A Nonparametric Spatio-temporal SDE Model

Cagatay Yildiz, Markus Heinonen and Harri Lähdesmäki
Department of Computer Science
Aalto University, Finland
{cagatay.yildiz, markus.o.heinonen, harri.lahdesmaki}@aalto.fi

Abstract

We propose a nonparametric spatio-temporal stochastic differential equation (SDE) model that can learn the underlying dynamics of arbitrary continuous-time systems without prior knowledge. We augment the input space of the drift function of an SDE with a temporal component to account for spatio-temporal patterns. The experiments demonstrate that the spatio-temporal model is better able to fit a real world data set that has complex dynamics than the spatial model, and can also reduce the forecasting error.

1 Introduction

Dynamical systems modeling is a cornerstone of experimental sciences. Modelers attempt to capture the dynamical behavior of a stochastic system or a phenomenon in order to improve its understanding and make predictions about its future state. Stochastic differential equations (SDEs) are an effective formalism for modelling systems with underlying stochastic dynamics, with wide range of applications [5]. The key problem in SDEs is estimation of the underlying deterministic driving function, and the stochastic diffusion component.

In this work, we are interested in a multivariate system governed by Markov process $\mathbf{x}_t$ described by an SDE

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + \sigma(\mathbf{x}_t, t)dW_t \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^D$ is the state vector of a D-dimensional dynamical system at continuous time $t \in \mathbb{R}$, $\mathbf{f}(\mathbf{x}, t) \in \mathbb{R}^D$ is a deterministic state evolution, $\sigma(\mathbf{x}_t, t) \in \mathbb{R}$ is a scalar magnitude of the stochastic multivariate Wiener process $W_t \in \mathbb{R}^D$. The Wiener process has zero initial state $W_0 = \mathbf{0}$, and the independent increments $W_{t+s} - W_t \sim \mathcal{N}(\mathbf{0}, sI)$ follow a Gaussian with standard deviation $\sqrt{s}$. The state solutions of SDE are given by the Itô integral [9]

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x}_\tau, \tau)d\tau + \int_0^t \sigma(\mathbf{x}_\tau, \tau)dW_\tau, \tag{2}$$

where we integrate the system state from an initial state $\mathbf{x}_0$ for time t forward, and where τ is an auxiliary time variable. We assume the states are observed with additive noise $\mathbf{y}(t) = \mathbf{x}(t) + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Omega)$ and $\Omega = \mathrm{diag}(\omega_1^2, \ldots, \omega_D^2)$. With $\sigma(\mathbf{x}_\tau, \tau) = 0$, we obtain the ordinary differential equation (ODE) solution.

There is a vast literature on inferring SDEs [5] where a parametric form for drift and diffusion functions are assumed to be known, and the parameters of those functions are optimized. Such methods cannot be readily applied to real world data sets, where the underlying dynamics are unknown. To tackle such data sets, non-parametric drift and diffusion functions have been proposed in several works [16, 3, 6, 13]. This work extends the method proposed in [16] to account for spatio-temporal patterns.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Nonparametric SDE Model

In this section, we review the sparse Gaussian process (GP) modeling of differential equations described in [16]. The model introduces two GP priors over the vector valued drift $\mathbf{f}(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^D$ and scalar valued diffusion $\sigma(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ functions

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, K_{\mathbf{f}}(\mathbf{x}, \mathbf{x}')), \qquad \sigma(\mathbf{x}) \sim \mathcal{GP}(0, k_\sigma(\mathbf{x}, \mathbf{x}')) \tag{3}$$

Drift and diffusion functions have zero mean, the drift kernel $K_{\mathbf{f}}(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^{D\times D}$ is matrix valued, and the diffusion kernel $k_\sigma(\mathbf{x}, \mathbf{x}') \in \mathbb{R}$ is univariate. Covariance matrices are defined using squared exponential kernel function

