Learning interpretable continuous-time models of latent stochastic dynamical systems

Lea Duncker [1] **Gergő Bohner** [1] **Julien Boussard** [2] **Maneesh Sahani** [1]

Abstract

We develop an approach to learn an interpretable semi-parametric model of a latent continuous-time stochastic dynamical system, assuming noisy high-dimensional outputs sampled at uneven times. The dynamics are described by a nonlinear stochastic differential equation (SDE) driven by a Wiener process, with a drift evolution function drawn from a Gaussian process (GP) conditioned on a set of learnt fixed points and corresponding local Jacobian matrices. This form yields a flexible nonparametric model of the dynamics, with a representation corresponding directly to the interpretable portraits routinely employed in the study of nonlinear dynamical systems. The learning algorithm combines inference of continuous latent paths underlying observed data with a sparse variational description of the dynamical process. We demonstrate our approach on simulated data from different nonlinear dynamical systems.

1. Introduction

Many dynamical systems with intrinsic noise may be modelled in continuous time using the framework of stochastic differential equations (SDE). However identifying a good SDE model from intermittent observations of the process is challenging, particularly if the dynamical process is nonlinear and the observations are indirect and noisy. A common response is to assume a latent process that operates in discretised time, often called a state-space model. This approach has been applied in contexts ranging from modelling human motion (Wang et al., 2006) to solving control problems (Eleftheriadis et al., 2017). However, it assumes that observations, and the critical phenomena of the dynamics, can be accurately modelled using a discrete time grid.

A further challenge when the goal is to gain insight into a physical or biological system whose parametric description is unknown, is to obtain an *interpretable* model of the dynamics from observed data, whether modelled in discrete or continuous time. State-space models that rely on nonparametric or flexibly parametrised descriptions of dynamics, for example using Gaussian process (GP) priors or recurrent neural networks (RNN), may be effective at prediction but inevitably leave interpretation to a second analytic stage, posing its own challenges.

In this paper, we consider continuous-time latent SDE models of the form

$$\begin{aligned} d\boldsymbol{x} &= \boldsymbol{f}(\boldsymbol{x})dt + \sqrt{\boldsymbol{\Sigma}}\, d\boldsymbol{w} \\ \mathbb{E}_{y|x}[\boldsymbol{y}(t_i)] &= g(\boldsymbol{C}\boldsymbol{x}(t_i) + \boldsymbol{d})\,, \qquad\qquad i = 1, \dots, T\,, \end{aligned} \tag{1}$$

where the temporal evolution of a latent variable $\boldsymbol{x} \in \mathbb{R}^K$ is described by a nonlinear SDE with dynamical evolution function $\boldsymbol{f} : \mathbb{R}^K \mapsto \mathbb{R}^K$ and incremental noise covariance $\boldsymbol{\Sigma}$ shaping the Wiener noise process $\boldsymbol{w}(t)$. Note that the nonlinear SDE induces a non-Gaussian prior on $\boldsymbol{x}(t)$ with no easy access to finite marginal distributions. The latent state is observed indirectly through noisy measurements $\boldsymbol{y}_i \in \mathbb{R}^N$ at unevenly spaced time points t_i. The measurements are distributed with a known parametric form and generalized linear dependence; that is the expected value is $g(\boldsymbol{C}\boldsymbol{x} + \boldsymbol{d})$ with inverse-link function g and parameters $\boldsymbol{C} \in \mathbb{R}^{N \times K}$ and $\boldsymbol{d} \in \mathbb{R}^N$. We seek to infer latent paths $\boldsymbol{x}(t)$ along with the dynamical parameters and an interpretable representation of the dynamical mapping $\boldsymbol{f}$.

What do we mean by interpretable? The properties of dynamical systems are frequently analyzed by characterizing dynamical fixed points and local behaviour near these points (Sussillo & Barak, 2013). When $\boldsymbol{f}$ is a learnt, general function, fixed points must be found numerically (Golub & Sussillo, 2018). This makes it difficult to propagate uncertainty about $\boldsymbol{f}$ to the number and location of fixed points, and to the local dynamics around them. Our approach is to develop a non-parametric Gaussian-process model for $\boldsymbol{f}$ *conditioned on* the learnt locations of fixed points and associated local Jacobians. Thus, we implicitly integrate out the details of

[1]Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom [2]Stanford University, Palo Alto, California, USA. Correspondence to: Lea Duncker <duncker@gatsby.ucl.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

$\boldsymbol{f}$, while optimising directly over the components of the intepretable dynamical portrait.

The paper is organised as follows: In section 2 we review background material on the related Gaussian Process State-Space Model (GP-SSM) and previous work on GP approximations to SDEs (Archambeau et al., 2007; 2008). We also briefly review the inducing point approach for GP models. In section 3 we make use of GP priors to represent the unknown nonlinear dynamics $\boldsymbol{f}$, incorporating interpretable structure by conditioning the GP on fixed points and local Jacobian matrices of the system. We derive a Variational Bayes algorithm for approximate inference and parameter learning in section 4. Finally, we demonstrate the performance of our algorithm on a number of nonlinear dynamical system examples in section 5.

2. Background

2.1. Gaussian Process State-Space-Model

A discrete-time analogue of the model in (1) is the GP-SSM, where the latent state evolution over a fixed step size is modelled as

$$\boldsymbol{x}_{\ell+1} = \boldsymbol{f}(\boldsymbol{x}_\ell) + \boldsymbol{\epsilon}_\ell \tag{2}$$

where $\boldsymbol{\epsilon}_\ell \sim \mathcal{N}(\boldsymbol{\epsilon}_\ell|0, D)$. There have been a range of approaches for performing approximate inference in this model, based on Assumed Density Filtering (Deisenroth et al., 2009; Ramakrishnan et al., 2011), Expectation Propagation (Deisenroth & Mohamed, 2012), variational inference (Frigola et al., 2014), or recurrent recognition networks (Eleftheriadis et al., 2017). The model in (1) requires a different treatment for latent path inference, as it maintains the continuous-time structure of the system of interest.

2.2. Gaussian Process Approximation to SDEs

The problem of performing approximate inference in continuous-time SDE models has been considered previously, with the two main approaches being Expectation Propagation (Cseke et al., 2016) and variational inference (Archambeau et al., 2007; 2008). We review the latter approach in this section, as our Variational Bayes algorithm in section 4 extends this work.

