Stochastic Differential Equations with Variational Wishart Diffusions

Martin Jørgensen [1]   **Marc Peter Deisenroth** [2]   **Hugh Salimbeni** [3]

Abstract

We present a Bayesian non-parametric way of inferring stochastic differential equations for both regression tasks and continuous-time dynamical modelling. The work has high emphasis on the *stochastic* part of the differential equation, also known as the diffusion, and modelling it by means of Wishart processes. Further, we present a semi-parametric approach that allows the framework to scale to high dimensions. This successfully lead us onto how to model both latent and autoregressive temporal systems with conditional heteroskedastic noise. We provide experimental evidence that modelling diffusion often improves performance and that this randomness in the differential equation can be essential to avoid overfitting.

1. Introduction

An endeared assumption to make when modelling multivariate phenomena with Gaussian processes (GPs) is that of independence between processes, i.e. every dimension of a multivariate phenomenon is modelled independently. Consider the case of a two-dimensional temporal process $\boldsymbol{x}_t$ evolving as

$$\boldsymbol{x}_t := f(\boldsymbol{x}_{t-1}) + \boldsymbol{\epsilon}_t, \tag{1}$$

where $f(\boldsymbol{x}_{t-1}) = (f_1(\boldsymbol{x}_{t-1}), f_2(\boldsymbol{x}_{t-1}))^\top$, f_1 and f_2 independent, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. This model is commonly used in the machine learning community and is easy to use and understand, but for many real-world cases the noise is too simplistic. In this paper, we will investigate the noise term $\boldsymbol{\epsilon}_t$ and also make it dependent on the state $\boldsymbol{x}_{t-1}$. This is also known as heteroskedastic noise. We will refer to the sequence of $\boldsymbol{\epsilon}_t$ as the *diffusion* or *process noise*.

[1]Department for Mathematics and Computer Science, Technical University of Denmark [2]Department of Computer Science, University College London [3]G-Research. Correspondence to: Martin Jørgensen <marjor@dtu.dk>.

Why model the process noise? Assume that in the example above, the two states represent meteorological measurements: rainfall and wind speed. Both are influenced by confounders, such as atmospheric pressure, which are not measured directly. This effect can in the case of the model in (1) only be modelled through the diffusion $\boldsymbol{\epsilon}$. Moreover, wind and rain may not correlate identically for all states of the confounders.

Dynamical modelling with focus in the noise-term is not a new area of research. The most prominent one is the *Auto-Regressive Conditional Heteroskedasticity* (ARCH) model (Engle, 1982), which is central to scientific fields like econometrics, climate science and meteorology. The approach in these models is to estimate large process noise when the system is exposed to a shock, i.e. an unforeseen significant change in states. Thus, it does not depend on the value of some state, but rather on a linear combination of previous states.

In this paper, we address this shortcoming and introduce a model to handle the process noise by the use of *Wishart processes*. Through this, we can sample covariance matrices dependent on the input state. This allows the system to evolve as a homogeneous system rather than independent sequences. By doing so, we can avoid propagating too much noise—which can often be the case with diagonal covariances—and potentially improve on modelling longer-range dependencies. Volatility modelling with GPs has been considered by Wu et al. (2014); Wilson & Ghahramani (2010); Heaukulani & van der Wilk (2019).

For regression tasks, our model is closely related to several recent works exploring *continuous-time* deep neural networks (E, 2017; Haber & Ruthotto, 2017; Chen et al., 2018). Here the notion of depth is no longer a discrete quantity (i.e. the number of hidden layers), but an interval on which a continuous flow is defined. In this view, continuous-time learning takes residual networks (He et al., 2016) to their infinite limit, while remaining computationally feasible. The flow, parameterized by a differential equation, allows for time-series modelling, even with temporal observations that are not equidistant.

This line of work has been extended with stochastic equivalents (Twomey et al., 2019; Tzen & Raginsky, 2019; Liu et al., 2019; Li et al., 2020), and the work by Andreas &

Kandemir (2019), who model the drift and diffusion of an SDE with Bayesian neural networks. These approaches make the framework more robust, as the original approach can fail even on simple tasks (Dupont et al., 2019).

The work that inspired our model most was by Hegde et al. (2019). They model the random field that defines the SDE with a Gaussian field. They consider regression and classification problems. To this end, they can take deep GPs (Damianou & Lawrence, 2013; Salimbeni & Deisenroth, 2017) to their 'infinite limit' while avoiding their degeneracy discussed by Duvenaud et al. (2014).

Our main focus throughout this paper lies on the *stochasticity* of the flow, what impact it has and to which degree it can be tamed or manipulated to improve overall performance. Contributions:

- A model that unifies theory from conditional heteroskedastic dynamics, stochastic differential equations (SDEs) and regression. We show how to perform variational inference in this model.

- A scalable approach to extend the methods to high-dimensional input without compromising with inter-dimensional independence assumptions.

2. Background

In this section, we give an overview of the relevant material on GPs, Wishart processes, and SDEs.

2.1. Gaussian Processes

A *Gaussian process* (GP) is a distribution over functions $f : \mathbb{R}^d \to \mathbb{R}^D$, satisfying that for any finite set of points $\boldsymbol{X} := \big(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\big)^\top \in \mathbb{R}^{N \times d}$, the outputs $\big(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)\big)^\top \in \mathbb{R}^{N \times D}$ are jointly Gaussian distributed. A GP is fully determined by a mean function $\mu : \mathbb{R}^d \to \mathbb{R}^D$ and a covariance function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{D \times D}$. This notation is slightly unorthodox, and we will elaborate.

The usual convention when dealing with *multi-output* GPs (i.e. $D > 1$) is to assume D i.i.d. processes that share the same covariance function (Álvarez & Lawrence, 2011), which equivalently can be done by choosing the covariance matrix $\boldsymbol{K} = k(\boldsymbol{X}, \boldsymbol{X}) \otimes \boldsymbol{I}_D$, where $\otimes$ denotes the Kronecker product and k is a covariance function for univariate output. For ease of notation we shall use $k^D(\boldsymbol{a}, \boldsymbol{b}) := k(\boldsymbol{a}, \boldsymbol{b}) \otimes \boldsymbol{I}_D$; that is, $k(\boldsymbol{a}, \boldsymbol{b})$ returns a kernel matrix of dimension number of rows in $\boldsymbol{a}$ times the number of rows in $\boldsymbol{b}$. This corresponds to the assumption of independence between output dimensions. Furthermore, we write $\boldsymbol{f} := f(\boldsymbol{X})$, $\boldsymbol{\mu} := \mathrm{vec}(\mu(\boldsymbol{X}))$ and denote by $\boldsymbol{K}$ the $ND \times ND$-matrix with $K_{i,j} = k^D(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then we can write in short $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$.