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d - x'_d)^2}{\ell_d^2}\right) \tag{4}$$

where $\theta_{\mathbf{f}} = \{\sigma_f, \ell_{\mathbf{f}1}, \ldots, \ell_{\mathbf{f}D}\}$ and $\theta_\sigma = \{\sigma_\sigma, \ell_{\sigma 1}, \ldots, \ell_{\sigma D}\}$ stand for the function-specific kernel parameters. Following [16], the identity decomposable kernel $K_{\mathbf{f}}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \cdot I_D$ is used for the matrix valued drift kernel [2], whereas $k_\sigma(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$. Given a set of states $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T \in \mathbb{R}^{N\times D}$, the values of the drift function $F = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_N))^T \in \mathbb{R}^{N\times D}$ and the diffusion function $\boldsymbol{\sigma} = (\sigma(\mathbf{x}_1), \ldots, \sigma(\mathbf{x}_N))^T \in \mathbb{R}^N$ follow normal prior distributions

$$p(F) = \mathcal{N}(\operatorname{vec} F | \mathbf{0}, \mathbf{K}_{\mathbf{f}}(X, X)), \qquad p(\boldsymbol{\sigma}) = \mathcal{N}(\boldsymbol{\sigma} | \mathbf{0}, K_\sigma(X, X)) \tag{5}$$

where $\mathbf{K}_{\mathbf{f}}(X, X) = (K_{\mathbf{f}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^N \in \mathbb{R}^{ND\times ND}$ is a block matrix of matrix-valued kernels $K_{\mathbf{f}}(\mathbf{x}_i, \mathbf{x}_j)$ and diffusion kernel is given by $K_\sigma(X, X) = (k_\sigma(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^N \in \mathbb{R}^{N\times N}$.

In standard GP regression, the posterior over the function values are obtained by conditioning the GP prior on the data [11]. On the other hand, in differential equation models, the conditionals $\mathbf{f}(\mathbf{x})|Y$ and $\sigma(\mathbf{x})|Y$ are intractable due to the integral mapping between observed states and differentials. To overcome this, two sets of *inducing variables* $U_{\mathbf{f}} = (\mathbf{u}_{\mathbf{f}1}, \ldots, \mathbf{u}_{\mathbf{f}M})^T \in \mathbb{R}^{M\times D}$ and $\mathbf{u}_\sigma = (u_{\sigma 1}, \ldots, u_{\sigma M})^T \in \mathbb{R}^M$ are introduced [10]. $U_{\mathbf{f}}$ and $\mathbf{u}_\sigma$ contain the values of the drift and diffusion functions at *inducing locations* $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$, which live in the same space as the states X. Finally, the drift and diffusion functions are interpolated from the locations and variables:

$$\mathbf{f}(\mathbf{x}) \triangleq \mathbf{K}_{\mathbf{f}}(\mathbf{x}, Z)\mathbf{K}_{\mathbf{f}}(Z, Z)^{-1}\mathbf{u}_{\mathbf{f}}, \qquad \sigma(\mathbf{x}) \triangleq K_\sigma(\mathbf{x}, Z)K_\sigma(Z, Z)^{-1}\mathbf{u}_\sigma \tag{6}$$

where $\mathbf{u}_{\mathbf{f}} = \operatorname{vec} U_{\mathbf{f}}$.

The SDE model is determined via the inducing locations Z, the inducing values $U_{\mathbf{f}}$ and $\mathbf{u}_\sigma$, the observation noise variance Ω, and the kernel parameters $\theta_{\mathbf{f}}$ and θ_σ of the drift and diffusion kernels. The model learns the underlying system to induce state distributions with high expected likelihood $p(\mathbf{y}_i|\mathbf{f}, \sigma, \Omega) = \mathbb{E}_{p(\mathbf{x}|t_i;\mathbf{f},\sigma)}[\mathcal{N}(\mathbf{y}_i|\mathbf{x}, \Omega)]$, which is intractable. The posterior of the model combines the likelihood $p(\mathbf{y}_i|\mathbf{f}, \sigma, \Omega)$ and the independent priors $p(U_{\mathbf{f}})$ and $p(\mathbf{u}_\sigma)$ using Bayes' theorem as

$$p(U_{\mathbf{f}}, \mathbf{u}_\sigma|Y) \propto p(U_{\mathbf{f}}, \mathbf{u}_\sigma)p(Y|U_{\mathbf{f}}, \mathbf{u}_\sigma) = p(U_{\mathbf{f}})p(\mathbf{u}_\sigma)\prod_{i=1}^{N}\mathbb{E}_{p(\mathbf{x}|t_i;\mathbf{f},\sigma)}[\mathcal{N}(\mathbf{y}_i|\mathbf{x}, \Omega)] \tag{7}$$