Archambeau et al. (2007; 2008) consider the model in (1) under linear Gaussian observations. The authors derive an approximate inference algorithm based on a variational Gaussian approximation to the posterior process on $\boldsymbol{x}(t)$ under the constraint that the approximate process has Markov structure, as is the case for the true posterior process. The most general way to construct such an approximation is via a linear time-varying SDE of the form

$$d\boldsymbol{x} = (-\boldsymbol{A}(t)\boldsymbol{x}(t) + \boldsymbol{b}(t))\,dt + \sqrt{\boldsymbol{\Sigma}}\,d\boldsymbol{w} \tag{3}$$

The instantaneous marginal distributions of this approximation at any time t are Gaussian, with means $\boldsymbol{m}_x(t)$ and covariances $\boldsymbol{S}_x(t)$ that evolve in time according to the ordinary differential equations (ODEs):

$$\begin{aligned}\frac{d\boldsymbol{m}_x}{dt} &= -\boldsymbol{A}(t)\boldsymbol{m}_x + \boldsymbol{b}(t)\\ \frac{d\boldsymbol{S}_x}{dt} &= -\boldsymbol{A}(t)\boldsymbol{S}_x - \boldsymbol{S}_x\boldsymbol{A}(t)^\mathsf{T} + \boldsymbol{\Sigma}\end{aligned} \tag{4}$$

Archambeau et al. (2007; 2008) derive a lower bound to the marginal log-likelihood – often called the variational free energy or evidence lower bound – whose maximisation with respect to q_x is equivalent to minimising the Kullback-Leibler (KL) divergence between the approximate and true posterior process. The free energy has the form

$$\mathcal{F} = \sum_i \langle \log p(\boldsymbol{y}_i|\boldsymbol{x}_i)\rangle_{q_x} - \mathsf{KL}[q_x(\boldsymbol{x})\|p(\boldsymbol{x})] \tag{5}$$

The first term is the expected log-likelihood under the approximation and only depends on the marginal distributions $q_x(\boldsymbol{x}(t))$. The second term is the KL-divergence between the continuous-time approximate posterior process and the prior process. Archambeau et al. (2007) show that this term can be written as

$$\mathsf{KL}[q_x(\boldsymbol{x})\|p(\boldsymbol{x})] = \int_\mathcal{T} dt\,\left\langle(\boldsymbol{f}-\boldsymbol{f}_q)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{f}_q)\right\rangle_q \tag{6}$$

where both $\boldsymbol{f}$ and $\boldsymbol{f}_q$ are evaluated at $\boldsymbol{x}(t)$, and $\boldsymbol{f}_q(\boldsymbol{x}(t)) = -\boldsymbol{A}(t)\boldsymbol{x}(t) + \boldsymbol{b}(t)$. Note that the noise covariance $\boldsymbol{\Sigma}$ is deliberately chosen to be equal for the SDEs in q_x and p, as this term would diverge otherwise.

To maximise $\mathcal{F}$ with respect to $\boldsymbol{m}_x(t)$ and $\boldsymbol{S}_x(t)$, subject to the constraint that the approximate posterior process has Markov structure according to equation (3), one can find the stationary points of the Lagrangian

$$\mathcal{L} = \mathcal{F} - \mathcal{C}_1 - \mathcal{C}_2 \tag{7}$$

with

$$\begin{aligned}\mathcal{C}_1 &= \int_\mathcal{T} dt\ \mathsf{Tr}\left[\Psi\left(\frac{d\boldsymbol{S}_x}{dt} + \boldsymbol{A}\boldsymbol{S}_x + \boldsymbol{S}_x\boldsymbol{A}^\mathsf{T} - \boldsymbol{\Sigma}\right)\right]\\ \mathcal{C}_2 &= \int_\mathcal{T} dt\ \boldsymbol{\lambda}^\mathsf{T}\left(\frac{d\boldsymbol{m}_x}{dt} + \boldsymbol{A}\boldsymbol{m}_x - \boldsymbol{b}\right)\end{aligned} \tag{8}$$

where Ψ and $\boldsymbol{\lambda}$ are Lagrange multipliers. Archambeau et al. (2007; 2008) derive a smoothing algorithm that involves iterating fixed point updates of this Lagrangian. These are either closed form, or require solving ODEs forward and backward in time, thus achieving linear time complexity. In section 4, we will modify this original algorithm in order to improve its numerical stability, and show how to incorporate it in an efficient Variational Bayes algorithm.

2.3. Sparse Gaussian Processes using inducing points

In later sections of the paper, we will make use of the sparse variational inducing point approach of Titsias (2009). The key idea of inducing point approaches is to condition a GP $\zeta(\boldsymbol{x}) \sim \mathcal{GP}(0, \kappa(\boldsymbol{x}, \boldsymbol{x}'))$ on so-called inducing variables $\boldsymbol{u} \in \mathbb{R}^M$. These can be thought of as pseudo-observations of the function at M locations $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M] \in \mathbb{R}^{K \times M}$. An augmented prior for the GP and inducing variables can be written as

$$\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{u}|0, \boldsymbol{K}_{zz}\right), \; \zeta|\boldsymbol{u} \sim \mathcal{GP}(\mu_{\zeta|u}(\boldsymbol{x}), \nu_{\zeta|u}(\boldsymbol{x}, \boldsymbol{x}') \tag{9}$$

The conditioned GP mean and covariance function are

$$\begin{aligned} \mu_{\zeta|u}(\boldsymbol{x}) &= \boldsymbol{K}_{\cdot z}(\boldsymbol{x})\boldsymbol{K}_{zz}^{-1}\boldsymbol{u} \\ \nu_{\zeta|u}(\boldsymbol{x}, \boldsymbol{x}') &= \kappa(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{K}_{\cdot z}(\boldsymbol{x})\boldsymbol{K}_{zz}^{-1}\boldsymbol{K}_{z\cdot}(\boldsymbol{x}) \end{aligned} \tag{10}$$

Where $[\boldsymbol{K}_{zz}]_{ij} = \kappa(\boldsymbol{z}_i, \boldsymbol{z}_j)$, and $[\boldsymbol{K}_{\cdot z}(\boldsymbol{x})]_i = \kappa(\boldsymbol{x}, \boldsymbol{z}_i)$. The computational complexity of building the mean and covariance in (10) is linear in the number of $\boldsymbol{x}$ input points and cubic only in the number of inducing points M. If we were to integrate over the inducing variables in this augmented prior, we would recover the original model. However, the inducing variables can also be kept in the model as auxiliary variables, which may be incorporated into approaches for variational inference (Titsias, 2009).

3. Interpretable priors on nonlinear dynamics

Similarly to the GP-SSM work, we wish to model $\boldsymbol{f}$ using the framework of GPs. GPs can represent a flexible class of nonlinear dynamics. However, it may be difficult to interpret the inferred function with respect to studying the underlying dynamical system that generated the observed data. As stated above, standard analysis approaches for nonlinear dynamical systems rely on identifying local fixed points $\boldsymbol{s}_i$, where $\boldsymbol{f}(\boldsymbol{s}_i) = \boldsymbol{0}$, and the locally-linearised dynamics around them, given by the Jacobians $\boldsymbol{\nabla}_x \boldsymbol{f}(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{s}_i}$ (Sussillo & Barak, 2013). This strategy motivates our approach to interpretability.

