Semester project:

Generalized method of moments for learning unknown parameters in SDEs.

January 18, 2023

Dept. of Mathematics
Scientific Computing and Uncertainty Quantification (CSQI) chair,
EPFL university

student: Konstantin Medyanikov (Master of applied mathematics)
supervisor: Prof. Fabio Nobile
cosupervisor: Andrea Zanoni

1 Introduction

In this research project, we propose a new technique called Method of Moments (MoM) for estimating unknown parameters in stochastic differential equations (SDEs) that are commonly used in various applications. The MoM makes use of either continuous-time trajectories $(X_t)_{t\in[0,T]}$ or discrete-time observations $\{X_j\}_{j=0}^{J}$ (as shown in the Figure 1). As it is often the case that only discrete data is available in practice, we will focus primarily on studying the MoM technique when applied to this type of data.

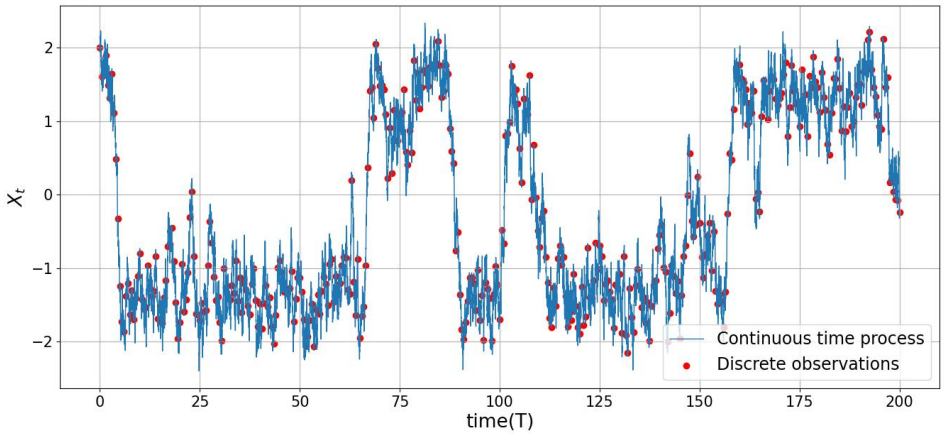


Figure 1: Example of process called Brownian motion in a bistable potential.

We remark, that we will consider a specific class of ergodic stochastic processes, which are characterized by the presence of a unique invariant measure. This assumption is crucial in the development of our MoM technique and enables us to utilize various statistics to estimate the unknown coefficients. The main idea behind our research is the consideration of the stationary Fokker-Planck equation. Through simple manipulations, as outlined in Section 2, we are able to derive an expression in terms of unknown parameters and moments of the underlying diffusion process. Afterwards, to estimate these moments and find the unknown parameters, we utilize the ergodic theorem. However, the class of processes we consider is restricted, it still has a range of applications, such as in the modeling of interest rates using the Cox-Ingersoll-Ross process in finance.

In this research, we will delve into the details of the MoM. To this end, we will begin by presenting the derivation of the MoM in Section 2. We will then proceed to compare the MoM technique to classical methods in Section 3. Additionally, we will derive error bounds for both the case where continuous time data is available and for the case with discrete observations in Section 4. We will also extend our approach to the setting where the actual model is unknown in Section 5 and investigate the application of MoM on processes with multiplicative noise in Section 6. Finally, we will derive multidimensional extension of the MoM in Section 7. Through this comprehensive analysis, we will gain a deep understanding of the capabilities and limitations of the MoM, and its performance in various contexts. Furthermore, in Section 8, we will discuss practical considerations and implementation details for applying the MoM in real-world scenarios.

2 Methodology

Settings

Let $\{P_n\}_{n=0}^{\infty}$ be a sequence of polynomials such that P_n has degree n (e.g. monomials, Chebyshev, Hermite polynomials). We focus on a one-dimensional stochastic process $(X_t)_{t\in[0,T]}$ that satisfies the following stochastic differential equation (SDE)

$$dX_t = f(X_t;\alpha)dt + \sqrt{2g(X_t;\sigma)}dW_t, \qquad X_0 = x, \tag{1}$$

where the drift and diffusion terms $f(x;\alpha)$ and $g(x;\sigma)$ are given by polynomials of the form

$$f(x;\alpha) = \sum_{k=0}^{K} \alpha_k P_k(x) \qquad \text{and} \qquad g(x;\sigma) = \sum_{e=0}^{E} \sigma_e P_e(x),$$

and where the parameters $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^K$ and $\sigma = (\sigma_1, \dots, \sigma_l) \in \mathbb{R}^E$ are unknown. The initial conditions in (1) can be taken to be either deterministic or random. We assume to know at least one realisation of the process X_t in the interval $[0, T]$. In order to apply our approach we have to make several assumption on the underlying stochastic process.

Assumption 2.1

- the function $g(;\sigma)$ is strictly positive, i.e., $g(x;\sigma) > 0$ for all $x \in \mathbb{R}$.[1]
- the process X_t is ergodic with the unique invariant measure μ which has density ρ with the respect to Lebesque measure.
- $\lim_{x\to\pm\infty} x^n\rho(x) = 0$ and $\lim_{x\to\pm\infty} x^n\rho'(x) = 0$, for $n = 1, \dots, L$.

Method of moments (MoM)

In order to solve the inference problem, we employ the generalised method of moments. We consider density ρ of the invariant measure μ for the process X_t (1), where, ρ has to satisfy the stationary Fokker-Planck equation

$$-\frac{d}{dx}(f(x)\rho(x)) + \frac{d^2}{dx^2}(g(x)\rho(x)) = 0.$$

Then, we multiply equation above by P_n with $n = 1, 2 \dots L$. We would get the system of equation

$$-P_n(x)\frac{d}{dx}(f(x)\rho(x)) + P_n(x)\frac{d^2}{dx^2}(g(x)\rho(x)) = 0. \qquad n = 1, 2, \dots, L.$$

Integrating both sides over $\mathbb{R}$

$$-\int_{\mathbb{R}} P_n(x)\frac{d}{dx}(f(x)\rho(x))dx + \int_{\mathbb{R}} P_n(x)\frac{d^2}{dx^2}(g(x)\rho(x))\,dx = 0 \qquad n = 1, 2, \dots, L,$$

[1] In diffusion term we consider square-root of the objective function $g(x;\sigma)$, whereas it is done to have more explicit and nicer expression for unknown parameters σ. And in general settings, this assumption can be removed.

We apply successfully integration by parts, in order to remove differential operator from density function ρ. As well, we use assumption on the $\rho(x)$, that $\lim_{x\to\pm\infty} x^n\rho(x) = 0$ and $\lim_{x\to\pm\infty} x^n\rho'(x) = 0$,

$$\int_{\mathbb{R}} P_n'(x)f(x)\rho(x)dx + \int_{\mathbb{R}} P_n''(x)g(x)\rho(x)dx = 0 \qquad n = 1, 2, \dots, L.$$

Using the definitions of $f(x)$ and $g(x)$, we obtain

$$\sum_{k=0}^{K} \alpha_k \mathbb{E}^{\mu}\left[P_k(X)P_n'(X)\right] + \sum_{e=0}^{E} \sigma_e \mathbb{E}^{\mu}\left[P_e(X)P_n''(X)\right] = 0 \qquad n = 1, 2, \dots, L,$$

where the superscript μ means that the random variable X is distributed according to measure μ. Now, let us look on the obtained linear system of equations for unknown parameters $\theta = (\alpha, \sigma)$

$$\begin{cases} \sum_{k=0}^{K} \alpha_k \mathbb{E}^{\mu}\left[P_k(X)P_1'(X)\right] = 0 & , n = 1 \\ \sum_{k=0}^{K} \alpha_k \mathbb{E}^{\mu}\left[P_k(X)P_2'(X)\right] + \sum_{e=0}^{E} \sigma_e \mathbb{E}^{\mu}\left[P_e(X)P_2''(X)\right] = 0 & , n = 2 \\ \vdots & \\ \sum_{k=0}^{K} \alpha_k \mathbb{E}^{\mu}\left[P_k(X)P_L'(X)\right] + \sum_{e=0}^{E} \sigma_e \mathbb{E}^{\mu}\left[P_e(X)P_L''(X)\right] = 0 & , n = L. \end{cases} \tag{2}$$

One can notice that the system has one trivial solution $\mathbf{0} = (0, 0 \dots, 0)$. Therefore, we propose an additional equation that can be derived using quadratic variation Q_T of the process $(X_t)_{t\in[0,T]}$.Indeed we have

$$\int_0^T g(X_t, \sigma)dt = \sum_{e}^{E} \sigma_e \int_0^T P_e(X_t)dt = \frac{Q_T}{2}. \tag{3}$$

By using ergodicity of the process, in particular, by the ergodic theorem we have for a function $h : \mathbb{R} \to \mathbb{R}$

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T h(X_t)dt = \mathbb{E}^{\mu}\left[h(X)\right] = \int_{\mathbb{R}} h(x)\rho(x)dx \qquad \text{a.s.},$$

which implies that if T is sufficiently large, we can apply ergodic theorem and rewrite equation (3)

$$\sum_{e}^{E} \sigma_e \mathbb{E}^{\mu}\left[P_e(X)\right] \approx \frac{Q_T}{2T}. \tag{4}$$

By adding additional constraint we get rid-off trivial zero solution, and solving system (2) would give us estimation for the unknown parameters $\theta = (\alpha^T, \sigma^T)$ of the SDE (1).

Application of MoM

Our goal is to determine unknown parameters from given observations. We assume to have an access to discrete observations or to the an entire path X_t, $t \in [0, T]$, where the length of the path is fixed, and equal to T. To apply MoM we have to estimate a number of expectations of the from $\mathbb{E}^{\mu}\left[P_n(X)P_m'(X)\right]$. Like in a case of

quadratic variation we use ergodic theorem to approximate the expectations. Given discrete sample of equidistant observations $\{X_j\}_{j=0}^J$ of the path $(X_t)_{t\in[0,T]}$, with sufficiently small time step $\Delta t = \delta$ and $J\delta = T$.

$$\mathbb{E}^\mu\left[P_n(X)P'_m(X)\right] \approx \frac{1}{T}\int_0^T P_n(X_t)P'_m(X_t)dt \approx \frac{1}{J}\sum_{j=0}^{J} P_n(X_j)P'_m(X_j) =: \widehat{M}'_{n,m}$$

$$\mathbb{E}^\mu\left[P_n(X)P''_m(X)\right] \approx \frac{1}{T}\int_0^T P_n(X_t)P''_m(X_t)dt \approx \frac{1}{J}\sum_{j=0}^{J} P_n(X_j)P''_m(X_j) =: \widehat{M}''_{n,m}$$

In order to estimate quadratic variation Q_T we use the following result.

$$Q_T = \lim_{J\to\infty}\sum_{j=0}^{J}\left|X_{t_{k+1}} - X_{t_k}\right|^2 \quad \text{in prob.}$$

$$\widehat{Q}_T = \sum_{j=0}^{J-1}\left(X_{j+1} - X_j\right)^2$$

Estimator $\widehat{Q}_T$ is called the high-frequency limit. It should be noted that T can be arbitrarily small,as well as the process X_t can be in non-stationary phase. This is in contrast to the estimation we use to approximate the expectations in the system (2). where, we assume the final time T to be large enough and process's measure to be close to the stationary distribution, in order to apply it's ergodic properties.
Recalling again the system (2), now we can rewrite it using estimators

$$\begin{pmatrix} \widehat{M}'_{K,1} & \widehat{M}'_{K-1,1} & \dots & 1 & 0 \\ \widehat{M}'_{K,2} & \widehat{M}'_{K-1,2} & \dots & \widehat{M}''_{2,2} & 1 \\ \widehat{M}'_{K,3} & \widehat{M}'_{K-1,3} & \dots & \widehat{M}''_{2,3} & \widehat{M}''_{1,3} \\ \vdots & & & & \\ \widehat{M}'_{1,L} & \widehat{M}'_{2,L} & \dots & \widehat{M}''_{E-1,L} & \widehat{M}''_{E,L} \\ 0 & 0 & \dots & \widehat{M}_{E-1,L} & \widehat{M}_{E,L} \end{pmatrix} \begin{pmatrix} \alpha_K \\ \alpha_{K-1} \\ \vdots \\ \alpha_0 \\ \sigma_E \\ \vdots \\ \sigma_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \frac{\widehat{Q}_T}{2T} \end{pmatrix},$$

or in compact form

$$\widehat{\mathbb{M}}_T\hat{\theta}_T = q_T. \tag{5}$$

Finally, to solve the system (5) we look for least square solution:

$$\hat{\theta} = \text{argmin}_{\theta\in\mathbb{R}^{K+E}}\|\widehat{\mathbb{M}}_T\theta - q_T\|_2$$

Remark 2.1 Important to note that proposed approach is not restricted only to one-dimensional stochastic process, but might be generalised and adapted to multi-dimensional case (7).
Remark 2.2 In the proposed settings (1), we consider polynomial expression for drift and diffusion terms. However, the actual model can be expressed with any sufficiently smooth functions, that we try to fit with polynomial expansion. We study such case in the section (5). Moreover, the problem could be seen in a more general framework if we replace the sequence of polynomials $\{P_n\}_{n=0}^\infty$ with a set of basis function of an appropriate Hilbert space.

3 Additive noise processes

In this section, we will be using the Method of Moments (MoM) to analyze stochastic processes with additive noise.[1] The drift term in these processes is a polynomial of the form $f(X_t) = \sum_{k=0}^{K} \alpha_k X_t^k$, while the diffusion term is a constant $g(X_t) = \sqrt{2\sigma_0}$. We will be assuming that we have knowledge of the exact forms of both the drift and diffusion functions.