As the number N of training data points gets large, the size of $\boldsymbol{K}$ becomes a challenge as well, due to a required inversion during training/prediction. To circumvent this, we consider *sparse* (or low-rank) GP methods. In this respect, we choose M auxiliary *inducing* locations $\boldsymbol{Z} = \big(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M\big)^\top \in \mathbb{R}^{M \times d}$, and define their function values $\boldsymbol{u} := f(\boldsymbol{Z}) \in \mathbb{R}^{M \times D}$. Since any finite set of function values are jointly Gaussian, $p(\boldsymbol{f}, \boldsymbol{u})$ is Gaussian as well, and we can write $p(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$, where $p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{K}})$ with

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \boldsymbol{\alpha}^\top \mathrm{vec}(\boldsymbol{u} - \mu(\boldsymbol{Z})), \tag{2}$$

$$\tilde{\boldsymbol{K}} = k^D(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\alpha}^\top k^D(\boldsymbol{Z}, \boldsymbol{Z})\boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{\alpha} = k^D(\boldsymbol{X}, \boldsymbol{Z})k^D(\boldsymbol{Z}, \boldsymbol{Z})^{-1}$. Here it becomes evident why this is computationally attractive, as we only have to deal with the inversion of $k^D(\boldsymbol{Z}, \boldsymbol{Z})$, which due to the structure, only requires inversion of $k(\boldsymbol{Z}, \boldsymbol{Z})$ of size $M \times M$. This is opposed to a matrix of size $ND \times ND$ had we not used the low-rank approximation and independence of GPs.

We will consider variational inference to marginalise $\boldsymbol{u}$ (Titsias, 2009). Throughout the paper, we will choose our variational posterior to be $q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})$, where $q(\boldsymbol{u}) := \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$, similar to Hensman et al. (2013). Further, q factorises over the dimensions, i.e. $q(\boldsymbol{u}) = \prod_{j=1}^D \mathcal{N}(\boldsymbol{m}_j, \boldsymbol{S}_j)$, where $\boldsymbol{m} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_D)$ and $\boldsymbol{S}$ is a block-diagonal $MD \times MD$-matrix, with block-diagonal entries $\{\boldsymbol{S}_j\}_{j=1}^D$. In this case, we can analytically marginalise $\boldsymbol{u}$ in (2) to obtain

$$q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u} = \mathcal{N}(\boldsymbol{\mu}_f^q, \boldsymbol{K}_f^q), \tag{4}$$

$$\boldsymbol{\mu}_f^q = \boldsymbol{\mu} + \boldsymbol{\alpha}^\top \mathrm{vec}(\boldsymbol{m} - \mu(\boldsymbol{Z})), \tag{5}$$

$$\boldsymbol{K}_f^q = k^D(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\alpha}^\top \big(k^D(\boldsymbol{Z}, \boldsymbol{Z}) - \boldsymbol{S}\big)\boldsymbol{\alpha}, \tag{6}$$

which resembles (2)–(3), but which is analytically tractable given variational parameters $\{\boldsymbol{m}, \boldsymbol{S}, \boldsymbol{Z}\}$.

Recall that a *vector field* is a mapping $f : \mathbb{R}^d \to \mathbb{R}^D$ that associates a point in $\mathbb{R}^d$ with a vector in $\mathbb{R}^D$. A *Gaussian (random) field* is a vector field, such that for any finite collection of points $\{\boldsymbol{x}_i\}_{i=1}^N$, their associated vectors in $\mathbb{R}^D$ are jointly Gaussian distributed, i.e. a Gaussian field *is* a GP. We shall use both terminologies, but when we refer to a Gaussian field, we will think of the outputs as having a *direction*.

2.2. Wishart Processes

The *Wishart distribution* is a distribution over symmetric, positive semi-definite matrices. It is the multidimensional generalisation of the χ^2-distribution. Suppose $\boldsymbol{F}_v$ is a D-variate Gaussian vector for each $v = 1, \ldots, \nu$ independently, say $\boldsymbol{F}_v \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A})$. Then $\boldsymbol{\Sigma} = \sum_{v=1}^\nu \boldsymbol{F}_v \boldsymbol{F}_v^\top$ is Wishart

distributed with ν degrees of freedom and scale matrix $\boldsymbol{A}$. We write for short $\boldsymbol{\Sigma} \sim \mathcal{W}_D(\boldsymbol{A}, \nu)$. By Bartlett's decomposition (Kshirsagar, 1959), this can also be represented as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{L}^\top$, where $\boldsymbol{F}$ is a $D \times \nu$-matrix with all entries unit Gaussian and $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$.

With this parametrization we define Wishart processes, as in (Wilson & Ghahramani, 2010):

Definition 1. *Let $\boldsymbol{L}$ be a $D \times D$ matrix, such that $\boldsymbol{L}\boldsymbol{L}^\top$ is positive semidefinite and $f_{d,v} \sim \mathcal{GP}(0, k_{d,v}(\boldsymbol{x}, \boldsymbol{x}'))$ independently for every $d = 1, \ldots, D$ and $v = 1 \ldots, \nu$, where $\nu \geq D$. Then if*

$$\Sigma(\boldsymbol{x}) = \boldsymbol{L}\left(\sum_{v=1}^{\nu} \boldsymbol{f}_v(\boldsymbol{x})\boldsymbol{f}_v^\top(\boldsymbol{x})\right)\boldsymbol{L}^\top \tag{7}$$

is Wishart distributed for any marginal $\boldsymbol{x}$, and if for any finite collection of points $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ the joint distribution $\Sigma(\boldsymbol{X})$ is determined through the covariance functions $k_{d,v}$, then $\Sigma(\cdot)$ is a Wishart process. We will write

$$\Sigma \sim \mathcal{WP}_D(\boldsymbol{L}\boldsymbol{L}^\top, \nu, \kappa), \tag{8}$$

where κ is the collection of covariance functions $\{k_{d,v}\}$.

If $\boldsymbol{\Sigma}$ follows a Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{L}\boldsymbol{L}^\top$ of size $D \times D$, then for some $\rho \times D$-matrix $\boldsymbol{R}$ of rank ρ, we have that $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^\top \sim \mathcal{W}_\rho(\boldsymbol{R}\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{R}^\top, \nu)$. That is, $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^\top$ is Wishart distributed on the space of $\rho \times \rho$ symmetric, positive semi-definite matrices.

The Wishart distribution is closely related to the Gaussian distribution in a Bayesian framework, as it is the conjugate prior to the precision matrix of a multivariate Gaussian. Furthermore, it is the distribution of the maximum likelihood estimator of the covariance matrix.

The Wishart process is a slight misnomer as the posterior processes are *not* marginally Wishart. This is due to the mean function not being constant 0, and a more accurate name could be *matrix-Gamma* processes. We shall not refrain from the usual terminology: a Wishart process is a stochastic process, whose *prior* is a Wishart process.

2.3. Stochastic Differential Equations

We will consider SDEs of the form

$$d\boldsymbol{x}_t = \mu(\boldsymbol{x}_t)dt + \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_t)}dB_t, \tag{9}$$

where the last term of the right-hand side is the Itô integral (Itô, 1946). The solution $\boldsymbol{x}_t$ is a stochastic process, often referred to as a *diffusion process*, and μ and $\boldsymbol{\Sigma}$ are the drift and diffusion coefficients, respectively. In (9), B_t denotes the Brownian motion.

The Brownian motion is the GP satisfying that all increments are independent in the sense that, for $0 \leq s_1 < t_1 \leq s_2 < t_2$, then $B_{t_1-s_1}$ is independent from $B_{t_2-s_2}$. Further, any increment has distribution $B_t - B_s \sim \mathcal{N}(0, t - s)$. Lastly, $B_0 = 0$. This is equivalent to the GP with constant mean function 0 and covariance function $(t, s) \mapsto \min\{s, t\}$ (Rasmussen & Williams, 2006).