3.1. A Gaussian Process prior for dynamics

In order to arrive at a modelling framework that makes fixed points and Jacobian matrices readily available for analysis, we introduce a GP prior conditioned directly on these parameters (see also Bohner & Sahani, 2018). The fixed point locations and Jacobians around them can be viewed as further hyperparameters specifying the prior mean and covariance function of the GP, which we will denote by $\boldsymbol{\theta} = \{\boldsymbol{f}_s^{(i)}, \boldsymbol{J}_s^{(i)}\}_{i=1}^L$, where L denotes the total number of fixed point locations. With $\boldsymbol{f}_s^{(i)} = \boldsymbol{f}(\boldsymbol{s}_i) = \boldsymbol{0}$ and $[\boldsymbol{J}_s^{(i)}]_{k,m} = \frac{\partial f_k(\boldsymbol{x})}{\partial x_m}|_{\boldsymbol{x}=\boldsymbol{s}_i}$. We can hence write a GP prior conditioned on the fixed points and Jacobians for each dimension in $\boldsymbol{f}$, using the fact that a GP and its derivative process are still jointly distributed as a GP.

The Variational Bayes approach in section 4 will make use of a sparse variational approximation for $\boldsymbol{f}$ using inducing variables, as in Titsias (2009). To make later notation more compact, we therefore directly introduce the augmented model including inducing variables drawn from the conditioned GP prior here. We denote the joint covariance matrix between inducing variables, fixed points and Jacobians as

$$\mathsf{K}_{zz}^{\theta} = \begin{bmatrix} \boldsymbol{K}_{zz} & \boldsymbol{K}_{zs} & \boldsymbol{K}_{zs}^{\nabla_2} \\ \boldsymbol{K}_{sz} & \boldsymbol{K}_{ss} & \boldsymbol{K}_{ss}^{\nabla_2} \\ \boldsymbol{K}_{sz}^{\nabla_1} & \boldsymbol{K}_{ss}^{\nabla_1} & \boldsymbol{K}_{ss}^{\nabla_1\nabla_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}_{zz} & \tilde{\boldsymbol{K}}_{zs} \\ \tilde{\boldsymbol{K}}_{sz} & \tilde{\boldsymbol{K}}_{ss} \end{bmatrix} \tag{11}$$

where the superscript ∇_i denotes the derivative of the covariance function with respect to its ith input argument such that $[\boldsymbol{K}_{zs}^{\nabla_2}]_{ij} = \frac{\partial}{\partial \boldsymbol{s}}\kappa(\boldsymbol{z}_i, \boldsymbol{s})|_{\boldsymbol{s}=\boldsymbol{s}_j}$. The conditional prior on the inducing variables given $\boldsymbol{\theta}$ can then be written as

$$\boldsymbol{u}_k|\boldsymbol{\theta} = \mathcal{N}\left(\boldsymbol{u}\middle|\, \tilde{\boldsymbol{K}}_{zs}\tilde{\boldsymbol{K}}_{ss}^{-1}\boldsymbol{v}_k^{\theta}, \boldsymbol{K}_{zz} - \tilde{\boldsymbol{K}}_{zs}\tilde{\boldsymbol{K}}_{ss}^{-1}\tilde{\boldsymbol{K}}_{sz}\right) \tag{12}$$

where $\boldsymbol{v}_k^{\theta} = [f_{s,k}^{(1)}, \ldots, f_{s,k}^{(L)}, \boldsymbol{J}_{k,:}^{(1)}, \ldots, \boldsymbol{J}_{k,:}^{(L)}]^{\mathsf{T}}$ collects the fixed-point and derivative observations relating to f_k. Finally, for the conditional prior on f_k, given the inducing variables and $\boldsymbol{\theta}$, we have

$$f_k|\boldsymbol{u}_k, \boldsymbol{\theta} \sim \mathcal{GP}\left(\mu_{f|u}^{\theta}(\boldsymbol{x}), \nu_{f|u}^{\theta}(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{13}$$

with

$$\begin{aligned} \mu_{f|u}^{\theta}(\boldsymbol{x}) &= \boldsymbol{a}_z^{\theta}(\boldsymbol{x}) \begin{bmatrix} \boldsymbol{u}_k \\ \boldsymbol{v}_k^{\theta} \end{bmatrix} \\ \nu_{f|u}^{\theta}(\boldsymbol{x}, \boldsymbol{x}') &= \kappa(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{a}_z^{\theta}(\boldsymbol{x})\mathsf{K}_{zz}^{\theta}\boldsymbol{a}_z^{\theta}(\boldsymbol{x})^{\mathsf{T}} \end{aligned} \tag{14}$$

$$\boldsymbol{a}_z^{\theta}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{K}_{\cdot z}(\boldsymbol{x}) & \boldsymbol{K}_{\cdot s}(\boldsymbol{x}) & \boldsymbol{K}_{\cdot s}^{\nabla_2}(\boldsymbol{x}) \end{bmatrix} {\mathsf{K}_{zz}^{\theta}}^{-1} \tag{15}$$

3.2. Automatic selection of the number of fixed points

When the generative SDE dynamics are unknown, so are the number of fixed points in the system. We therefore take the general approach of introducing more fixed points than expected, and 'pruning' by hyperparameter optimisation. In particular, we include noise variance parameters for each fixed point, representing uncertainty about the zero-value of the function at the fixed point location. We hence have

$$\boldsymbol{f}_i^s = \boldsymbol{f}(\boldsymbol{s}_i) + \alpha_i\boldsymbol{\epsilon} = \boldsymbol{0} + \alpha_i\boldsymbol{\epsilon} \tag{16}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. The variance parameters α_i will enter our model simply via an added diagonal matrix to the $\boldsymbol{K}_{ss}$ block in (11). When the α_i are optimised, the uncertainty for superfluous fixed points will grow, while that of the fixed points the system is actually using will shrink. When the uncertainty for a fixed point is large, conditioning on it in the GP prior for $\boldsymbol{f}$ will have negligible effect on prediction.

