OUTPUT-DEPENDENT GAUSSIAN PROCESS STATE-SPACE MODEL

Zhidi Lin[†‡]   *Lei Cheng*[⋄]   *Feng Yin*[†](✉)   *Lexi Xu*[∘]   *Shuguang Cui*[†‡]

† School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
‡ Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen, China
⋄ College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
∘ Research Institute, China United Network Communications Corporation, Beijing, China

ABSTRACT

Gaussian process state-space model (GPSSM) is a fully probabilistic state-space model that has attracted much attention over the past decade. However, the outputs of the transition function in the existing GPSSMs are assumed to be independent, meaning that the GPSSMs cannot exploit the inductive biases between different outputs and lose certain model capacities. To address this issue, this paper proposes an output-dependent and more realistic GPSSM by utilizing the well-known, simple yet practical linear model of coregionalization (LMC) framework to represent the output dependency. To jointly learn the output-dependent GPSSM and infer the latent states, we propose a variational sparse GP-based learning method that only gently increases the computational complexity. Experiments on both synthetic and real datasets demonstrate the superiority of the output-dependent GPSSM in terms of learning and inference performance.

Index Terms— Gaussian process state-space model, linear model of coregionalization, variational inference, sparse Gaussian process.

1. INTRODUCTION

A well-established probabilistic tool for modeling the underlying dynamical system of sequential data is the state-space model (SSM), which has been successfully applied in many areas of engineering, statistics, computer science, and economics [1]. For the case when the system dynamics are fairly known, a plethora of out-of-the-box learning and inference methods have been developed over the past decades, such as the Kalman filter (KF) for linear Gaussian dynamic systems, and the particle filter (PF) for nonlinear dynamic systems [1]. However, in some harsh scenarios, such as model-based reinforcement learning, or disease epidemic propagation, the underlying system dynamics cannot be well determined *a priori* [2]. Thus, the dynamics need to be learned from the observed noisy measurements, leading to the emergence of data-driven state-space models [3, 4, 5, 6, 7, 8, 9, 10, 11].

Gaussian processes (GPs), being an eminent Bayesian nonparametric models for machine learning [12, 13, 14], can be adopted as function priors in classical SSM, giving rise to the Gaussian process state-space model (GPSSM) [7]. A carefully selected GP prior provides not only meaningful uncertainty calibration in low data regime but also automatic scaling of model complexity based upon data volume [15, 16]. Due to these appealing properties, GPSSM and its variants have been applied to various applications, such as human motion capture and pedestrian tracking and navigation [17, 18, 19, 20].

Despite the ever-increasing popularity of GPSSM, accurate, simultaneous learning and inference in GPSSM remains a challenging problem. Much progress has been made over the past decade along different paths [6, 7, 8, 9, 10, 11, 21, 22]. Concretely, the first fully

---

✉ The corresponding author is Feng Yin (*yinfeng@cuhk.edu.cn*).

Bayesian learning of GPSSM was proposed in [7] using particle Markov chain Monte Carlo. Variational inference methods were developed based upon the mean-field (MF) assumption to reduce the heavy computational load of the sampling methods [6, 8, 22]. More recent works have been devoted to overcoming the MF assumption for enhanced learning and inference performance [10, 11, 21, 23]. However, all the existing methods utilize independent GPs to model the multi-outputs of the transition function while ignoring their dependencies, which can cause inductive bias between the outputs that cannot be transferred to improve the model generalization [24]. The inference performance can be significantly degraded, especially when the latent states are only partially observed (see Section 4.1). Moreover, high-dimensional data features nowadays are often entangled. For instance, in disease progression prediction application, the disease state of a patient comprises a series of mutually influencing physiological metrics [4]. Therefore, assuming the independence of outputs is simplifying but unrealistic.

In this paper, we aim to address the above-mentioned issues by explicitly modeling the output dependency without sacrificing much computational complexity. The main contributions are summarized as follows. First, we resort to a simple yet practical framework, namely the linear model of coregionalization (LMC) [25, 26] to encode dependency among outputs of the GP-based transition function in GPSSM. To the best of our knowledge, this is the first study on output-dependent GPSSMs. Second, we propose a variational learning method based upon the sparse GP [27], in which learning and inference only gently increase the computational complexity. Third, experimental results obtained using real and synthetic datasets corroborate that the proposed output-dependent GPSSM outperforms various benchmark methods, including the output-independent GPSSM [10] and the deep state-space model (DSSM) [3].

