Modeling the drift function in stochastic differential equations using reduced rank Gaussian processes

Roland Hostettler, Filip Tronarp, and Simo Särkkä

DOI:

10.1016/j.ifacol.2018.09.137

Modeling the Drift Function in Stochastic Differential Equations using Reduced Rank Gaussian Processes ⋆

Roland Hostettler * **Filip Tronarp** * **Simo Särkkä** *

* *Department of Electrical Engineering and Automation*
Aalto University, Finland
(e-mail: firstname.lastname@aalto.fi)

Abstract: In this paper, we propose a Gaussian process-based nonlinear, time-varying drift model for stochastic differential equations. In particular, we combine eigenfunction expansion of the Gaussian process' covariance kernel in the spatial input variables with spectral decomposition in the time domain to obtain a reduced rank state space representation of the drift model, which avoids the growing complexity (with respect to time) of the full Gaussian process solution. The proposed approach is evaluated on two nonlinear benchmark problems, the Bouc–Wen and the cascaded tanks systems.

Keywords: Nonlinear system identification, nonparametric methods, Bayesian methods, filtering and smoothing, estimation and filtering, Gaussian processes

1. INTRODUCTION

Stochastic differential equations (SDEs) are a powerful tool for modeling time series (Øksendal, 2010). They have been used to model different types of systems such as heat dynamics in buildings (Madsen and Holst, 1995), forecasting of solar irradiation (Iversen et al., 2014), or in financial statistics (Lindström et al., 2015).

In this article, we consider nonlinear SDEs of the form

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t)\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}_t, \tag{1}$$

where $\boldsymbol{x}_t \triangleq \boldsymbol{x}(t) \in \mathbb{R}^{N_{\boldsymbol{x}}}$ is the state vector, $\boldsymbol{u}_t \triangleq \boldsymbol{u}(t) \in \mathbb{R}^{N_{\boldsymbol{u}}}$ is a deterministic input, $\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t) = [f_1(\boldsymbol{x}_t, \boldsymbol{u}_t, t)\ f_2(\boldsymbol{x}_t, \boldsymbol{u}_t, t)\ \dots\ f_{N_{\boldsymbol{x}}}(\boldsymbol{x}_t, \boldsymbol{u}_t, t)]^{\mathsf{T}}$ is the nonlinear, time-varying drift function, and $\boldsymbol{\beta}_t \triangleq \boldsymbol{\beta}(t)$ is Brownian motion with diffusion matrix $\boldsymbol{Q}$. Given the SDE model (1), our aim is then to infer the drift function $\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t)$ and the diffusion matrix $\boldsymbol{Q}$ from system identification data obtained as

$$\boldsymbol{y}_n = \boldsymbol{H}\boldsymbol{x}_{t_n} + \boldsymbol{v}_n, \tag{2}$$

where $\boldsymbol{y}_n \in \mathbb{R}^{N_{\boldsymbol{y}}}$ is the measurement, $\boldsymbol{H} \in \mathbb{R}^{N_{\boldsymbol{y}} \times N_{\boldsymbol{x}}}$, and $\boldsymbol{v}_n$ is a white noise sequence with covariance $\boldsymbol{R}$.

If the system under consideration is well known, for example from physical relationships, then a first principles-based model or parametric grey-box model may be most suitable and the system identification problem reduces to estimating the model parameters (Ljung, 1997; Kristensen et al., 2004). However, when the knowledge of the system is not sufficient to infer a suitable model structure, a non-parametric black-box approach may be required. This paper discusses the latter, where the drift function is assumed to be an outcome of a Gaussian process (GP) (Rasmussen and Williams, 2006). A similar approach has been considered in Ruttor et al. (2013) and Batz et al. (2017), where the time-invariant drift was modeled as a GP and measurements of the full state vector were assumed. Furthermore, the problem was solved using the full batch formulation for small datasets and sparse approximations for large datasets (Ruttor et al., 2013; Batz et al., 2017). Modeling the state transition function as GPs has also been considered in the context of discrete time state space models, where basis function expansions were used to mitigate the problem of increasing computational complexity with time (Svensson et al., 2016; Svensson and Schön, 2017). However, modeling the state transition function of discrete time state space models is not equivalent to modeling the drift. In fact, if the discrete time model originates from a discretized continuous time model, the resulting discrete time GP state space model may actually be more complex than the corresponding drift GP model.

In contrast to these approaches, in this paper, we use eigenfunction expansion of the covariance kernel in its spatial arguments $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$, similar to Svensson et al. (2016), together with spectral decomposition in t to obtain a reduced rank GP model for the time-varying drift function (Hartikainen and Särkkä, 2010; Hostettler et al., 2017). The resulting model is suitable for recursive Bayesian estimation using, for example, Kalman or particle filtering and scales well with respect to the temporal domain. Furthermore, estimating the hyperparameters can readily be achieved by using standard approaches such as maximizing the marginal likelihood, which is used in this paper, or particle Markov chain Monte Carlo (Svensson et al., 2016; Hostettler et al., 2017). The proposed model is evaluated on two nonlinear benchmark problems, the Bouc–Wen oscillator and the cascaded tanks systems (Schoukens and Noël, 2017).

⋆ Financial support of the Academy of Finland under grants no. #266940 and #295080 is hereby gratefully acknowledged.

2. GAUSSIAN PROCESSES

In this section we briefly review GP regression, eigenfunction expansion of the covariance function, as well as spectral decomposition to obtain a GP formulation suitable for sequential inference.