Given some initial condition (e.g. $\boldsymbol{x}_0 = \boldsymbol{0}$), we can generate *sample paths* $[0, T] \to \mathbb{R}^D$ by the Euler-Maruyama method. Euler-Maruyama (Kloeden & Platen, 2013) finely discretizes the temporal dimension $0 = t_0 < t_1 < \ldots < t_l = T$, and *pushes* $\boldsymbol{x}_{t_i}$ along the vector field $\boldsymbol{x}_{t_{i+1}} = \boldsymbol{x}_{t_i} + \boldsymbol{\mu}(\boldsymbol{x}_{t_i})\Delta_i + \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_{t_i})\Delta_i}\boldsymbol{N}$, where $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ and $\Delta_i = t_{i+1} - t_i$.

3. Model and variational inference

We consider a random field $f : \mathbb{R}^D \times [0, T] \to \mathbb{R}^D$ and a GP $g : \mathbb{R}^D \to \mathbb{R}^\eta$. Their priors are

$$f \sim \mathcal{GP}(0, k_f(\cdot, \cdot) \otimes \boldsymbol{I}_D), \quad g \sim \mathcal{GP}(0, k_g(\cdot, \cdot) \otimes \boldsymbol{I}_\eta). \tag{10}$$

We also have a Wishart process $\Sigma : \mathbb{R}^D \times [0, T] \to \mathcal{G}$, where $\mathcal{G}$ is the set of symmetric, positive semi-definite $D \times D$ matrices; the specific prior on this will follow in Section 3.1. We will approximate the posteriors of f, g and Σ with variational inference, but first we will formalise the model.

We propose a continuous-time deep learning model that can propagate noise in high-dimensions. This is done by letting the diffusion coefficient $\boldsymbol{\Sigma}(\boldsymbol{x}_t)$ of an SDE be governed by a Wishart process. The model we present factorises as

$$\begin{aligned} p(\boldsymbol{y}, \Theta) = & p(\boldsymbol{y}|\boldsymbol{g})p(\boldsymbol{g}|\boldsymbol{x}_T, \boldsymbol{u}_g)p(\boldsymbol{u}_g)p(\boldsymbol{x}_T|\boldsymbol{f}) \\ & \cdot p(\boldsymbol{f}|\boldsymbol{\Sigma}, \boldsymbol{u}_f)p(\boldsymbol{u}_f)p(\boldsymbol{\Sigma}|\boldsymbol{u}_\Sigma)p(\boldsymbol{u}_\Sigma), \end{aligned} \tag{11}$$

where $\Theta := \{\boldsymbol{g}, \boldsymbol{u}_g, \boldsymbol{x}_T, \boldsymbol{f}, \boldsymbol{u}_f, \boldsymbol{\Sigma}, \boldsymbol{u}_\Sigma\}$ denotes all variables to be marginalised. We assume that data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ is i.i.d. given the process, such that $p(\boldsymbol{y}|\boldsymbol{g}) = \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{g}_i)$. We approximate the posterior of g with the variational distribution as in (4), i.e.

$$q(\boldsymbol{g}_i) = \int p(\boldsymbol{g}_i|\boldsymbol{u}_g)q(\boldsymbol{u}_g)d\boldsymbol{u}_g \tag{12}$$

$$= \mathcal{N}(\tilde{\mu}_g(\boldsymbol{x}_i), \tilde{k}_g(\boldsymbol{x}_i, \boldsymbol{x}_i)), \tag{13}$$

where

$$\tilde{\mu}_g(\boldsymbol{x}_i) = \boldsymbol{\alpha}_g^\top(\boldsymbol{x}_i)\text{vec}(\boldsymbol{m}_g), \tag{14}$$

$$\tilde{k}_g(\boldsymbol{x}_i, \boldsymbol{x}_i) = k_g^\eta(\boldsymbol{x}_i, \boldsymbol{x}_i) \tag{15}$$
$$+ \boldsymbol{\alpha}_g^\top(\boldsymbol{x}_i)(k_g^\eta(\boldsymbol{Z}_g, \boldsymbol{Z}_g) - \boldsymbol{S}_g)\boldsymbol{\alpha}_g(\boldsymbol{x}_i),$$

where $\boldsymbol{\alpha}_g(\boldsymbol{x}_i) := k_g^\eta(\boldsymbol{x}_i, \boldsymbol{Z}_g)k_g^\eta(\boldsymbol{Z}_g, \boldsymbol{Z}_g)^{-1}$. Here $\boldsymbol{m}_g$ is an $M \times \eta$ matrix, and $\boldsymbol{S}_g$ is an $M\eta \times M\eta$-matrix, constructed as η different $M \times M$-matrices $\boldsymbol{S}_g = \{\boldsymbol{S}_j\}_j^\eta$. During inference (Quiñonero Candela & Rasmussen, 2005), we

(a) Graphical model based on Eq. (11)

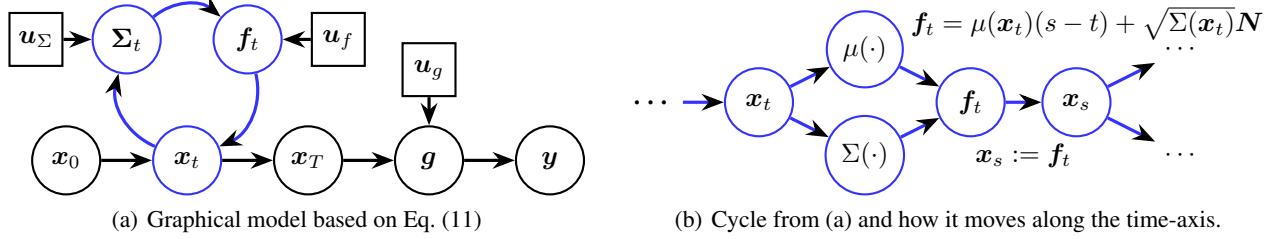(b) Cycle from (a) and how it moves along the time-axis.

Figure 1. (a) Graphical model based on the factorisation in Eq. (11); (b) The cycle from (a), which represents the *field* f, and how it moves along the time-axis. Here $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, (s - t)\boldsymbol{I})$. Blue represents the flow/SDE, square nodes are variational variables.

additionally assume that the marginals $g_i = g(\boldsymbol{x}_i)$ are independent when conditioned on $\boldsymbol{u}_g$. This is an approximation to make inference computationally easier.

The inputs to g are given as the state distribution of an SDE at a fixed time point $T \geq 0$. We construct this SDE from the viewpoint of a random field. Consider the random walk with step size Δ on the simplest Gaussian field, where any state has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For any time point t, the state distribution is tractable, i.e. $p(\boldsymbol{x}_t) = \boldsymbol{x}_0 + \sum_{s=1}^{S} \mathcal{N}(\Delta_s \boldsymbol{\mu}, \Delta_s \boldsymbol{\Sigma})$, where $\sum \Delta_s = t$ and S is any positive integer.