The remainder of this paper is organized as follows. Some preliminaries related to GPSSM are provided in Section 2. In Section 3, we introduce our proposed output-dependent GPSSM and detail the learning and inference algorithm. Numerical results are provided in Section 4. Finally, we conclude the paper in Section 5.

2. PRELIMINARIES

2.1. Gaussian Process

Gaussian process (GP) defines a collection of random variables indexed by $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, such that any finite collection of these variables follows a joint Gaussian distribution [12]. With this definition, a Gaussian process is typically used to define a distribution over functions $f(\mathbf{x}) : \mathbb{R}^{d_x} \mapsto \mathbb{R}$,

$$f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}),\ k_{\boldsymbol{\theta}_{gp}}\left(\mathbf{x}, \mathbf{x}'\right)\right), \qquad (1)$$

where $\mu(\mathbf{x})$ is a mean function, usually set to be zero in practice; $k_{\boldsymbol{\theta}_{gp}}\left(\mathbf{x}, \mathbf{x}'\right)$ is a covariance function/kernel function; $\boldsymbol{\theta}_{gp}$ is a set of hyperparameters to be tuned for model selection. Following Bayes' theorem, the function prior is combined with new data to obtain an

analytical posterior distribution. More specifically, given a noise-free training dataset $\mathcal{D} = \{X, \mathbf{f}\} = \{\mathbf{x}_i, \mathbf{f}_i\}_{i=1}^n$, the posterior distribution $p(f(\mathbf{x}_*)|\mathbf{x}_*, \mathcal{D})$ at any test input $\mathbf{x}_* \in \mathcal{X}$ is Gaussian, fully characterized by the posterior mean ξ and the posterior variance Ξ:

$$\xi(\mathbf{x}_*) = \mu(\mathbf{x}_*) + \boldsymbol{K}_{\mathbf{x}_*,X} \boldsymbol{K}_{X,X}^{-1} \left(\mathbf{f} - \mu(X)\right), \qquad (2a)$$

$$\Xi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \boldsymbol{K}_{\mathbf{x}_*,X} \boldsymbol{K}_{X,X}^{-1} \boldsymbol{K}_{\mathbf{x}_*,X}^{\top}, \qquad (2b)$$

where $\boldsymbol{K}_{X,X}$ denotes the covariance matrix evaluated on the training input X, and each entry is $[\boldsymbol{K}_{X,X}]_{i,j} = k_{\boldsymbol{\theta}_{gp}}(\mathbf{x}_i, \mathbf{x}_j)$; $\boldsymbol{K}_{\mathbf{x}_*,X}$ denotes the cross covariance matrix between the test input $\mathbf{x}_*$ and the training input X; $\mu(X) = \{\mu(\mathbf{x}_i)\}_{i=1}^n$ denotes the prior mean function evaluated on X.

2.2. Gaussian Process State-Space Model

A generic state-space model (SSM) describes the probabilistic dependence between latent state $\mathbf{x}_t \in \mathbb{R}^{d_x}$ and observation $\mathbf{y}_t \in \mathbb{R}^{d_y}$. Mathematically, it can be written as

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t) + \mathbf{v}_t, \qquad (3a)$$

$$\mathbf{y}_t = g(\mathbf{x}_t) + \mathbf{e}_t, \qquad (3b)$$

where $\mathbf{v}_t$ and $\mathbf{e}_t$ are additive noise terms; $f(\cdot)$ and $g(\cdot)$ are *transition function* and *emission function*, respectively.

