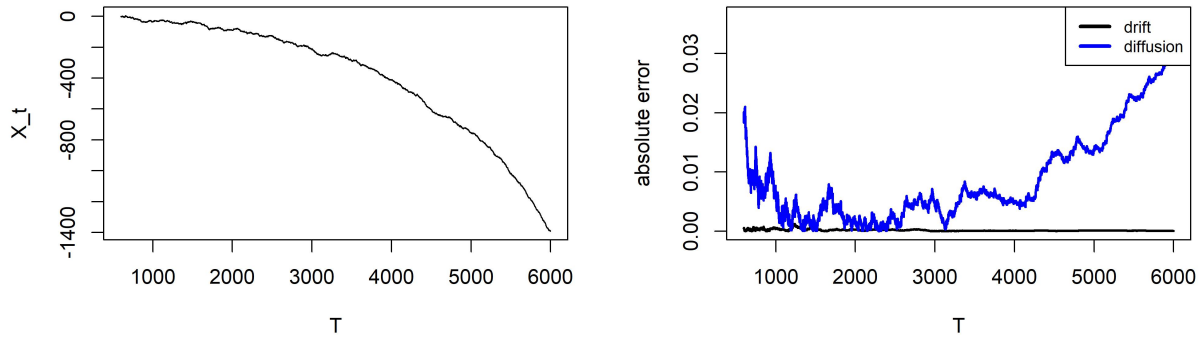

Figure 3: Left: one trajectory of a process following Equation 2 with $\alpha = -0.0006$ and $\sigma = 1$. Right: absolute drift and diffusion error versus T.

It is also possible to estimate the drift of nonlinear SDEs using the same MLE methodology from Section 3.2. We illustrate one example below from Pavliotis [10], and note that a similar procedure may be extended for general polynomial drift $b(X_t)$, provided that the solutions exist and the related integrals are well defined.

Example 1. *(Stationary bistable process) Let*

$$dX_t = (\alpha X_t - \beta X_t^3)dt + dW_t. \tag{23}$$

Then, by equation (21)*, the log-likelihood function is given by*

$$\ell(\{X_i\}_{i=0}^{N-1}|\theta) = \int_0^T (\alpha X_t - \beta X_t^3)dX_t - \frac{1}{2}\int_0^T (\alpha X_t - \beta X_t^3)^2 dt \tag{24}$$

$$= \alpha B_1 - \beta B_3 - \frac{1}{2}\alpha^2 M_2 - \frac{1}{2}\beta^2 M_6 + \alpha\beta M_4, \tag{25}$$

where $\theta = (\alpha, \beta)$ *and*

$$M_n = \int_0^T X_t^n dt \approx \sum_{i=0}^{N-1} X_i^n \Delta t$$

$$B_n = \int_0^T X_t^n dX_t \approx \sum_{i=0}^{N-1} X_i^n \Delta X_i.$$

We must have $\frac{\partial \ell}{\partial \alpha}(\hat{\alpha}, \hat{\beta}) = \frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\beta}) = 0$ *and hence,*

$$\begin{bmatrix} M_2 & -M_4 \\ M_4 & -M_6 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} B_1 \\ B_3 \end{bmatrix}$$

Direct computation shows that this linear system is solved with

$$\hat{\alpha} = \frac{B_1 M_6 - B_3 M_4}{M_2 M_6 - M_4^2}$$

$$\hat{\beta} = \frac{B_1 M_4 - B_3 M_2}{M_2 M_6 - M_4^2}$$

Remark 9. *It is well known that the existence and uniqueness of SDE* (1) *is ensured by* $b(X_t, t), \sigma(X_t, t)$ *obeying the Lipschitz condition in the space variable and linear growth in the time variable. These conditions are also commonly assumed for parameter estimation [11, 7]. Parameter estimation for superlinear drift has nevertheless gained considerable research interest [10, 3].*

4.2 Experiment for stationary bistable processes

We reproduce the experiment from Pavliotis [10] for stationary bistable processes for $\alpha = \beta = 1$. As shown in Figure 4, as the observation period $[0, T]$ increases, we can estimate both drift parameters α, β with more stability and accuracy.