For a state-dependent Gaussian field, we define the random walk

$$\boldsymbol{x}_{t+\Delta} = \boldsymbol{x}_t + \mu(\boldsymbol{x}_t)\Delta + \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_t)\Delta}\boldsymbol{N}, \qquad (16)$$

with $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Given an initial condition $\boldsymbol{x}_0$, the state $\boldsymbol{x}_S$ after S steps is given by

$$\boldsymbol{x}_S = \boldsymbol{x}_0 + \sum_{s=0}^{S-1} \Big(\mu(\boldsymbol{x}_s)\Delta + \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_s)\Delta}\boldsymbol{N}\Big). \qquad (17)$$

In the limit $\Delta \to 0$, this random walk dynamical system is given by the diffusion process (Durrett, 2018)

$$\boldsymbol{x}_T - \boldsymbol{x}_0 = \int_0^T \mu(\boldsymbol{x}_t)dt + \int_0^T \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_t)}dB_t, \qquad (18)$$

where B is a Brownian motion. This is an SDE in the Îto-sense, which we numerically can solve by the Euler-Maruyama method. We will see that by a particular choice of variational distribution that $\boldsymbol{\Sigma}(\boldsymbol{x}_t)$ will be the realisation of a Wishart process. The coefficients in (18) are determined as the mean and covariance of a Gaussian field f. The posterior of f is approximated with a Gaussian $q(\boldsymbol{f}_i) = \mathcal{N}(\mu_f^q(\boldsymbol{x}_i), k_f^q(\boldsymbol{x}_i, \boldsymbol{x}_i))$, where

$$\mu_f^q(\boldsymbol{x}_i) = \boldsymbol{\alpha}_f^\top(\boldsymbol{x}_i)\text{vec}(\boldsymbol{m}_f), \qquad (19)$$
$$k_f^q(\boldsymbol{x}_i, \boldsymbol{x}_i) = k_f^D(\boldsymbol{x}_i, \boldsymbol{x}_i) \qquad (20)$$
$$+ \boldsymbol{\alpha}_f^\top(\boldsymbol{x}_i)\big(k_f^D(\boldsymbol{Z}_f, \boldsymbol{Z}_f) - \boldsymbol{S}_f\big)\boldsymbol{\alpha}_f(\boldsymbol{x}_i), \qquad (21)$$

and $\boldsymbol{\alpha}_f(\cdot) = k_f^D(\cdot, \boldsymbol{Z}_f)k_f^D(\boldsymbol{Z}_f, \boldsymbol{Z}_f)^{-1}$.

So far, we have seen how we move a data point $\boldsymbol{x}_0$ through the SDE (18) to $\boldsymbol{x}_T$, and further through the GP g, to make a prediction. However, each coordinate of $\boldsymbol{x}$ moves independently. By introducing the Wishart process, we will see how this assumption is removed.

3.1. Wishart-priored Gaussian random field

We are still considering the Gaussian field f, whose posterior is approximated by the variational distribution $q(\boldsymbol{f})$. To regularise (or learn) the noise propagated through this field into g, while remaining within the Bayesian variational framework, we define a hierarchical model as

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u}_f, \boldsymbol{\Sigma})p(\boldsymbol{u}_f)p(\boldsymbol{\Sigma}|\boldsymbol{u}_\Sigma)p(\boldsymbol{u}_\Sigma)d\{\boldsymbol{\Sigma}, \boldsymbol{u}_f, \boldsymbol{u}_\Sigma\}, \qquad (22)$$

where Σ is a Wishart process. Specifically, its prior is

$$\boldsymbol{\Sigma} \sim \mathcal{WP}_D(\boldsymbol{L}\boldsymbol{L}^\top, \nu, k_f), \qquad (23)$$

that is any marginal $\boldsymbol{\Sigma}(\boldsymbol{x}_t) = \boldsymbol{L}\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{L}^\top$, where $\boldsymbol{J}$ is the $D \times \nu$-matrix with all independent entries $j_{d,v}(\boldsymbol{x}_t)$ drawn from GP's that share the same prior $j_{d,v}(\cdot) \sim \mathcal{GP}(0, k_f(\cdot, \cdot))$. To approximate the posterior of the Wishart process we choose a variational distribution

$$q(\boldsymbol{J}, \boldsymbol{u}_\Sigma) = q(\boldsymbol{J}|\boldsymbol{u}_\Sigma)q(\boldsymbol{u}_\Sigma) := p(\boldsymbol{J}|\boldsymbol{u}_\Sigma)q(\boldsymbol{u}_\Sigma), \qquad (24)$$

where $q(\boldsymbol{u}_\Sigma) = \prod_{d=1}^{D} \prod_{v=1}^{\nu} \mathcal{N}(\boldsymbol{m}_{d,v}^\Sigma, \boldsymbol{S}_{d,v}^\Sigma)$. Here, $\boldsymbol{m}_{d,v}^\Sigma$ is $M \times 1$ and $\boldsymbol{S}_{d,v}^\Sigma$ is $M \times M$ for each pair $\{d, v\}$. Notice the same kernel is used for the Wishart process as is used for the random field f, that is: only one kernel *controls* the vector field f. The posterior of $\boldsymbol{\Sigma}$ is naturally defined through the posterior of J. Given our choice of kernel, this approximate posterior is identical to Eqs. (19)-(21), only changing the variational parameters to $\boldsymbol{m}_\Sigma$ and $\boldsymbol{S}_\Sigma$, and D changes to $D\nu$.

What remains to be defined in (11) is $p(\boldsymbol{f}|\boldsymbol{\Sigma}, \boldsymbol{u}_f)$. Since

$\Sigma(\boldsymbol{x}_t)$ is a $D \times D$-matrix we define

$$p\big(\boldsymbol{f}|\{\boldsymbol{\Sigma}(\boldsymbol{x}_i)\}_{i=1}^N, \boldsymbol{u}_f\big) = \mathcal{N}\big(\tilde{\mu}(\boldsymbol{X}), \tilde{k}_f^\Sigma(\boldsymbol{X},\boldsymbol{X})\big), \quad (25)$$

$$\tilde{\mu}(\boldsymbol{x}_i) = \boldsymbol{\alpha}_f^\top(\boldsymbol{x}_i)\mathrm{vec}(\boldsymbol{u}_f), \quad (26)$$

$$\tilde{k}_f^\Sigma(\boldsymbol{x}_i,\boldsymbol{x}_j) = \big(\boldsymbol{\Sigma}(\boldsymbol{x}_i) - \boldsymbol{h}_{ij}\big)\delta_{ij} + \boldsymbol{h}_{ij}, \quad (27)$$

where $\boldsymbol{h}_{ij} = \boldsymbol{\alpha}_f(\boldsymbol{x}_i)^\top k_f^D(\boldsymbol{Z}_f,\boldsymbol{Z}_f)\alpha_f(\boldsymbol{x}_j)$ and δ_{ij} is Kronecker's delta. Notice this, conditioned on the Wishart process, constitutes a FITC-type model (Snelson & Ghahramani, 2006).