2.1 Gaussian Process Regression

A GP is a random function defined as (Rasmussen and Williams, 2006)

$$f(\boldsymbol{x}, t) \sim \mathcal{GP}(m(\boldsymbol{x}, t), k(\boldsymbol{x}, t, \boldsymbol{x}', t')) \tag{3}$$

where $\boldsymbol{x}$ and t are the function's inputs (t is considered to be the time here), $m(\boldsymbol{x}, t)$ is the process' mean function (without loss of generality assumed to be zero for the remainder of this paper), and $k(\boldsymbol{x}, t, \boldsymbol{x}', t')$ is the covariance function. Both $m(\boldsymbol{x}, t)$ and $k(\boldsymbol{x}, t, \boldsymbol{x}', t')$ are parametrized by a set of hyperparameters $\boldsymbol{\theta}$. This parametrization is implicit and we will return to it in Section 4.

Assume now that we are given the observations of the function $\boldsymbol{f}_{1:N} = [f_1 \; f_2 \; \dots \; f_N]^\mathsf{T}$ (where $f_n \triangleq f(\boldsymbol{x}_n, t_n)$ and the observations are assumed to be noise free for simplicity of presentation) as well as the corresponding inputs $x_{1:N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ and $t_{1:N} = \{t_1, t_2, \dots, t_N\}$. Then, the predictive distribution of f_k ($k \notin \{1, \dots, N\}$) given the test inputs $\{\boldsymbol{x}_k, t_k\}$ is

$$p(f_k \mid \boldsymbol{f}_{1:N}) = \mathcal{N}(f_k; \mu_{k|1:N}, \sigma^2_{k|1:N}) \tag{4}$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian probability density function with mean vector μ and covariance matrix $\boldsymbol{\Sigma}$. Furthermore, the predictive moments are

$$\mu_{k|1:N} = \boldsymbol{K}_{k,1:N} \boldsymbol{K}^{-1}_{1:N,1:N} \boldsymbol{f}_{1:N} \tag{5a}$$

$$\sigma^2_{k|1:N} = K_{k,k} - \boldsymbol{K}_{k,1:N} \boldsymbol{K}^{-1}_{1:N,1:N} \boldsymbol{K}_{1:N,k} \tag{5b}$$

with $\boldsymbol{K}_{k,1:N} \triangleq k(\boldsymbol{x}_k, t_k, x_{1:N}, t_{1:N})$ and $\boldsymbol{K}_{1:N,1:N} = k(x_{1:N}, t_{1:N}, x_{1:N}, t_{1:N})$. Hence, the GP allows us to make principled statistical predictions of the function value f_k for the test inputs $\{\boldsymbol{x}_k, t_k\}$, based on a set of training inputs $\{x_{1:N}, t_{1:N}\}$ and outputs $\boldsymbol{f}_{1:N}$. Please see Rasmussen and Williams (2006) for a more thorough introduction to GPs.

For the remainder of this paper, we will make use of a particular class of covariance functions, namely covariance functions that are separable in $\boldsymbol{x}$ and t as well as stationary in t, that is, covariance functions of the form

$$k(\boldsymbol{x}, t, \boldsymbol{x}', t') = k_S(\boldsymbol{x}, \boldsymbol{x}') k_T(\tau), \tag{6}$$

where $\tau = t - t'$. This assumption is not strictly necessary and does not impose significant restrictions, but it simplifies the further derivations. The choice of covariance function is generally up to the user, and thus, any such choice is valid and has also been made before (Hartikainen and Särkkä, 2010; Carron et al., 2016).

2.2 Basis Function Expansion

Unfortunately, the prediction in (5) scales according to $\mathcal{O}(N^3)$ due to the inversion of the $N \times N$ covariance matrix $\boldsymbol{K}_{1:N,1:N}$, which is prohibitive for large N. This will be particularly challenging when we use GPs to model the drift function in Section 3, since in this case, the computational complexity will grow with time (Ruttor et al., 2013). Approaches to mitigate this problem include, among others, kernel approximations using the generalized power spectral density and random feature approximation (Zorzi and Chiuso, 2017) or basis function expansions (Rasmussen and Williams, 2006; Solin and Särkkä, 2014). In this work, we consider a time-varying Karhunen–Loève expansion for $f(\boldsymbol{x}, t)$ of the form

$$f(\boldsymbol{x}, t) = \sum_{j=0}^{\infty} \alpha_{j,t} \psi_j(\boldsymbol{x}) \tag{7}$$

where $\psi_j(\boldsymbol{x})$ is the jth orthonormal eigenfunction for the integral operator defined by the kernel $k_S(\boldsymbol{x}, \boldsymbol{x}')$. That is (Mercer, 1909),

$$\langle \psi_i(\boldsymbol{x}), \psi_j(\boldsymbol{x}) \rangle = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases} \tag{8a}$$

$$\langle k_S(\boldsymbol{x}, \boldsymbol{x}'), \psi_j(\boldsymbol{x}') \rangle = \lambda_j \psi_j(\boldsymbol{x}), \tag{8b}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and λ_j is the eigenvalue associated with the eigenfunction $\psi_j(\boldsymbol{x})$. Furthermore, the time-varying coefficients $\alpha_{j,t}$ in (7) are given by

$$\alpha_{j,t} = \langle f(\boldsymbol{x}, t), \psi_j(\boldsymbol{x}) \rangle. \tag{9}$$

Note that

$$\begin{aligned} \operatorname{Cov}\{\alpha_{i,t}, \alpha_{j,t'}\} &= \operatorname{Cov}\{\langle f(\boldsymbol{x}, t), \psi_i(\boldsymbol{x}) \rangle, \langle f(\boldsymbol{x}, t'), \psi_j(\boldsymbol{x}) \rangle\} \\ &= k_T(t, t') \langle \psi_i(\boldsymbol{x}), \langle k(\boldsymbol{x}, \boldsymbol{x}'), \psi_j(\boldsymbol{x}') \rangle \rangle \\ &= k_T(t, t') \langle \psi_i(\boldsymbol{x}), \lambda_j \psi_j(\boldsymbol{x}) \rangle \\ &= k_T(t, t') \lambda_j \delta_{ij}, \end{aligned}$$

where δ_{ij} is the Kronecker delta function. Thus, the coefficients $\alpha_{j,t}$ are iid Gaussian processes of the form

$$\alpha_{j,t} \sim \mathcal{GP}(0, k_{\alpha,j}(t, t')) \tag{10}$$

with

$$k_{\alpha,j}(t, t') = \lambda_j k_T(t, t'), \tag{11}$$

which follows from the fact that $\psi_j(\boldsymbol{x})$ are eigenfunctions associated to the space kernel $k_S(\boldsymbol{x}, \boldsymbol{x}')$. Note that for general basis expansions the coefficients are indeed correlated.