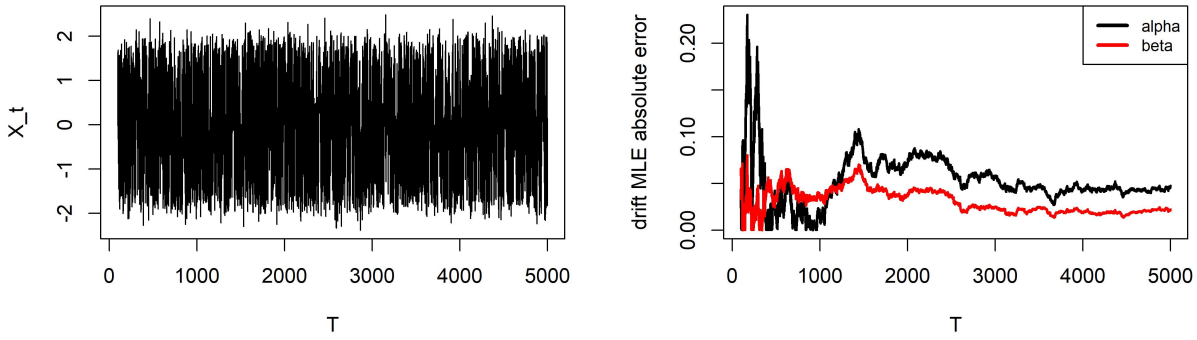

Figure 4: Left: one trajectory of a stationary bistable process with $\alpha = 1$ and $\beta = 1$. Right: absolute error for two drift parameters versus T.

5 Multivariate processes with additive or multiplicative noise

In this section, we follow Wang et al. [15], which considers multidimensional processes $X_t \in \mathbb{R}^d$ with either additive noise

$$dX_t = AX_t dt + GdW_t, \qquad A \in \mathbb{R}^{d\times d}, G \in \mathbb{R}^{d\times m}, X_0 = x_0, \tag{26}$$

or multiplicative noise

$$dX_t = AX_t dt + G(X_t)dW_t, \qquad A \in \mathbb{R}^{d\times d}, X_0 = x_0 \tag{27}$$

$$G(X_t) = \begin{bmatrix} G_1 X_t & G_2 X_t & \dots & G_m X_t \end{bmatrix}, G_i \in \mathbb{R}^{d\times d}, \tag{28}$$

where each process is driven by m components of independent Brownian motion $W_t = \begin{bmatrix} W_t^{(1)} & \dots & W_t^{(m)} \end{bmatrix}^T$

Wang et al. identify rank conditions on A and G such that the generator for both SDEs are identifiable. Since these rank conditions hold with probability 1 for randomly generated matrices, this implies that parameter estimation (A and GG^T for Equation (26) and A and $G(x)G(x)^T$ for Equation (28)) can almost surely be performed for a randomly parameterized process, given the observational distribution P_X.

5.1 Estimation for $\sigma = G$ for additive noise multidimensional SDEs

Motivated by the quadratic variation-based estimate for $\hat{\sigma}$ in the one-dimensional additive noise case, we derive an analagous estimate for arbitrary dimensions.

Theorem 3. *Let* X_t *solve the multidimensional additive noise SDE* (26). *The diagonal entries of* GG^T *are given by*

$$(GG^T)_{ii} = \frac{[X^{(i)}]_T}{T} \tag{29}$$

The other entries of GG^T are given by

$$(GG^T)_{ij} = \frac{[X^{(i)}, X^{(j)}]_T}{T}. \tag{30}$$

Furthermore, the estimate $\widehat{GG}^T_{ij} = \frac{1}{T}\sum_{k=0}^{N-1}(\Delta X_k^{(i)})(\Delta X_k^{(j)})$ is asymptotically unbiased in each entry i, j as $N \to \infty$.

