One shot learning of stochastic differential equations with Gaussian processes

M. Darcy[1] A. Gualandi[5] B. Hamzi[1,2] G. Livieri[3] H. Owhadi[1] P. Tavallali[4]

[1]California Institute of Technology

[2]Alan Turing Institute [3]Scuola Normale Superiore [4]JPL, NASA [5]University of Cambridge

DEDS 2024

Table of Contents

Caltech

This talk focuses on past and ongoing work on the learning of stochastic differential equations from data using Gaussian processes.

1. Matthieu Darcy, Boumediene Hamzi, Giulia Livieri, Houman Owhadi, and Peyman Tavallali. "One-shot learning of stochastic differential equations with data adapted kernels". In: *Physica D: Nonlinear Phenomena* 444 (2023)
2. Ongoing work with Adriano Gualandi on learning and predicting earthquakes, and extensions of previous work.

Motivation

The general objective of this line of work is to learn the unknown drift f and the diffusion σ of a generic SDE:

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t$$

where $X_t \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}^d, \sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$.

Objective

Given samples $X := (X_{t_n})_{n=1}^N$ from the SDE, learn the drift f and diffusion σ.

Motivation

The general objective of this line of work is to learn the unknown drift f and the diffusion σ of a generic SDE:

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t$$

where $X_t \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}^d, \sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$.

Objective

Given samples $X := (X_{t_n})_{n=1}^N$ from the SDE, learn the drift f and diffusion σ.

This problem is challenging:

- The observations X come from a single (non-ergodic) trajectory.
- We make few assumptions on f and σ.
- Because of the inherent stochasticity of W_t, the observations X only provide indirect information on f and σ.
- The sampling time-steps Δt can introduce a discretization error.

Table of Contents

Problem statement

We[1] considered the case where $d = 1$ and N is small (a few hundred data points).

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t$$

where

$$\begin{aligned} f &: \mathbb{R} \to \mathbb{R} \quad \text{drift} \\ \sigma &: \mathbb{R} \to \mathbb{R} \quad \text{diffusion.} \end{aligned}$$

[1] Matthieu Darcy, Boumediene Hamzi, Giulia Livieri, Houman Owhadi, and Peyman Tavallali. "One-shot learning of stochastic differential equations with data adapted kernels". In: *Physica D: Nonlinear Phenomena* 444 (2023)

Method summary

Our method is summarized

1. Assume a Euler-Maruyama discretization of the dynamics.
2. Place Gaussian Process priors on f and σ and recover them using Maximum A Posteriori (MAP) estimation.
3. Optimize the covariance/kernel functions of the Gaussian processes using randomized cross-validation.

Step 1 : Modeling Assumption

Let $X_n := X_{t_n}$. We assume the following discretization, based on the Euler-Maruyama model:

$$X_{n+1} = X_n + f(X_n)\Delta t + \sigma(X_n)\sqrt{\Delta t}\xi_n + \varepsilon_n$$

where

$$\xi_n \overset{d}{\sim} \mathcal{N}(0,1) \quad \text{dynamics noise}$$
$$\varepsilon_n \overset{d}{\sim} \mathcal{N}(0,\lambda) \quad \text{modeling noise}$$

are independent.

Defining $Y_n := X_{n+1} - X_n$, our model can be restated as

$$\boxed{Y_n = f(X_n)\Delta t + \sigma(X_n)\sqrt{\Delta t}\xi_n + \varepsilon_n}$$

Step 2: Gaussian process prior

We assume that f and σ are distributed according to **independent** Gaussian processes:

$$f \overset{d}{\sim} \mathcal{GP}(\mathbf{0}, \boldsymbol{K})$$
$$\sigma \overset{d}{\sim} \mathcal{GP}(\mathbf{0}, \boldsymbol{G}).$$

We recover $\bar{f} \in \mathbb{R}^N$ and $\bar{\sigma} \in \mathbb{R}^N$, the function values at the observed data points:

$$\bar{f}_n := f(X_n)$$
$$\bar{\sigma}_n := \sigma(X_n).$$

Once we have recovered, we can predict future values of f and σ.

Step 2: MAP estimation

By Baye's rule

$$p(\bar{f},\bar{\sigma}|Y,X) \propto p(Y|\bar{f},\bar{\sigma})\overbrace{p(\bar{f}|X)p(\bar{\sigma}|X)}^{\text{independence}}.$$

Step 2: MAP estimation

By Baye's rule

$$p(\bar{f}, \bar{\sigma} | Y, X) \propto p(Y | \bar{f}, \bar{\sigma}) \overbrace{p(\bar{f} | X) p(\bar{\sigma} | X)}^{\text{independence}} .$$

Using our model and our prior on $\bar{f}$ and $\bar{\sigma}$:

$$-\ln p(\bar{f}, \bar{\sigma} | Y, X) \propto \mathcal{L}(\bar{f}, \bar{\sigma}) := \overbrace{(Y - \Delta t \bar{f})^T (\Delta t \Sigma + \lambda I)^{-1} (Y - \Delta t \bar{f}) + \sum_{n=1}^{N} \ln(\bar{\sigma}_n^2 \Delta t + \lambda)}^{-\ln p(Y|\bar{f},\bar{\sigma})}$$

$$+ \underbrace{\bar{f}^T K(X, X)^{-1} \bar{f}}_{-\ln p(\bar{f}|X)} + \underbrace{\bar{\sigma}^T G(X, X)^{-1} \bar{\sigma}}_{-\ln p(\bar{\sigma}|X)} .$$

where Σ is a diagonal matrix with entries $\bar{\sigma}_n^2$.