Ornstein-Uhlenbeck proces

The Ornstein-Uhlenbeck process with unknown drift and diffusion coefficients $\alpha_1 < 0$, $\sigma_0 > 0$:

$$dX_t = \alpha_1 X_t dt + \sqrt{2\sigma_0} dW_t. \tag{6}$$

In order to illustrate the methodology in details, for the Ornstein-Uhlenbeck process (6) we are going to derive the whole scheme, describing each step. We start by considering the stationary Fokker-Planck equation:

$$-\frac{d}{dx}\left(f\left(x\right)\rho(x)\right) + \frac{d^2}{dx^2}\left(g\left(x\right)\rho(x)\right) = 0.$$

As we mentioned before, we assume to know the exact form of the drift and diffusion term. In our case it is: $f\left(x, \alpha_1\right) = \alpha_1 x$ and $g\left(x, \sigma_0\right) = \sigma_0$, Therefore, we can rewrite the equation:

$$-\alpha_1 \frac{d}{dx}(x\rho(x)) + \sigma_0 \frac{d^2}{dx^2}(\rho(x)) = 0,$$

- Multiply both sides of the equation by monomials x^n with $n = 1, 2 \dots k$:

$$-\alpha_1 x^n \frac{d}{dx}(x\rho(x)) + \sigma_0 x^n \frac{d^2}{dx^2}(\rho(x)) = 0$$

- Integrating both sides:

$$-\alpha_1 \int_{\mathbb{R}} x^n \frac{d}{dx}(x\rho(x))dx + \sigma_0 \int_{\mathbb{R}} x^n \frac{d^2}{dx^2}\left(\rho(x)\right)dx = 0,$$

- Integrating by parts, combined with the assumption on the invariant measure $\lim_{x\to\pm\infty} x^n\rho(x) = 0$, as well as $\lim_{x\to\pm\infty} x^n\rho'(x) = 0$[2], $n = 1, \dots, L$, we obtain:

$$n\alpha_1 \int_{\mathbb{R}} x^n \rho(x)dx - n\sigma_0 \int_{\mathbb{R}} x^{n-1} \frac{d}{dx}(\rho(x))dx = 0,$$

- One more integration by parts:

$$n\alpha_1 \int_{\mathbb{R}} x^n \rho(x)dx + \sigma_0 n(n-1) \int_{\mathbb{R}} x^{n-2} \rho(x)dx = 0,$$

[1] When the amplitude of the noise does not depend on the state of the system, we say that noise is additive, otherwise multiplicative.

[2] For additive noise processes, the derivation of the invariant density is relatively simple and given by so-called Gibbs distribution: $\rho(x) = \frac{1}{Z} e^{\frac{V(x)}{\sigma}}$, where Z is normalizing constant and $V'(x) = f(x;\alpha)$ (see Prop.4.2 Pavliotis 2014, pages: 109-111). And due to the exponential nature of the invariant distribution assumption is always satisfied.

- Recalling the expectation with respect to invariant measure μ:

$$\alpha_1 \mathbb{E}^{\mu}\left[X^n\right] + \sigma_0(n-1)\mathbb{E}^{\mu}\left[X^{n-2}\right] = 0,$$

-By applying the above procedure using different monomials, we get a system of equations:

$$\begin{cases} \alpha_1 \mathbb{E}[X] = 0 & , n = 1 \\ \alpha_1 \mathbb{E}\left[X^2\right] + \sigma_0 = 0 & , n = 2 \\ \alpha_1 \mathbb{E}\left[X^3\right] + 2\sigma_0 \mathbb{E}[X] = 0 & , n = 3 \\ \vdots & \\ \alpha_1 \mathbb{E}\left[X^k\right] + (k-1)\sigma_0 \mathbb{E}[X^{k-2}] = 0 & , n = k \end{cases} \tag{7}$$

-In in matrix form and using the notation $\mathbb{E}\left[X^k\right] =: M_k$, we have:

$$\begin{pmatrix} M_1 & 0 \\ M_2 & 1 \\ M_3 & 2M_1 \\ \vdots & \\ M_k & (k-1)M_{k-2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \sigma_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It is worth noting that one of the solutions to the system in equation (7) is to set $\alpha_1 = \sigma_0 = 0$. To avoid this trivial solution, an additional constraint is needed. As previously mentioned, we can obtain an additional equation by considering the quadratic variation Q_T of the stochastic process $(X_t)_{t\in[0,T]}$. Noting that $\sigma 0 = \frac{Q_T}{2T}$we get the following system to solve:

$$\begin{pmatrix} M_1 & 0 \\ M_2 & 1 \\ M_3 & 2M_1 \\ \vdots & \\ M_k & (k-1)M_{k-2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \sigma_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \frac{Q_T}{2T} \end{pmatrix}. \tag{8}$$

Brownian motion in a bistable potential

Another simple model we are going to consider is a Brownian motion in a bistable potential (9) with unknown coefficients $\alpha_3 < 0$, $\alpha_1 > 0$, $\sigma_0 > 0$:

$$dY_t = \left(\alpha_3 Y_t^3 + \alpha_1 Y_t\right) dt + \sqrt{2\sigma_0} dW_t \tag{9}$$

By following similar steps to the ones above, one obtain the following system:

$$\begin{pmatrix} M_3 & M_1 & 0 \\ M_4 & M_2 & 1 \\ M_5 & M_3 & 2M_1 \\ \vdots & & \\ M_{k+2} & M_k & (k-1)M_{k-2} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_2 \\ \alpha_1 \\ \sigma_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \frac{Q_T}{2T} \end{pmatrix} \tag{10}$$

Results and comparison to classical techniques

To assess the effectiveness of our new methodology, we simulated both processes X_t and Y_t using the Euler-Maruyama scheme with a time step of h=0.01. We use the resulting paths to determine the accuracy of the estimated parameters relative to the true value, as well as to compare the MoM to the classical approaches3, in particular we consider the Likelihood technique that is quite popular when dealing with an additive framework. Figures (2) and (3) shows the estimated coefficients over different final time T, in addition we plot the dynamics of the error, calculated using the euclidean norm $\|\theta\|_2$.

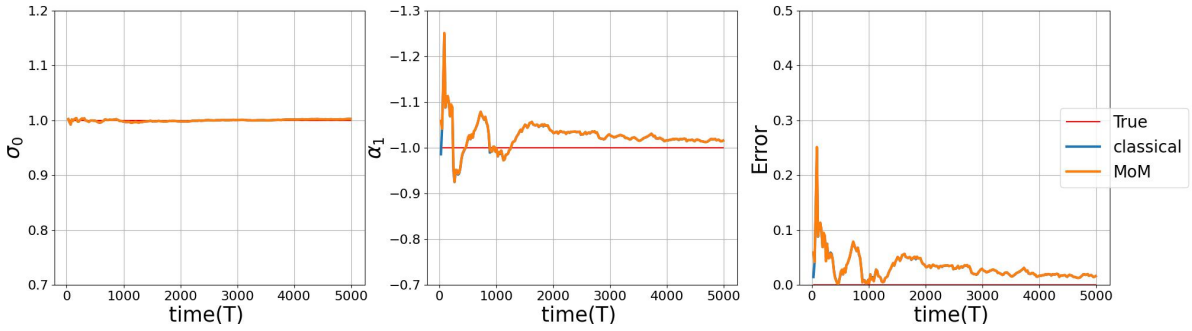


Figure 2: MoM and Classical approach applied for OU process X_t with $\alpha_1 = -1$ and $\sigma_0 = 1$, $X_0 = 0$

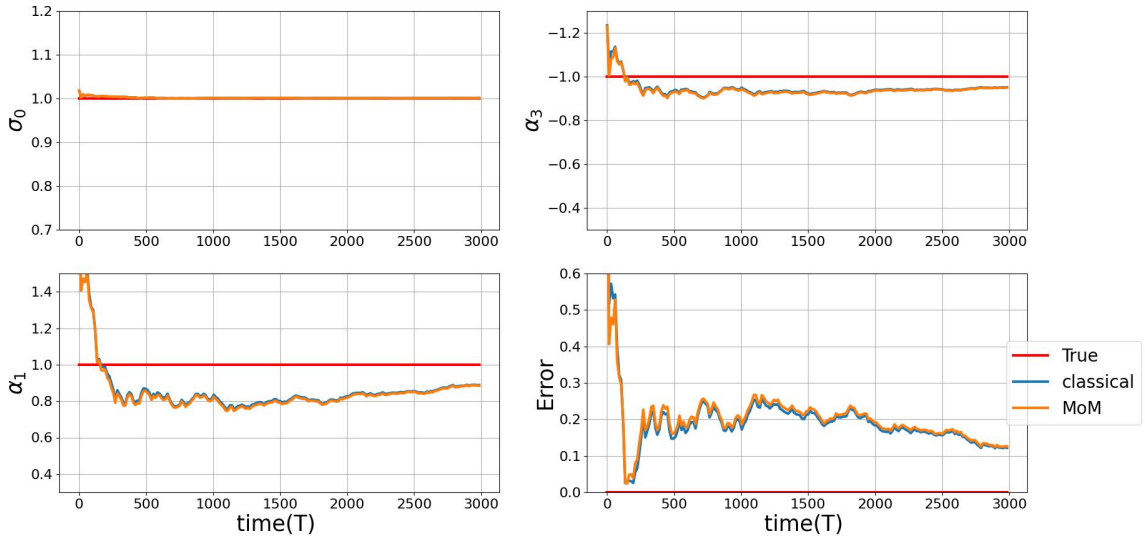


Figure 3: MoM and Classical approach applied for bistable potential process Y_t with $\alpha_3 = -1$, $\alpha_1 = +1$ and $\sigma_0 = 1$, $Y_0 = 0$

We observe that for both the OU process and the bistable potential process, the two approaches exhibit similar error dynamics. This can be attributed to the fact that both approaches utilize similar statistics. Additionally, both methods demonstrate good convergence to values that are relatively close to the actual ones.
Important to mention that, despite exhibiting similar dynamics, the two techniques

are fundamentally different. In fact, we could say that the MoM is more restricted, as it relies on time averages to estimate moments, while the Likelihood method directly uses time averages. Therefore, for the MoM to be effective, we not only need a large number of observations, but also a sufficiently large final time T. In the next section, we will be estimating the error bound associated with our approach, and also examining the role of time in the process. However, it is worth noting that both approaches have their own set of advantages and limitations, in particular the MoM can be applied to estimate parameters in non-trivial diffusion function, whereas the likelihood method works only with the drift term.

Remark 3.1
We use approach presented in the book 'Stochastic processes and applications'(see Section 5.3 Parameter Estimation for Stochastic Differential Equations Pavliotis 2014, p. 156-162). Where estimation of the diffusion and drift term are done separately, which is not a case for the MoM. Diffusion coefficient estimated using quadratic variation: $\widehat{\sigma}_T^2 = \frac{1}{T}\sum_{j=0}^{J-1}(X_{j+1} - X_j)^2$. For the drift term we use MLE estimate, where likelihood function is given by the Radon–Nikodym derivative (12) of the stochastic process X_T (11) with respect to the browien motion:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \tag{11}$$

$$\frac{d\mathbb{P}_{X_T}}{d\mathbb{P}_B} = \exp\left(\int_0^T b\left(X_s;\theta\right)dX_s - \frac{1}{2}\int_0^T \left(b\left(X_s;\theta\right)\right)^2 ds\right), \tag{12}$$

where, in order to approximate integrals in equation (12) we use the same statistics as in a case of MoM.

4 Error bound

The drawback of our approach is that we assume to have sufficiently enough time to estimate moments via time-average, in other words to apply ergodic theorem. In this section we want to estimate the error bound related to the error produced by time-average estimator of moments. We start by considering following proposition.

Proposition 4.1 Let $\{X_t\}_{t\geqslant 0}$ be a second-order stationary process on a probability space with mean μ and covariance $C(t)$, and assume that $C(t) \in L^1(0, +\infty)$. Then

$$\mathbb{E}\left|\frac{1}{T}\int_0^T X_s ds - \mu_1\right|^2 \leqslant \frac{2}{T}\int_0^{+\infty} C_{X_t}(u)du,$$

Which implies

$$\lim_{T\to+\infty} \mathbb{E}\left|\frac{1}{T}\int_0^T X_s ds - \mu_1\right|^2 = 0.$$

We will note μ_n the nth moment of the process X_t under invariant measure μ

Proof. (See Proposition 1.3 Pavliotis 2014, pages: 8-9) □

Definition 4.1: A stochastic process $X_t \in L^2$ is called second-order stationary, if the first moment $\mathbb{E}X_t$ is a constant and the covariance function $\mathbb{E}\left[(X_t - \mu)(X_s - \mu)\right]$ depends only on the difference $t - s$:

$$\mathbb{E}[X_t] = \mu_1, \quad \mathbb{E}\left[(X_t - \mu_1)(X_s - \mu_1)\right] = C(t - s).$$

The idea would be to consider underlying stochastic process X_t to start from it's invariant distribution $X_0 \sim \mu$[1]. In such framework, we can say that our process is strictly stationary. In particular, we have that

$$\mathbb{E}^{X_0\sim\rho}\left[X^n_{t+s}\right] = \mathbb{E}^{\mu}\left[X^n\right] = \mathbb{E}^{X_0\sim\rho}\left[X^n_t\right],$$

which would imply that $\mathbb{E}[X_t]$ is constant. As well we know that at any time $s \in [0, T]$ process follows invariant distribution $X_s \sim \rho$, therefore

$$\mathbb{E}^{X_0\sim\rho}\left[\left(X^n_{t_1+s} - \mu_n\right)\left(X^n_{t_2+s} - \mu_n\right)\right] = \mathbb{E}^{X_s\sim\rho}\left[\left(X^n_{t_1+s} - \mu_n\right)\left(X^n_{t_2+s} - \mu_n\right)\right] =$$
$$= \mathbb{E}^{X_0\sim\rho}\left[\left(X^n_{t_1} - \mu_n\right)\left(X^n_{t_2} - \mu_n\right)\right],$$