Placing a GP prior over the transition function $f(\cdot)$ and assuming a parametric emission function $g(\cdot)$ in SSM leads to the well-known Gaussian process state-space model (GPSSM)[1] [9], which is depicted in Fig. 1 and expressed mathematically as:

$$\mathbf{f}_t = f(\mathbf{x}_{t-1}), \;\; f(\cdot) \sim \mathcal{GP}\left(\boldsymbol{\mu}(\cdot), \boldsymbol{k}_{\boldsymbol{\theta}_{gp}}(\cdot, \cdot)\right), \;\; \mathbf{x}_0 \sim p(\mathbf{x}_0), \quad (4a)$$

$$\mathbf{x}_t \mid \mathbf{f}_t \sim \mathcal{N}\left(\mathbf{x}_t \mid \mathbf{f}_t, \mathbf{Q}\right), \qquad \mathbf{y}_t \mid \mathbf{x}_t \sim \mathcal{N}\left(\mathbf{y}_t \mid \boldsymbol{C}\mathbf{x}_t, \boldsymbol{R}\right), \quad (4b)$$

where the emission model is assumed to be known and linear with the coefficient matrix, $\boldsymbol{C} = [\boldsymbol{I}_{d_y}, \mathbf{0}] \in \mathbb{R}^{d_y \times d_x}$, to reduce the system non-identifiability [9]. In the case of $d_x > d_y$, we say that the latent states are *partially observable*. The state transitions and observations are corrupted by zero-mean Gaussian noise with covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively. If the state dimension $d_x > 1$, the transition $f(\cdot): \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ is typically modeled with d_x mutually independent GPs. More concretely, independent GP priors are placed on each dimension-specific function $f_d(\cdot): \mathbb{R}^{d_x} \to \mathbb{R}$, and

$$\mathbf{f}_t = f(\mathbf{x}_{t-1}) \triangleq \{f_d(\mathbf{x}_{t-1})\}_{d=1}^{d_x} \triangleq \{\mathbf{f}_{t,d}\}_{d=1}^{d_x}, \qquad (5)$$

where each independent GP has its own mean function, kernel function, and hyperparameters. The challenging task in GPSSM is to learn the transition function and noise models, i.e., $\boldsymbol{\theta} = [\boldsymbol{\theta}_{gp}, \boldsymbol{Q}, \boldsymbol{R}]$, and infer the latent states of interest simultaneously.

3. OUTPUT-DEPENDENT GPSSM

In this section, we first point out the issues existing in the GPSSM literature, then propose our output-dependent GPSSM and explain why it is able to overcome these issues. Lastly, we detail the proposed variational learning method for the output-dependent GPSSM.

3.1. Problem Statement and Output-Dependent GPSSM

As depicted in Fig. 2a and described in Section 2.2 (see Eq. (5)), the existing GPSSM works assume the transition function outputs are independent when modeling high-dimensional latent dynamics. The adverse effects of independent modeling are twofold. First, model mismatch can occur, especially when there are strong correlations among the outputs. In fact, high-dimensional latent states in various

---

[1]The GPSSM considered in this paper keeps the same model capacity as the *ones with both transition and emission GPs* while avoiding the severe *non-identifiability* issue. One can refer to [9] (Section 3.2.1) for more details.

Fig. 1: Graphical model of GPSSM. The thick horizontal bar represents a set of fully connected nodes, i.e., the GP.

applications tend to be inherently dependent. For example, in navigation and tracking applications, the latent states comprise physical quantities such as displacement, acceleration, and velocity [1] that are strongly correlated according to physic law. In such applications, the independent outputs assumption will degenerate the learning and inference performance. Second, the inductive bias cannot be transferred between outputs, which limits the model learning capacity [24], especially when the latent states are only partially observed.

To explicitly model the dependency among outputs of the transition function, we propose to apply the linear model of coregionalization (LMC) [25, 26], which is a well-known, simple, yet practical multiple-output GP framework that linearly mixes Q independent latent GPs for modeling multiple dependent outputs simultaneously. More specifically, as depicted in Fig. 2b, the d_x transition outputs $\{\mathbf{f}_{t,d}\}_{d=1}^{d_x}$ are obtained by the linear combinations of Q independent latent GPs, $\mathbf{h}_t = \{\mathbf{h}_{t,q}\}_{q=1}^Q$, where $\mathbf{h}_{t,q} = h_q(\mathbf{x}_{t-1}), h_q(\cdot) \sim \mathcal{GP}(0, k_q(\cdot, \cdot))$, i.e.,