This goes beyond the assumption of independent output dimensions, and instead makes the model learn the inter-dimensional dependence structure through the Wishart process Σ. This structure shall also be learned in the variational inference setup. The posterior of conditional $\boldsymbol{f}$ is approximated by

$$q(\boldsymbol{f},\boldsymbol{u}_f|\{\boldsymbol{\Sigma}(\boldsymbol{x}_i)\}_{i=1}^N) = q(\boldsymbol{f}|\{\boldsymbol{\Sigma}(\boldsymbol{x}_i)\}_{i=1}^N, \boldsymbol{u}_f)q(\boldsymbol{u}_f)$$
$$= p(\boldsymbol{f}|\{\boldsymbol{\Sigma}(\boldsymbol{x}_i)\}_{i=1}^N, \boldsymbol{u}_f)q(\boldsymbol{u}_f), \quad (28)$$

where $q(\boldsymbol{u}_f) := \mathcal{N}(\boldsymbol{m}_f, k_f^D(\boldsymbol{Z}_f,\boldsymbol{Z}_f))$. At first, this might seem restrictive, but covariance estimation is already in Σ and the variational approximation is the simple expression

$$q(\boldsymbol{f}|\{\boldsymbol{\Sigma}(\boldsymbol{x}_i)\}_{i=1}^N) = \prod_{i=1}^N \mathcal{N}\big(\alpha_f^\top(\boldsymbol{x}_i)\boldsymbol{m}_f, \boldsymbol{\Sigma}(\boldsymbol{x}_i)\big). \quad (29)$$

The marginalisation can then be computed with Jensen's inequality

$$\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y},\Theta)d\Theta$$
$$\geq \int \log\Big(\frac{p(\boldsymbol{y},\Theta)}{q(\Theta)}\Big)q(\Theta)d\Theta$$
$$= \int \log p(\boldsymbol{y}|\boldsymbol{g})q(\boldsymbol{g}|\Theta\backslash\{\boldsymbol{g}\})d\Theta \quad (30)$$
$$- \mathrm{KL}\big(q(\boldsymbol{u}_g)\|p(\boldsymbol{u}_g)\big)$$
$$- \mathrm{KL}\big(q(\boldsymbol{u}_f)\|p(\boldsymbol{u}_f)\big) - \mathrm{KL}\big(q(\boldsymbol{u}_\Sigma)\|p(\boldsymbol{u}_\Sigma)\big),$$

or, in a more straightforward language,

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(g)}[\log p(\boldsymbol{y}|\boldsymbol{g})] - \mathrm{KL}\big(q(\boldsymbol{u}_g)\|p(\boldsymbol{u}_g)\big) \quad (31)$$
$$- \mathrm{KL}\big(q(\boldsymbol{u}_f)\|p(\boldsymbol{u}_f)\big) - \mathrm{KL}\big(q(\boldsymbol{u}_\Sigma)\|p(\boldsymbol{u}_\Sigma)\big).$$

The right-hand side in (31) is the so-called *evidence lower bound* (ELBO). The first term, the expectation, is analytically intractable, due to $q(g)$ being non-conjugate to the likelihood. Therefore, we determine it numerically with Monte Carlo (MC) or with Gauss-Hermite quadrature (Hensman et al., 2015). With MC, often a few samples are enough for reliable inference (Salimans & Knowles, 2013).

The KL-terms in (31) can be computed analytically as they all involve multivariate Gaussians. Still, due to some of the modelling constraints, it is helpful to write them out, which yields

$$\mathrm{KL}\big(q(\boldsymbol{u}_g)\|p(\boldsymbol{u}_g)\big) = \sum_{d=1}^\eta \mathrm{KL}\big(q(\boldsymbol{u}_{g_d})\|p(\boldsymbol{u}_{g_d})\big), \quad (32)$$

$$\mathrm{KL}\big(q(\boldsymbol{u}_\Sigma)\|p(\boldsymbol{u}_\Sigma)\big) = \sum_{d=1}^D \sum_{v=1}^\nu \mathrm{KL}\big(q(\boldsymbol{u}_{\Sigma_{d,v}})\|p(\boldsymbol{u}_{\Sigma_{d,v}})\big), \quad (33)$$

where in both instances we used the independence between the GPs. The remaining one is special. Since both distribution share the same covariance it reduces to

$$\mathrm{KL}\big(q(\boldsymbol{u}_f)\|p(\boldsymbol{u}_f)\big) = \frac{1}{2}\sum_{d=1}^D \boldsymbol{m}_{f_d}^\top k_f^D(\boldsymbol{Z}_f,\boldsymbol{Z}_f)^{-1}\boldsymbol{m}_{f_d}. \quad (34)$$

Here, $k_f^D(\boldsymbol{Z}_f,\boldsymbol{Z}_f)^{-1}$ is already known from the computation of (33), as the kernel and inducing locations are shared.

Summarising this section, we have inputs $\boldsymbol{x}_0 := \boldsymbol{x}$ that are warped through an SDE (governed by a random field f) with drift μ and diffusion Σ that is driven by *one* kernel k_f^D. The value of this SDE, at some given time T, is then used as input to a final layer g, i.e. $g(\boldsymbol{x}_T)$ predicts targets $y(\boldsymbol{x})$. All this is inferred by maximising the ELBO (31).

3.2. Complexity and scalability

The computational cost of estimating $\boldsymbol{\Sigma}$ with a Wishart, as opposed to a diagonal matrix, can be burdensome. For the diagonal, the cost is $\mathcal{O}(DNM^2)$ since we need to compute (3) D times. Sampling $D\nu$ GP values and then matrix-multiplying it with a $D \times \nu$ matrix is of complexity $\mathcal{O}(D\nu NM^2 + D\nu D)$. Hence, if we, for simplicity, let $\nu = D$, we have overhead cost of $\mathcal{O}(D^2NM^2 + D^3)$. Note this is only the computational budget associated with the diffusion coefficients of the random field; the most costly one.

On this inspection, we propose a way to overcome a too heavy burden if D is large. Naturally this involves an approximation; this time a low-rank approximation on the dimensionality-axis. Recall that, if $\boldsymbol{\Sigma}_\rho \sim \mathcal{WP}_\rho(\boldsymbol{I},\nu,\kappa)$, then $\boldsymbol{\Sigma}_D := \boldsymbol{L}\boldsymbol{\Sigma}_\rho\boldsymbol{L}^\top \sim \mathcal{WP}_D(\boldsymbol{L}\boldsymbol{L}^\top,\nu,\kappa)$. The matrices naturally are of rank $\rho \ll D$. The computational overhead is reduced to $\mathcal{O}(\rho^2 NM^2 + D\rho^2)$ if $\nu = \rho$. This same structure was introduced by Heaukulani & van der Wilk (2019) for time-series modelling of financial data; and it reminisces the structure of Semiparametric Latent Factor Models (SLFM) (Seeger et al., 2005). That is, we have ρ GPs, and the D-dimensional outputs are all linear combinations of these. For clarity, we need only to compute/sample $\sqrt{\boldsymbol{\Sigma}_D} = \boldsymbol{L}\boldsymbol{J}$,

where $\boldsymbol{J}$ is a $\rho \times \nu$ matrix, with GP values according to the approximate posterior $q(\boldsymbol{J})$, where D replaced by ρ.

3.3. Further model specifications

If ρ is too small it can be difficult to identify a good diffusion coefficient as the matrix is too restricted by the low rank. One possible way to overcome this is too add 'white noise' to the matrix

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{L}^\top + \boldsymbol{\Lambda}, \tag{35}$$

where $\boldsymbol{\Lambda}$ is a diagonal $D \times D$-matrix. In many situations, this will ensure that the diffusion is full rank, and this provides more freedom in estimating the marginal variances. However, if the values on the diagonal of $\boldsymbol{\Lambda}$ are estimated by maximum likelihood, we have to be cautious. If $\boldsymbol{\Lambda}$ becomes to 'dominant', inference can turn off the Wishart-part, potentially leading to overfitting.