Finally, it follows that

$$\begin{aligned} k(\boldsymbol{x}, t, \boldsymbol{x}', t') &= \operatorname{Cov}\{f(\boldsymbol{x}, t), f(\boldsymbol{x}', t')\} \\ &= \operatorname{Cov}\left\{ \sum_{i=0}^{\infty} \alpha_{i,t} \psi_i(\boldsymbol{x}), \sum_{j=0}^{\infty} \alpha_{j,t} \psi_j^*(\boldsymbol{x}') \right\} \\ &= \sum_{j=0}^{\infty} \lambda_j k_T(t, t') \psi_j(\boldsymbol{x}) \psi_j^*(\boldsymbol{x}') \\ &= k_T(t, t') \sum_{j=0}^{\infty} \lambda_j \psi_j(\boldsymbol{x}) \psi_j^*(\boldsymbol{x}') \end{aligned} \tag{12}$$

where the superscript $*$ denotes the complex conjugate. Note that in practice, the infinite sum over j can not be realized and thus, it has to be truncated at some upper limit $J \ll \infty$, which introduces an approximation error.

Example 1 below demonstrates one possible eigenfunction decomposition for stationary covariance functions.

Example 1. Consider a GP

$$f(x, t) \sim \mathcal{GP}(0, k_S(x, x') k_T(t, t'))$$

with the scalar inputs x and t and stationary covariance function in x, that is, $k_S(x, x') = k(x - x')$. In this case, the eigenfunctions on the domain $x \in [-\gamma, \gamma]$ with respect to the Lebesque measure are given by the unitary Fourier basis functions (Rasmussen and Williams, 2006)

$$\psi_j(x) = \frac{1}{\sqrt{\gamma}} \exp\left(\frac{\mathrm{i}\, j2\pi x}{\gamma}\right). \tag{13}$$

Furthermore, the eigenvalues are found from (8) as follows:

$$\begin{aligned}
\lambda_j \psi_j(x) &= \langle k_S(x - x'), \psi_j(x') \rangle \\
&= \int_{-\gamma}^{\gamma} k_S(x - x'), \psi_j(x') \mathrm{d}x' \\
&= \int_{-\gamma}^{\gamma} k_S(x - x') \frac{1}{\sqrt{\gamma}} \exp\left(\frac{\mathrm{i}\, j2\pi x'}{\gamma}\right) \mathrm{d}x' \\
&= \frac{1}{\sqrt{\gamma}} \exp\left(\frac{\mathrm{i}\, j2\pi x}{\gamma}\right) \int_{-\gamma}^{\gamma} k_S(x) \exp\left(-\frac{\mathrm{i}\, j2\pi x}{\gamma}\right) \mathrm{d}x.
\end{aligned}$$

Thus, the eigenvalues are given by

$$\lambda_j = \int_{-\gamma}^{\gamma} k_S(x) \exp\left(-\frac{\mathrm{i}\, j2\pi x}{\gamma}\right) \mathrm{d}x,$$

that is, the jth Fourier series coefficient.

2.3 Spectral Decomposition

In addition to the rank reduction in the input $\boldsymbol{x}$ using basis function expansion, we also propose to use spectral decomposition of the resulting GPs $\alpha_{j,t}$ in (10) (Hartikainen and Särkkä, 2010). This requires that the resulting covariance function $k_{\alpha,j}(t, t')$ is in fact stationary, that is, $k_{\alpha,j}(t, t') = k_{\alpha,j}(\tau)$ with $\tau = t - t'$.

In this case, the spectral density $S_{\alpha,j}(\omega)$ of $k_{\alpha,j}(\tau)$ can be decomposed (either exactly or approximately) into a white process with spectral density $q_{\alpha,j}$ and a linear system with frequency response function $H(\mathrm{i}\,\omega_t)$ such that (Papoulis, 1984; Hartikainen and Särkkä, 2010)

$$S_{\alpha,j}(\omega) = q_{\alpha,j} H(\mathrm{i}\,\omega) H(\mathrm{i}\,\omega)^*. \tag{14}$$

Thus, $\alpha_{j,t}$ can equivalently be written as the output of a linear system that is driven by a white process. This can be written as the stochastic differential equation (Papoulis, 1984; Hartikainen and Särkkä, 2010)

$$\begin{aligned}
\mathrm{d}\boldsymbol{z}_{j,t} &= \boldsymbol{A}_j \boldsymbol{z}_{j,t} \mathrm{d}t + \boldsymbol{B}_j \mathrm{d}\varepsilon_{j,t}, \\
\alpha_{j,t} &= \boldsymbol{C}_j \boldsymbol{z}_{j,t},
\end{aligned} \tag{15}$$

where the matrix $\boldsymbol{A}_j$, vectors $\boldsymbol{B}_j$ and $\boldsymbol{C}_j$, as well as the representation in the vector $\boldsymbol{z}_{j,t}$ are completely defined by the linear system $H(\mathrm{i}\,\omega_t)$ and the particular state space representation (e.g., control canonical form or companion form), and $\varepsilon_{j,t}$ denotes Brownian motion with diffusion coefficient $q_{\alpha,j}$. A simple example for how this is achieved in practice is shown in Example 2 below, please refer to Hartikainen and Särkkä (2010) for more details.