4. Variational inference and learning

We can derive an efficient Variational Bayes (VB) algorithm (Attias, 2000) for variational inference and learning in the model in (1) by maximising a variational free energy. We assume that our full variational distribution factorises as

$$q(\boldsymbol{x}, \boldsymbol{f}, \boldsymbol{u}) = q_x(\boldsymbol{x}) q_{f,u}(\boldsymbol{f}, \boldsymbol{u}) \tag{17}$$

Following Titsias (2009), we choose $q_{f,u}(\boldsymbol{f}, \boldsymbol{u}) = \prod_{k=1}^{K} p(f_k|\boldsymbol{u}_k, \boldsymbol{\theta}) q_u(\boldsymbol{u}_k)$. The variational approximation of the posterior over the inducing variables are chosen to be of the form $q_u(\boldsymbol{u}_k) = \mathcal{N}\left(\boldsymbol{u}_k|\boldsymbol{m}_u^k, \boldsymbol{S}_u^k\right)$. The marginal variational distribution $q_f(\boldsymbol{f}) = \prod_k \int d\boldsymbol{u}_k p(f_k|\boldsymbol{u}_k, \boldsymbol{\theta}) q_u(\boldsymbol{u}_k)$ is also a GP. The resulting expression for the variational free energy is of the form:

$$\mathcal{F}^* = \langle \mathcal{F} \rangle_{q_f} - \sum_{k=1}^{K} \mathsf{KL}[q_u(\boldsymbol{u}_k) \| p(\boldsymbol{u}_k|\boldsymbol{\theta})] \tag{18}$$

The VB algorithm iterates over an inference step, where the distribution q_x over the latent path is updated, a learning step, where $q_{f,u}$ and the parameters in the output mapping are updated, and a hyperparameter learning step, where the kernel hyperparameters and fixed point locations are updated.

4.1. Inference

Our inference approach extends the work of Archambeau et al. (2007; 2008) to a wider class of observation models and to a nonparametric Bayesian treatment of the dynamics $\boldsymbol{f}$ under the conditioned sparse GP prior introduced in section 3.

After using integration by parts on the Lagrangian in (7) (exchanging $\mathcal{F}$ for $\mathcal{F}^*$), we take variational derivatives with respect to $\boldsymbol{m}_x(t)$ and $\boldsymbol{S}_x(t)$. Since our model has a rotational non-identifiability with respect to the latents $\boldsymbol{x}$, we fix $\boldsymbol{\Sigma} = I$ without loss of generality. We arrive at the following set of fixed point equations:

$$\frac{d\boldsymbol{\Psi}}{dt} = \boldsymbol{A}(t)^{\mathsf{T}} \boldsymbol{\Psi}(t) + \boldsymbol{\Psi}(t) \boldsymbol{A}(t) - \frac{\partial \mathcal{F}^*}{\partial \boldsymbol{S}_x} \odot \mathbb{P} \tag{19}$$

$$\frac{d\boldsymbol{\lambda}}{dt} = \boldsymbol{A}(t)^{\mathsf{T}} \boldsymbol{\lambda}(t) - \frac{\partial \mathcal{F}^*}{\partial \boldsymbol{m}_x} \tag{20}$$

$$\boldsymbol{A}(t) = \left\langle \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \right\rangle_{q_x q_f} + 2\boldsymbol{\Psi}(t) \tag{21}$$

$$\boldsymbol{b}(t) = \langle \boldsymbol{f}(\boldsymbol{x}) \rangle_{q_x q_f} + \boldsymbol{A}(t) \boldsymbol{m}_x(t) - \boldsymbol{\lambda}(t) \tag{22}$$

with $\mathbb{P}_{ij} = \frac{1}{2}$ for $i \neq j$ and 1 otherwise and $\odot$ denotes the Hadamard product. In contrast to previous work, we explicitly take the symmetric variations of $\boldsymbol{S}_x(t)$ into account, which leads to slightly modified equations in (19) compared to the work in Archambeau et al. (2007; 2008), and thus to improved numerical stability in practice. As a result, we can work with the fixed point updates (21) and (22) directly, without introducing a learning rate parameter that blends the updates with the previous value of the variational parameters $\boldsymbol{A}$ and $\boldsymbol{b}$, as was done by Archambeau et al. (2007; 2008).

The inference algorithm involves solving the set of coupled ODEs in (4) and (19)-(22) using the conditions $\boldsymbol{m}_x(0) = \boldsymbol{m}_{x,0}, \boldsymbol{S}_x(0) = \boldsymbol{S}_{x,0}$ and $\boldsymbol{\lambda}(T) = 0, \boldsymbol{\Psi}(T) = 0$. In principle, it is possible to use any ODE solver to do this. In this work, we choose to solve (4) using the forward Euler method with fixed step size Δt to obtain $\boldsymbol{m}_x$ and $\boldsymbol{S}_x$ evaluated on an evenly spaced grid. Similarly, we then solve (19) and (20) backwards in time to obtain evaluations of $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$. The solutions from the ODEs can then be used with equations (21) and (22) to obtain evaluations of $\boldsymbol{A}$ and $\boldsymbol{b}$ on the same time-grid used for solving the ODEs.

Evaluating the expectations of the terms involving $\boldsymbol{f}$ with respect to q_x and q_f only involves computing Gaussian expectations of covariance functions and their derivatives. These can be computed analytically for choices such as an exponentiated quadratic covariance function. We update the initial state values $\boldsymbol{m}_{x,0}$ and $\boldsymbol{S}_{x,0}$ using the same procedure as that described by Archambeau et al. (2008). Given the function evaluations on the inference time-grid, we use linear interpolation to obtain function evaluations of $\boldsymbol{m}_x$ and $\boldsymbol{S}_x$ at arbitrary time points. Further details on the inference algorithm are given in the supplementary material.

4.2. Learning

4.2.1. Dynamics

The only terms in (18) that depend on parameters in $\boldsymbol{f}$ are the expected KL-divergence between the prior and approximate posterior processes and the KL-divergence relating to the inducing variables for $\boldsymbol{f}$, which are jointly quadratic in the inducing variables and Jacobians. Thus, given $\boldsymbol{m}_x(t)$, $\boldsymbol{S}_x(t)$, $\boldsymbol{A}(t)$ and $\boldsymbol{b}(t)$, we can find closed form updates for the Jacobians and variational parameters relating to $\boldsymbol{f}$. For $\boldsymbol{S}_u^k$ the update is of the form

$$\boldsymbol{S}_u^k = \left({\boldsymbol{\Omega}_u}^{-1} + \int_{\mathcal{T}} dt [\langle \boldsymbol{a}_z^{\theta}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{a}_z^{\theta}(\boldsymbol{x}) \rangle_{q_x}]_{[u,u]} \right)^{-1} \tag{23}$$

with $\boldsymbol{\Omega}_u = \boldsymbol{K}_{zz} - \tilde{\boldsymbol{K}}_{zs} \tilde{\boldsymbol{K}}_{ss}^{-1} \tilde{\boldsymbol{K}}_{sz}$ and where the operation $[X]_{[u,u]}$ selects the first $M \times M$ block of X. The inducing variable means and Jacobians around the fixed-point locations can be updated jointly as