Proof. Note that for each component $X_t^{(i)}$, we have

$$dX_t^{(i)} = A_{i,\cdot}X_t dt + G_{i,\cdot}dW_t. \tag{31}$$

Thus, similarly to the one-dimensional case,

$$\begin{aligned}
[X^{(i)}, X^{(j)}]_T &= \lim_{N\to\infty}\sum_{k=0}^{N-1}\Delta X_k^{(i)}\Delta X_k^{(j)} \\
&= \lim_{N\to\infty}\sum_{k=0}^{N-1}\left(A_{i,\cdot}X_k\Delta t + G_{i,\cdot}\Delta W_k\right)\left(A_{j,\cdot}X_k\Delta t + G_{j,\cdot}\Delta W_k\right) \\
&= \lim_{N\to\infty}\sum_{k=0}^{N-1}\left(A_{i,\cdot}X_k\Delta t + G_{i,\cdot}\Delta W_k\right)\left(A_{j,\cdot}X_k\Delta t + \Delta W_k^T(G_{j,\cdot})^T\right) \qquad (G_{j,\cdot}\Delta W_k \in \mathbb{R}) \\
&= \lim_{N\to\infty}\sum_{k=0}^{N-1}G_{i,\cdot}\Delta W_k\Delta W_k^T(G_{j,\cdot})^T \qquad (dW_t dt = dt dt = 0) \\
&= \lim_{N\to\infty}\sum_{k=0}^{N-1}\Delta t G_{i,\cdot}(G_{j,\cdot})^T \qquad (\forall i \neq j\ dW_k^{(i)} \perp\!\!\!\perp dW_k^{(j)},\ dW_t^{(i)}dW_t^{(i)} = dt) \\
&= T(GG^T)_{ij}.
\end{aligned}$$

This directly gives (29) and (30). The asymptotic convergence of the estimate can be shown by applying the arguments from Theorem (1) for each component, using (31). In particular, for all $i, j \in 1, ..., d$,

$$|\mathbb{E}(\widehat{GG^T})_{ij} - (GG^T)_{ij}| \leq C\left(\Delta t + \Delta t^{1/2}\right). \tag{32}$$

Indeed, if we let

$$\begin{aligned}
B_k^{(i)} &= \int_{k\Delta t} A_{i,\cdot}X_s ds \\
B_k^{(j)} &= \int_{k\Delta t} A_{j,\cdot}X_s ds.
\end{aligned}$$

Then, Cauchy-Schwarz yields

$$\begin{aligned}
\mathbb{E}[B_k^{(i)}B_k^{(j)}] &\leq K(\Delta t)^2 \\
\mathbb{E}[(B_k^{(i)})^2] &\leq K(\Delta t)^2 \\
\mathbb{E}[(B_k^{(j)})^2] &\leq K(\Delta t)^2
\end{aligned}$$

with $K \geq \sup_{s\in[0,t]}|AX_s|^2$. These estimates, along with $\mathbb{E}[G_{i,\cdot}\Delta W_k G_{j,\cdot}\Delta W_k] = \sum_{l=1}^m G_{i,l}G_{j,l}\Delta t$, directly yield (32) using the same argument as in the proof of Theorem (1).

Remark 10. *This closed form estimate for the additive noise matrix G is not used by Wang et al. [15]. Instead, they formulate a likelihood function and optimize simultaneously for drift and diffusion, shown in the next section.*

Remark 11. *Similarly to remark* (3)*, the process is distributionally invariant for all functions σ such that $\sigma(X_t, t)^T\sigma(X_t, t)$.*

Remark 12. *More sophisticated estimators are required for estimating $G(x)G(x)^T$ for the multiplicative noise SDE, as the diffusion is not constant in this case.*

5.2 Maximum Likelihood for multidimensional SDEs

In this section, we derive the density functions for X_t from the additive noise model and the multiplicative noise model, as well as the corresponding likelihood functions, given by the Radon-Nikodym derivative with respect to Brownian motion.

Theorem 4. *1. Let X_t solve (26) and $\theta = (A, GG^T)$. The discretized density of X_t is given by*

$$p_X^N = \left(\frac{1}{(2\pi\Delta t)^{d/2}|G|}\right)^N \exp\Bigg(\frac{-1}{2\Delta t}\sum_{i=0}^{N-1}\big(\Delta X_i (GG^T)^{-1}\Delta X_i - 2\Delta t (AX_i)^T(GG^T)^{-1}\Delta X_i + (AX_i)^T(GG^T)^{-1}AX_i\Delta t^2\big)\Bigg). \quad (33)$$

The likelihood function given observations $\{X_i\}_{i=0}^{N-1}$ is

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) = \lim_{N\to\infty}\frac{dP_X^N}{dP_W^N} = \exp\left(\int_0^T (AX_t)^T(GG^T)^{-1}dX_t - \frac{1}{2}\int_0^T (AX_t)^T(GG^T)^{-1}AX_t dt\right) \quad (34)$$

and admits the discrete approximation

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) \approx \exp\left(\sum_{i=0}^{N-1}(AX_i)^T(GG^T)^{-1}\Delta X_i - \frac{1}{2}\sum_{i=0}^{N-1}(AX_i)^T(GG^T)^{-1}AX_i\Delta t\right) \quad (35)$$