Example 2. Consider a Gaussian process (10)

$$\alpha_{j,t} \sim \mathcal{GP}(0, k_{\alpha,j}(t, t'))$$

with

$$k_{\alpha,j}(t, t') = \lambda_j k_{\mathrm{OU}}(\tau),$$

where $k_{\mathrm{OU}}(\tau) = \sigma^2 \exp(-|\tau|/\ell)$ is the Ornstein–Uhlenbeck covariance function. The spectral density of this covariance function can be decomposed as (with $\kappa = 1/\ell$)

$$\begin{aligned}
S_{\alpha,j}(\omega) &= \lambda_j \sigma^2 \frac{2\kappa}{\kappa^2 + \omega^2} \\
&= 2\lambda_j \kappa \sigma^2 \frac{1}{(\kappa + \mathrm{i}\,\omega)(\kappa - \mathrm{i}\,\omega)},
\end{aligned}$$

and thus, $H(\mathrm{i}\,\omega) = 1/(\kappa + \mathrm{i}\,\omega)$. Converting $H(\mathrm{i}\,\omega)$ to companion form yields $A = \kappa$, $B = 1$, and $C = 1$, and the diffusion coefficient of the Brownian motion is $2\lambda_j \kappa \sigma^2$.

3. REDUCED RANK GAUSSIAN PROCESS DRIFT MODEL

Having developed the general reduced rank Gaussian process model in the previous section, we now turn our attention to modeling the drift function in (1) in this section. The basic idea is to model the individual drift functions $f_l(\boldsymbol{x}_t, \boldsymbol{u}_t, t)$ in (1) as reduced rank separable GPs as introduced in Section 2, that is, such that

$$f_l(\boldsymbol{x}_t, \boldsymbol{u}_t, t) \sim \mathcal{GP}(0, k_S(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{x}_t', \boldsymbol{u}_t') k_T(\tau)), \tag{16}$$

where the inputs are the time-varying state $\boldsymbol{x}_t$, the deterministic control input $\boldsymbol{u}_t$, and time t.

Using the results from Section 2.2 and (16), the lth component of (1) can be written as

$$\mathrm{d}x_l = \boldsymbol{\Psi}(\boldsymbol{x}_t, \boldsymbol{u}_t) \boldsymbol{\alpha}_{l,t} \mathrm{d}t + \mathrm{d}\beta_{l,t}. \tag{17}$$

Here,

$$\begin{aligned}
\boldsymbol{\Psi}(\boldsymbol{x}_t, \boldsymbol{u}_t) &= \left[\psi_1(\boldsymbol{x}_t, \boldsymbol{u}_t)\ \psi_2(\boldsymbol{x}_t, \boldsymbol{u}_t)\ \ldots\ \psi_J(\boldsymbol{x}_t, \boldsymbol{u}_t)\right], \\
\boldsymbol{\alpha}_{l,t} &= \left[\alpha_{1,t}\ \alpha_{2,t}\ \ldots\ \alpha_{J,t}\right]^{\mathsf{T}},
\end{aligned}$$

and $\alpha_{j,t}$ is as in (10).

Furthermore, using the spectral decomposition in Section 2.3 for each $\alpha_{j,t}$ we obtain

$$\mathrm{d}\tilde{\boldsymbol{z}}_{l,t} = \boldsymbol{A} \tilde{\boldsymbol{z}}_{l,t} \mathrm{d}t + \boldsymbol{B} \mathrm{d}\varepsilon_{l,t}, \tag{18a}$$

$$\boldsymbol{\alpha}_{l,t} = \boldsymbol{C} \tilde{\boldsymbol{z}}_{l,t}, \tag{18b}$$

where

$$\begin{aligned}
\tilde{\boldsymbol{z}}_{l,t} &= \left[\boldsymbol{z}_{1,t}^{\mathsf{T}}\ \boldsymbol{z}_{2,t}^{\mathsf{T}}\ \ldots\ \boldsymbol{z}_{J,t}^{\mathsf{T}}\right]^{\mathsf{T}}, \\
\boldsymbol{\varepsilon}_{l,t} &= \left[\varepsilon_{1,t}\ \varepsilon_{2,t}\ \ldots\ \varepsilon_{J,t}\right]^{\mathsf{T}}, \\
\boldsymbol{A} &= \mathrm{blkdiag}(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_J), \\
\boldsymbol{B} &= \mathrm{blkdiag}(\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_J), \\
\boldsymbol{C} &= \mathrm{blkdiag}(\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_J),
\end{aligned}$$

and $\boldsymbol{Q}_{\boldsymbol{\varepsilon},l} = \mathrm{diag}(q_{\alpha,1}, q_{\alpha,2}, \ldots, q_{\alpha,J})$ is the diffusion matrix for $\boldsymbol{\varepsilon}_{l,t}$.