We see that covariance function only depends on the difference. The assumption of $X_0 \sim \rho$ let us apply the proposition on the stochastic process X_t. Moreover, we can as well consider a version in power of n, since process X^n_t is still second-order stationary. Finally, we have following bound:

$$\mathbb{E}\left|\hat{\mu}_n - \mu_n\right|^2 \leqslant \frac{2}{T}\int_0^{+\infty} C_{X^n_t}(u)du, \tag{13}$$

[1] In practice, the initial source point of a process may not be close to the stationary distribution, i.e., $\mathbb{P}_\mu(X_0 = x) \approx 0$. However, if the process is ergodic, we expect it to quickly reach the stationary phase. Where, after a certain period of time Δt we can say that $X_{\delta t}$ is uncorrelated with X_0 and $X_{\Delta t} \sim \mu$. In the case of additive noise, the correlation function decreases exponentially, makings Δt to be small.

$$\hat{\mu}_n = \frac{1}{T}\int_0^T X_s^n ds$$

In order to learn unknown parameters, we have to solve least square problem: $\boldsymbol{M}\theta = q_T$ (2). Therefore we have to analyse the sensitivity of our system to small perturbation. We note $\widehat{\boldsymbol{M}} = \boldsymbol{M} + \Delta\boldsymbol{M}$ perturbed system and we denote $\widehat{\theta}$ solution to perturbed problem. When, $\mathbf{M}$ is full-rank, parameters θ can be found by pseudo-inverse: $\theta = \boldsymbol{M}^+ q_T$[1]. Actually we are solving :

$$\hat{\boldsymbol{M}}\hat{\theta} = (\boldsymbol{M} + \Delta\boldsymbol{M})(\theta + \Delta\theta) = q_T,$$

where, q_T we assume to know exactly[2]. By opening brackets and rearranging terms we obtain:

$$\boldsymbol{M}\theta - q_T + \Delta\boldsymbol{M}\theta + \boldsymbol{M}\Delta\theta + \Delta\boldsymbol{M}\Delta\theta = 0,$$

where $\Delta\boldsymbol{M}\Delta\theta$ is negligible compared to the other terms and $\boldsymbol{M}\theta - q_T = 0$:

$$\boldsymbol{M}\Delta\theta = -\Delta\boldsymbol{M}\theta.$$

Finally, solving for $\Delta\theta$ we get:

$$\Delta\theta = -\boldsymbol{M}^+\Delta\boldsymbol{M}\theta.$$

Now we can establish the boundary on error $\Delta\theta$ using euclidean norm:

$$\|\Delta\theta\|_2 \leq \left\|\boldsymbol{M}^+\right\|_2 \|\Delta\boldsymbol{M}\|_2\|\theta\|_2 = \left\|\boldsymbol{M}^+\right\|_2 \|\boldsymbol{M}\|_2 \frac{\|\Delta\boldsymbol{M}\|_2}{\|\boldsymbol{M}\|_2}\|\theta\|_2 = \kappa(\boldsymbol{M})\frac{\|\Delta\boldsymbol{M}\|_2}{\|\boldsymbol{M}\|_2}\|\theta\|_2,$$

where $\kappa(\boldsymbol{M})$ is the conditional number[3]. Since, our process is stochastic we consider L^2-norm of $\|\Delta\theta\|_2$

$$\|(\sum_i(\hat{\theta}_i - \theta_i)^2)^{\frac{1}{2}}\|_{L^2} = \mathbb{E}\left[\|\Delta\theta\|_2^2\right]^{\frac{1}{2}} \leq \kappa(\boldsymbol{M})\frac{\mathbb{E}[\|\Delta\boldsymbol{M}\|_2^2]^{\frac{1}{2}}}{\|\boldsymbol{M}\|_2}\|\theta\|_2. \tag{14}$$

Now, we will bound the error term $\|\Delta\boldsymbol{M}\|_2^2$ by considering the result we got in equation (13).
In general, for any matrix $\boldsymbol{A} \in \mathbb{R}^{L\times m}$, with $L >= m$ one can establish following inequality

$$\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_{\mathrm{F}}^2 = \sum_i^L\sum_j^m |A_{ij}|^2 \leq 2L(max_{i,j}(A_{i,j}^2)),$$

Which in our case, would be

$$\mathbb{E}\|\Delta\boldsymbol{M}\|_2^2 \leq 2L\max_{n_i}\mathbb{E}[(\hat{\mu}_{n_i} - \mu_{n_i})^2] \approx 2L\mathbb{E}[(\hat{\mu}_{n_{max}} - \mu_{n_{max}})^2]$$

Where n_{max} maximal moment[4] we have deal with. By using result (13):

$$\mathbb{E}[\|\Delta\boldsymbol{M}\|_2^2]^{\frac{1}{2}} \leq \frac{(\sqrt{2L})2}{\sqrt{T}}\left(\int_0^{+\infty} C_{X_t^{n_{max}}}(u)du\right)^{\frac{1}{2}}$$

[1] $\boldsymbol{M}^+ = (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T$

[2] Such assumption is convenient, since the perturbation we get in the core matrix is way greater compare to one due to the estimation of quadratic variation.

[3] $\kappa(\boldsymbol{M}) = \left|\left|\boldsymbol{M}^+\right|\right|_2\|\boldsymbol{M}\|_2$

[4] We assume to get maximal error/perturbation in the term with the highest moment.

Finally we obtain:

$$\mathbb{E}\left[\|\Delta\theta\|_2^2\right]^{\frac{1}{2}} \leq \frac{\kappa(\boldsymbol{M})}{\|\boldsymbol{M}\|_2}\frac{(\sqrt{8L})}{\sqrt{T}}\left(\int_0^{+\infty} C_{X_t^{nmax}}(u)du\right)^{\frac{1}{2}}\|\theta\|_2 \tag{15}$$

Therefore, by analysing equation (15) we distinguish three main factors that represent error bound :

- Model we use - $\int_0^{+\infty} C_{X_t^{nmax}}(u)du$, θ
- Final time - $\frac{1}{\sqrt{T}}$
- System we have to solve - $\left(\frac{\kappa(\boldsymbol{M})(\sqrt{2L})}{\|\boldsymbol{M}\|_2}\right)$.

Figure (4) represent the numerical support, that demonstrate dependency of error $\|\Delta\theta\|_2$ over time. From the log-log plot, we see that actual error tendency is pretty close to the expected rate.

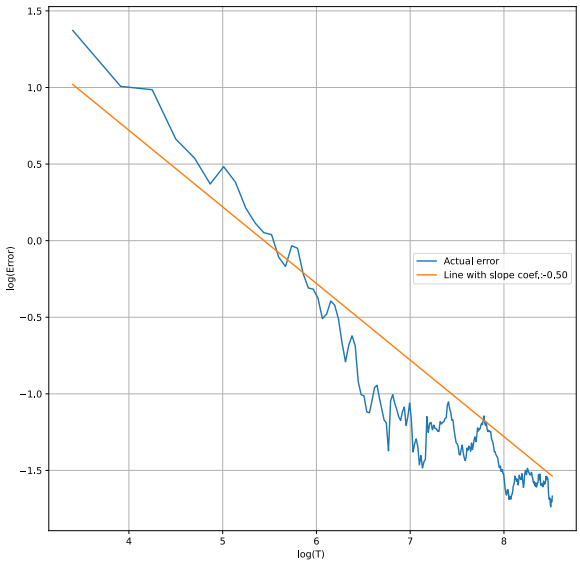


Figure 4: Log-log plot of Error with respect to different final time of O-U process.

It is important to mention, that result we just have derived uses

$$\hat{\mu}_n = \frac{1}{T}\int_0^T X_s^n ds,$$

as estimator. Whereas, in practice, we have to consider the discrete observations $\{X_j\}_{j=0}^J$.

$$\hat{\mu}_n = \frac{1}{J}\sum_{J=0}^{J} X_J^n$$

That could further increase the potential error. In particular, if the trajectory (observed data) is produced by the first-order scheme, like Euler–Maruyama method

that we use to simulate most of the processes in this work. One can establish similar result[1] to one in Proposition (4.1)

$$\mathbb{E}\left[(\hat{\mu}_n - \mu_n)^2\right] \leq C\left(h^2 + \frac{1}{T}\right) \tag{16}$$

With time-step h and $C > 0$ positive constant that do not depend on time and time-step h. But time we have one more term. This would make appear the h - step term in our final bound (15).

$\widehat{\mathbf{Q}}_T$ estimator

Now we are going to consider the error bound for our estimator of quadratic variation. It should be mention that if the actual time continuous path is available, one can exactly estimate the quadratic variation. Therefore, we directly pass to application framework, where only the discrete observations are available.

$$\frac{\widehat{Q}_T}{2T} = \frac{1}{2hJ}\sum_{j=0}^{J}\left(X_{(j+1)} - X_j\right)^2$$

$$|\mathbb{E}\left[\frac{\widehat{Q}_T}{2T}\right] - \mathbb{E}\left[\frac{Q_T}{2T}\right]| \leqslant C\left(h + h^{1/2}\right)$$

Where $C > 0$ positive constant. As we mentioned before, the error is not greatly affected by the final time, but rather it depends on the time step h (the equidistant space between observed data). This means that the accuracy of the estimator is primarily determined by how finely we divide the time interval. In order to derive the error bound, we fist will make assumption that initial value follow invariant distribution , i.e. , $X_0 \sim \mu$. As well, we assume that observed data is obtained from actual continuous path, nevertheless the same bound can be obtained for data obtained by EM scheme.
We consider

$$X_{j+1} - X_j = \int_{jh}^{(j+1)h} f(X_t, \alpha)dt + \int_{jh}^{(j+1)} \sqrt{2g(X_t, \sigma)}dW_t = I_j + M_j,$$

where, $\quad I_j := \int_{jh}^{(j+1)h} f(X_t, \alpha)dt, \quad M_j := \int_{jh}^{(j+1)} \sqrt{2g(X_t, \sigma)}dW_t.$

We substitute

$$\frac{\widehat{Q}_T}{2T} = \frac{1}{2hJ}\sum_{j=0}^{J}\left(M_j\right)^2 + \frac{2}{2hJ}\sum_{j=0}^{J} I_j M_j + \frac{1}{2hJ}\sum_{j=0}^{J} I_j^2,$$

consequently,

$$\mathbb{E}[\frac{\widehat{Q}_T}{2T}] = \frac{1}{2hJ}\sum_{j=0}^{J}\mathbb{E}[M_j^2] + \frac{2}{2hJ}\sum_{j=0}^{J}\mathbb{E}[I_j M_j] + \frac{1}{2hJ}\sum_{j=0}^{J}\mathbb{E}[I_j^2],$$

We look term by term:

[1] (See Theorem 5.2. Mattingly, Stuart, and Tretyakov 2010, p. 11)

- By Ito isometry, we get

$$\frac{1}{2hJ}\sum_{j=0}^{J}\mathbb{E}[M_j^2] = \frac{1}{2hJ}\sum_{j=0}^{J}\mathbb{E}[\int_{jh}^{(j+1)} 2g(X_t,\sigma)dt] = \mathbb{E}[\frac{Q_T}{2T}]$$

- We use the fact that process is ergodic, making drift term $f(X_t,\alpha)$ bound and using the Cauchy-Schwarz inequality, we get

$$\mathbb{E}I_j^2 \leqslant h\int_{jh}^{(j+1)h}\mathbb{E}\left(f\left(X_s;\theta\right)\right)^2 ds \leqslant C_1 h^2.$$

  Where $C>0$ positive constant. We have

$$\frac{1}{2hJ}\sum_{j=0}^{J}\mathbb{E}[I_j^2] \leqslant \frac{C_1}{2}h.$$

- For last term we as well apply Cauchy-Schwarz inequality to get

$$\mathbb{E}[I_j M_j] \leqslant \mathbb{E}[I_j^2]^{1/2}\mathbb{E}[M_j^2]^{1/2} = C_2 h^{\frac{3}{2}}.$$

$$\frac{2}{2hJ}\sum_{j=0}^{J}\mathbb{E}[I_j M_j] \leqslant C_2 h^{\frac{1}{2}}$$

Finally, we obtain our bound

$$|\mathbb{E}[\frac{\widehat{Q}_T}{2T}] - \mathbb{E}[\frac{Q_T}{2T}]| \leqslant C\left(h + h^{1/2}\right) \tag{17}$$

Equation (17) itself is not interesting since in application we would like to get bound for Q_T and not for it's expectation. Actually, if we consider additive noise , i.e. , $g(X_t,\sigma)=\sigma$, in this framework our bound is

$$|\mathbb{E}[\frac{\widehat{Q}_T}{2T}] - \sigma]| \leqslant C\left(h + h^{1/2}\right).$$

As well, if we consider time T sufficiently big enough, due to ergodic theorem one can say that quadratic variation term gets deterministic

$$\frac{Q_T}{2T} \approx \mathbb{E}^{\mu}[g(X)] =: q.$$

Regarding results obtained in previous section

$$|\mathbb{E}[\frac{\widehat{Q}_T}{2T}] - q]| \leqslant |\mathbb{E}[\frac{\widehat{Q}_T}{2T}] - \mathbb{E}[\frac{Q_T}{2T}]| + |\mathbb{E}[\frac{Q_T}{2T}] - q| \leqslant C\left(h + h^{1/2}\right) + C'\frac{1}{\sqrt{T}}$$

Therefore, in a case when noise term is constant or time is sufficiently big, our estimator in average have good performance that mostly depends on the time step h. In practice, this estimator always shows high accuracy, even when the diffusion term have complex expression. However, in terms of our methodology we are interested in the precise estimation of the constraint equation (3), that with the discrete observation takes the form of

$$\sum_{e}^{E}\sigma_e\left(\frac{1}{J}\sum_{j=0}^{J}X_j^e\right)\underbrace{=}_{?}\frac{\widehat{Q_T}}{2T} \tag{18}$$