$$\mathbf{f}_{t,d} = f_d(\mathbf{x}_{t-1}) = \sum_{q=1}^Q \boldsymbol{a}_{d,q} \cdot h_q(\mathbf{x}_{t-1}) = \boldsymbol{a}_d^{\top} \mathbf{h}_t, \; d = 1, ..., d_x, \quad (6)$$

where $\boldsymbol{a}_d = [\boldsymbol{a}_{d,1}, \boldsymbol{a}_{d,2}, ..., \boldsymbol{a}_{d,Q}]^{\top} \in \mathbb{R}^Q$ are the dimension-specific coefficients that form the coregionalization matrix $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_{d_x}]^{\top}$. In this way, the Q latent GPs will learn a shared knowledge (inductive bias) of the underlying dynamics, and $\boldsymbol{a}_d$ will adapt the behaviours for the dimension-specific output. Under the LMC framework and the assumption of Q independent latent GPs, the vector-valued transition function follows a Gaussian process prior $f(\mathbf{x}) \sim \mathcal{GP}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{k}_{\boldsymbol{\theta}_{gp}}(\mathbf{x}, \mathbf{x}'))$, where the mean function is $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{0}$, and the multi-output kernel function $\boldsymbol{k}_{\boldsymbol{\theta}_{gp}}(\mathbf{x}, \mathbf{x}')$ has d_x^2 entries, specifically, $[\boldsymbol{k}_{\boldsymbol{\theta}_{gp}}(\mathbf{x}, \mathbf{x}')]_{i,j} = \sum_{q=1}^Q \boldsymbol{a}_{i,q} \boldsymbol{a}_{j,q} k_q(\mathbf{x}, \mathbf{x}'), i, j = 1, 2, ..., d_x$. Compared with the classic (independent) GP modeling, the GP prior with LMC explicitly models the correlation between different outputs (with extra matrix multiplication operations). Thus, the inductive biases among outputs can be transferred to improve the overall model learning and inference capacity, which is beneficial to, e.g., partially observable state inference. Note that the new set of $\boldsymbol{\theta}_{gp}$ includes the coregionalization matrix $\boldsymbol{A}$ and the hyperparameters from all Q independent latent GPs. It is also noteworthy that even though modeling the output dependency improves the GPSSM flexibility, it also brings model identifiability issues (i.e., given $\mathbf{f}_t$ and $\boldsymbol{A}$, the underlying $\mathbf{h}_t$ may not be inferred uniquely). An identifiable model is critical to state inference. The following corollary indicates that severe non-identifiability can be eliminated by carefully selecting the Q parameter.

Corollary 1. *The proposed output-dependent GPSSM does not compromise the model identifiability if $Q \le d_x$ and $\text{rank}(\boldsymbol{A}) = Q$.*

Proof. When $Q \le d_x$ and $\boldsymbol{A}$ is full column rank, given $\mathbf{f}_t$ and $\boldsymbol{A}$ for the underlying linear system $\boldsymbol{A}\mathbf{h}_t = \mathbf{f}_t$, the estimate $\hat{\mathbf{h}}_t = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\mathbf{f}_t$ gives an exact solution if true $\mathbf{h}_t$ exists. Then, by following the theorem in Section 3.2.1, [9], severe model idenfiability issue can be solved. □

(a) Independent GPs   (b) Dependent GPs using LMC

Fig. 2: Output-independent GPSSMs vs. output-dependent GPSSMs

Remark 1. *When $Q > d_x$, the output-dependent GPSSM is more flexible, and is potentially beneficial to sequence forecasting (see Section 4.2). However, for latent state inference purposes, Corollary 1 suggests setting $Q \leq d_x$, which can help avoid the model non-identifiability and improve state inference performance.*