Then, combining (17) and (18) yields the complete reduced rank GP drift model

$$\mathrm{d}x_{l,t} = \boldsymbol{\Psi}(\boldsymbol{x}_t, \boldsymbol{u}_t) \boldsymbol{C} \tilde{\boldsymbol{z}}_{l,t} \mathrm{d}t + \mathrm{d}\beta_{l,t}, \tag{19a}$$

$$\mathrm{d}\tilde{\boldsymbol{z}}_{l,t} = \boldsymbol{A} \tilde{\boldsymbol{z}}_{l,t} \mathrm{d}t + \boldsymbol{B} \mathrm{d}\varepsilon_{l,t}. \tag{19b}$$

Finally, let $\boldsymbol{\xi}_t = \left[\boldsymbol{x}_t^{\mathsf{T}}\ \boldsymbol{z}_t^{\mathsf{T}}\right]^{\mathsf{T}}$ denote the augmented system where $\boldsymbol{z}_t$ is a vector of all $\tilde{\boldsymbol{z}}_{l,t}$ (for $l = 1, \ldots, N_x$), then $\boldsymbol{\xi}_t$ is governed by the following SDE

$$\mathrm{d}\boldsymbol{\xi}_t = \boldsymbol{g}(\boldsymbol{\xi}_t, \boldsymbol{u}_t, t) \mathrm{d}t + \mathrm{d}\boldsymbol{w}_t, \tag{20}$$

where the combined drift $\boldsymbol{g}(\boldsymbol{\xi}_t, \boldsymbol{u}_t, t)$ is given by

$$\boldsymbol{g}(\boldsymbol{\xi}_t, \boldsymbol{u}_t, t) = \begin{bmatrix} (\boldsymbol{I}_{N_{\boldsymbol{x}}} \otimes \boldsymbol{\Psi}(\boldsymbol{x}_t, \boldsymbol{u}_t) \boldsymbol{C}) \boldsymbol{z}_t \\ \boldsymbol{I}_{N_{\boldsymbol{x}}} \otimes A \boldsymbol{z}_t \end{bmatrix},$$

and $\boldsymbol{w}_t$ is a Wiener process with instantaneous covariance, $\boldsymbol{Q}_{\boldsymbol{w}}$, given by

$$\boldsymbol{Q}_{\boldsymbol{w}} = \mathrm{blkdiag}(\boldsymbol{Q}, \boldsymbol{Q}_{\boldsymbol{\varepsilon},1}, \ldots, \boldsymbol{Q}_{\boldsymbol{\varepsilon},N_{\boldsymbol{x}}}).$$

4. ESTIMATION

In order to fit the model to a given data set $y_{1:N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$, the Gaussian process hyperparameters, $\boldsymbol{\theta}$, need to be estimated. Here, a maximum a posteriori approach is taken along with the prediction error decomposition (Ljung, 1997; Särkkä, 2013),

$$p(y_{1:N} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{y}_n \mid y_{1:n-1}, \boldsymbol{\theta}), \tag{21}$$

where $y_{k:n} = \emptyset$ for $k > n$ by convention. The posterior distribution for $\boldsymbol{\theta}$ is then

$$p(\boldsymbol{\theta} \mid y_{1:N}) \propto p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\boldsymbol{y}_n \mid y_{1:n-1}, \boldsymbol{\theta}). \tag{22}$$

The prediction densities in (21) can be approximated with an extended Kalman filter, similar to the approach in Kristensen et al. (2004). That is, let $\hat{\boldsymbol{\xi}}_{t|s}(\boldsymbol{\theta})$ and $\boldsymbol{P}_{t|s}(\boldsymbol{\theta})$ be the mean and covariance matrix of the state $\boldsymbol{\xi}_t$ conditioned on $\boldsymbol{\theta}$ (in the sequel the dependency on $\boldsymbol{\theta}$ is omitted for notational convenience). The mean and covariance of $\boldsymbol{\xi}_t$ then satisfy the following differential equation on the interval $[t_n, t_{n-1}[$

$$\frac{\mathrm{d}\hat{\boldsymbol{\xi}}_{t|t_{n-1}}}{\mathrm{d}t} \approx \boldsymbol{g}(\hat{\boldsymbol{\xi}}_{t|t_{n-1}}) \tag{23a}$$

$$\frac{\mathrm{d}\boldsymbol{P}_{t|t_{n-1}}}{\mathrm{d}t} \approx \boldsymbol{G}_{\boldsymbol{\xi},t}\boldsymbol{P}_{t|t_{n-1}} + \boldsymbol{P}_{t|t_{n-1}}\boldsymbol{G}_{\boldsymbol{\xi},t}^{\mathsf{T}} + \boldsymbol{Q}_{\boldsymbol{w}} \tag{23b}$$

where $\boldsymbol{G}_{\boldsymbol{\xi},t}$ is the Jacobian of $\boldsymbol{g}(\boldsymbol{\xi}_t, \boldsymbol{u}_t, t)$ evaluated at $\hat{\boldsymbol{\xi}}_{t|t_{n-1}}$. For high frequency data, an Euler discretization is sufficient to solve (23). Furthermore, note that (23) is also the method for predicting future data for a given $\boldsymbol{\theta}$.

The measurement $\boldsymbol{y}_n$ then follows a Gaussian distribution,

$$\boldsymbol{y}_n \mid y_{1:n-1} \sim \mathcal{N}(\hat{\boldsymbol{y}}_{n|1:n-1}, \boldsymbol{S}_{n|1:n-1}), \tag{24}$$

where

$$\boldsymbol{H}_{\boldsymbol{\xi}} = [\boldsymbol{H} \ \boldsymbol{0}_{N_{\boldsymbol{z}}}], \tag{25a}$$

$$\hat{\boldsymbol{y}}_{n|1:n-1} = \boldsymbol{H}_{\boldsymbol{\xi}}\boldsymbol{\xi}, \tag{25b}$$

$$\boldsymbol{S}_{n|1:n-1} = \boldsymbol{H}_{\boldsymbol{\xi}}\boldsymbol{P}_{t_n|t_{n-1}}\boldsymbol{H}_{\boldsymbol{\xi}}^{\mathsf{T}} + \boldsymbol{R}. \tag{25c}$$

The state estimate at t_n given measurements up to t_n is given by

$$\boldsymbol{K}_n = \boldsymbol{P}_{t_n|t_{n-1}}\boldsymbol{H}_{\boldsymbol{\xi}}^{\mathsf{T}}\boldsymbol{S}_{n|1:n-1}^{-1}, \tag{26a}$$

$$\hat{\boldsymbol{\xi}}_{t_n|t_n} = \hat{\boldsymbol{\xi}}_{t_n|t_{n-1}} + \boldsymbol{K}_n(\boldsymbol{y}_n - \hat{\boldsymbol{y}}_{n|1:n-1}), \tag{26b}$$

$$\boldsymbol{P}_{t_n|t_N} = \boldsymbol{P}_{t_n|t_{n-1}} - \boldsymbol{K}_n\boldsymbol{S}_{n|1:n-1}\boldsymbol{K}_n^{\mathsf{T}}. \tag{26c}$$

The posterior for $\boldsymbol{\theta}$ can then be evaluated (up to a normalization constant) in a recursive manner and parameter estimation can be done by standard optimization methods.