Constraint equation

When selecting a constraint equation,it is important to consider the impact that it will have on the overall system. Ideally, the constraint equation (18) should not introduce any additional perturbations or should only introduce negligible perturbations to the core system. By looking on the system 2 and on the constraint 3, we notice, that proposed constraint equation does not require the estimation of moments. Instead, it directly relies on the estimation of time averages through equidistant discrete observations. This means that the error produced by the constraint equation should be relatively lower compared to one produced in the core system. Moreover, in the constraint equation error is dependent on the time distance between observations rather than on the final time, as is the case with moment estimation in the core matrix. Therefore, if time step h is sufficiently small the additional perturbations are negligible. In practice, it is often useful to use data with small time steps for constraint estimation, if mixed time step data is available, while using all data for the core system. This can help to minimize errors and improve the accuracy of the results. Now, we are going to consider the model (19) with a complex diffusion term and will investigate the relative error produced over time for different time steps in the constraint equation.

$$dX_t = \left(\alpha_3 X_t^3 + \alpha_1 X_t\right) dt + \sqrt{2(\sigma_0 + \sigma_2 X_t^2)^2} dW_t \tag{19}$$

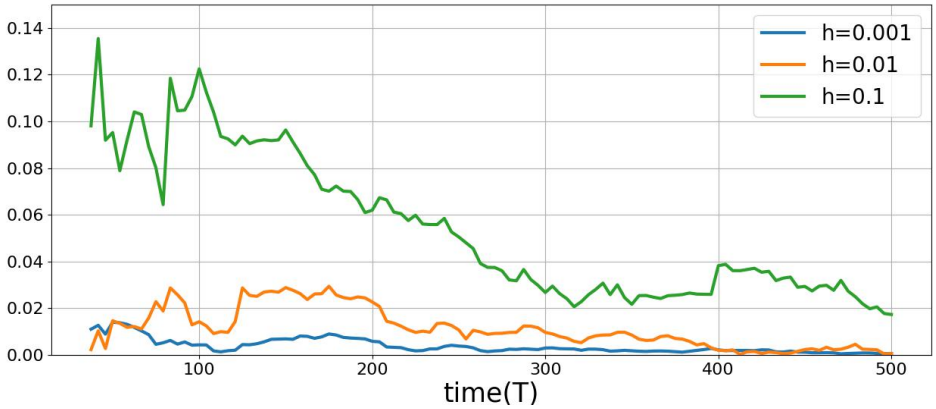


Figure 5: Plot represents the relative error: $\left|\sum_e^E \sigma_e \left(\frac{1}{J}\sum_{j=0}^J X_j^e\right) - \widehat{\frac{Q_T}{2T}}\right| / \frac{Q_T}{2T}$

From the Figure (5) we see that decreasing time step leads to the better accuracy. We can say that for low time step, the choice of the constraint is more then convenient. In addition, we would like to mention, that the main source error is estimation of time average and not estimator of a quadratic variation. Due to this reason, if we consider the same model (19) but with additive noise

$$dX_t = \left(\alpha_3 X_t^3 + \alpha_1 X_t\right) dt + \sqrt{2\sigma_0} dW_t,$$

we expect to get better performance compare to multiplicative settings. Figure (6) demonstrate the relative error for case of an additive noise. We notice that performance is much better when it comes to the relatively big time step. Actually,

this time the produced error has only one source, which is due to the quadratic variation estimation.

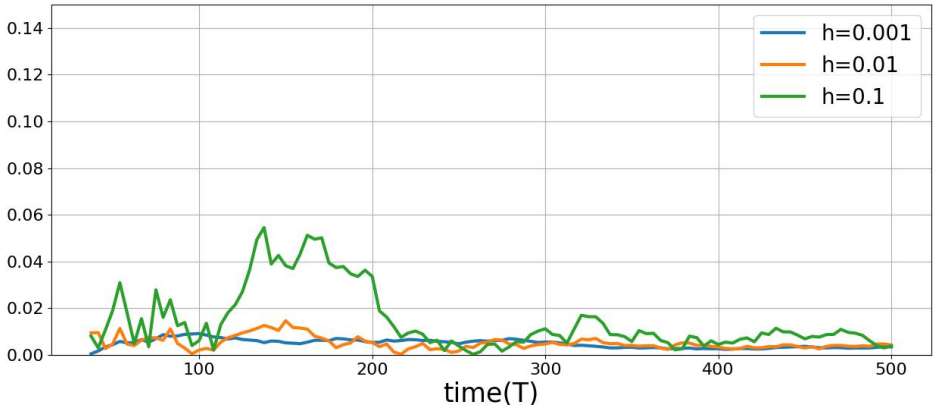


Figure 6: Plot represents the relative error $|\sigma - \widehat{\frac{Q_T}{2T}}| \ / \ \sigma$ over different final time.

Remark 4.1
It should be as well noted, that in a case when the initial source point X_0 is far from the stationary distribution, the domination in terms of accuracy in additive case compare to multiplicative one is much more visible. Such performance in the additive noise framework of our constraint is one of the reason, why in section of 'Multiplicative noise' (6) we are going to consider one technique that would let us to remap the initial process with multiplicative noise to one with additive noise and increase the accuracy of the MoM approach.

5 Non-parametric

In this section, we will remove the assumption of knowing the exact form of the model and allowing our technique (MoM) to discover the shape of the drift function on its own. The goal of this section is not to obtain the exact unknown parameters of the model, but rather to approximate the objective drift term through the use of a polynomial expansion. This involves fitting the drift function with a polynomial of a specified degree, and using the resulting model to make predictions about the behavior of the system. While this approach may not yield the most accurate results in all cases, it can provide a useful starting point for further analysis and modeling efforts.
So we have our true model (20) that drive the dynamic of underlying process and produce observed data $\{X_j\}_{j=0}^J$

$$dX_t = b(X_t)dt + \sqrt{2\sigma}dW_t, \qquad X_0 = x. \tag{20}$$

The idea consist in the consideration of the Fokker-Planck equation with drift term being polynomial expansion $f_K(x,\alpha) = \sum_k^K \alpha_k x^k$ of chosen degree. Afterwards, we proceed as usual to obtain the system of equation (2). Solving the system give us parameters $\hat{\theta} = (\hat{\alpha}, \hat{\sigma})$ which, should be able explain behavior of the drift function , i.e., $\mathbb{E}^\mu[|b(X) - f(X,\hat{\alpha})|] < \epsilon$, for some reasonably small positive ϵ. As a result, we use new norm to measure the error. We replace the euclidean norm of the coefficients with weighted L^2 norm

$$\text{Error } (\hat{f}) = \mathbb{E}^\mu\left[(b-\hat{f})^2\right] = \int_{\mathbb{R}} \left(b(x) - f(x,\hat{\alpha})\right)^2 \rho(x)dx,$$

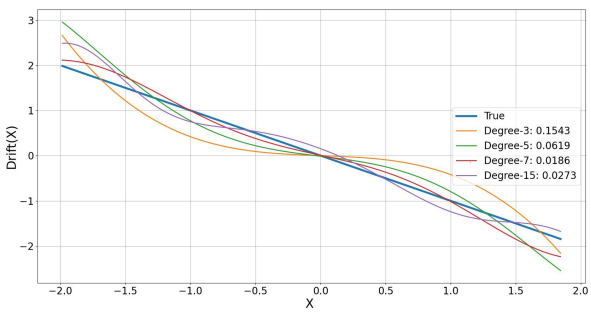
where ρ is the dencity of the invariant measure μ.

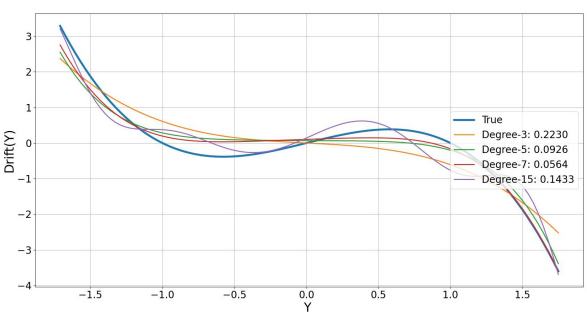
Examples

We consider three examples, where one can examine the performance of such approach. In particular, we try to estimate the drift term for following processes

- OU (Ornstein-Uhlenbeck) process X_t (6).
- Bistable (Brownian motion in a bistable potentia) process Y_t (9).
- New process, which contains non-polynomial drift term Z_t

$$dZ_t = \left(\sin(2Z_t) - Z_t\right)dt + \sqrt{2\sigma}dW_t, \qquad Z_0 = 0.$$





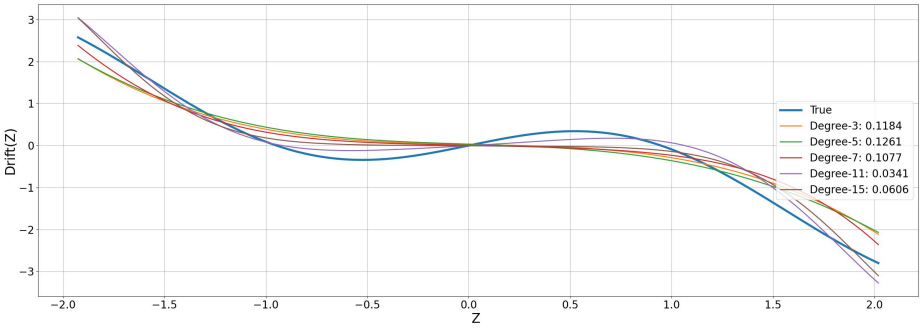
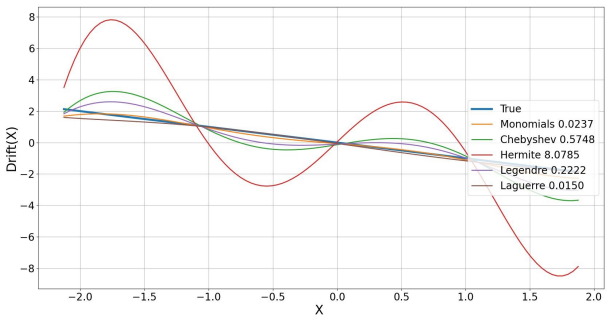


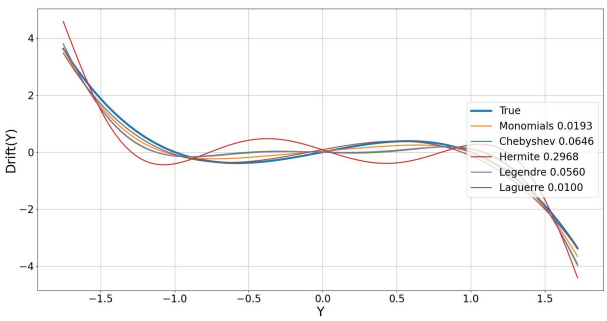
Figure 7: The three plots shows the actual drift function used in the model and polynomial expansions of different degree K, i.e., $f_K(x, \hat{\lambda}_K)$, where the parameters $\hat{\lambda}_K$ we obtain through MoM technique.

Figure (7) consist of three plots. Where on the first two, one can see fitting of the drift term with different polynomials (of degree $n \in [3, 5, 7, 15]$) for the processes X_t (left) and Y_t (right). The last plot corresponds to approximation of the Z_t drift function. In all three cases, we notice the general trend, that increasing of the degree until some point, would lead to the better fit. Nevertheless, high degree of $n = 15$ shows good performance in all the three cases. Therefore, one can always try to fit underlying drift term with the high degree polynomial expansion, that would already give an idea of the actual shape of the drift term. However, we should keep in mind that higher the degree we use, larger the error we get in the moment approximation. In a case, when only small amount of observations is available, usage of high degree, will defiantly lead to inappropriate results.

Different Polynomial basis

As mentioned in the beginning of this project, we can use different polynomial bases and in general sense, different orthogonal bases of an appropriate Hilbert space. The derivation of the approach is the same as in a case of monomials with only one difference, which is this time we multiply Fokker-Planck equation with a polynomial P_n of the corresponding bases. The following small section presents the results of applying different polynomial bases to the three processes previously described. In all three cases we will try to fit drift term with the polynomial of the seventh degree.





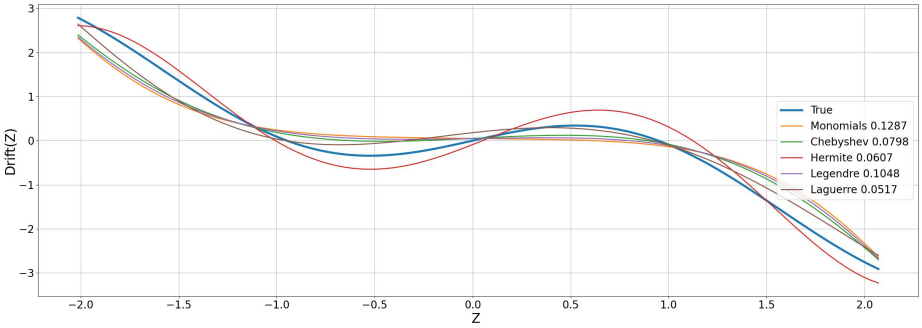


Figure 8: The three plots shows the actual drift function used in the model and fitted seventh degree polynomial that correspond to different bases.

A review of the three plots, on the Figure (8), suggests that monomials and Laguerre polynomials are the most accurate bases for the proposed processes. These two bases consistently outperform other options in terms of accuracy and stability when tested across various models (which have polynomial expression for the drift term, as well as drift term with a basic trigonometric functions). In particular, they contribute to a lower conditional number for the system. It is worth noting, however, that there is no one-size-fits-all basis for all types of processes.