$$\begin{bmatrix} \boldsymbol{m}_u^1 & \dots & \boldsymbol{m}_u^K \\ \boldsymbol{J}_1 & \dots & \boldsymbol{J}_K \end{bmatrix} = \boldsymbol{B}_1^{-1} (\boldsymbol{B}_2 - \boldsymbol{B}_3) \tag{24}$$

with

$$\boldsymbol{B}_1 = \left(\tilde{\boldsymbol{\Omega}} + \int_{\mathcal{T}} dt \left[\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})^\top \boldsymbol{a}_z^\theta(\boldsymbol{x})\rangle_{q_x}\right]_{[uj,uj]}\right)$$
$$\boldsymbol{B}_2 = \int_{\mathcal{T}} dt \left[\left\langle \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,uj]}^\top \langle \boldsymbol{f}_q\rangle_{q_x}^\top$$
$$\boldsymbol{B}_3 = \int_{\mathcal{T}} dt \left[\left\langle \nabla_x \boldsymbol{a}_z^\theta(\boldsymbol{x})\right\rangle_{q_x}\right]_{[:,uj]}^\top \boldsymbol{S}_x \boldsymbol{A}^\top$$
$$\tilde{\boldsymbol{\Omega}} = \begin{bmatrix} \boldsymbol{\Omega}_u{}^{-1} & -\boldsymbol{\Omega}_u{}^{-1}\boldsymbol{G} \\ -\boldsymbol{G}^\top \boldsymbol{\Omega}_u{}^{-1} & \boldsymbol{G}^\top \boldsymbol{\Omega}_u{}^{-1}\boldsymbol{G} \end{bmatrix}, \boldsymbol{G} = \left[\tilde{\boldsymbol{K}}_{zs}\tilde{\boldsymbol{K}}_{ss}^{-1}\right]_{[j,j]}$$

where $[X]_{[uj,uj]}$ selects the first $M \times M$ and last $LK \times LK$ block of X, $[X]_{[:,uj]}$ selects the first M and last LK columns of X, and $[X]_{[j,j]}$ selects the last $LK \times LK$ block of X. The one-dimensional integrals can be computed efficiently using, for instance, Gauss-Legendre quadrature. Detailed derivations are given in the supplementary material, where we also provide closed form updates for the sparse variational GP approach for modelling $\boldsymbol{f}$ without further conditioning on fixed points and Jacobians.

4.2.2. Output Mapping and Hyperparameters

The only term that depends on the parameters $\boldsymbol{C}$ and $\boldsymbol{d}$ in (18) is the expected log-likelihood. Depending on the choice of likelihood, the optimal update may be available in closed form. Otherwise, parameter updates may be found by direct optimisation of the variational free energy. Similarly, the covariance function hyperparameters and fixed point locations are learnt by maximising the variational free energy. The inducing point locations can also be included here, though we chose to hold them fixed on a grid for all examples shown in this paper.

4.3. Computational Complexity

The main costs of the algorithm come from evaluating GP predictions at a set of input points, and solving the ODEs (19)-(20). Computing the GP predictions using the basic sparse inducing-point approach scales cubically in the number of inducing points, the number of fixed points and the number of entries in the Jacobians, but scales linearly in the number of input points. Solving the ODEs using simple forward Euler integration achieves linear time complexity. In principle, adaptive ODE solvers could achieve a lower cost. Similarly, recent advances in scalable sparse Gaussian Process methods could improve on the cubic dependence on the inducing points. Hence, the cubic cost relating to the candidate number of fixed points could be viewed as the intrinsic cost of our description of the dynamics, while the costs relating to integrating ODEs and computing sparse Gaussian Process predictions could be improved.

5. Experiments

In this section, we apply our algorithm to data generated from different nonlinear dynamical systems. In all experiments, we choose an exponentiated quadratic covariance function in the prior over the dynamics $\boldsymbol{f}$ and initialise the inducing point means and Jacobian matrices at zero. Each fixed point observation's uncertainty is initialised with a standard deviation of 0.1. We generate $\boldsymbol{C}$ and $\boldsymbol{d}$ by drawing their entries from Gaussian distributions unless otherwise stated, and initialise our algorithm at these parameter values. For inference, we solve the ODEs (19)-(22) using the forward Euler method with $\Delta t = 1$ms. Unless stated otherwise, the link function is the indentity $g(z) = z$.

5.1. Double-well dynamics

We first demonstrate our method on the classic one-dimensional double-well example, where the latent SDE evolves with drift $f(x) = 4x(1 - x^2)$. We simulate data on 20 trials with multivariate Gaussian outputs of dimensionality $N = 15$ with unknown variances 0.25, and observe the output process at 20 randomly sampled time-points per trial. We chose 8 evenly spaced inducing points in $(-3, 3)$ for f. While the true dynamics have three fixed points, we condition the prior on f on four fixed points and use the method outlined in section 3.2 to automatically select the correct number. The results are summarised in Figure 1, demonstrating that our algorithm can successfully perform inference and interpretable learning of the SDE path and dynamics, respectively, and does not move away from the good initial location for the model parameters $\boldsymbol{C}$ and $\boldsymbol{d}$.

5.2. Van der Pol's oscillator

Our next example examines a two-dimensional system where the dynamics contain a limit cycle around an unstable fixed point. The dynamics are given by

$$f_1(\boldsymbol{x}) = \rho\tau\left(x_1 - \frac{1}{3}x_1^3 - x_2\right), \quad f_2(\boldsymbol{x}) = \frac{\tau}{\rho}x_1 \qquad (25)$$

with a time constant τ. We generate data from (1) using these dynamics with $\rho = 2, \tau = 15$, $N = 20$ output dimensions and Gaussian measurement noise with unknown variances 2.25 on 20 repeated trials. We use 5×5 inducing points evenly spaced in $(-2, 2)$. The results are summarised in Figure 2, demonstrating that our description of the dynamics successfully captures the limit cycle of the generative dynamics.