$$\approx \mathcal{N}(\mathbf{u}_{\mathbf{f}}|\mathbf{0}, \mathbf{K}_{\mathbf{f}}(Z, Z))\mathcal{N}(\mathbf{u}_\sigma|\mathbf{0}, K_\sigma(Z, Z)) \times \prod_{i=1}^{N}\frac{1}{N_s}\sum_{s=1}^{N_s}\mathcal{N}(\mathbf{y}_i|\mathbf{x}_i^{(s)}, \Omega) \tag{8}$$

where $\mathbf{x}^{(s)} \sim p(\mathbf{x}_{0\ldots t}|U_{\mathbf{f}}, \mathbf{u}_\sigma, Z)$ denotes a path sample $\mathbf{x}_t^{(s)}$ that is drawn from the time dependent state distribution $p(\mathbf{x}_{0\ldots t}|U_{\mathbf{f}}, \mathbf{u}_\sigma, Z)$ by sampling a Brownian motion path $W_t^{(s)}$. The true expected likelihood is approximated by unbiased Monte Carlo averaging. The likelihood estimate with N_s samples turns out to be a kernel density estimator with Gaussian bases.

We draw the sample paths using Euler-Maruyama (EM) [9]:

$$\mathbf{x}_{i+1}^{(s)} = \mathbf{x}_i^{(s)} + \mathbf{f}(\mathbf{x}_i^{(s)})\Delta t + \sigma(\mathbf{x}_i^{(s)})\Delta W_i^{(s)}, \tag{9}$$

where we discretise time into N_T subintervals $t_0, t_1, \ldots, t_{N_T}$ of width $\Delta t = t_{N_T}/N_T$, and sample the Wiener coefficients as $\Delta W_i^{(s)} \sim \mathcal{N}(\mathbf{0}, \Delta t \cdot I)$ with standard deviation $\sqrt{\Delta t}$. We set $\mathbf{x}_0^{(s)}$ to the initial observation and use (9) to compute state path $\mathbf{x}^{(s)} \equiv (\mathbf{x}_0^{(s)}, \mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{T_N}^{(s)})$. The number of time steps $N_T > N$ is often higher than the number of observed timepoints to achieve sufficient path resolution. Finally, the maximum a posteriori (MAP) estimate of the model parameters are found using sensitivity equation based gradients (see [16] for details).

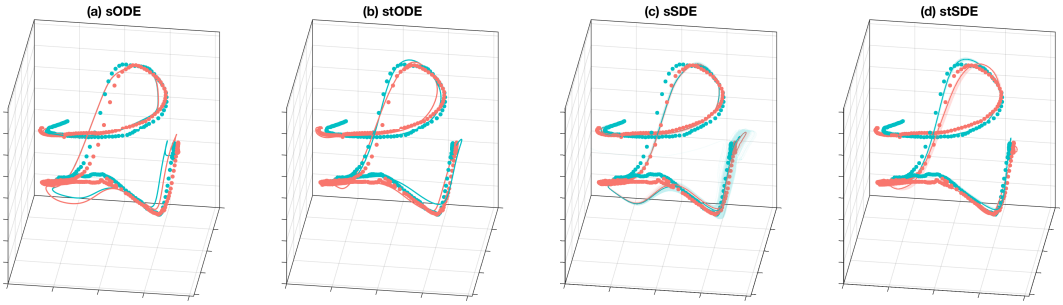

Figure 1: Four different training fits attaining the highest posterior values. For visualization purposes, only two input sequences are drawn with red and green dots. The solid lines in ODE models show the inferred trajectories with two different initial values. In SDE models, solid lines are single random state paths, and the colored regions contain 1000 state paths.