6 Multiplicative noise

The Method of Moments (MoM) has several advantages compared to other methods such as Maximum likelihood estimation (2), Least squares optimization, Bayesian inference and Implied volatility. One major advantage of MoM is that it has direct access to the estimation of parameters in diffusion terms of the stochastic differential equation based on the observed data $\{X_j\}_{j=0}^{J}$, without relying on any assumptions about the prior distribution, transition density, or knowledge of the exact solution of the SDE. This makes MoM a flexible and robust method that can be applied to a wide range of problems in finance and other fields. In addition to its flexibility and robustness, the Method of Moments (MoM) has the advantage of having low computational cost. This makes it an efficient and practical method for estimating parameters in situations where computational resources are limited. The low computational cost of MoM is due to the fact that it does not require the optimization of complex objective functions, like approaches based on martingale estimating functions (for exemple: approach proposed by Kessler and Sørensen 1999), and only requires estimation of moments via time average. Finally, MoM can often provide more accurate estimates of parameters than other methods, making it a valuable tool for statistical analysis and decision making.
Let us consider Cox–Ingersoll–Ross process driven by (21) SDE

$$dX_t = (\alpha_1 X_t + \alpha_0)\, dt + \sqrt{2\sigma X_t} dW_t. \tag{21}$$

This process is quite popular, since under some condition[1] it stays positive a.s and could be used to model interest rate. We are going to estimate performance of MoM on this process with relatively simple diffusion term. Results presented on Figure (9) shows the high accuracy of our approach, that is obtained with relatively low final time $T \approx 200$.

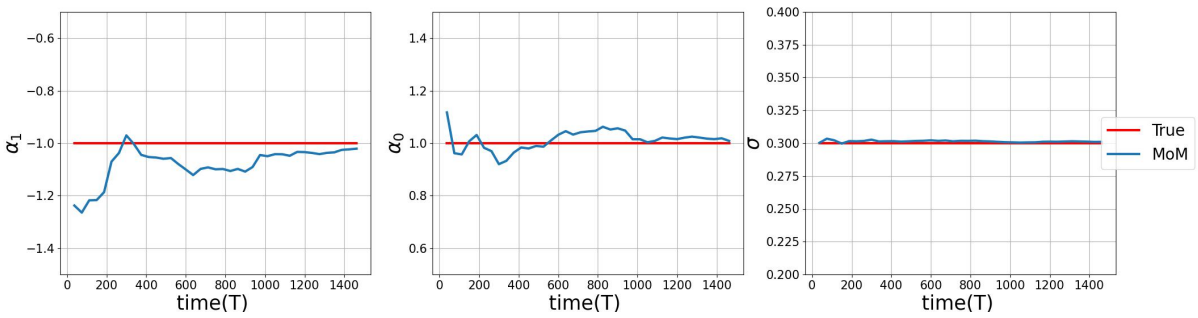


Figure 9: Plot of estimated unknown parameters using MoM over different final time. Where trajectory was produced using E-M scheme with h=0.001.

Interesting to note, that with MoM we do not need even to know the exact SDE of the process and diffusion function could be estimated approximately as we did it with drift function in previous section. However, the drawback when dealing with

[1]Cox–Ingersoll–Ross is highly popular diffusion processes in financial literature, which is used in order to model interest rate. In particular, it can be shown that by setting $2\alpha_1 > (2\sigma)^4$, $\alpha_0 > 0$ one can ensure the process to be a.s. positive (see lecuture notes Mishura 2021, pages: 33-40).

multiplicative noise is that it increases complexity of the model, making moment estimation harder. Moreover, in order to achieve satisfactory accuracy, it requires a longer final time T and a larger number of observations. Let us consider the process with more complex expression in diffusion term.

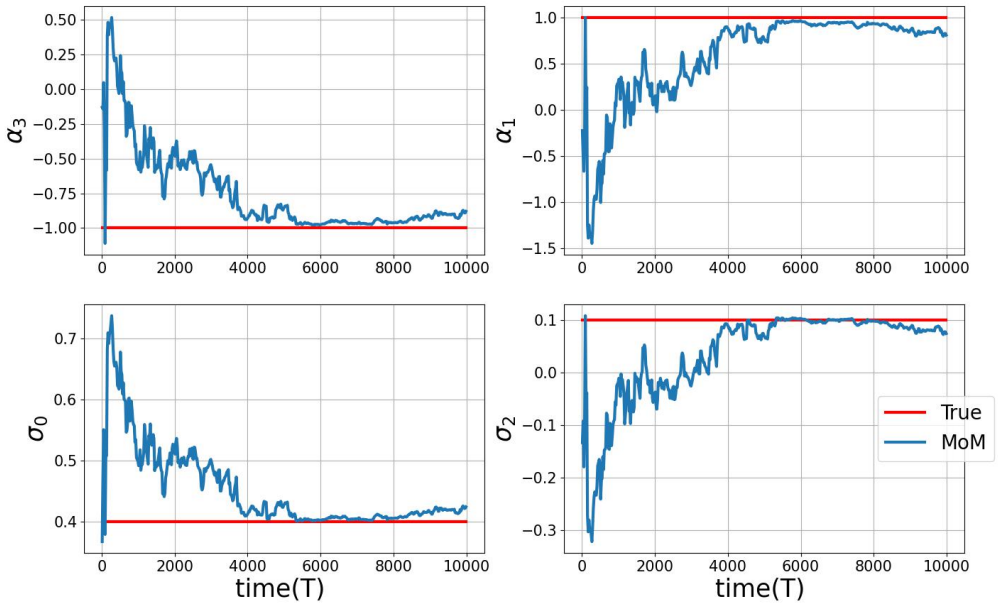


Figure 10: Plot of estimated unknown parameters using MoM over different final time. Where trajectory was produced using E-M sheme with h=0.001.

Figure (10) shows performance of MoM in parameter estimation of the process

$$dX_t = \left(\alpha_3 X_t^3 + \alpha_1 X_t\right) dt + \sqrt{2(\sigma_0 + \sigma_2 X_t^2)^2} dW_t \tag{22}$$

Which is actually bistable process to which we added multiplicative noise. By comparing figures (3) and (10) we see that with addative noise we require around $T \approx 1000$ final time to reach satisfactory level of error, whereas in multiplicative case we require $T > 4000$.

Lamperti transformation

After experimenting with different kind of processes, we could notice that working with additive noise leads to better performance compared to multiplicative noise. We therefore propose the following technique, that maps the initial process with multiplicative noise to the one with additive noise.In particular, we search for the map $h : \mathbb{R} \to \mathbb{R}$ such that $(h(X_t))_{t\in[0,T}$ has constant diffusion. Actually, it would create a new system without heavy terms for the diffusion and straight forward constraint equation. Therefore, it would already remove a part of the perturbation, that occur due to the additional constraint equation, described in the end of the section four (4). It is noteworthy to mention that this technique does not always

result in improved performance, as the transformed drift function can potentially lead to a more complex system. However, it is possible to evaluate whether the resulting system has improved or not, in terms of the error approximation. In this section we are going consider two cases (21, 22) where such transformation would greatly improve accuracy of the obtained results. In order to find the desire function h, one can use Ito formula

$$h(X_T) = h(X_0) + \int_0^T h'(X_t) \underbrace{dX_t}_{f(X_t)dt + \sqrt{2g(X_t,\sigma)}dW_t} + \frac{1}{2}\int_0^T h''(X_t) \underbrace{d\langle X, X\rangle_t}_{2g(X_t,\sigma)dt},$$

$$h(X_T) = h(X_0) + \int_0^T \left(h'(X_t)f(X_t) + \frac{(2g(X_t,\sigma)}{2} h''(X_t) \right) dt + \int_0^T h'(X_t)\sqrt{2g(X_t,\sigma)}dW_t,$$

or, in SDE form

$$dh(X_t) = \left(h'(X_t)f(X_t) + \frac{(2g(X_t))}{2} h''(X_t) \right) dt + h'(X_t)\sqrt{2g(X_t,\sigma)}dW_t.$$

We can rewrite in more compact form by considering the Markov generator operator of the initial process $L = f(x)\frac{\partial}{\partial x} + g(x)\frac{\partial^2}{\partial x^2}$

$$dh(X_t) = L(h(X_t))dt + h'(X_t)\sqrt{2g(X_t,\sigma)}dW_t. \tag{23}$$

In order to get constant diffusion in equation (23), we have to set

$$h(x,\sigma) = \int_{x_0}^{x} \frac{1}{\sqrt{2g(x,\sigma)}} dx.$$

Such map h is known as Lampetri transformation. Unfortunately, we are not able to directly apply Lamperti transformation, since expression of h contains unknown parameters $\sigma = (\sigma_0, \ldots, \sigma_E)$. However, we can apply similar a technique on the process where in the diffusion term we have only one parameter $\sigma \in \mathbb{R}$ to estimate, in other words, where we can rewrite objective function as: $g(x,\sigma) = \sigma g(x)$. In this case, one could apply following transformation

$$h(x) = \int_{x_0}^{x} \frac{1}{\sqrt{g(x)}} dx,$$

and get the desired SDE with additive noise

$$dh(X_t) = L(h(X_t))dt + \sqrt{2\sigma}dW_t. \tag{24}$$

Finally, one can consider MoM applied on the $\{h(X_j)\}_{j=0}^{J}$, together with equation (24) To demonstrate the effect of the modified Lamperti transformation, we can consider a variant of the process (22) described by the following equation

$$dX_t = \alpha \left(X_t^3 + X_t \right) dt + \sqrt{2\sigma(1+X_t^2)^2}dW_t, \tag{25}$$

with $\alpha < 0$ and $\sigma > 0$. And compare relative error for the MoM used on initial data and on the transformed data points. First of all, we are going to derive the map h and new SDE for $h(X_t)$ (25).

$$h(x) = \int_{x_0}^{x} \frac{1}{\sqrt{g(x)}} dx = \int_{x_0}^{x} \frac{1}{(1+X_t^2)} dx = \arctan(x).$$

$$dh(X_t) = L(h(X_t))dt + \sqrt{2\sigma}dW_t =$$
$$= \alpha\left(\left(X_t^3 + X_t\right)\frac{\partial}{\partial x}(\arctan(x)) + \sigma(1+X_t^2)^2\frac{\partial^2}{\partial x^2}(\arctan(x))\right)dt + \sqrt{2\sigma}dW_t =$$
$$= (\alpha - 2\sigma)X_t dt + \sqrt{2\sigma}dW_t.$$

We substitute $h(X_t) = Y_t$, $X_t = \tan(Y_t)$ and $\mu = (\alpha - 2\sigma)$

$$dY_t = \mu\tan(Y_t)dt + \sqrt{2\sigma}dW_t. \tag{26}$$

We apply MoM on the process Y_t (26) with mapped data $\{h(X_j)\}_{j=0}^J$, to estimate $\theta = (\mu, \sigma)$, from which we can extract alpha $\alpha = \mu - 2\sigma$. We consider new system to solve

$$\begin{pmatrix} (\mathbb{E}^\mu[X^3] + \mathbb{E}^\mu[X]) & 0 \\ \vdots & \\ (\mathbb{E}^\mu[X^{n+2}] + \mathbb{E}^\mu[X^n]) & (n-1)(\mathbb{E}^\mu[X^{n-1} + 2\mathbb{E}^\mu[X^{n+1}] + \mathbb{E}^\mu[X^{n+3}]) \\ 0 & (\mathbb{E}^\mu[X^0 + 2\mathbb{E}^\mu[X^2] + \mathbb{E}^\mu[X^4]) \end{pmatrix}\begin{pmatrix} \alpha \\ \sigma \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\widehat{Q}_T}{2T} \end{pmatrix}$$
$$\downarrow h$$
$$\begin{pmatrix} \mathbb{E}^\mu[XY] & 0 \\ \vdots & \\ \mathbb{E}^\mu[XY^{n-1}] & (n-1)\mathbb{E}^\mu[Y^{n-2}] \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\widehat{Q}_T}{2T} \end{pmatrix}.$$

The transformed system is preferred due to its reduced number of required moment estimations and lower degree. As a result, the mapped process not only eliminates errors in constraint estimation, but also yields improved overall performance by simplifying the system.

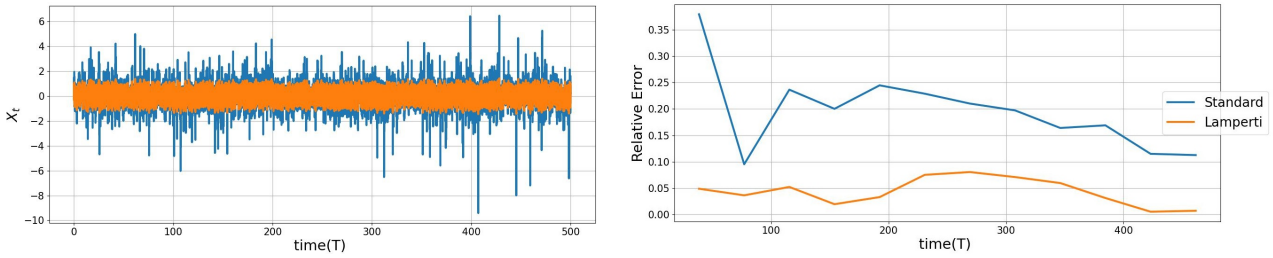


Figure 11: (Right)Plot of the estimated unknown parameters using MoM over different final time. Where trajectory was produced using E-M scheme with h=0.001. (Left) The actual produces path is in blue and the transformed version in orange.

On the Figure (11), it is evident that the orange curve, which represents the accuracy of the MoM method applied on the transformed data, dominates over the other curve. This can be seen by the orange curve consistently having higher accuracy values throughout the plot and reaching almost zero error toward $T = 400$ final time. This indicates that the transformation of the data prior to applying the MoM method greatly improves the accuracy of the results.