2. Let X_t solve (28) and $\theta = (A, G(x)G(x)^T)$. The discretized density of X_t is given by

$$p_X^N = \prod_{i=0}^{N-1}\frac{1}{(2\pi\Delta t)^{d/2}|G(X_i)|}\exp\Bigg(\frac{-1}{2\Delta t}\sum_{i=0}^{N-1}\big(\Delta X_i (G(X_i)G(X_i)^T)^{-1}\Delta X_i \quad (36)$$

$$-2\Delta t (AX_i)^T(G(X_i)G(X_i)^T)^{-1}\Delta X_i + (AX_i)^T(G(X_i)G(X_i)^T)^{-1}AX_i\Delta t^2\big)\Bigg). \quad (37)$$

Similarly, the likelihood function given observations $\{X_i\}_{i=0}^{N-1}$ is

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) = \exp\Big(\int_0^T (AX_t)^T(G(X_t)G(X_t)^T)^{-1}dX_t - \frac{1}{2}\int_0^T (AX_t)^T(G(X_t)G(X_t)^T)^{-1}AX_t dt\Big). \quad (38)$$

This admits the discrete approximation

$$\mathcal{L}(\{X_i\}_{i=0}^{N-1}|\theta) \approx \exp\Big(\sum_{i=0}^{N-1}(AX_i)^T(G(X_i)G(X_i)^T)^{-1}\Delta X_i - \frac{1}{2}\sum_{i=0}^{N-1}(AX_i)^T(G(X_i)G(X_i)^T)^{-1}AX_i\Delta t\Big). \quad (39)$$

Proof. Recall that a multivariate normal $\mathcal{N}(\mu, \Sigma)$ has p.d.f

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

We use the Euler-Maruyama discretization to approximate the transition densities. In the additive noise case, the transition obeys $X_{i+1}|X_i \sim \mathcal{N}(X_i + A\Delta t, \Delta t GG^T)$. We directly compute the discretized density(33)

$$p_X^N = \prod_{i=0}^{N-1} p(X_{i+1}|X_i) \qquad \text{(Markov property)} \quad (40)$$

$$= \frac{1}{(2\pi\Delta t)^{d/2}|G|}\exp\left(-\frac{1}{2\Delta t}\sum_{i=0}^{N-1}(X_{i+1}-X_i-A\Delta t)^T(GG^T)^{-1}(X_{i+1}-X_i-A\Delta t)\right) \quad (41)$$

$$= \frac{1}{(2\pi\Delta t)^{d/2}|G|}\exp\left(-\frac{1}{2\Delta t}\sum_{i=0}^{N-1}(\Delta X_i)^T(GG^T)^{-1}\Delta X_i - 2\Delta t(AX_i)^T(GG^T)^{-1}\Delta X_i + (AX_i)^T(GG^T)^{-1}AX_i\Delta t^2\right), \quad (42)$$

where we have used the fact that $(AX_i)^T(GG^T)^{-1}\Delta X_i = (\Delta X_i)^T(GG^T)^{-1}AX_i$. This follows from $(ABC)^T = (C^T B^T A^T)$ and $(GG^T)^{-1} = ((GG^T)^{-1})^T$. For the multiplicative noise case, the discretized density (37) is computed in the same way, but with $X_{i+1}|X_i \sim \mathcal{N}(X_i + A\Delta t, \Delta t G(X_i)G(X_i)^T)$.

To derive the likelihood function, we informally compute $\lim_{N\to\infty} \frac{P_X^N}{P_W^N}$, where the density of discretized multidimensional Brownian motion is given by

$$p_W^N = \frac{1}{(2\pi\Delta t)^{d/2}} \exp\left(-\frac{1}{2\Delta t}\sum_{i=0}^{N-1}(\Delta W_i)^T \Delta W_i\right) \tag{43}$$

We then apply the asymptotic approximation $(\Delta X_i)^T \Delta X_i = GG^T(\Delta W_i)^T \Delta W_i$ based on the stochastic calculus laws. Computing $\frac{dP_X^N}{dP_W^N} = \frac{p_X^N dX_t^N}{p_W^N dW_t^N}$ therefore yields the discrete approximation (35), which converges to the likelihood function (34) as $N \to \infty$. The same analysis produces the analagous quantities in the multiplicative noise setting. □

Remark 13. *Wang et al. [15] consider the law of X_t (33) as the likelihood function for the additive noise case. In particular, they use the BFGS optimizer to minimize the negative log-likelihood with respect to $\theta = (A, GG^T)$, such that each term in the log-likelihood represents the log of one of X_t's (estimated) transition densities. An analogous procedure is used for the multiplicative noise case.*