5.3. Neural population dynamics

This example demonstrates our algorithm under multivariate point-process observations. We model the intensity functions of the nth output process as $\eta_n(t) =$

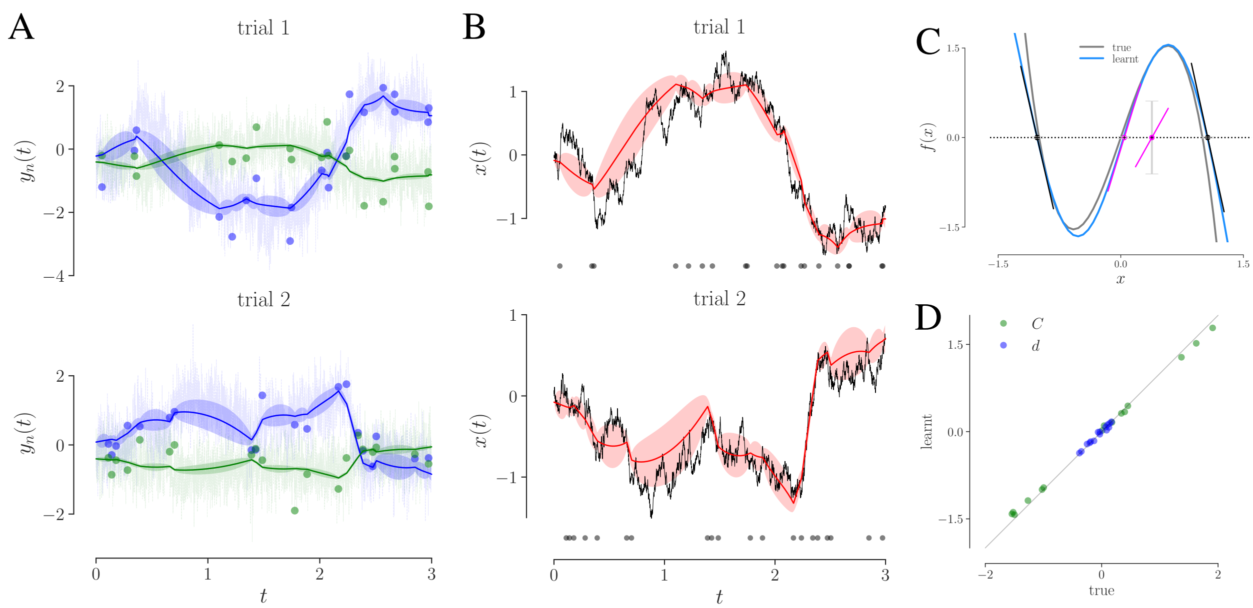

Figure 1. Double-well dynamics. A: Two example dimensions of the output process on two different trials. The dots represent the observed data-points of the noisy output processes plotted in faint lines. The solid blue/green traces are the inferred posterior means with ± 1 posterior standard deviation tubes around them. B: True and inferred latent SDE trajectory for the same example trials as in A. The red traces represent the posterior means with ± 1 posterior standard deviation tubes around them, black traces show the true latent SDE path. The black dots indicate the times when observations of $\boldsymbol{y}$ were made. C: True and learnt dynamics together with the learnt fixed-point locations and tangent lines. Stable fixed points are shown in black, unstable ones in magenta. The uncertainty about the fixed point observation is illustrated using grey error bars representing ± 1 standard deviation. Only the additional fourth fixed point is associated with high uncertainty. D: True vs. learnt model parameters $\boldsymbol{C}$ and $\boldsymbol{d}$.

$\exp(\sum_{k=1}^{K} C_{nk} x_k(t) + d_n)$. Conditioned on the intensity function, the $\phi(n)$ observed event-times $\boldsymbol{t}^{(n)}$ are generated by a Poisson process with log-likelihood

$$\log p(\boldsymbol{t}^{(n)}|\eta_n) = -\int_{\mathcal{T}} \eta_n(t)dt + \sum_{i=1}^{\phi(n)} \log \eta_n(t_i^{(n)}) \quad (26)$$

In contrast to the Gaussian observation case, the first term in the log-likelihood above is continuous in $\eta_n(t)$ and the absence of events is also informative towards the underlying intensity of the process.

An interesting application for this setting lies in neural data analysis, where data may be available as a set of spike times of a population of simultaneously recorded neurons jointly embedded in a circuit involved in performing a computation. In fact, studying neural population activity as a dynamical system has gained increasing traction in the field of neuroscience in recent years (Macke et al., 2011; Shenoy et al., 2013; Pandarinath et al., 2018), and data analysis methods that can obtain such descriptions are thus of great interest.

We simulate a two-dimensional latent SDE using the dynamics $f_k(\boldsymbol{x}) = -x_k + \sigma_k(w_{k1}x_1 - w_{k2}x_2 - z_k)$ for $k = 1, 2$, where $\sigma_k(x) = (1 + \exp(-b_k x))^{-1}$. Depending on the choice of parameters b_k, w_{kj} and z_k the dynamical system will exhibit different properties. We explore the two regimes where the system either has two stable and one unstable fixed points (Figure 3C left) or exhibits a single stable spiral (Figure 3C right).

We simulate data from 50 neurons on 25 trials for each of the two parameter regimes for b_k, w_{kj} and z_k. Figure 3A shows example neural spike trains under the two regimes. Figure 3B illustrates sample paths through the latent space under the different dynamical regimes, together with the density of latent locations visited across all trials. In both settings, we initialise our algorithm with three fixed points and inducing points placed on an evenly spaced 4×4 grid in $(-0.25, 1.25)$, and hold the parameters relating to the output mapping constant. Figure 3D shows the estimated flow fields in both settings, together with the location of the fixed points and their stability as indicated by the eigenvalues of the Jacobian matrices. In both settings, our method successfully recovers the main qualitative distinguishing features of the dynamics. In the regime where the dynamics are conditioned on three fixed points but the generative system only contains one, the two additional fixed points will either be associated with higher uncertainty or move to regions where no or little data was observed, as indicated by the superimposed density plots.