5. RESULTS

We evaluate the proposed approach on two nonlinear system identification benchmark datasets, namely the Bouc–Wen and the cascaded tanks benchmarks (Noël and Schoukens, 2016; Schoukens et al., 2016; Schoukens and Noël, 2017). In both examples, the hyperparameters of the GP priors are estimated by maximizing the marginal log-posterior of the training data as discussed in Section 4 and the predictive performance of the model is evaluated on the validation datasets. The performance is measured using the root mean squared error (RMSE) of the one-step ahead prediction

$$e_{\mathrm{RMS}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_{n|n-1})^2} \tag{27}$$

where $\hat{y}_{n|n-1}$ is the mean of the predictive distribution. Since the RMSE (27) is scale-dependent, we also evaluate and compare the proposed approach in terms of the coefficient of determination defined as

$$R^2 = 1 - \frac{e_{\mathrm{RMS}}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2}}, \tag{28}$$

Table 1. Comparison of the RMSE and R^2 values for the Bouc–Wen benchmark.

Model	Multisine		Swept sine	
	RMSE	R^2	RMSE	R^2
AR(1)	21.6×10^{-5}	0.676	20.9×10^{-5}	0.685
BLA[1]	1.13×10^{-5}	0.983	0.698×10^{-5}	0.989
Volterra[1]	0.895×10^{-5}	0.986	0.347×10^{-5}	0.994
GP drift	0.580×10^{-5}	0.991	0.096×10^{-5}	0.998

[1]Schoukens and Griesing-Scheiwe (2016)

where $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ is the sample mean.

The method is implemented as m-code in Matlab and run on Linux desktop machine with a 3.4 GHz Intel Xeon E3 central processing unit and 16 GB random access memory.

5.1 Bouc–Wen System

The Bouc–Wen system is a nonlinear oscillator with hysteretic behavior and thus, a system with time-varying nonlinearity. The model, together with the governing differential equation and further details about the benchmark problem can be found in Noël and Schoukens (2016) and Schoukens and Noël (2017).

In this case, we incorporate domain knowledge (i.e. the system being a second order mechanical system) and only model the drift of the second state as a GP such that

$$\mathrm{d}\begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} x_{2,t} \\ f(\boldsymbol{x}_t, u_t, t) \end{bmatrix}\mathrm{d}t + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\mathrm{d}\beta_t.$$

We choose the covariance kernel of $f(\boldsymbol{x}_t, u_t, t)$ as a sum of squared exponential kernels in the variables $\boldsymbol{x}_t$ and u_t times an Ornstein–Uhlenbeck kernel in t, such that

$$k(\boldsymbol{x}_t, t, \boldsymbol{x}_t', t') = (k_{\mathrm{SE}}(\boldsymbol{x}_t, \boldsymbol{x}_t') + k_{\mathrm{SE}}(u_t, u_t'))k_{\mathrm{OU}}(t, t'),$$

where the squared exponential kernel is given by

$$k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^{\mathsf{T}}\Lambda^{-1}(\boldsymbol{x} - \boldsymbol{x}')\right)$$

with $\Lambda = \mathrm{diag}(\ell_1^2, \ell_2^2, \ldots, \ell_{N_{\boldsymbol{x}}}^2)$. (Note that this yields a model similar to the one used in Nielsen et al. (2000).) In this case, the hyperparameters are the length scales in $\boldsymbol{x}_t$ (two), the length scale in u_t, as well as the temporal variance and length scale, and the diffusion coefficient of β_t, which yields a total of six parameters to optimize. The number of eigenfunctions was chosen based on a pre-study such that the covariance function approximation is well-supported with few negligible eigenvalues. This led to 25 eigenfunctions of the form (13) for expanding $k_{\mathrm{SE}}(\boldsymbol{x}_t, \boldsymbol{x}_t')$ and 5 for $k_{\mathrm{SE}}(u_t, u_t')$, and $\gamma = 5$ in both cases.

The training data is generated using the simulator provided by the nonlinear benchmark data set (Noël and Schoukens, 2016). We use a random phase multisine excitation with amplitude $A = 50$, $N = 4096$ samples per period, and $P = 1$ periods (Pintelon and Schoukens, 2001). The predictive performance of the model is evaluated on the two test datasets, one with a multisine excitation and one with a swept sine excitation, as provided by the benchmark dataset, see Noël and Schoukens (2016) for details.