Remark 14. *If we do not have additive noise, then the solution to the SDE will no longer be a Gaussian process. Using the Euler-Maruyama discretization approximates each increment as Gaussian, with varying diffusion parameters, and hence introduces additional error to the approximation.*

□

5.3 Experiments with multivariate SDEs

In this section, we aim to recreate the experiments of Wang et al. [15] while also contributing minor extensions of some of their experiments. We follow their experimental setup by considering two cases, which have identifiable drift and diffusion terms (the matrices obey the necessary rank conditions). The first case is the additive noise model following Equation 26 with:

$$X_0 = \begin{bmatrix} 1.87 \\ -0.98 \end{bmatrix}, \; A = \begin{bmatrix} 1.76 & -0.1 \\ 0.98 & 0 \end{bmatrix} \; \textit{and } G = \begin{bmatrix} -0.11 & -0.14 \\ -0.29 & -0.22 \end{bmatrix}.$$

The second case is the multiplicative noise model following Equation 28 with:

$$X_0 = \begin{bmatrix} 1.87 \\ -0.98 \end{bmatrix}, \; A = \begin{bmatrix} 1.76 & -0.1 \\ 0.98 & 0 \end{bmatrix}, \; G_1 = \begin{bmatrix} -0.11 & -0.14 \\ -0.29 & -0.22 \end{bmatrix} \; \textit{and } G_2 = \begin{bmatrix} -0.17 & 0.59 \\ 0.81 & 0.18 \end{bmatrix}.$$

Across all our evaluations, we use Wang et al's [15] parameter estimation method of using MLE with a BFGS solver, to simultaneously estimate all matrices in the drift and diffusion. The log-likelihood function is the negative log of the product of (estimated) transition densities of X_t, given in (33) and (37). We found that other estimation methods performed less well.

Our first set of experiments on multivariate processes with additive noise aim to extend the experimental results found in the Section 3. Figure 5 shows the corresponding drift and diffusion errors as we change dt while T is static (left) and change T while N is constant (right). Importantly, the conclusions gathered from both plots align nicely with the theoretical and experimental conclusions in Section 3. Specifically, we see that drift error depends on T, while diffusion error depends on dt.

In our second set of experiments, we directly recreate the experiments done by [15] by increasing the number of trajectories, while supplying only 50 data points per trajectory on the time interval $[0, 1]$, each with the initial starting point. We record the mean squared error of estimating the drift and diffusion parameters as the number of trajectories is increased. We note that increasing the number of trajectories decreases diffusion and drift error for both the additive and multiplicative noise scenarios (Figure 6).

There may be scenarios under which we cannot expect each trajectory to have the same starting point. Thus, our next experiment simulates random trajectory starting points given our standard additive noise experiment. We find that this makes the estimation of drift more difficult (Figure 7), but the negative trend

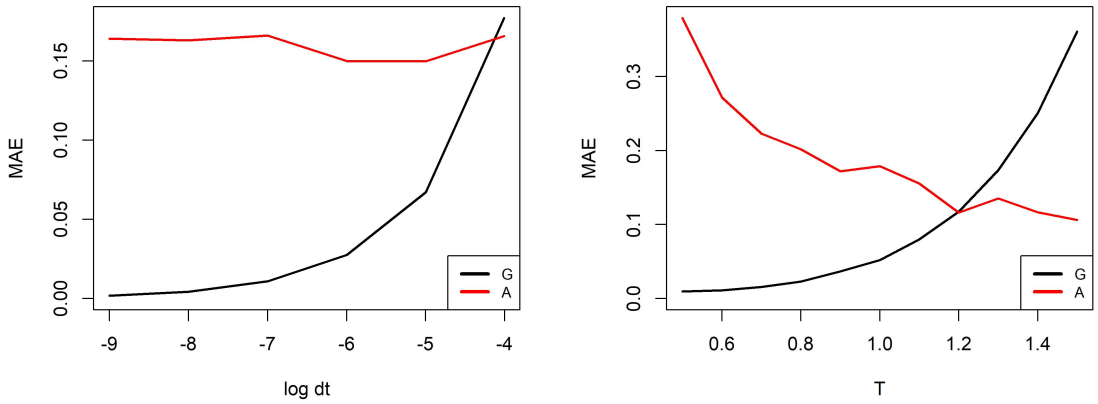

Figure 5: Left: Mean Absolute Error (MAE) after estimating drift (red) and diffusion (black) plotted against different levels of log dt. Right: MAE after estimating drift (red) and diffusion (black) plotted against different levels of T. The number of observations N is kept constant.