3.2. Variational Learning and Inference

Performing inference for the GPSSM is generally intractable. Instead of using Monte Carlo-based methods, we resort to variational inference methods for computational tractability and efficiency. For the convenience of discussion, we denote $\vec{\mathbf{x}} = \{\mathbf{x}_t\}_{t=0}^T$, $\vec{\mathbf{y}} = \{\mathbf{y}_t\}_{t=1}^T$, and $\vec{\mathbf{f}} = \{\mathbf{f}_t\}_{t=1}^T$. To alleviate the computation cost of GP models, we use the sparse GP method [27], which introduces a small set of inducing points $\vec{\mathbf{z}} = \{\mathbf{z}_i\}_{i=1}^m$ and $\vec{\mathbf{u}} = \{\mathbf{u}_{i,q}\}_{i,q=1}^{m,Q}$ to serve as the surrogate of the associated GPs, where $\mathbf{u}_{i,q} = h_q(\mathbf{z}_i) \in \mathbb{R}$ and $\mathbf{z}_i \in \mathbb{R}^{d_x}$. Here, we assume the Q inducing outputs $\{\mathbf{u}_{i,q}\}_{q=1}^Q$ share the inducing inputs $\mathbf{z}_i$, since all the latent GPs take the shared input $\mathbf{x}_{t-1}$ at any time step t, as depicted in Fig. 2b. Sharing inducing inputs between different latent GPs will also reduce the number of variational parameters. Based on these settings, the joint distribution of the output-dependent GPSSM augmented with inducing points is

$$p(\vec{\mathbf{y}}, \vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}}) = p(\mathbf{x}_0) \prod_{t=1}^T p(\mathbf{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{f}_t) p(\mathbf{f}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) p(\vec{\mathbf{u}}), \quad (7)$$

where $p(\vec{\mathbf{u}}) = \prod_{q=1}^Q p(\{\mathbf{u}_{i,q}\}_{i=1}^m)$ due to the independence assumption. The distribution of the transition function outputs $p(\mathbf{f}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}})$ are determined by the latent GPs $p(\mathbf{h}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}})$, where

$$p(\mathbf{h}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) = \prod_{q=1}^Q \mathcal{N}(\mathbf{h}_{t,q}|\boldsymbol{\xi}_{\mathbf{h}_{t,q}}, \boldsymbol{\Xi}_{\mathbf{h}_{t,q}}), \quad (8)$$

and $\mathcal{N}(\boldsymbol{\xi}_{\mathbf{h}_{t,q}}, \boldsymbol{\Xi}_{\mathbf{h}_{t,q}})$ is the q-th GP posterior distribution with $\mathbf{x}_{t-1}$ as test input while $(\vec{\mathbf{z}}, \{\mathbf{u}_{i,q}\}_{i=1}^m)$ as training data, see Eqs. (2a) and (2b). Therefore,

$$p(\mathbf{f}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) = \int_{\mathbf{h}_t} p(\mathbf{f}_t|\mathbf{h}_t) \, p(\mathbf{h}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) = \mathcal{N}(\mathbf{f}_t|\boldsymbol{\xi}_{\mathbf{f}_t}, \boldsymbol{\Xi}_{\mathbf{f}_t}), \quad (9)$$

where $\boldsymbol{\xi}_{\mathbf{f}_t} = \boldsymbol{A}\boldsymbol{\xi}_{\mathbf{h}_t}$, $\boldsymbol{\Xi}_{\mathbf{f}_t} = \boldsymbol{A}\boldsymbol{\Xi}_{\mathbf{h}_t}\boldsymbol{A}^\top$ and $\boldsymbol{\xi}_{\mathbf{h}_t} = [\boldsymbol{\xi}_{\mathbf{h}_{t,1}}, .., \boldsymbol{\xi}_{\mathbf{h}_{t,Q}}]^\top$, $\boldsymbol{\Xi}_{\mathbf{h}_t} = \mathrm{diag}(\boldsymbol{\Xi}_{\mathbf{h}_{t,1}}, ..., \boldsymbol{\Xi}_{\mathbf{h}_{t,Q}})$.