Cox–Ingersoll–Ross exemple

In our next example, we will be using the CIR (Cox–Ingersoll–Ross) process 21 to further evaluate the effectiveness of the Lamperti transformation. Although the MoM method demonstrated impressive accuracy with the initial data points (9), we hope to achieve greater accuracy within modified Lamperti transformation.
Again, we start by searching for map h

$$h(x) = \int_{x_0}^{x} \frac{1}{\sqrt{g(x)}} dx = \int_{x_0}^{x} \frac{1}{(\sqrt{x})} dx = 2\sqrt{x}.$$

$$dh(X_t) = L(h(X_t))dt + \sqrt{2\sigma} dW_t =$$

$$= \left((\alpha_1 X_t + \alpha_0) \frac{\partial}{\partial x}(2\sqrt{X}) + \sigma X_t \frac{\partial^2}{\partial x^2}(\sqrt{X}) \right) dt + \sqrt{2\sigma} dW_t =$$

$$= \left(\frac{\alpha_0 + \alpha_1 X_t}{\sqrt{X_t}} - \frac{\sigma X}{2X_t^{3/2}} \right) dt + \sqrt{2\sigma} dW_t.$$

We substitute $2\sqrt{X_t} = Y_t$ and $(\mu_1, \mu_2) = (2\alpha_0 - \sigma, \frac{\alpha_1}{2})$

$$dY_t = \left(\frac{2\alpha_0}{Y_t} + \frac{\alpha_1}{2} Y_t - \frac{\sigma}{Y_t} \right) dt + \sqrt{2\sigma} dW_t.$$

$$dY_t = \left(\frac{\mu_1}{Y_t} + \mu_2 Y_t \right) dt + \sqrt{2\sigma} dW_t. \tag{27}$$

We apply MoM on the process Y_t[1] with mapped data $\{2\sqrt{X_j}\}_{j=0}^{J}$, to estimate $\theta = (\mu_1, \mu_2, \sigma)$, from which we can extract alpha $(\alpha_0, \alpha_1) = (\frac{\mu_1 + \sigma}{2}, 2\mu_2)$. We consider new system to solve

$$\begin{pmatrix} \mathbb{E}^{\mu}[X] & 1 & 0 \\ \vdots & & \\ \mathbb{E}^{\mu}[X^n] & \mathbb{E}^{\mu}[X^{n-1}] & (n-1)\mathbb{E}^{\mu}[X^n] \\ 0 & 0 & \mathbb{E}^{\mu}[X] \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_0 \\ \sigma \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\widehat{Q}_T}{2T} \end{pmatrix}$$

$$\downarrow h$$

$$\begin{pmatrix} 1 & \mathbb{E}^{\mu}[Y^2] & 1 \\ \vdots & & \\ \mathbb{E}^{\mu}[Y^{n-2}] & \mathbb{E}^{\mu}[Y^n] & (n-1)\mathbb{E}^{\mu}[Y^{n-2}] \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\widehat{Q}_T}{2T} \end{pmatrix}.$$

For the new system, we skipped the first line because it contained a term of the form $\mathbb{E}^{\mu}[\frac{1}{Y}]$ that could potentially lead to a high error. In general, we expect that the new system will have a lower error since Y is equal to $2\sqrt{X}$ and the new system uses lower degrees. This should result in a more accurate and reliable system overall.

[1] Actually, the obtained process Y_t so-called, radial Ornstein-Uhlenbeck process.

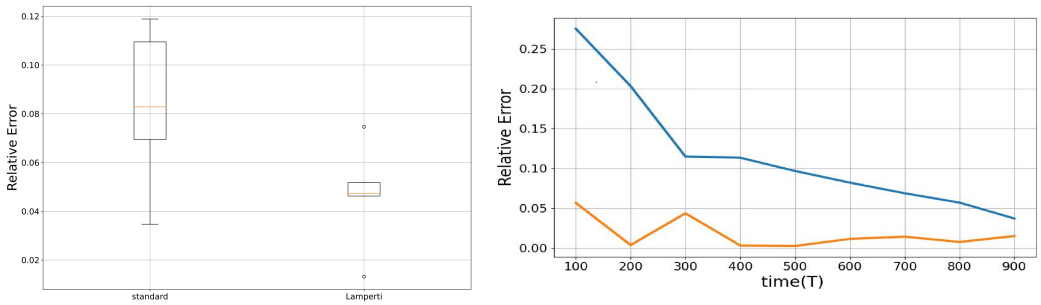


Figure 12: (Right)Plot of estimated unknown parameters using MoM over different final time. Where trajectory was produced using E-M scheme with $h = 0.001$ and observations we use correspond to the $h' = 0.1$ time step. (Left) Statistics of the relative error, corresponding to the final time $T = 300$.

On the Figure (12), the right plot, we present relative error due to MoM estimation, where this time we use bigger time step then one the figure (9). This is done in purpose to increase the bias due to the additional constraint equation estimation. In such settings the impact of the Lamperti transformation is clearly visible.

7 Multidimensional extension

As we have already mentioned in remark (2), MoM can be generalized to multi-dimensional stochastic processes. In this section we will show how to adapt our methodology to general multi-dimensional framework, and we illustrate derivation of one concrete example.

Settings

We consider a stochastic process X_t on $\mathbb{R}^d$, with initial value x_0, where we have m-dimensional Wiener process $\mathbf{W}_t \in \mathbb{R}^m$, drift vector function $f(x) = (f_1(x), f_2(x), \dots, f_d(x)) \in \mathbb{R}^d$ and diffusion matrix $(g(x))_{i,j} = g_{i,j}(x) \in \mathbb{R}^{d\times m}$.

$$dX_t = f(X_t)dt + \sqrt{2}g(X_t)d\mathbf{W}_t, \qquad X_0 = x_0,$$

or, component-wise,

$$dX^i(t) = f_i(X_t)dt + \sum_{j=1}^{m} \sqrt{2}g_{ij}(X_t)dW_j(t), \quad i = 1, \dots d, \qquad X_0 = x_0,$$

$$f_i(x;\alpha_i) = \sum_{k=0}^{K} \alpha_{i,k}P_k(x) \qquad \text{and} \qquad g_{i,j}(x;\sigma_{i,j}) = \sum_{e=0}^{E} \sigma_{i,j,e}P_e(x), \tag{28}$$

where $P_i \in \{P_n\}_{n=0}^{\infty}$ is a sequence of polynomials of the form

$$P_i(\mathbf{x}) = x_1^{n_1^{(i)}} x_2^{n_2^{(i)}} x_3^{n_3^{(i)}} \dots x_{d-1}^{n_{d-1}^{(i)}} x_d^{n_{d^{(i)}}}.$$

As in 1-dimensional case we assume the ergodicity of the process and existence of unique invariant measure μ and invariant density function $\rho(x)$, that satisfy the stationary Fokker-Planck equation

$$\begin{aligned} 0 &= \nabla \cdot (-f(x)\rho(x) + \nabla \cdot (\Sigma(x)\rho(x)) \\ &= -\sum_{j=1}^{d} \frac{\partial}{\partial x_j} \left(f_j(x)p\right) + \sum_{i=1}^{d}\sum_{j=1}^{m} \frac{\partial^2}{\partial x_i \partial x_j} \left(\Sigma_{ij}(x)p\right), \end{aligned}$$

where, $\Sigma(x) = g(x)g(x)^T$.
We multiply both sides of F-P by polynomials $P_{n_1,\dots,n_d}(x) = x_1^{n_1}x_2^{n_2}\dots x_d^{n_d}$ and integrate over $\mathbb{R}^d$

$$\begin{aligned} 0 = -\sum_{j=1}^{d} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} P_{n_1,\dots,n_d}(x) \frac{\partial}{\partial x_j} \left(f_j(x)\rho(x)\right) dx_i dx_{\hat{i}} + \\ + \sum_{i=1}^{d}\sum_{j=1}^{m} \int_{\mathbb{R}^{d-2}} \int_{\mathbb{R}} \int_{\mathbb{R}} P_{n_1,\dots,n_d}(x) \frac{\partial^2}{\partial x_i \partial x_j} \left(\Sigma_{ij}(x)\rho(x)\right) dx_i dx_j dx_{\hat{i},\hat{j}}, \end{aligned}$$

applying the integration by part with respect to x_j for the first and second terms, followed by one more integration by parts for the second term with respect to x_i, we

get

$$0 = \sum_{j=1}^{d} \int_{\mathbb{R}^d} \frac{\partial}{\partial x_j} \left(P_{n_1,\dots,n_d}(x)\right) f_j(x)\rho(x)dx + \\ + \sum_{i=1}^{d}\sum_{j=1}^{m} \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_i \partial x_j} \left(P_{n_1,\dots,n_d}(x)\right) \Sigma_{ij}(x)\rho(x)dx.$$

In both terms one can recognise the exact definition of the expectation with respect to invariant measure μ. Therefore we can rewrite the equation above as follow

$$0 = \sum_{j=1}^{d} \mathbb{E}^{\mu}\left[\frac{\partial}{\partial x_j} \left(P_{n_1,\dots,n_d}(x)\right) f_j(X)\right] + \\ + \sum_{i=1}^{d}\sum_{j=1}^{m} \mathbb{E}^{\mu}\left[\frac{\partial}{\partial x_j} \left(P_{n_1,\dots,n_d}(x)\right) \Sigma_{i,j}(X)\right]. \tag{29}$$

In equation (29) by using different polynomials $P_{n_1,\dots,n_d}(x)$ and replacing drift and diffusion term with their polynomial definition (28) we would get a system of equation similar to the one we got in the 1-dimensional case (2). The only difference would be that the expectations are computed for mixed X_i terms: $\mathbb{E}^{\mu}\left[X_1^{n_1} X_2^{n_2} X_3^{n_3} \dots X_{d-1}^{n_{d-1}} X_d^{n_d}\right]$. For illustration purposes we will consider simpler class of stochastic processes. In particular we will continue our discussion by analysing the processes driven by linear stochastic differential equations with diagonal matrix in the diffusion, i.e., a multidimensional version of OU.

Linear SDE

Let $A, D \in R^{d\times d}$ be symmetric positive definite matrices, and let $W(t)$ be a Wiener process in $\mathbb{R}^d$. As we mentioned before, we could as well consider Brownian motion in more general $\mathbb{R}^m$ space, but for simplicity we take $d = m$. Let X_t,$t \in [0, T]$, be a solution to the linear multidimensional SDE

$$dX(t) = -AX(t)dt + \sqrt{2D}dW(t), \tag{30}$$

$$dX_i(t) = -\underbrace{\sum_{j=1}^{d} \alpha_{ij} X_j(t)\, dt}_{f_i(X_t)} + \sum_{j=1}^{d} \underbrace{\sqrt{2D_{ij}}}_{g_{ij}(X_t)} dW_j(t), \quad i = 1, \dots d$$

$$A = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,d} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,d} \\ \vdots & & & \\ \alpha_{d,1} & \alpha_{d,2} & \dots & \alpha_{d,d} \end{pmatrix} \qquad D = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \sigma_d \end{pmatrix},$$

In a case of a diagonal diffusion matrix we have $\Sigma = \sqrt{DD^T} = D$, and now we recall the stationary F-P equation

$$0 = \nabla \cdot (Ax\rho(x) + \nabla \cdot (\Sigma\rho(x))$$

$$0 = \sum_{i=1}^{d} \frac{\partial}{\partial x_i}((AX(t))_i \rho(x)) + \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}(\sigma_j \rho(x)).$$

This time we will consider the particular case where we multiply both sides by monomials $X_k^{n_k}$. As usual, we integrate over $\mathbb{R}^d$ and successfully apply integration by parts

$$0 = \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{\partial}{\partial x_i}(X_k^{n_k})(AX(t))_i \rho(x) dx + \sum_{j=1}^{d} \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_j^2}(X_k^{n_k})\sigma_j \rho(x) dx,$$

$$0 = \sum_{i=1}^{d} n_k \int_{\mathbb{R}^d} \delta_{i=k} X_k^{n_k-1}(AX(t))_i \rho(x) dx + \sum_{j=1}^{d} n_k(n_k-1) \int_{\mathbb{R}^d} \delta_{j=k} X_k^{n_k-2}\sigma_j \rho(x) dx,$$

$$0 = \sum_{i=1}^{d} \mathbb{E}^{\mu}\left[\delta_{i=k} X_k^{n_k-1}(AX)_i\right] + (n_k-1)\sum_{j=1}^{d} \sigma_j \mathbb{E}^{\mu}\left[\delta_{j=k} X_k^{n_k-2}\right],$$

$$0 = \sum_{i=1}^{d} \alpha_{k,i}\mathbb{E}^{\mu}\left[X_k^{n_k-1}X_i\right] + (n_k-1)\sigma_k \mathbb{E}^{\mu}\left[X_k^{n_k-2}\right]. \tag{31}$$

if we multiply by $X_k^{n_k} X_m^{n_m}$, we would get

$$0 = \sum_{i=1}^{d} \alpha_{k,i} n_k \mathbb{E}^{\mu}\left[(X_k^{n_k-1}X_m^{n_m})X_i\right] + \sum_{i=1}^{d} \alpha_{m,i} n_m \mathbb{E}^{\mu}\left[(X_k^{n_k}X_m^{n_m-1})X_i\right] +$$
$$+(n_k-1)n_k\sigma_k \mathbb{E}^{\mu}\left[X_k^{n_k-2}X_m^{n_m}\right] + (n_m-1)n_m\sigma_m \mathbb{E}^{\mu}\left[X_k^{n_k}X_m^{n_m-2}\right]. \tag{32}$$

By considering different polynomials in equations (31) and (32) we get a system, that we can solve applying lest squares approach. Like in 1-dimensional case, we have to find an additional constraint to assure non-triviality of the solution. We use the same idea by estimating quadratic variation matrix $Q_T \in \mathbb{R}^{d\times d}$, forcing σ_i to be non-zero. Let us find the expression of Q_T for the process (30).

$$(Q_T)_{ij} = \langle X_i, X_j\rangle_T = \int_0^T d\langle X_i, X_j\rangle_t = \int_0^T \sum_{k=1}^{d}\sum_{m=1}^{d} \sqrt{2D_{ik}}\sqrt{2D_{jm}} d\langle W_k, W_m\rangle_t,$$

where,

$$\langle W_k, W_m\rangle_t = \delta_{k=m} t, \qquad \text{and} \qquad D_{ij} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

Finally we obtain

$$\frac{(Q_T)_{ij}}{2T} = \frac{\delta_{i=j}}{T}\int_0^T \sqrt{\sigma_i\sigma_j} dt = \delta_{i=j}\sqrt{\sigma_i\sigma_j} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