Consider the matrix $\boldsymbol{L}$, that makes up the scale matrix of the Wishart process. It is fully inferred by maximum likelihood, hence there is no KL-term to regularise it. Effectively, this can turn off the stochasticity of the flow by making some matrix norm of $\boldsymbol{L}$ be approximately zero. Then the flow is only determined by its drift and overfitting is a likely scenario.

To alleviate this concern we propose to regularise $\boldsymbol{L}$ by its rownorms. That is,

$$\forall d = 1, \dots, D: \quad \sum_{r=1}^{\rho} L_{d,r}^2 = 1, \tag{36}$$

where $L_{d,r}$ denotes the entries of $\boldsymbol{L}$. First of all, this ensures that the prior variance for all dimensions is determined by the kernel hyperparameters, as it makes the diagonal of the scale matrix $\boldsymbol{L}\boldsymbol{L}^\top$ equal to 1. This way the variance in each dimension is a 'fair' linear combination of the ρ GPs that control the Wishart.

3.4. Extending to time series

The specified model can be specified to model temporal data $\mathcal{D} = \{\boldsymbol{y}_i, t_i\}_{i=1}^{N}$ in a straightforward way. In a few lines, see also Figure 1, we write

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \int_0^t \mu(\boldsymbol{x}_s)ds + \int_0^t \sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_s)}dB_s, \tag{37}$$

$$f(\cdot)|\boldsymbol{\Sigma}(\cdot), \mathcal{D} \sim \mathcal{GP}(\mu(\cdot), \boldsymbol{\Sigma}(\cdot)), \tag{38}$$

$$\boldsymbol{\Sigma}(\cdot) \sim \mathcal{WP}(\cdot|\mathcal{D}), \tag{39}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}(g(\boldsymbol{x}_t), \boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{x}_t)\boldsymbol{A}^\top + \boldsymbol{\Lambda}). \tag{40}$$

If g is not the identity mapping, we can define a *latent dynamical* model. Say g is a GP mapping from $\mathbb{R}^D$ to $\mathbb{R}^\eta$.

This is similar to GP state space models (GPSSM) where the dynamics, or transitions, are defined $\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + \boldsymbol{\epsilon}_x$ and $\boldsymbol{y}_t = g(\boldsymbol{y}_t) + \boldsymbol{\epsilon}_y$, for GPs f and g and some noise variables $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$, usually Gaussian with zero mean and diagonal covariance matrix (Deisenroth et al., 2012; Eleftheriadis et al., 2017).

The latent dynamics defined in (37)–(39) are not restricted to have equi-temporal measurements and model non-diagonal covariance structure both in the latent states $\boldsymbol{x}$ and in the observed states $\boldsymbol{y}$ through the matrix $\boldsymbol{A}$, which is an $\eta \times D$-matrix. Adding the diagonal $\eta \times \eta$-matrix $\boldsymbol{\Lambda}$ is necessary to avoid singularity. Even though $\boldsymbol{\Sigma}(\cdot)$ is a $D \times D$-matrix, we can still lower-rank approximate with a ρ-rank matrix, as described in Section 3.2. The log-likelihood we compute is

$$\log p(\boldsymbol{y}_t|\boldsymbol{g}_t, \boldsymbol{\Sigma}(\boldsymbol{x}_t)) = \frac{\eta}{2}\log(2\pi) - \log(\det(\boldsymbol{B}))$$
$$- \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{g}_t)^\top \boldsymbol{B}^{-1}(\boldsymbol{y}_t - \boldsymbol{g}_t), \tag{41}$$

where $\boldsymbol{B} := \boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{x}_t)\boldsymbol{A}^\top + \boldsymbol{\Lambda}$. As a consequence of the matrix-determinant lemma and the Woodbury identity, we can evaluate the likelihood cheaply, because of $\boldsymbol{B}$'s structure. The ELBO that we optimise during training is similar to (31), only the likelihood term is different: it is swapped for a variational expectation over (41). We assume independence between all temporal observations, i.e. $p(\mathcal{D}) = \prod_{i=1}^{N} p(\{\boldsymbol{y}_i, t_i\})$.

4. Experiments

We evaluate the presented model in both regression and a dynamical setup. In both instances, we use baselines that are similar to our model to easier distinguish the influence the diffusion has on the experiments. We evaluate on a well-studied regression benchmark and on a higher-dimensional dynamical dataset.

4.1. Regression

We compare our model, which we will dub *Wishart-priored GP flow* (*diffWGP*), to three baseline models in order to shed light on some properties of the diffWGP.

GP flows Reproducing the model from Hegde et al. (2019) will give indications, if it is possible to increase overall performance by modelling the randomness in the flow. This model has a diagonal matrix $\boldsymbol{\Sigma}$ with entries determined solely by the chosen covariance function. We will refer to this model with *diffGP*.

No noise flows We also evaluate the model, where $\boldsymbol{\Sigma} = \boldsymbol{0}$, i.e. the situation where the flow is *deterministic*. The remaining part of the flow is still as in (19) to make fair
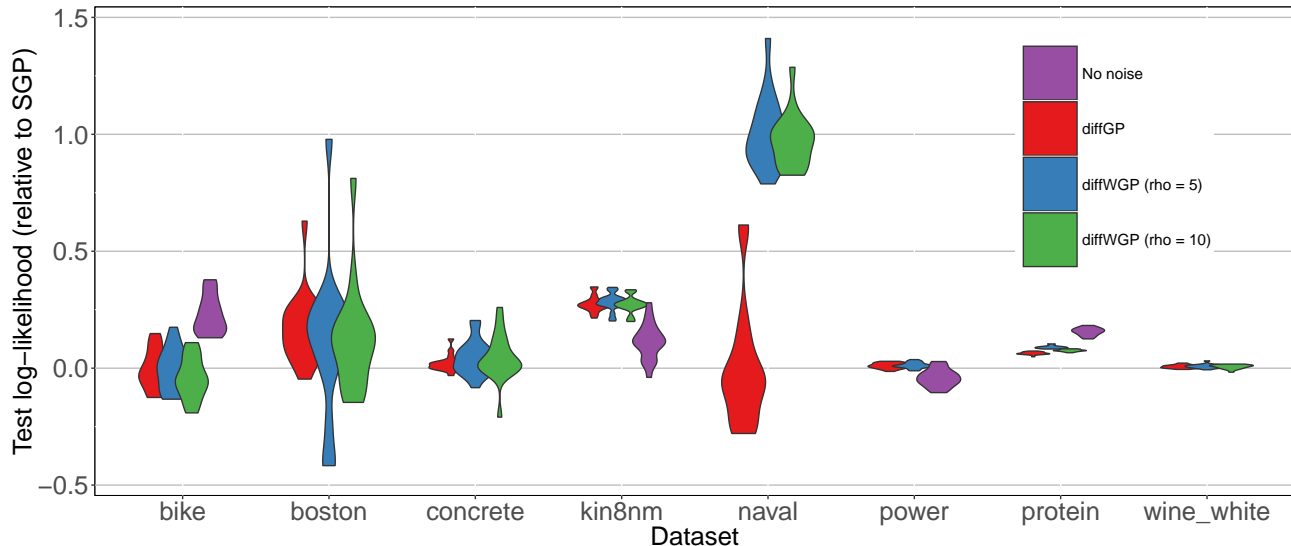
Figure 2. Test-set log-likelihood values on eight UCI regression datasets. The violin plots show the test-set log (likelihood-ratio) of baseline diffusion models with respect to the SGP baseline. Values greater than 0 indicate an improvement over SGP. Key findings are that No noise can overfit heavily (boston, concrete, naval), and diffWGP performs best on most datasets. The figure has been cut for readability—this explain why occasionally purple violins are missing.

comparisons. All the relevant KL-terms are removed from the ELBO (31). We refer to this as *No noise*.