3 Spatio-Temporal SDE Model

So far, the formulation relies on SDE states $\mathbf{x}(t)$ and the inducing locations living in the same space, $\mathbb{R}^D$, which remains valid as long as the input to the drift function is only the state. In order to define drift functions explicitly parameterized by time, we augment the space to include a temporal component, that is, $\mathbf{f}(\mathbf{x}, t) : \mathbb{R}^{D+} \to \mathbb{R}^D$ where $\mathbb{R}^{D+} = (\mathbb{R}^D \times \mathbb{R}^+)$. The GP prior over the drift function then becomes

$$\mathbf{f}(\mathbf{x}, t) \sim \mathcal{GP}(\mathbf{0}, K_{\mathbf{f}}((\mathbf{x}, t), (\mathbf{x}', t'))). \tag{10}$$

The matrix valued kernel is again defined to be the identity decomposable kernel $K_{\mathbf{f}}((\mathbf{x}, t), (\mathbf{x}', t')) = k((\mathbf{x}, t), (\mathbf{x}', t')) \cdot I_D$, where we redefine the kernel function of the drift as

$$k((\mathbf{x}, t), (\mathbf{x}', t')) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d - x'_d)^2}{\ell_{\mathbf{f}d}^2} - \frac{(t-t')^2}{2\ell_t^2}\right). \tag{11}$$

New kernel function turns out to be the product of squared exponential kernel functions over space and time and requires one additional parameter ℓ_t modeling the smoothness in temporal domain. In the limit $\ell_t \to \infty$, we recover the spatial model. Because the kernel function is defined in the extended space, inducing locations Z must have an additional component: $\mathbf{z}_i \in \mathbb{R}^{D+}$. Time variant diffusion functions can also be defined in a similar way, which we do not investigate in this work.

Spatio-temporal drift function boosts the model power and flexibility without any additional parameter other than temporal lengthscale. In spatial systems, state trajectory follows the same dynamics (up to random perturbations due to diffusion) when a particular state is visited at different points in time. With the proposed spatio-temporal formulation, temporal proximity is also taken into account when defining state dynamics, meaning that different representations can be learned over time.

The model is implemented in Tensorflow [1]. We find the MAP estimates of the inducing variables $U_{\mathbf{f}}, \mathbf{u}_\sigma$ and the noise variance Ω using the Adam [8] optimizer with learning rate 0.001 and staircase (every 20th iteration) exponential decay with rate 0.99. Kernel parameters in complex GP models are usually hard to optimize since they appear in the covariance matrix. In our case, the covariance matrix itself is connected to the likelihood through an integral, further complicating the optimization. Therefore, we perform a grid search to optimize lengthscales $\{\ell_{\mathbf{f}1} \dots \ell_{\mathbf{f}D}, \ell_t, \ell_{\sigma 1} \dots \ell_{\sigma D}\}$, and set the signal variance terms to one, $\sigma_f = \sigma_g = 1$. The inducing points are scattered around the data and kept fixed during optimization.

4 Experiments

The experiments aim at demonstrating that the spatio-temporal model is better able to fit the data than the spatial model, and also better at capturing the data generating dynamics. The experiments are performed on a benchmark dataset of human motion capture data from the Carnegie Mellon University motion capture (CMU mocap) database. We followed the data preprocessing technique in [14], which results in 50-dimensional pose measurements from a person swinging a golf club where

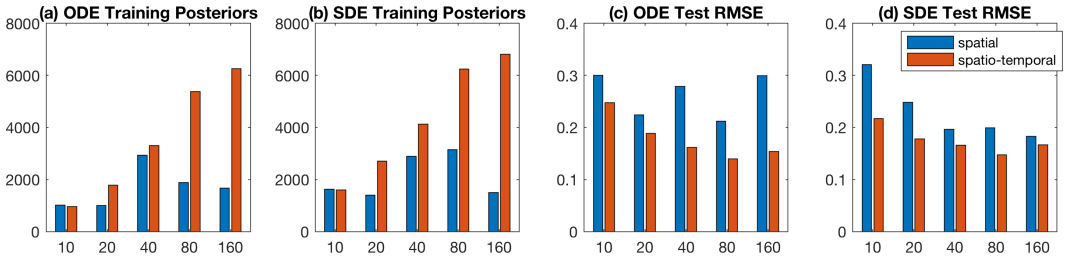

Figure 2: **(a-b)** Training log posteriors, **(c-d)** RMSEs computed after hyperparameter selection and retraining. Numbers on the x-axis denote the number of inducing points.

each pose dimension records a measurement in different parts of the body during the action[1]. We consider four input sequences, each having 350 data points to train the model. In order to tackle the problem of dimensionality, we project the original dataset with PCA to a three dimensional latent space where the system is specified, following [4] and [15].