Estimation

Let $\{X_j\}_{j=0}^J$ be a sampled equidistant observations of the trajectory $(X_t)_{t\in[0,T]}$, that it assumed to be given. In order to estimate expectations we use ergodic theorem and we find moments via time average

$$\mathbb{E}^\mu\left[X_{i_1}^{n_{i_1}} X_{i_2}^{n_{i_2}} X_{i_3}^{n_{i_3}}\right] \approx \frac{1}{T}\int_0^T X(t)_{i_1}^{n_{i_1}} X(t)_{i_2}^{n_{i_2}} X(t)_{i_3}^{n_{i_3}} dt \approx \frac{1}{J}\sum_{j=1}^J (X_{i_1}^j)^{n_{i_1}} (X_{i_2}^j)^{n_{i_2}} (X_{i_3}^j)^{n_{i_3}}.$$

To estimate Q_T we use similar estimator to one in 1-D case

$$(Q_T)_{ii} \approx (\sum_{j=1}^J \Delta X_j (\Delta X_j)^T)_{ii} = \sum_{j=1}^J (X_j^i - X_{j-1}^i)^2.$$

Example on $\mathbb{R}^2$

Figure (13) shows the estimation of the parameters using MoM, that was applied on the data collected from simulated trajectory according to the process

$$dX(t) = \begin{pmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{pmatrix} \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} dt + \sqrt{2\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}} dW(t), \tag{33}$$

With five unknown coefficients to estimate $\theta = (\alpha_1, \alpha_2, \alpha_3, \sigma_1, \sigma_2)$. Let us recall equations (31) and (32) with following polynomials: X_1^2, X_2^2, X_1X_2. That would bring us to the system

$$\begin{pmatrix} \mathbb{E}^\mu[X_1^2] & \mathbb{E}^\mu[X_1X_2] & 0 & 1 & 0 \\ 0 & \mathbb{E}^\mu[X_1X_2] & \mathbb{E}^\mu[X_2^2] & 0 & 1 \\ \mathbb{E}^\mu[X_1X_2] & \mathbb{E}^\mu[X_2^2] + \mathbb{E}^\mu[X_1^2] & \mathbb{E}^\mu[X_1X_2] & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \sigma_1 \\ \sigma_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{\widehat{Q}_{11}}{2T} \\ \frac{\widehat{Q}_{22}}{2T} \end{pmatrix}.$$

Figure (13) demonstrates the solution to the system above according to the final time T used in estimation.

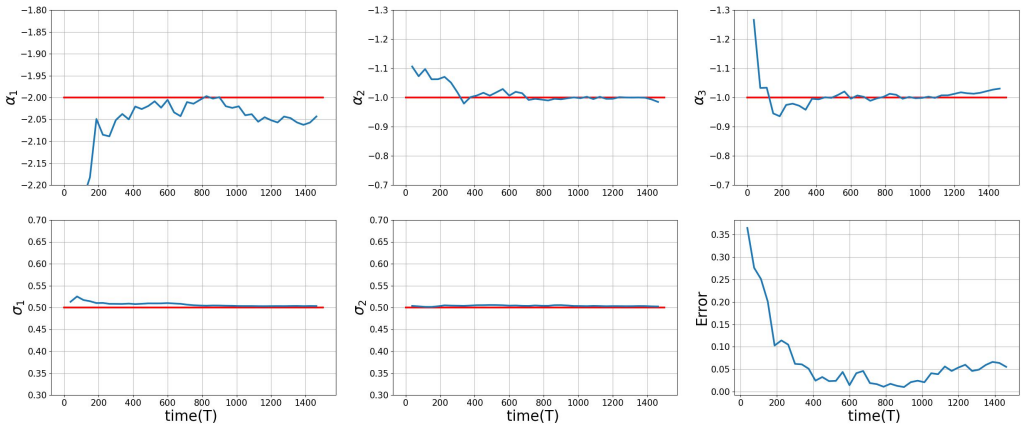


Figure 13: Estimation of unknown parameters over different final time T. Where trajectory was produced using E-M scheme with $h = 0.001$.

8 Some implementation details

8.1 Number of equations to use

In the previous sections we had to find the solutions of the various systems , but we did not specify how many rows of equations the system should contain. In this section, we will try to answer this question.

Source of error

Our main error source is lying in moment approximation. Hence, higher the moment we want to estimate, greater is the error. In fact, each line of equation added to the system requires a higher moment for the approximation.Therefore, it is quite reasonable to use a minimum number of constraints. For example, for the system (8), we use the first two lines plus one extra to avoid the zero solution and for the system (10), the first three lines plus one extra. On Figure (3), we plot the average error, observed for different numbers of equations (we use notation L to indicate the total number of rows) in the system. We can notice the tendency, that increasing the number of equations, contributes to a bigger error. On the other hand, in the case of a bistable process, the use of only three lines also contributes to a relatively high error. In fact, both OU and bistable processes have a symmetric invariant measure, which makes all odd moments equal to zero,and the systems (8) and (10) have lines containing only odd moments. These lines contain no relevant information, and the high error observed is the result of the absence of the required constraints.

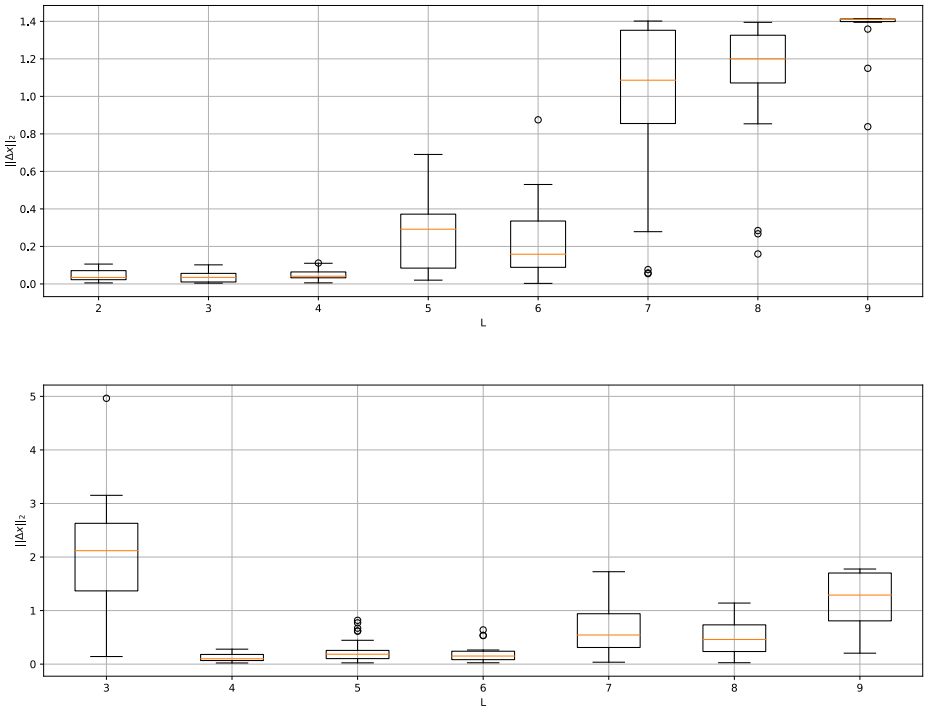


Figure 14: Two plots represent the average error over different simulations (50) of MoM applied on O-U process (on top) and bistable potential process (one below), with respect to different number of equations used in the systems.

We will call these rows **zero lines**. Since in MoM we are trying to approximate the moments, these zero lines are close but not equal to zero, so they contain irrelevant information about the system. In the figure (3) for the bistable process, we can see the following pattern: for every odd number of rows, we have a small jump in the error, which corresponds to the addition of an extra zero line to the system.

Zero lines

In many cases, we encounter processes with symmetric invariant distributions, which can lead to the so-called "zero line problem." To address this issue, we propose a simple approach that involves analyzing the SDE of the process to determine if the drift term is a symmetric function. Alternatively, if the model is unknown, one can evaluate various statistics, such as checking if all odd moments are close to zero and whether the sampled path exhibits symmetric dynamics. Based on these observations, we can set all odd moments to zero. As illustrated in Figure (15), the distribution of X_t and Y_t exhibits symmetry with respect to the 0-axis and the odd moments are nearly zero, supporting the validity of this approach.

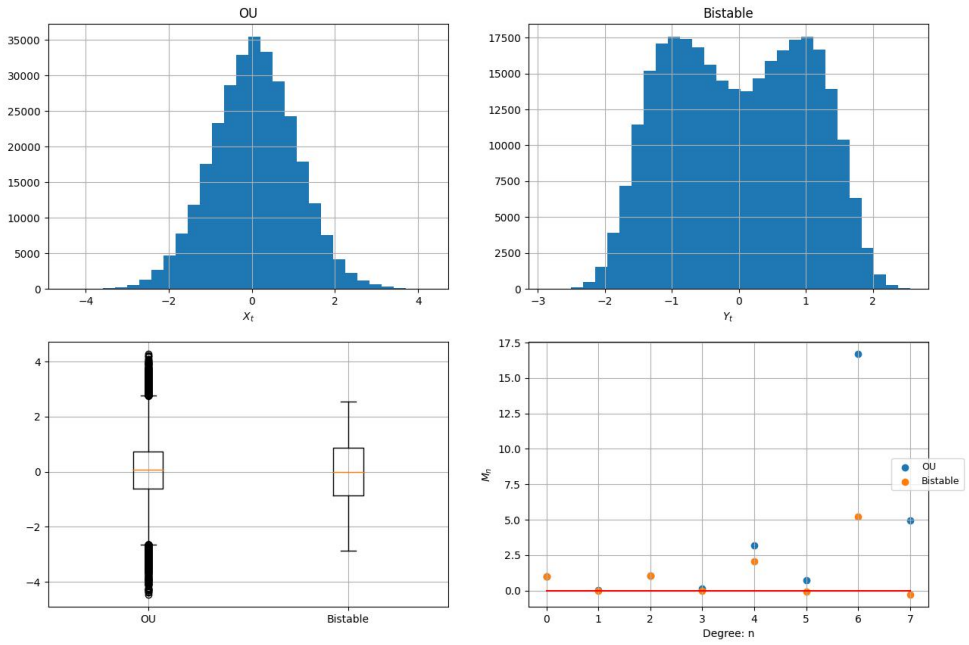


Figure 15: Examples of statistics one can use to detect symmetric nature of the stationary measure in the diffusion process.

Once zero lines are identified in the system, we can remove them to improve the accuracy of our analysis. In matrices (8) and (10), we remove every odd row to eliminate the influence of zero lines. As shown in Figure (16), this procedure leads to a significant decrease in error for both processes when comparing the full system (left) to the system without zero lines (right).

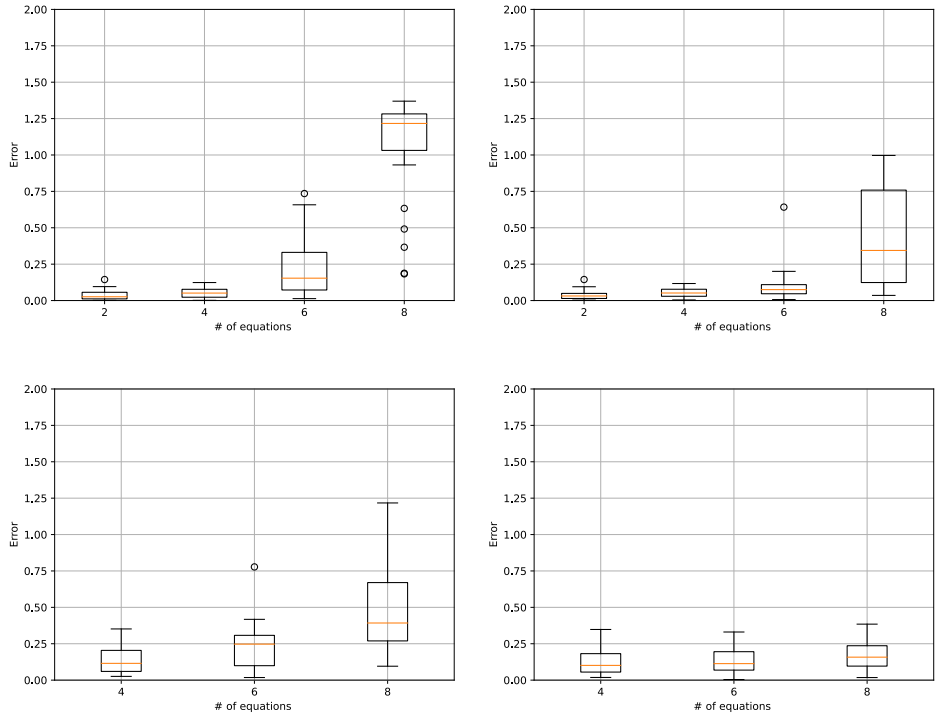


Figure 16: This time, two plots represent the average error over different simulations (50) of MoM applied on O-U process (on top) and bistable potential process (one below), with respect to different number of equations used in the systems. The results on the right figures are obtained using restricted system.