Sparse GPs Also in the variational setup we shall compare to vanilla sparse GPs, i.e. the model introduced by Titsias (2009). We will refer to this as *SGP*.

4.1.1. EXPERIMENTAL SETUP

In all experiments, we choose 100 inducing points for the variational distributions, all of which are Gaussians. All models are trained for 50000 iterations with a mini-batch size of 2000, or the number of samples in the data if smaller. In all instances, the first 10000 iterations are *warm-starting* the final layer GP g, keeping all other parameters fixed. We use the Adam-optimiser with a step-size of 0.01. After this all flows (this excludes SGP) are initialised with a constant mean 0 and covariance functions chosen as RBF with automatic relevance determination (ARD), initialised with tiny signal noise to ensure $\boldsymbol{x}_0 \approx \boldsymbol{x}_T$. The time variable T is always 1.

The remaining 40000 iterations (SGP excluded) are updating again with Adam with a more cautious step-size of 0.001. For the diffWGP, the first 4000 of these are warm-starting the KL-terms associated with the flow to speed up convergence. Note that this model fits more parameters than the baseline models. For the diffWGP, we update the ELBO

$$
\mathbb{E}_{q(g)}[\log p(y|g)] - \mathrm{KL}\big(q(\boldsymbol{u}_g)\|p(\boldsymbol{u}_g)\big) \\
- c^2 \mathrm{KL}\big(q(\boldsymbol{u}_f)\|p(\boldsymbol{u}_f)\big) - c\mathrm{KL}\big(q(\boldsymbol{u}_\Sigma)\|p(\boldsymbol{u}_\Sigma)\big), \tag{42}
$$

|	*diffGP vs. SGP*	*diffWGP vs. diffGP*
BIKE (14)	0.8695	0.2262
BOSTON (13)	**<0.0001**	0.9867
CONCRETE (8)	**0.0042**	**0.0348**
KIN8NM (8)	**<0.0001**	**0.0164**
NAVAL (26)	0.8695	**<0.0001**
POWER (4)	**<0.0001**	0.1387
PROTEIN (9)	**<0.0001**	**<0.0001**
WINE_WHITE (11)	**0.0003**	0.3238

Table 1. Wilcoxons paired signed rank-test. Listed are the p-values of the hypothesis of equal median versus alternative that location shift is negative. Bold highlights the significant ones at a 0.05 confidence level. In parenthesis are the input dimensionality of the datasets. Results are for $\rho = 5$.

where $c = \min(1, \frac{iteration}{4000})$, i.e. we warm-start the regularising KL-terms.

4.1.2. UCI REGRESSION BENCHMARK

Figure 2 shows the results on eight UCI benchmark datasets over 20 train-test splits (90/10). On the y-axis we see the distribution of the test-set log-likelihood subtracted by the SGP log-likelihood on the same split. Values greater than 0 are improvements over the baseline SGP. An analogous plot with RMSE is supplied in the supplementary material. In Table 1, we use Wilcoxon's paired rank test to evaluate whether the more advanced models perform better.

Key observations are: *not* having noise in the flow (No
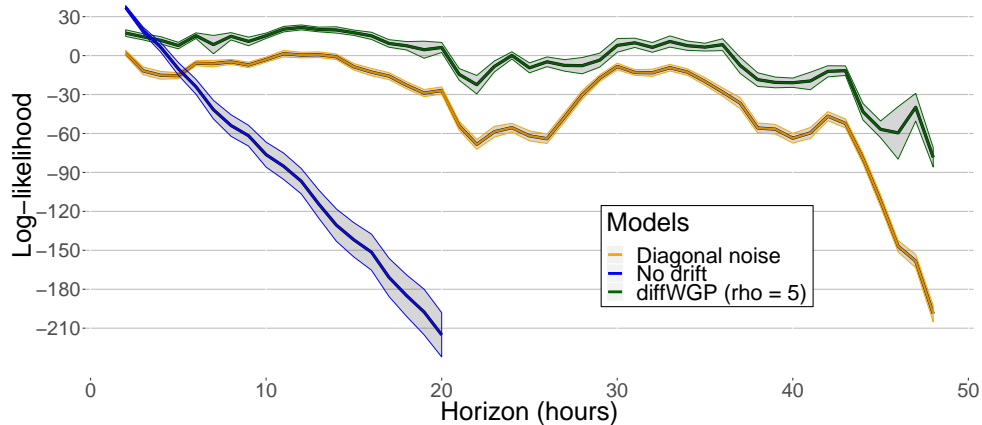
(a) The log-likelihood of *forecasted* measurements up to 48 hours. The bold lines mark the average log-likelihood at a given hour based on 50 simulations. The associated shaded areas span twice the standard error.
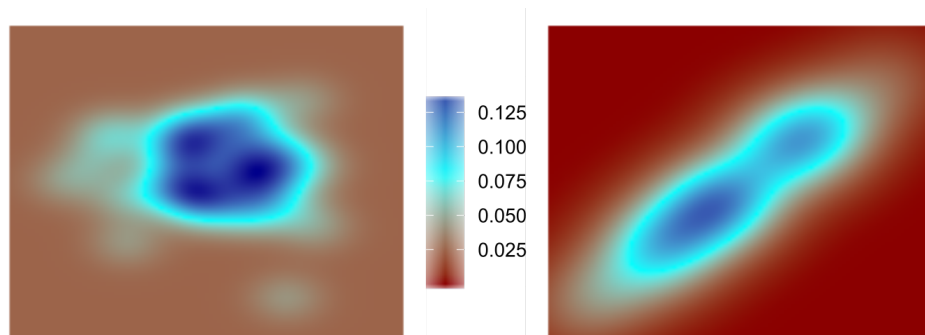
(b) The density (colour) of the 48-hour horizon predictions of temperature measurement in Tiantan (x-axis) and Dongsi (y-axis). These locations are within a few kilometres of each other. *Left*: diagonal noise case; *Right*: Wishart noise. The Wishart detects a correlation between these two temperature measurements, as we would expect for such nearby locations.