The resulting validation RMSEs are 0.580×10^{-5} and 0.096×10^{-5} for the multisine and swept sine excitation

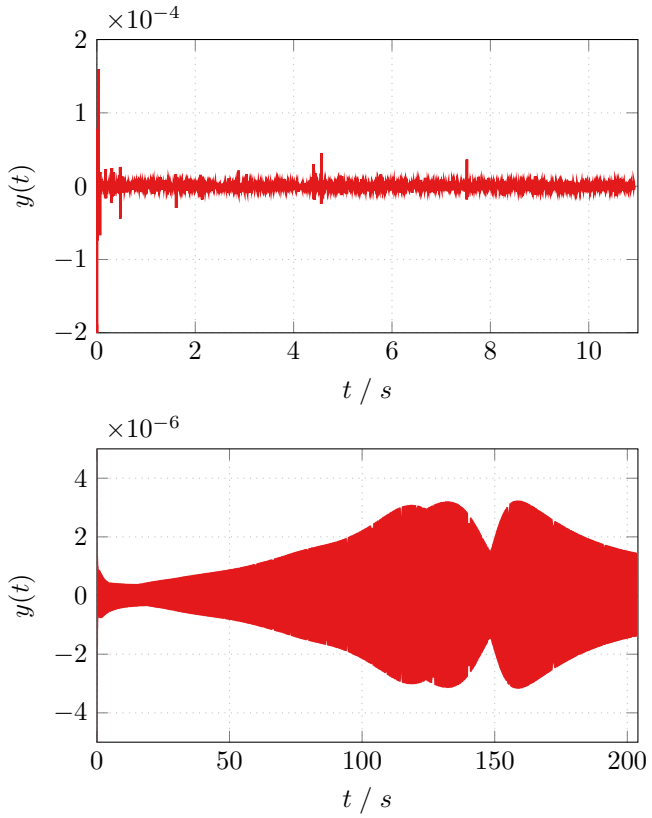

Fig. 1. Prediction error for the multisine (top) and swept sine (bottom) excitations for the Bouc–Wen validation datasets.

test datasets, respectively. Table 1 shows a comparison of the proposed model (GP drift, last row) to the naïve first order autoregressive predictor $\hat{y}_{n|n-1} = y_{n-1}$ (AR(1)), the best linear approximation (BLA; Schoukens and Griesing-Scheiwe (2016)), and the Volterra feedback model (Schoukens and Griesing-Scheiwe, 2016). As it can be seen, the proposed method outperforms the approaches from the literature in this example. Furthermore, Figure 1 shows the prediction error for the two validation datasets. In this example, estimation of the hyperparameters by maximizing the marginal log-posterior takes about 310 s.

The Gaussian process prior for the nonlinear drift function imposes a smoothness assumption on the model. In this example, this smoothness captures the nonlinearity well and thus, good performance is achieved. Furthermore, allowing for time-varying drift, the hysteretic behavior of the benchmark system is captured by the model, too.

5.2 Cascaded Tanks

The second example is the cascaded tanks system which is a very common instructional example, in, for example, control engineering. The system consists of two tanks where the first (upper) tank is fed water from a basin through a pump (control input). The upper tank has an outlet at the bottom, which feeds water into the second (lower) tank. The lower tank again has an outlet at the bottom leading to the reservoir from where the pump is fed. In this benchmark problem, both training and validation datasets are provided. For a more detailed description of the benchmark problem

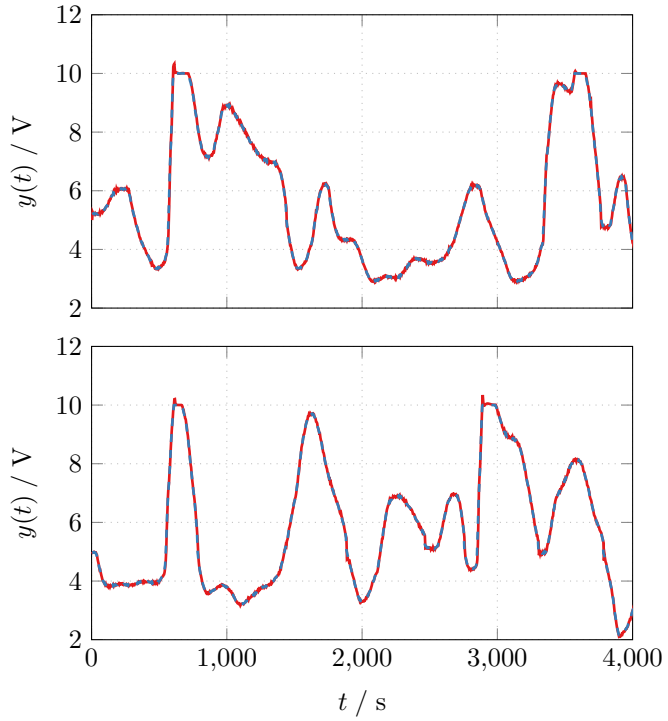

Fig. 2. Cascaded tanks benchmark one step ahead predictions (——) together with the measured level (- - -) for the training set (top) and validation set (bottom).

and the associated identification challenges see Schoukens et al. (2016) or Schoukens and Noël (2017).

In this case, we model the system as a coupled system as

$$\mathrm{d}x_{1,t} = f_1(x_{1,t}, u_t, t)\mathrm{d}t + \mathrm{d}\beta_{1,t} \tag{29a}$$
$$\mathrm{d}x_{2,t} = f_2(x_{1,t}, x_{2,t}, t)\mathrm{d}t + \mathrm{d}\beta_{2,t} \tag{29b}$$

where $f_1(x_{1,t}, u_t, t)$ and $f_2(x_{1,t}, x_{2,t}, t)$ are Gaussian processes as discussed in Sections 2–3, and $\beta_{1,t}$ and $\beta_{2,t}$ are Brownian motions with diffusion coefficients q_1 and q_2, respectively. We use the product of a squared exponential and Ornstein–Uhlenbeck kernels as the covariance function, that is,

$$k(\boldsymbol{x}, \boldsymbol{x}', t, t') = k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}')k_{\mathrm{OU}}(t, t'). \tag{30}$$

Furthermore, a two-dimensional Fourier basis with 25 basis functions is used, and $\lambda = 5$.