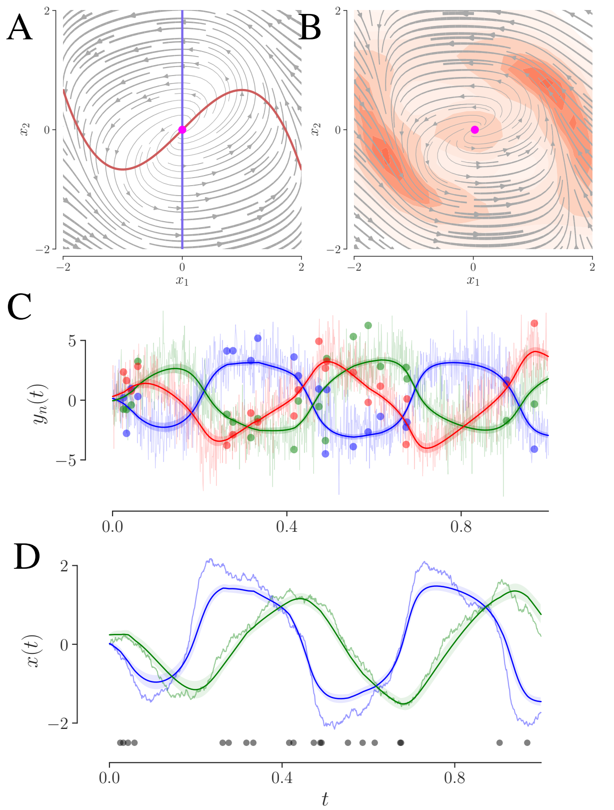

Figure 2. Van der Pol's oscillator. A: Streamline plot of the true dynamics together with nullclines and the unstable fixed point. B: Density plot of the locations visited by the latents across all trials used for learning in red, and streamline plot of the learnt dynamics with the location of the learnt fixed point. The eigenvalues of the learnt Jacobian matrix indicate that the fixed point is unstable. C: Three example dimensions of the output process. The dots represent the observed data-points of the noisy output process. The solid traces show the the posterior means with $\pm$ 1 standard deviation tubes around them. D: The true latent SDE path together with the posterior mean $\pm$ 1 posterior standard deviation of each latent dimension. Black dots represent the locations where the 20 measurements of the output process were made.

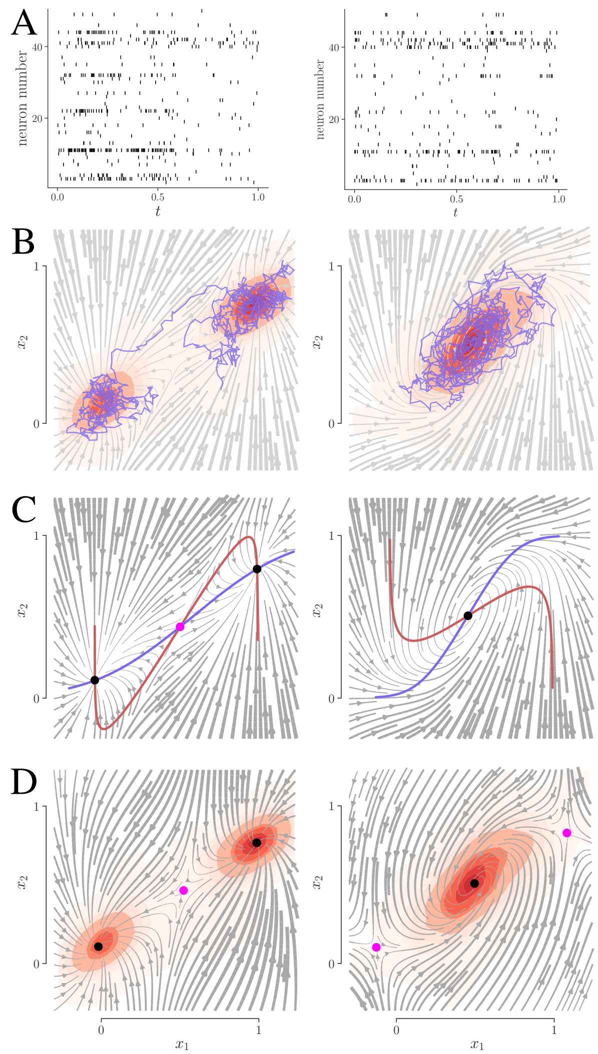

Figure 3. Neural population dynamics. Left: simulations with parameter settings $b_1 = 1.9, b_2 = 0.5, z_1 = 3, z_2 = 3.9, w_{11} = 10, w_{12} = 5, w_{21} = 9, w_{22} = 3$. Right: simulations with parameter settings $b_1 = 0.4, b_2 = 0.6, z_1 = 1.7, z_2 = 7, w_{11} = 20, w_{12} = 16, w_{21} = 21, w_{22} = 6$. A: Raster plot of the observed spike times for a population of 50 neurons for an example trial. B: Example paths through the two-dimensional latent space on the same trial as A, together with a density plot of latent locations visited across all trials that were used for learning the dynamics, shown in red. C: Streamline plots of the true dynamics together with their fixed points and nullclines for each latent dimension. Stable fixed points are black, unstable ones are magenta. D: Same density plots as in B together with streamline plots of the learnt dynamics and learnt fixed points. The fixed point stability is shown as indicated by the eigenvalues of the learnt Jacobian matrices.

5.4. Multistable chemical reaction dynamics

This example is based on the dynamical system in Ganapathisubramanian (1991), which describes nonlinear dynamics of two species of iodione in the iodate-AS(III) system under imperfect mixing by coupled first-order ODEs. We use these ODEs to describe $\boldsymbol{f}$ and generate data according to (1) with high-dimensional Gaussian observations representing spectroscopic measurements, which can approximately be described as a linear mapping from concentrations based on the I^- and IO_3^- absorption spectra provided in Kireev & Shnyrev (2015). We simulate data on 20 trials with different initial conditions, collecting 50 unevenly spaced samples from 13 spectroscopy measurements on each trial. Figure 4 shows observed data and the latent SDE path on an example trial, as well as the true and estimated dynamical portraits.

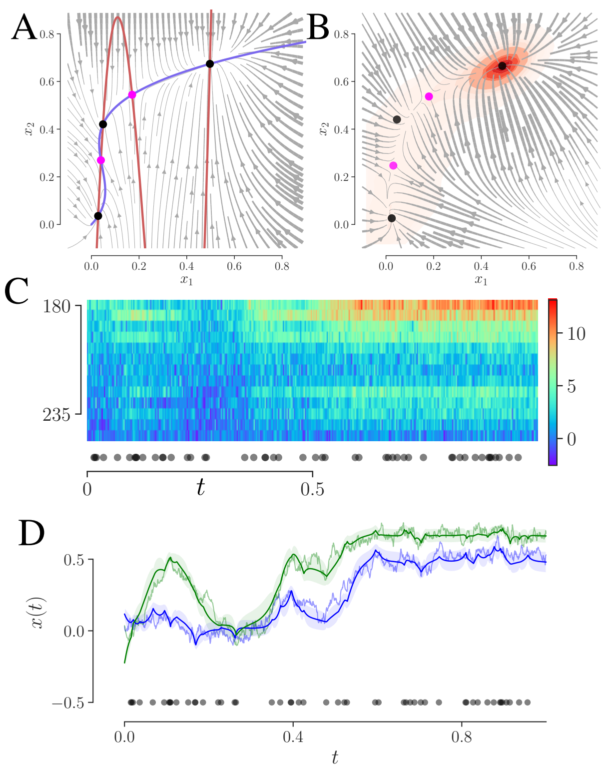

Figure 4. Multistable chemical reaction dynamics. A: Streamline plot of the concentration dynamics for two species of iodine, together with nullclines and fixed points. Stable fixed points are black, unstable ones are magenta. B: Learnt dynamics and fixed points with stability determined by eigenvalues of learnt Jacobians. The red contour plot illustrates the density of latent path locations across all trials used for training. C: Example spectroscopy measurements (output process) across light wavelengths (nm). D: Example true latent path together with the inferred posterior mean and ± 1 standard deviation tubes for each latent dimension on the same trial as C. The black dots indicate the time points at which measurements were taken.