The main idea behind the variational inference method is to approximate the intractable posterior distribution $p(\vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}}|\vec{\mathbf{y}})$ using a variational distribution $q(\vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}})$, leading to the evidence lower bound (ELBO), $\mathcal{L} \triangleq \mathbb{E}_{q(\vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}})} \left[\log \frac{p(\vec{\mathbf{y}}, \vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}})}{q(\vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}})} \right] \leq \log p(\vec{\mathbf{y}})$. Different choices of the variational distribution induce different EL-BOs, hence different learning algorithms for GPSSM [9]. In this paper, we choose the variational distribution $q(\vec{\mathbf{x}}, \vec{\mathbf{f}}, \vec{\mathbf{u}})$ in the form of $q(\mathbf{x}_0) \prod_{t=1}^T p(\mathbf{x}_t|\mathbf{f}_t) p(\mathbf{f}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) q(\vec{\mathbf{u}})$, where

$$q(\vec{\mathbf{u}}) = \prod_{q=1}^Q \mathcal{N}(\{\mathbf{u}_{i,q}\}_{i=1}^m|\mathbf{m}_q, \mathbf{S}_q) = \mathcal{N}(\vec{\mathbf{u}} \mid \vec{\mathbf{m}}, \mathbf{S}), \quad (10)$$

and moreover the mean vector $\vec{\mathbf{m}} = [\mathbf{m}_1^\top, ..., \mathbf{m}_Q^\top]^\top \in \mathbb{R}^{mQ}$, and the covariance matrix $\mathbf{S} = \mathrm{diag}(\mathbf{S}_1, ..., \mathbf{S}_Q) \in \mathbb{R}^{mQ \times mQ}$, are free variational parameters. The variational distribution for the initial state is parameterized by a recognition network with input, $\vec{\mathbf{y}}$, and parameters, $\boldsymbol{\zeta}$, i.e., $q_{\boldsymbol{\zeta}}(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0|\mathbf{m}_{\mathbf{x}_0}, \boldsymbol{S}_{\mathbf{x}_0})$, where $\mathbf{m}_{\mathbf{x}_0}, \boldsymbol{S}_{\mathbf{x}_0}$ are the outputs of the recognition network [22]. Note that the form of variational distribution selected in this paper is similar to the one used in the probabilistic SSM [10]. However, instead of assuming independent outputs for the transition function, the GPSSM prior as well as the approximated posterior considered in this paper explicitly construct the dependency among outputs by mixing Q latent GPs, thus making the proposed model more realistic and flexible. After some algebraic calculations, the corresponding ELBO becomes

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^T \mathbb{E}_{q(\mathbf{x}_t)} \left[\log p(\mathbf{y}_t|\mathbf{x}_t) \right] - \mathrm{KL} \left[q(\vec{\mathbf{u}}) \| p(\vec{\mathbf{u}}) \right] \\ - \mathrm{KL} \left[q(\mathbf{x}_0) \| p(\mathbf{x}_0) \right], \quad (11)$$

where the first term encourages decent samples drawn from the variational distribution $q(\mathbf{x}_t)$ to fit the emission model well, while the second and third terms regularize the initial state, and the posterior distributions of the Q latent GPs thus the posterior of $f(\cdot)$, respectively. The two latter terms can be computed analytically, however, the expectation terms, $\mathbb{E}_{q(\mathbf{x}_t)} \left[\log p(\mathbf{y}_t|\mathbf{x}_t) \right], \forall t$, need to be evaluated by sampling method and reparametrization trick [28] due to the intractability of $q(\mathbf{x}_t)$ [10]. The sampling steps are described as follows. By conditioning on the latent state $\mathbf{x}_{t-1}$, we have

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \int_{\mathbf{f}_t, \vec{\mathbf{u}}} p(\mathbf{x}_t|\mathbf{f}_t) p(\mathbf{f}_t|\mathbf{x}_{t-1}, \vec{\mathbf{u}}) q(\vec{\mathbf{u}}) \\ = \mathcal{N}(\mathbf{x}_t \mid \mathbf{m}_{t|t-1}, \mathbf{S}_{t|t-1}), \quad (12)$$

where $\mathbf{m}_{t|t-1} = \boldsymbol{A}\mathbf{m}_{\mathbf{h}_t}$, $\mathbf{S}_{t|t-1} = \boldsymbol{A}\mathbf{S}_{\mathbf{h}_t}\boldsymbol{A}^\top + \boldsymbol{Q}$, and $\mathbf{m}_{\mathbf{h}_t} = [\mathbf{m}_{\mathbf{h}_{t,1}}, ..., \mathbf{m}_{\mathbf{h}_{t,Q}}]^\top$, $\mathbf{S}_{\mathbf{h}_t} = \mathrm{diag}(\mathbf{S}_{\mathbf{h}_{t,1}}, ..., \mathbf{S}_{\mathbf{h}_{t,Q}})$ with