In this example, the validation RMSE is 57.6×10^{-3} and a comparison of the RMSE and R^2 values of the proposed method to the performance of methods from the literature is given in Table 2. In this case, the proposed approach (last row) performs worse than the compared models except for the naïve AR(1) model. Figure 2 shows the measured tank level for the lower tank together with the predicted level for both the training data (top) as well as the validation data (bottom). From this, it appears that the fitted model is capable of accurately predict the tank level in most regimes. However, as discussed before, an important property of the proposed model is the smoothness assumption of the drift function by the GP prior. This assumption is, however, violated in this example as the tanks benchmark does not only include smooth nonlinearities (the outflow of the tanks) but also includes hard nonlinearities (tank overflows). This can be seen by the overshooting predictions in both the training data (around $t \approx 750$ s and $t \approx 3500$ s) as well

Table 2. Comparison of the RMSE and R^2 values for the cascaded tanks benchmark.

Model	RMSE	R^2
AR(1)	185.9×10^{-3}	0.911
BLA[1]	55.6×10^{-3}	0.974
Volterra[1]	49.4×10^{-3}	0.991
GP drift	57.6×10^{-3}	0.972

[1]Schoukens and Griesing-Scheiwe (2016)
[2]Svensson and Schön (2017)

as the validation data (around $t \approx 750\,\mathrm{s}$ and $t \approx 2900\,\mathrm{s}$). These errors contribute the most to the prediction RMSE. Finally, it should also be noted that linear models tend to perform quite well in this benchmark in most regimes, see, for example, Svensson and Schön (2017).

6. CONCLUSIONS

In this paper, we proposed modeling the nonlinear time-varying drift function in SDE models using reduced rank Gaussian processes. The approach is suitable for online inference using Bayesian filtering methods and can readily be extended to nonlinear observations of the state. The numerical illustrations showed that the model performs well when the underlying assumptions such as smoothness of the drift hold. On the other hand, if these assumptions are violated, for example by hard nonlinearities as in the cascaded tanks example, the model is less appropriate.

It is important to point out that the proposed eigenfunction expansion can be challenging itself. In particular, high-dimensional inputs may require a large number of basis functions if the higher order eigenfunctions are the Cartesian product of the lower dimensional ones. This can, in part, be mitigated by incorporating domain knowledge to reduce the model complexity. For example, in many mechanical problems, the system is governed by an $N_{\boldsymbol{x}}$th order SDE, which eliminates the need of modeling the drift for the $N_{\boldsymbol{x}} - 1$ first states.

REFERENCES

Batz, P., Ruttor, A., and Opper, M. (2017). Approximate Bayes learning of stochastic differential equations. *ArXiv e-prints*. ArXiv:1702.05390v1.

Carron, A., Todescato, M., Carli, R., Schenato, L., and Pillonetto, G. (2016). Machine learning meets Kalman filtering. In *55th IEEE Conference on Decision and Control (CDC)*, 4594–4599.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 379–384.

Hostettler, R., Särkkä, S., and Godsill, S.J. (2017). Rao–Blackwellized particle MCMC for parameter estimation in spatio-temporal Gaussian processes. In *27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Tokyo, Japan.

Iversen, E.B., Morales, J.M., Møller, J.K., and Madsen, H. (2014). Probabilistic forecasts of solar irradiance using stochastic differential equations. *Environmetrics*, 25(3), 152–164.

Kristensen, N.R., Madsen, H., and Jørgensen, S.B. (2004). Parameter estimation in stochastic grey-box models. *Automatica*, 40(2), 225–237.

Lindström, E., Madsen, H., and Nielsen, J.N. (2015). *Statistics for Finance*. Chapman and Hall/CRC.

Ljung, L. (1997). *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ, USA.

Madsen, H. and Holst, J. (1995). Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1), 67–79.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209(441-458), 415–446.

Nielsen, H.A., Nielsen, T.S., Joensen, A.K., Madsen, H., and Holst, J. (2000). Tracking time-varying-coefficient functions. *International Journal of Adaptive Control and Signal Processing*, 14(8), 813–828.

Noël, J.P. and Schoukens, M. (2016). Hysteretic benchmark with a dynamic nonlinearity. In *Workshop on Nonlinear System Identification Benchmarks*, 7–14. Brussels, Belgium.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Pintelon, R. and Schoukens, J. (2001). *System Identification: A Frequency Domain Approach*. Wiley-IEEE Press, Piscataway, NJ, USA.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Ruttor, A., Batz, P., and Opper, M. (2013). Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems 26*, 2040–2048.

Schoukens, M. and Griesing-Scheiwe, F. (2016). Modeling nonlinear systems using a volterra feedback model. In *Workshop on Nonlinear System Identification Benchmarks*.

Schoukens, M., Mattsson, P., Wigren, T., and Noël, J.P. (2016). Cascaded tanks benchmark combining soft and hard nonlinearities. In *Workshop on Nonlinear System Identification Benchmarks*, 20–23. Brussels, Belgium.

Schoukens, M. and Noël, J.P. (2017). Three benchmarks addressing open challenges in nonlinear system identification. In *20th IFAC World Congress*, 446–451.

Solin, A. and Särkkä, S. (2014). Hilbert space methods for reduced-rank Gaussian process regression. ArXiv:1401.5508.

Svensson, A. and Schön, T.B. (2017). A flexible state-space model for learning nonlinear dynamical systems. *Automatica*, 80, 189–199.

Svensson, A., Solin, A., Särkkä, S., and Schön, T. (2016). Computationally efficient Bayesian learning of Gaussian process state space models. In *19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, 213–221. Cadiz, Spain.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Zorzi, M. and Chiuso, A. (2017). The harmonic analysis of kernel functions. *ArXiv e-prints*. ArXiv:1703.05216.

Øksendal, B. (2010). *Stochastic Differential Equaitons: An Introduction with Applications*. Springer, 6 edition.