6. Discussion

We have introduced a flexible and general variational Bayesian framework for the interpretable modelling of a continuous-time latent dynamical process from intermittent observations. Using a suitable GP prior, we integrate over a nonparametric description of the system dynamics, conditioned on its fixed points and associated local Jacobian matrices, thus both avoiding the need to assume a specific parametric dynamical form and directly obtaining a meaningful portrait of the dynamical structure. The approach applies to a variety of multivariate observation models, with many updates available in closed form.

The effectiveness of the approach is demonstrated using data simulated from a number of realistic but known nonlinear dynamical systems describing physical, biological and chemical phenomena. In each case, it was possible to recover a meaningful description of fixed points and nearby dynamics even when data were sparse; and an inferred dynamical model that approximated the true systems well over large regions of the state space.

A similar prior over dynamics could be adopted within a discrete-time model such as the GP-SSM, albeit with a less natural interpretation of the local Jacobians. However, real-world systems evolve in continuous time, and in some contexts available observations do not arrive at discrete sample times. Retaining a continuous-time model means that the variational posterior over latents can be described by a system of coupled ODEs. While the solution of these may incur a discretisation error, this is a numerical issue related to the choice of ODE solver, rather than the assumption of a discretised model. Indeed, the ODE solution can exploit an adaptive step size in a way that would be impractical within a discrete-time model.

Our work also differs from other GP-based approaches to time series modelling, where each dimension of the process $x_k(t)$ is modelled via an independent GP (Damianou et al., 2011; Duncker & Sahani, 2018). In this case, the prior on $\boldsymbol{x}(t)$ evaluated at any finite set of points can be described by a multivariate Gaussian distribution, which greatly simplifies the inference. However, this cannot capture correlations across the dimensions of the latent process and thus comes at a loss of the descriptive power.

The variational inference approach for SDEs from Archambeau et al. (2007; 2008) relied on a Gaussian observation model and known dynamics (Archambeau et al., 2007), or a known parameterisation of the dynamics (Archambeau et al., 2008), both of which are restrictive. Here, we have extended the inference approach to handle a wider class of observation models, as well as a nonparametric GP description of the dynamics. Batz et al. (2018) also use a GP to model the drift function of an SDE. However, they consider the setting where dense or sparse observations of the SDE path are directly available, while we treat the entire SDE output as latent. Furthermore, the interpretable nonparametric representation of the SDE dynamics in terms of their fixed points and local Jacobian matrices is novel.

While we have demonstrated our algorithm in the setting of unevenly sampled multivariate Gaussian and multivariate point process observations, the inference approach extends readily to other stochastic processes typically considered challenging to model, such as marked point processes. We therefore expect this approach to have diverse applications, ranging from neuroscience to chemistry and finance.

Acknowledgements

This work was funded by the Simons Foundation (SCGB 323228, 543039; MS) and the Gatsby Charitable Foundation.

References

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. Gaussian Process Approximations of Stochastic Differential Equations. In *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 2007.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. Variational Inference for Diffusion Processes. In *Advances in Neural Information Processing Systems*, volume 20, pp. 17–24, 2008.

Attias, H. A Variational Baysian Framework for Graphical Models. In *Advances in Neural Information Processing Systems 12*, pp. 209–215, 2000.

Batz, P., Ruttor, A., and Opper, M. Approximate Bayes learning of stochastic differential equations. *Phys. Rev. E*, 98:022109, 2018.

Bohner, G. and Sahani, M. Empirical fixed point bifurcation analysis. *arXiv preprint arXiv:1807.01486*, 2018.

Cseke, B., Schnoerr, D., Opper, M., and Sanguinetti, G. Expectation propagation for continuous time stochastic processes. *Journal of Physics A: Mathematical and Theoretical*, 49(49):494002, 2016.

Damianou, A., Titsias, M. K., and Lawrence, N. D. Variational Gaussian Process Dynamical Systems. In *Advances in Neural Information Processing Systems 24*, pp. 2510–2518, 2011.

Deisenroth, M. and Mohamed, S. Expectation Propagation in Gaussian Process Dynamical Systems. In *Advances in Neural Information Processing Systems 25*, pp. 2609–2617, 2012.

Deisenroth, M. P., Huber, M. F., and Hanebeck, U. D. Analytic moment-based Gaussian process filtering. In *Proceedings of the 26th annual International Conference on Machine Learning*, pp. 225–232, 2009.

Duncker, L. and Sahani, M. Temporal alignment and latent Gaussian process factor inference in population spike trains. In *Advances in Neural Information Processing Systems 31*, pp. 10445–10455. 2018.

Eleftheriadis, S., Nicholson, T., Deisenroth, M., and Hensman, J. Identification of Gaussian Process State Space Models. In *Advances in Neural Information Processing Systems 30*, pp. 5309–5319, 2017.

Frigola, R., Chen, Y., and Rasmussen, C. E. Variational Gaussian Process State-Space Models. In *Advances in Neural Information Processing Systems 27*, pp. 3680–3688, 2014.

Ganapathisubramanian, N. Tristability in the iodate–As (III) chemical system arising from a model of stirring and mixing effects. *The Journal of Chemical Physics*, 95(4): 3005–3008, 1991.

Golub, M. and Sussillo, D. Fixedpointfinder: A Tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks. *Journal of Open Source Software*, 3(31):1003, 2018.

Kireev, S. and Shnyrev, S. Study of molecular iodine, iodate ions, iodide ions, and triiodide ions solutions absorption in the UV and visible light spectral bands. *Laser Physics*, 25(7):075602, 2015.

Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., and Sahani, M. Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems 24*, pp. 1350–1358, 2011.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, pp. 1, 2018.

Ramakrishnan, N., Ertin, E., and Moses, R. L. Assumed density filtering for learning Gaussian process models. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 257–260, 2011.

Shenoy, K. V., Sahani, M., and Churchland, M. M. Cortical Control of Arm Movements: a Dynamical Systems Perspective. *Annual Review of Neuroscience*, 36(1):337–359, 2013.

Sussillo, D. and Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

Titsias, M. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Wang, J., Hertzmann, A., and Fleet, D. J. Gaussian Process Dynamical Models. In *Advances in Neural Information Processing Systems 18*, pp. 1441–1448, 2006.