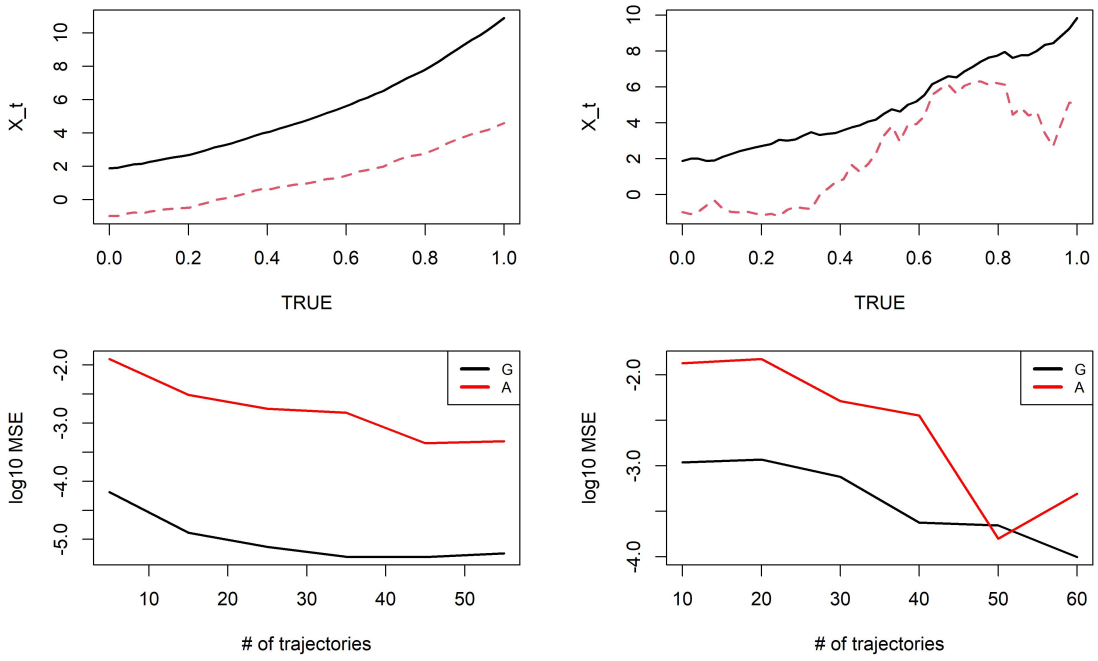

Figure 6: Top Left: an example trajectory of the 2D additive noise process tested by [15]. Bottom Left: log MAE for estimating the drift matrix (red) and diffusion matrix GG^T versus the number of trajectories in the additive noise case. Top Right: an example trajectory of the 2D multiplicative noise process texted by [15]. Bottom right: log MAE for estimating the drift matrix (red) and diffusion matrix $G(x)G(x)^T$ versus the number of trajectories in the multiplicative noise case.

of error with increasing number of trajectories is consistent with the experiments from [15] (Figure 6). The estimation of diffusion does not seem to be affected by the starting point distribution.

Our final experiment in this section tries to infer the relationship between the dimension of the process and estimation error given the same amount of observational data. For each dimension $d = 2, 3, 4$, additive noise processes are constructed with randomly generated drift and diffusion matrices, and then simulated via the Euler-Maruyama discretization from Equation (31). As shown on the right side of Figure 7, the

number of dimensions does not seem to have a huge effect on accuracy, though we noticed that finding the global minimum of the likelihood becomes more difficult and the optimization becomes more computationally expensive.