Step 2: MAP estimation

By Baye's rule

$$p(\bar{f}, \bar{\sigma}|Y, X) \propto p(Y|\bar{f}, \bar{\sigma}) \overbrace{p(\bar{f}|X)p(\bar{\sigma}|X)}^{\text{independence}} .$$

Using our model and our prior on $\bar{f}$ and $\bar{\sigma}$:

$$-\ln p(\bar{f}, \bar{\sigma}|Y, X) \propto \mathcal{L}(\bar{f}, \bar{\sigma}) := \overbrace{(Y - \Delta t\bar{f})^T(\Delta t\Sigma + \lambda I)^{-1}(Y - \Delta t\bar{f}) + \sum_{n=1}^{N} \ln(\bar{\sigma}_n^2 \Delta t + \lambda)}^{-\ln p(Y|\bar{f},\bar{\sigma})}$$
$$+ \underbrace{\bar{f}^T K(X, X)^{-1}\bar{f}}_{-\ln p(\bar{f}|X)} + \underbrace{\bar{\sigma}^T G(X, X)^{-1}\bar{\sigma}}_{-\ln p(\bar{\sigma}|X)} .$$

where Σ is a diagonal matrix with entries $\bar{\sigma}_n^2$.

The recovery of f, σ is reduced to the minimization of $\mathcal{L}(\bar{f}, \bar{\sigma})$.

Caltech

Step 2: Alternative minimization

Representer theorem

For any given $\bar{\sigma}$, the minimizer in $\bar{f}$ of $\mathcal{L}(\bar{f}, \bar{\sigma})$ is

$$\bar{f}^*(\sigma) := \arg\min_{\bar{f}} \mathcal{L}(\bar{f}, \bar{\sigma}) = K(X, X)\Delta t \Big(\Delta t^2 K(X, X) + \Delta t \Sigma + \lambda I\Big)^{-1} Y$$

Using the representer theorem, and plugging $\bar{f}^*(\sigma)$ into the original objective, the minimization in σ is given by:

$$\mathcal{L}(\bar{f}^*(\sigma), \sigma).$$

The objective function is non-convex in σ and its minimization is done through a gradient descent method.

Step 3: Hyper-parameter optimization

The kernel functions $\boldsymbol{K}, \boldsymbol{G}$ are parameterized by some parameter $\boldsymbol{\theta}$. We find that in the low-data regime, learning $\boldsymbol{\theta}$ is critical to good performance.
We use a variant of randomized cross-validation[2] to select $\boldsymbol{\theta}$ which is based on two principles:

- **Cross validation:** optimize the model on a subset $\mathcal{D}_{\Pi}$ of the data and measure the performance on a withheld subset $\mathcal{D}_{\Pi^c}$.
- **Randomly** sample subsets $(\mathcal{D}_{\Pi}, \mathcal{D}_{\Pi^c})$ randomly and use this noisy loss to optimize the hyperparameters $\boldsymbol{\theta}$.

[2] Houman Owhadi and Gene Ryan Yoo. "Kernel Flows: From learning kernels from data into the abyss". In: *Journal of Computational Physics* 389 (2019), Boumediene Hamzi and Houman Owhadi. "Learning dynamical systems from data: A simple cross-validation perspective, part I: Parametric kernel flows". In: *Physica D: Nonlinear Phenomena* 421 (2021), p. 132817

Example: Exponential decay volatility

$$\boxed{dX_t = \mu X_t dt + b \exp(-X_t^2) dW_t \quad \text{Exponential decay volatility.}}$$