Figure 3. (a): The performance of predictions plotted over the forecast horizon. (b): The joint development of two temperature measurements over the forecasted time-horizon for two different models.

noise) seem to lead to overfitting, except in two cases, where a more expressive model is preferred. In one of these cases (protein) Wishart modelling improves both the RMSE and the log-likelihood. In one case (boston), overfitting was absurdly large: on this dataset we were not able to reproduce the results from Hegde et al. (2019) either. In four cases (concrete, kin8nm, power, wine_white), *No noise* overfitted mildly. In two of these cases, *diffWGP* improved over *diffGP*. The two cases, where no improvement is significant, are simple cases, wine_white and power, which are almost linear or low-dimensional. On the naval dataset, the *No noise* model could not run due to numerical issues. Here *diffWGP* outperforms *diffGP* in the log-likelihood. We conjecture this is because of the high dimensionality and the fact that almost no observation noise is present. We found no substantial influence of the parameter ρ; if any then it actually seems to prefer lower-rank approximations. This

emphasises that training Wishart processes is difficult, and further research in this area is needed.

4.2. Auto-regressive modelling of air quality

We evaluate our dynamical model on atmospheric air-quality data from Beijing (Zhang et al., 2017). We pre-processed the data for three locations in the city (Shunyi, Tiantan, Dongsi), which each have hourly observation of ten features over the period of 2014–2016. Explicitly, the ten features are: the concentration of PM2.5, PM10, SO2, NO2, CO, O3, the temperature and dew point temperature, air pressure and amount of precipitation.[1]

---

[1]The full data set is available at `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`.

We use the first two years of this dataset for training and aim to forecast into the first 48 hours of 2016. Including the variables year, month, day and hour, we have in total 34 features for the three cities and 17520 temporal observations for training. Missing values were linearly interpolated. All features were standardised.

To analyse properties of our proposed model, we perform an ablation study with the following models:

diffWGP The model proposed in the paper to model the diffusion with Wisharts.

Diagonal noise The drift term remains as in the diffWGP model, but the diffusion is restricted to diagonal, i.e. correlated diffusion cannot be modelled. This becomes the model

$$\boldsymbol{x}_t = \boldsymbol{x}_s + \mu(\boldsymbol{x}_s)(t-s) + \sqrt{\boldsymbol{\Lambda}(t-s)}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{43}$$

No drift The drift is constantly zero, and the diffusion is modelled by a Wishart, which results in the model

$$\boldsymbol{x}_t = \boldsymbol{x}_s + \sqrt{\left(\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{x}_t)\boldsymbol{A}^\top + \boldsymbol{\Lambda}\right)(t-s)}\boldsymbol{\epsilon}_t. \tag{44}$$

This model is a continuous-time version of the model presented by Heaukulani & van der Wilk (2019).

In all instances, we train by minibatching shorter sequences, and we use the Adam optimiser (Kingma & Ba, 2014) with a learning rate 0.01. Due to the large amount of temporal observation compared to small batches we ease off on momentum.

Figure 3(a) shows how the different models *forecast* future observations by reporting the log-likelihood traces of individual models at test time. The figure shows the mean and two times the standard error, which we obtain from 50 simulations. At first, we see that having no drift starts off better, but quickly drops in performance. This is not unexpected, as the data has structure in its evolution. The difference between the models with drift, but different diffusions, are more interesting for this dataset. Overall, Wishart diffusions perform best, and it seems to be resilient and take only few and relatively small 'dips'.

We expect this dataset to have highly correlated features. The three locations in Beijing are, in distance, close to each other; naturally the different air measurements are similar in their evolution over time. Figure 3(b) illustrates how a model with diagonal noise is incapable of learning this joint development of temperature measurements. Here, the Wishart learns that when the temperature in Dongsi is high, it is also high in Tiantan. This behaviour is seen in many pairs of the features considered, and it suggests *diffWGP* has dynamics moving on a manifold of smaller dimension than if diagonal noise was considered. This supports the hypothesis that *diffWGP* moves as *one* dynamical systems, opposed to 34.

5. Conclusion

In a non-parametric Bayesian way, we presented a scalable approach to continuous-time learning with high emphasis on correlated process noise. This noise is modelled with a Wishart process, which lets high-dimensional data evolve as a single system, rather than D independent systems. We presented a way to scale this to high dimensions. We found that it is never worse taking the dependence structure in the process noise into account. However, with certain types of data, it can mitigate overfitting effects and improve performance.

Code is publicly available at: `https://github.com/JorgensenMart/Wishart-priored-SDE`.

References

Álvarez, M. A. and Lawrence, N. D. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12:1459–1500, 2011.

Andreas, L. and Kandemir, M. Differential Bayesian neural nets. *arXiv:1912.00796*, 2019.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Damianou, A. and Lawrence, N. D. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, 2013.

Deisenroth, M. P., Turner, R., Huber, M., Hanebeck, U. D., and Rasmussen, C. E. Robust filtering and smoothing with Gaussian processes. *IEEE Transactions on Automatic Control*, 57(7):1865–1871, 2012.

Dupont, E., Doucet, A., and Teh, Y. W. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, 2019.

Durrett, R. *Stochastic calculus: a practical introduction*. CRC press, 2018.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, 2014.

E, W. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5 (1):1–11, 2017.

Eleftheriadis, S., Nicholson, T. F. W., Deisenroth, M. P., and Hensman, J. Identification of Gaussian process state space models. In *Advances in Neural Information Processing Systems*, 2017.

Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

Haber, E. and Ruthotto, L. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Heaukulani, C. and van der Wilk, M. Scalable Bayesian dynamic covariance modeling with variational Wishart and inverse Wishart processes. In *Advances in Neural Information Processing Systems*, 2019.

Hegde, P., Heinonen, M., Lähdesmäki, H., and Kaski, S. Deep learning with differential Gaussian process flows. In *Artificial Intelligence and Statistics*, 2019.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, 2013.

Hensman, J., Matthews, A. G. d. G., Filippone, M., and Ghahramani, Z. MCMC for variationally sparse Gaussian processes. In *Advances in Neural Information Processing Systems*, 2015.

Itô, K. On a stochastic integral equation. *Proceedings of the Japan Academy*, 22(2):32–35, 1946.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Kloeden, P. E. and Platen, E. *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer Science & Business Media, 2013.

Kshirsagar, A. M. Bartlett Decomposition and Wishart distribution. *The Annals of Mathematical Statistics*, 30 (1):239–241, 1959.

Li, X., Wong, T.-K. L., Chen, R. T. Q., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *Artificial Intelligence and Statistics*, 2020.

Liu, X., Xiao, T., Si, S., Cao, Q., Kumar, S., and Hsieh, C.-J. Neural SDE: Stabilizing neural ODE networks with stochastic noise. *arXiv:1906.02355*, 2019.

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

Rasmussen, C. E. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Salimans, T. and Knowles, D. A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

Salimbeni, H. and Deisenroth, M. P. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

Seeger, M., Teh, Y.-W., and Jordan, M. Semiparametric latent factor models. Technical report, 2005.

Snelson, E. and Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, 2006.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, 2009.

Twomey, N., Kozłowski, M., and Santos-Rodríguez, R. Neural ODEs with stochastic vector field mixtures. *arXiv:1905.09905*, 2019.

Tzen, B. and Raginsky, M. Neural stochastic differential equations: deep latent Gaussian models in the diffusion limit. *arXiv:1905.09883*, 2019.

Wilson, A. G. and Ghahramani, Z. Generalised Wishart processes. *arXiv:1101.0240*, 2010.

Wu, Y., Hernández-Lobato, J. M., and Ghahramani, Z. Gaussian process volatility model. In *Advances in Neural Information Processing Systems*. 2014.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473, 2017.