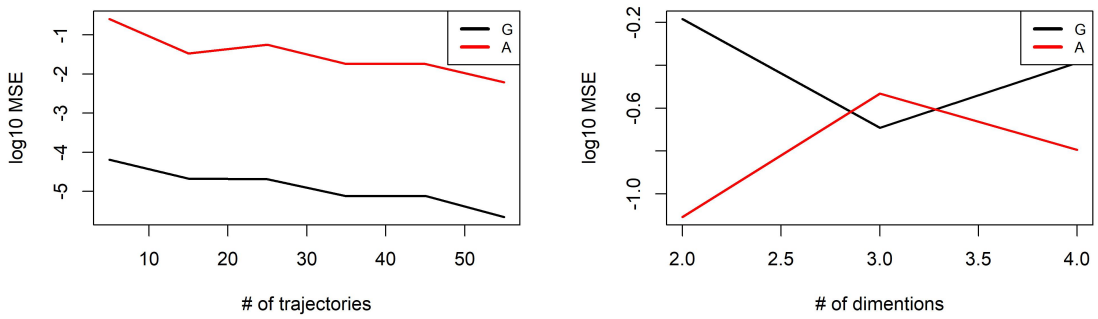

Figure 7: Left: Mean Absolute Error (MAE) after estimating drift (red) and diffusion (black) plotted against various numbers of trajectories with random starting points X_0. Right: MAE after estimating drift (red) and diffusion (black) plotted against the process dimension. Note that both figures correspond to the additive noise model.

6 Distributed delay stochastic differential equations

A causal inference research group, Manten et al. [9] recently came out with work on path-dependent SDEs:

$$d\mathbf{X}_t = \mu(\mathbf{X}_{[0,t]})dt + \sigma(\mathbf{X}_{[0,t]})dW_t, \qquad \mathbf{X}_0 = \mathbf{x}_0, \tag{44}$$

which are also known as distributed delay SDEs [12, 2, 14].

The primary difference between this and previous SDEs is the loss of the Markov property, with the previous history having an important effect. One special case of Equation 44 that we are especially interested in is the functional convolutional models with autocorrelated noise [6]:

$$y(t) = \int_0^D \beta(s)x(t-s)ds + \epsilon(t). \tag{45}$$

Remark 15. *It turns out that some path dependent SDEs can be embedded in higher-dimensional linear SDEs [9]. Consider the linear 3D SDE with additive noise:*

$$dX_t = AX_t dt + GdW_t, \tag{46}$$

where

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } G = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Consequently, the second variable evolves based on the current value of the third variable:

$$dX_t^{(2)} = X_t^{(3)}dt + 0.3dW_t^{(2)} \tag{47}$$

and the third variable evolves based on the current value of the first variable:

$$dX_t^{(3)} = X_t^{(1)}dt \implies X_t^{(3)} = \int_0^t X_s^{(1)}ds. \tag{48}$$

Thus, we also see that the second variable evolves based on the full history of the first variable:

$$dX_t^{(2)} = \left(\int_0^t X_s^{(1)}ds\right)dt + 0.3dW_t^{(2)} \tag{49}$$

Remark 16. *We may also generalize the previous example to accommodate a broad class of SDEs with distributed delay, by introducing time-inhomogeneity in the coefficients of the matrix $A(t)$. For example, consider*

$$dX_t = A(t)X_t dt + G dW_t, \tag{50}$$

where

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha(t) \\ \beta(t) & 0 & 0 \end{bmatrix} \text{ and } G = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This yields something similar to the convolutional model that we are interested in (Equation 45):

$$dX_t^{(2)} = \alpha(t)\left(\int_0^t \beta(s)X_s^{(1)} ds\right) dt + 0.3 dW_t^{(2)} \tag{51}$$

6.1 Experiments on path-dependent SDEs

To test parameter estimation on path-dependent SDEs, we consider the simple time-homogeneous linear 3D example from remark (15), and again apply parameter estimation via MLE with the BFGS optimizer on the negative log of the product of transition densities.

For this experiment, we set the number of time steps per trajectory to be 1000, and we consider up to 20 trajectories (# of trajectories $\in \{2, 3, 5, 10, 20\}$). The observation duration is fixed at $[0, 1]$, which fixes the time step to be $\Delta t = \frac{1}{1000}$. The substantial increase in data predictably boosts the estimation of the diffusion GG^T, as mean-squared errors are on the order of 0.0001 in Figure 8 rather than 0.1 as was shown in Figure (7) for time steps of size $\Delta t = \frac{1}{50}$ on $d = 3$.

Figure 8: Mean squared error (MSE) of estimating drift (red) and diffusion (black) versus the number of trajectories for the path dependent SDE considered in [9] and Remark 15.

Although increasing the number of trajectories has a noticeable effect on decreasing the MSE of the estimate of the drift matrix A (Figure 8), the effect is more moderate than in previous experiments from Section 5. In all settings, the MSE for estimating A was at least 0.24. Moreover, we strangely observed increasing MSE going from 10 trajectories to 20. This suggests that the task of parameter estimation may be too difficult given the provided data. In particular, the observation period may need to be significantly extended to succeed in this setting.