Sensitivity of the solution

Another factor to consider when selecting the number of rows is the stability of the obtained results. As shown in Figure (3), the error can vary significantly depending on the number of equations used, with some configurations leading to a much wider spread of errors than others. For instance, when using nine equations ($L = 9$) for the OU process, the variance is even lower than for configurations with fewer equations ($L = 2, 3, 4, ...$). This may be due to the fact that each realization of the stochastic process and the corresponding moments obtained from the simulations are slightly different. To understand this behavior, it is helpful to analyze the sensitivity of our methodology to small perturbations in moment estimation, which can be quantified using the conditional number of the system, $\kappa := \kappa(\boldsymbol{M})$. It is worth noting that κ also appears in the expression for the error bound

$$\|\Delta\theta\|_2 \leq \kappa(\boldsymbol{M}) \frac{\|\Delta \boldsymbol{M}\|_2}{\|\boldsymbol{M}\|_2} \|\theta\|_2, \tag{34}$$

$$\mathbb{E}\left[\|\Delta\theta\|_2^2\right]^{\frac{1}{2}} \leq \frac{\kappa(\boldsymbol{M})}{\|\boldsymbol{M}\|_2} \frac{(\sqrt{8L})}{\sqrt{T}} \left(\int_0^{+\infty} C_{X_t^{n_{max}}}(u) du \right)^{\frac{1}{2}}$$

The conditional number of the matrix, $\kappa(\boldsymbol{M})$, can be thought of as a factor by which the error term $\|\Delta \boldsymbol{M}\|_2$ may be multiplied. However, in practical applications, we

do not have access to the exact moments and cannot calculate $\kappa(\boldsymbol{M})$ directly. As a workaround, we can take advantage of the symmetry of the problem and consider $\widehat{\boldsymbol{M}}$ to be the desired, unperturbed matrix and $\boldsymbol{M}$ to be the perturbed version. With this approach, equation (34) can be rewritten as follows

$$\|\Delta x\|_2 \leq \kappa(\widehat{\boldsymbol{M}}) \frac{\|\Delta \boldsymbol{M}\|_2}{\|\widehat{\boldsymbol{M}}\|_2} \|\widehat{x}\|_2,$$

To evaluate the sensitivity of the solution, we propose using $\kappa(\widehat{\boldsymbol{M}})$ as a reference value. In most cases, this value is similar to $\kappa(\boldsymbol{M})$ if T is large enough. Figure (17) illustrates the conditional number for the exact system and for the approximated version for OU process. However, there may be instances where these values differ significantly, particularly in the presence of zero lines. As an example, consider Figure (18), in which zero lines have not been removed from the system. We can see that there is a significant deviation from the exact result, especially for odd numbers of rows. Despite this, we still recommend using $\kappa(\widehat{\boldsymbol{M}})$ as a reference value, as it reflects the sensitivity of the underlying system that we are actually dealing with. Additionally, by analyzing Figures (18) and (3), we observe that a smaller reference value corresponds to a smaller variation in the solution. In the case of results using nine equations, we see that the reference value is relatively low, in contrast to the exact conditional number.

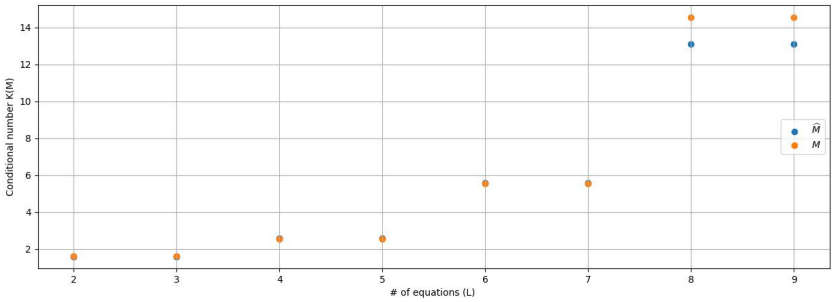


Figure 17: Conditional number of the system for OU process (with removed zero lines).

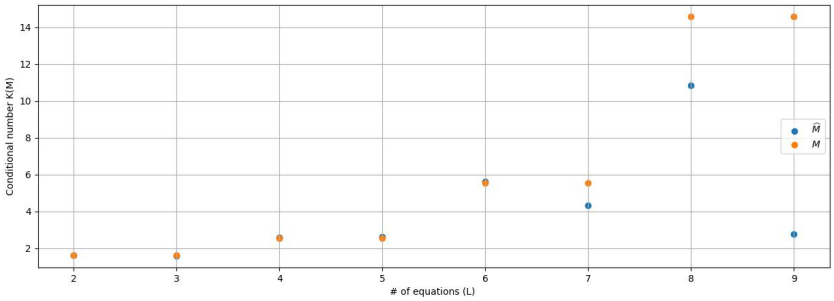


Figure 18: Conditional number of the system for OU process (with zero lines).

We also mention that we are not restricted to consider rows in increasing manner, i.e., until $L = 3, 4....$ One can actually choose the rows (equations) that will be integrated to the system $\mathbf{M}$. For example, in multidimensional setting, one can choose the

row corresponding to the equation multiplied by polynomial $X_1X_2^2$, alternatively to $X_1^2X_2$. In order to build the optimal system, zero line problem and conditional number are two important indicators which we can take into account.

8.2 Source point bias

In practice, it is not always possible to ensure that the initial state,X_0, follows the invariant distribution. If X_0 is fixed at a point that has a low probability of occurring according to the stationary distribution, i.e., $\mathbb{P}_\mu(X_0 = x) \approx 0$, this can lead to bias in the estimations. Even if the diffusion process converges quickly to the stationary distribution, the bias introduced by the initial state can still have a significant effect on the results. To mitigate this bias and ensure good performance, it is necessary to remove the initial data that may be affected by this bias and only consider data that has been collected after the stationary distribution has been reached. We propose quite simple approach, that consist in cut off the begging of the stream, which is outside of the red lines bound, see figure (19), where these red lines bound the 90% of all collected data.

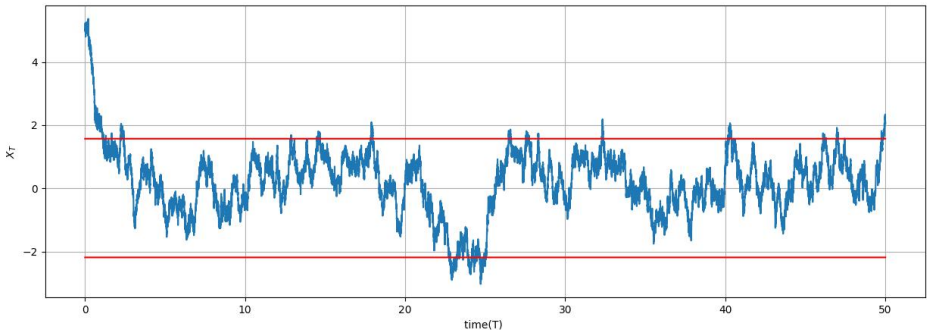


Figure 19: Plot represent a part of the stream corresponding to the OU process.

Following this simple principal, results in significant increase of the accuracy. Figure (20) represents the error over time for OU process with whole data steam (left) and its removed bias version (right). In particular, we see the significant decrease in error, for processes with more extreme initial conditions.

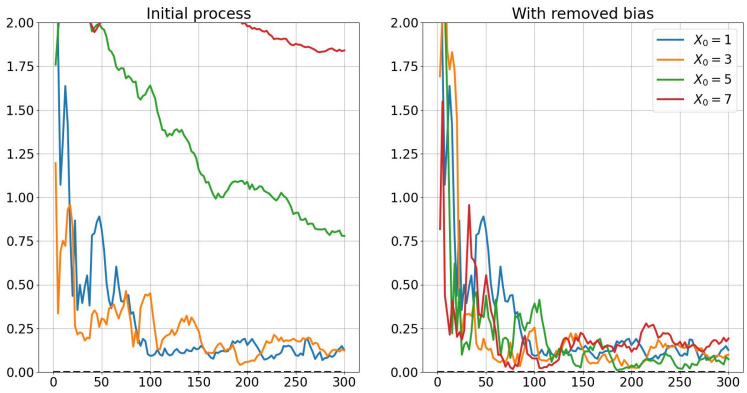


Figure 20: Figure represent error plot according to the euclidean norm, i.e., $\|(\theta - \hat{\theta})\|_2$, over different final time T.

8.3 Several trajectories

When there are available discrete data from multiple independent trajectories, there is often a question of whether it is better to average over the moments $\bar{\mu}_n = \sum_i^I \hat{\mu}_n^i$ or to average the obtained parameters $\bar{\theta} = \sum_i^I \hat{\theta}^i$. Through experimentation with different models, we have found that it is always better to average over moments. In Figure (21), we show an example of a bistable process with a final time of $T = 100$. On the left plot, we take the average over time points, with the x-axis representing the number of independent trajectories used to average. On the right, we take the average over the obtained parameters. We can see that the first approach produces much better results. However, when the final time is sufficiently large (see Figure (22), $T = 1000$), both approaches produce similar results and simply reduce the variance of the obtained results.

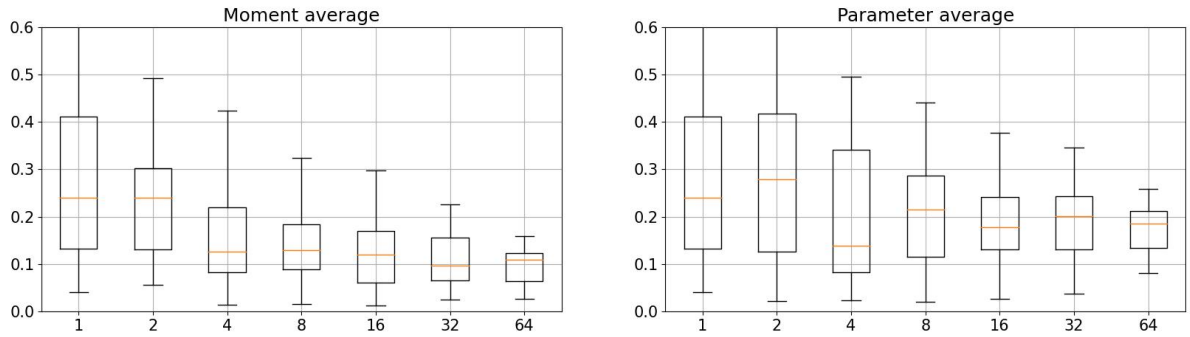


Figure 21: Both plot represent error plot according to the euclidean norm, i.e., $\|(\theta - \hat{\theta})\|_2$, for different number of independent trajectories with $T = 100$. Statistics collected over 50 trials.

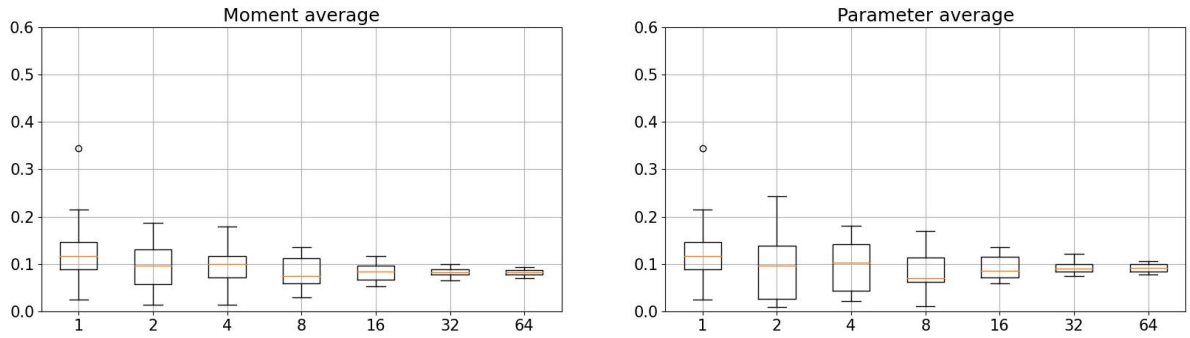


Figure 22: Both plot represent error plot according to the euclidean norm, i.e., $\|(\theta - \hat{\theta})\|_2$, for different number of independent trajectories with $T = 1000$. Statistics collected over 50 trials.

9 Conclusion and further discussions

The Method of Moments (MoM) is a simple and effective technique for estimating unknown parameters in diffusion processes. In cases where the final time T is sufficiently large, MoM is accurate and can be compared to other more classical techniques, such as the likelihood estimator used in Section 3. MoM can be used in the setting of multiplicative noise and can be easily generalized to multidimensional diffusion processes. In addition, as we saw in Section 5, it can be used to derive the forms of the drift and diffusion terms when the actual model is unknown, and can even be able to approximate non-polynomial drift functions. It is important to note that, in more general settings, we are not limited to considering only polynomial expansions for the drift and diffusion terms and we can apply the same approach to any drift/diffusion term of the form $\sum_k^K \alpha_k f_k(x)$, where $f_k(x)$ are sufficiently regular functions. However, since most models have polynomial terms, we chose to focus on this particular case in our research project. Nevertheless, in Section 6, we analyzed processes with non-polynomial terms, where the MoM performed as expected. Overall, MoM is a useful and versatile tool for analyzing diffusion processes.

There are still many aspects of this technique that we were unable to explore due to the limited time available. In particular, we note that we are not limited to multiplying the Fokker-Planck equation by monomials or polynomials and we can use bases functions from an appropriate Hilbert space instead. One example of this is the use of eigenfunctions of the generator of the underlying diffusion process. Interesting to note, that for some processes eigenfunctions are polynomials. In particular, the Hermit polynomials (Pavliotis 2014, see section 4.4) are eigenfunctions for the Ornstein-Uhlenbeck process or the Laguerre polynomials (see Karlin and Taylor 1981, page: 333) for the radial Ornstein-Uhlenbeck process.

Another avenue for future research is the adaptation of the MoM to stochastic models with fractional Brownian motion and jump processes. These types of processes pose unique challenges and it would be interesting to see how MoM can be applied in these contexts. Overall, there is still much to learn about the capabilities and limitations of the MoM approach, and we look forward to further investigating this technique in the future.

References

Karlin, Samuel and Howard E Taylor (1981). *A second course in stochastic processes*. Elsevier.

Kessler, Mathieu and Michael Sørensen (1999). "Estimating equations based on eigenfunctions for a discretely observed diffusion process". In: *Bernoulli*, pp. 299–314.

Mattingly, Jonathan C, Andrew M Stuart, and Michael V Tretyakov (2010). "Convergence of numerical time-averaging and stationary measures via Poisson equations". In: *SIAM Journal on Numerical Analysis* 48.2.

Mishura, Yuliya (2021). *Linear stochastic differential equation. Cox-Ingersoll-Ross equation*.

Pavliotis, Grigorios A (2014). *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*. Vol. 60. Springer.