We also investigated whether incorporating temporal information into the drift function enhances non-parametric ordinary differential equation (ODE) model presented in [7], which is tested by discarding the diffusion component from (1) and fitting the same input sequences. Overall four models are examined, which we call in short sODE, stODE, sSDE and stSDE, lower case letters representing if the system is spatial or spatio-temporal. Each model is optimized 750 times using different lengthscales and inducing locations. In order to select the hyperparameters that prevent overfitting, we compute the root mean square error (RMSE) between 3 holdout input sequences and state paths (at observed time points) that are estimated using the EM integration from the measured initial states. The models are then retrained using 4 training and 3 holdout sequences, and finally RMSE between 3 test input sequences and the estimated state paths are reported.

Figure 1 visualizes fits for four models on the same training data. In both ODE and SDE settings, we see that inferred spatio-temporal trajectories are closer to the data than the spatial counterparts. For the same reason, stochastic state trajectory cloud needs to be more voluminous in sSDE to achieve a greater posterior whereas stSDE can better fit individual trajectories. Also, the bottom left corners in the figures show that temporal information helps capturing rapidly changing spatial fields.

Quantitative results are presented in Figure 2. Unsurprisingly, time-variant drift functions boost the training performance with more than 10 inducing points, which is compatible with the plots in Figure 1. We also observe that increasing the number of inducing points yields better training posteriors in SDE model. Figures 2c-d illustrate that time-variant drift function can also reduce prediction error, at least for golf swing trajectories that span approximately the same part of the state space. We also see that the error consistently decreases as the number of inducing points M is increased, and reaches the minimum at $M = 80$.

5 Conclusion and Future Work

We propose an approach for learning non-parametric spatio-temporal drift and diffusion functions of stochastic differential equation (SDE) systems such that the resulting simulated state distributions match data. The experiment on a real world data set shows that our model can better fit complex dynamics than the spatial counterpart. This increase in model capacity, however, results in larger data set requirements and makes the model more vulnerable to overfitting, which could be better accounted for using e.g. variational inference. An interesting future research direction is the study of various vector field kernels, such as divergence-free, curl-free or spectral kernels [12]. The model could be extended to have an observation model, e.g., GPLVM or deep neural network, rather than PCA. Including inputs or controls to the system would allow precise modelling in interactive settings, such as robotics.

[1]We use the files `64_01-64_04.amc` for training, `64_05-64_07.amc` for cross validation, and `64_08-64_10.amc` for testing

Acknowledgments

The data used in this project was obtained from `mocap.cs.cmu.edu`. The database was created with funding from NSF EIA-0196217. This work has been supported by the Academy of Finland grants no. 260403, 299915, 275537, 311584.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] M. Alvarez, L. Rosasco, and N. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 2012.

[3] P. Batz, A. Ruttor, and M. Opper. Approximate bayes learning of stochastic differential equations. *arXiv:1702.05390*, 2017.

[4] Andreas Damianou, Michalis K Titsias, and Neil D Lawrence. Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2011.

[5] R. Friedrich, J. Peinkeb, M. Sahimic, and R. Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. *Phys. reports*, 506:87–162, 2011.

[6] C. García, A. Otero, P. Felix, J. Presedo, and D. Marquez. Nonparametric estimation of stochastic differential equations with sparse Gaussian processes. *Physical Review E*, 96(2):022104, 2017.

[7] Markus Heinonen, Cagatay Yildiz, Henrik Mannerström, Jukka Intosalmi, and Harri Lähdesmäki. Learning unknown ODE models with Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1959–1968. PMLR, 10–15 Jul 2018.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2014.

[10] J. Quiñonero-Candela and C.E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[11] C.E. Rasmussen and K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[12] S. Remes, M. Heinonen, and S. Kaski. Non-stationary spectral kernels. *NIPS*, 2017.

[13] A. Ruttor, P. Batz, and M. Opper. Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems*, pages 2040–2048, 2013.

[14] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. on pattern analysis and machine intelligence*, 30:283–298, 2008.

[15] Jack Wang, Aaron Hertzmann, and David M Blei. Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448, 2006.

[16] Cagatay Yildiz, Markus Heinonen, Jukka Intosalmi, Henrik Mannerström, and Harri Lähdesmäki. Learning stochastic differential equations with gaussian processes without gradient matching. *arXiv preprint arXiv:1807.05748*, 2018.