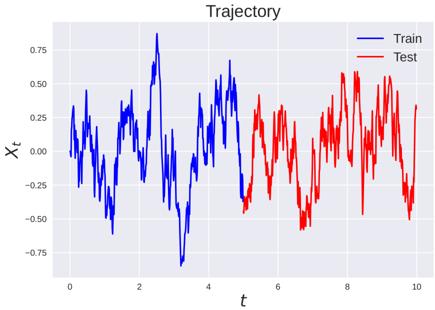

Figure: Exponential decay volatility process

Example: Exponential decay volatility

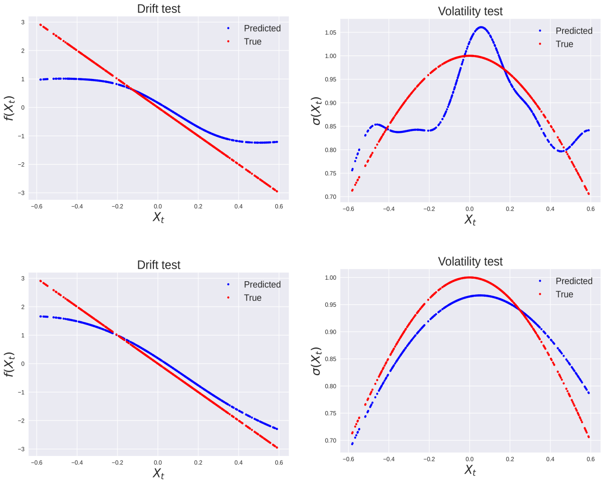

Figure: Forecast: non-learned kernel (top) and learned kernel (bottom).

Caltech

Summary:

We propose a general method to learn the drift and diffusion of general SDEs from one sample trajectory.

- We can capture a broad class of f and σ thanks to the generality of Gaussian processes.
- We can address some level of misspecification due to a coarse Δt.
- We provide a method learn the hyper-parameters of the GPs, which is critical for a good performance in the low data setting.
- We can leverage the theory of kernels/GPs to obtain theoretical guarantees and uncertainty quantification.

However, learning the diffusion σ is generally expensive when N or d is large.

Table of Contents

SDEs for labquakes

A recent study [3] has found that laboratory earthquakes can be accurately modeled by a 4 dimensional stochastic differential equations:

$$
\begin{aligned}
dx_t &= \left(\frac{e^x\left((\beta_1 - 1)x(1+\lambda u) + y\right) - u + \kappa\left(\frac{v_0}{v_*} - e^x\right) - \frac{du_t + \lambda xy}{1+\lambda u} + \nu e^x}{1 + \lambda u + \nu e^x} \right) dt \\
dy_t &= \kappa\left(\frac{v_0}{v_*} - e^x\right) dt - \nu e^x dx_t + \varepsilon_y dW_t^y \\
dz_t &= -\rho e^x(\beta_2 x + z)dt \\
du_t &= (-\alpha - \gamma u)dt + dz_t + \varepsilon_u dW_t^u
\end{aligned}
$$

However, this system depends on parameters that are very hard to estimate in practice, and it is unknown if this model is accurate for real earthquakes.

[3] A. Gualandi, D. Faranda, C. Marone, M. Cocco, and G. Mengaldo. "Deterministic and stochastic chaos characterize laboratory earthquakes". In: *Earth and Planetary Science Letters* 604 (2023)

SDEs for labquakes

This is an SDE of the form:

$$dX_t = f(X_t)dt + \sqrt{\Gamma}dW_t$$

where Γ is diagonal and constant. Compared to the previous section:

- The system is in higher dimensions $d = 4$ and we have more data points $N = 10k - 20k$.
- The noise is additive and there are good estimates for $\sqrt{\Gamma}$ (no need to learn σ).
- The system is characterized by areas of high acceleration and sharp drops.

Because of the structure of the diffusion, we can apply our method to each dimension of X_t independently.

Example

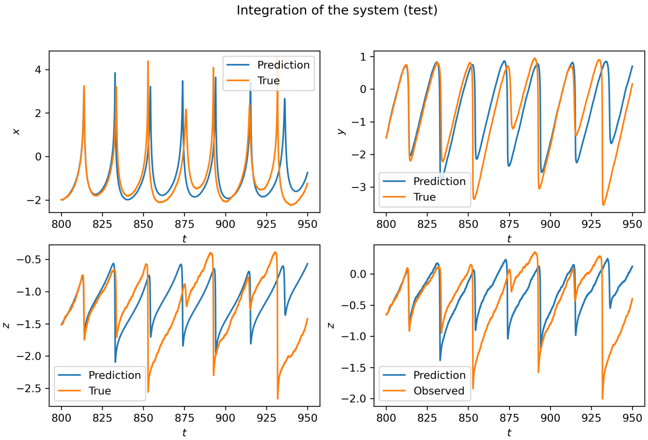

Figure: We find good recovery of the dynamics even without kernel learning.

Caltech

Conclusion and future work

1. We propose a general framework to learn the drift and diffusion of SDEs which is effective in the low data regime.
2. We apply this framework to the prediction of earthquakes under some simplifying assumptions on the noise.

Future work focuses on two questions:

1. How can we provide rigorous theoretical guarantees and effective uncertainty quantification of the prediction?
2. Can we learn the matrix $\sqrt{\Gamma}$? Can we extend this to "simple" $\sigma(X_t)$?

Matthieu Darcy, Boumediene Hamzi, Giulia Livieri, Houman Owhadi, and Peyman Tavallali. "One-shot learning of stochastic differential equations with data adapted kernels". In: *Physica D: Nonlinear Phenomena* 444 (2023)