Furthermore, it seems that the scale of the diffusion is critical in inhibiting the parameter estimation of A, even though GG^T is precisely estimated. Indeed, when we reduced the scale of Brownian motion by an order of 10, such that the non-zero entries of G are 0.03 rather than 0.3, the MSE for A is less than 0.025 for just 2 trajectories (not shown).

7 Conclusions and future work

In this work, we have overviewed some of the basic methods for estimating the drift and diffusion parameters of an SDE from discrete observations. For one-dimensional SDEs, we saw that constant diffusion (additive noise) may be estimated directly using quadratic variation analysis, and that linear drift may be estimated using maximum likelihood estimation, which has a closed-form solution. We then proved that higher dimensional processes have a similar quadratic variation estimate for diffusion and an analogous likelihood function. An analyst may also use a likelihood function, either the Radon-Nikodym derivative with respect to Brownian motion, or the density of the process, to optimize over diffusion and drift parameters. Our experimental evaluation demonstrates the important dependence on parameter estimation accuracy and the richness of the data, namely the observation period $[0, T]$, the time granularity Δt, and the number of trajectories.

Although parameter estimation has been an active area of research for decades, we have observed that most evaluations have been limited to simple settings, featuring low dimensionality and simple functional models, such as OU processes [10, 11, 15], though a recent paper uses deep learning to estimate time-inhomogeneous diffusion [7]. These empirical limitations may be attributed to the demand for high-quality data in order to adequately estimate transition densities, and the computation required to optimize over large parameter spaces.

In future work, we aim to extend the aforementioned experiments to high dimensions, and to impose sparsity, such that only causal features have non-zero components. In particular, we would like to further explore parameter estimation for path-dependent SDEs to learn lagged relationships.

Code for all experiments can be found at `https://github.com/HydroML/SDEParameterEstimation`.

References

[1] Jaya PN Bishwal. *Parameter estimation in stochastic differential equations*. Springer, 2007.

[2] Evelyn Buckwar. *Euler-Maruyama and Milstein approximations for stochastic functional differential equations with distributed memory term*. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, 2005.

[3] Christa Cuchiero, Sara Svaluto-Ferro, and Josef Teichmann. "Signature SDEs from an affine and polynomial perspective". In: *arXiv preprint arXiv:2302.01362* (2023).

[4] Flávio B Gonçalves and Pedro Franklin. "On the definition of likelihood function". In: *arXiv preprint arXiv:1906.10733* (2019).

[5] Niels Hansen and Alexander Sokol. "Causal interpretation of stochastic differential equations". In: (2014).

[6] Joseph Janssen et al. "Learning from limited temporal data: Dynamically sparse historical functional linear models with applications to Earth science". In: *arXiv preprint arXiv:2303.06501* (2023).

[7] Andrzej Kałuża et al. "Deep learning-based estimation of time-dependent parameters in Markov models with application to nonlinear regression and SDEs". In: *arXiv preprint arXiv:2312.08493* (2023).

[8] Alain Le Breton. "Parameter estimation in a linear stochastic differential equation". In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians: held at Prague, from August 18 to 23, 1974*. Springer. 1977, pp. 353–366.

[9] Georg Manten et al. "Signature Kernel Conditional Independence Tests in Causal Discovery for Stochastic Processes". In: *arXiv preprint arXiv:2402.18477* (2024).

[10] Grigorios A Pavliotis. "Stochastic processes and applications". In: *Texts in Applied Mathematics* 60 (2014).

[11] Asger Roer Pedersen. "A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations". In: *Scandinavian journal of statistics* (1995), pp. 55–71.

[12] Alexandre René and André Longtin. "Mean, covariance, and effective dimension of stochastic distributed delay dynamics". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 27.11 (2017).

[13] R Sh Liptser-AN Shiryayev. *Statistics of Random Processes*. Springer-Verlag, New York, 1977.

[14] Shahab Torkamani and Eric A Butcher. "Stochastic parameter estimation in nonlinear time-delayed vibratory systems with distributed delay". In: *Journal of Sound and Vibration* 332.14 (2013), pp. 3404–3418.

[15] Yuanyuan Wang et al. "Generator Identification for Linear SDEs with Additive and Multiplicative Noise". In: *Advances in Neural Information Processing Systems* 36 (2024).