$$\begin{cases} \mathbf{m}_{\mathbf{h}_{t,q}} = \boldsymbol{K}_{\mathbf{x}_{t-1}, \vec{\mathbf{z}}} \, \boldsymbol{K}_{\vec{\mathbf{z}}, \vec{\mathbf{z}}}^{-1} \, \mathbf{m}_q, \\ \mathbf{S}_{\mathbf{h}_{t,q}} = \boldsymbol{K}_{\mathbf{x}_{t-1}, \mathbf{x}_{t-1}} - \boldsymbol{K}_{\mathbf{x}_{t-1}, \vec{\mathbf{z}}} \, \boldsymbol{K}_{\vec{\mathbf{z}}, \vec{\mathbf{z}}}^{-1} [\boldsymbol{K}_{\vec{\mathbf{z}}, \vec{\mathbf{z}}} - \mathbf{S}_q] \, \boldsymbol{K}_{\vec{\mathbf{z}}, \vec{\mathbf{z}}}^{-1} \, \boldsymbol{K}_{\mathbf{x}_{t-1}, \vec{\mathbf{z}}}^\top, \end{cases}$$

for any q. Note that here we omit the subscript q in the kernel matrix of the q-th latent GP for notation brevity. Using Eq. (12) we can recursively sample latent states $\mathbf{x}_t, t = 1, 2, ..., T$, by starting from sampling $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, so that the ELBO can be numerically evaluated. Together with all, we apply stochastic gradient ascent and use the Adam optimizer to maximize the lower bound $\mathcal{L}(\boldsymbol{\theta})$ over parameters $\boldsymbol{\theta} = [\boldsymbol{\zeta}, \vec{\mathbf{z}}, \vec{\mathbf{m}}, \mathbf{S}, \boldsymbol{\theta}_{gp}, \boldsymbol{Q}, \boldsymbol{R}]$. The gradient can be propagated back through time owing to the chain rule sampling and reparametrization trick [28], and the parameters will converge to a stationary point.

Remark 2 (Computational Complexity). *Typically, the number of GP inducing points, m, is larger than the number of latent GPs, Q, and the state dimension, d_x. For a data sequence with the length of T and assuming $T \gg m > Q \geq d_x$, the computational complexity of evaluating the ELBO (Eq. (11)) scales as $\mathcal{O}(TQm^2 + TQ^2 d_x)$. Compared with the output-independent GPSSM [10] that scales as $\mathcal{O}(Td_x m^2)$, we can observe that only gentle computational complexity increases in the output-dependent GPSSM (especially in the case of $Q = d_x$).*

4. EXPERIMENTAL RESULTS

In this section, we show the performance of the proposed *output-dependent* GPSSM (termed *ODGPSSM*) on one synthetic dataset and five real system identification datasets. For comparison, we choose two *output-independent* baseline models: 1) probabilistic recurrent SSM (*PRSSM*) [10]; and 2) deep state-space model (*DSSM*) [3].
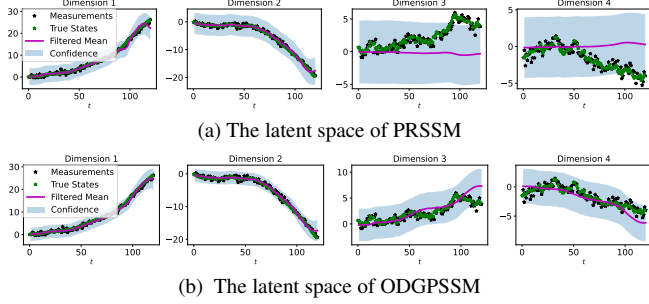
(a) The latent space of PRSSM

(b) The latent space of ODGPSSM

Fig. 3: The latent space learned by two different GPSSMs

4.1. Synthetic Dataset

We first use a simple and stylized numerical 2-dimensional (2-D) car tracking example provided in [1] (Example 4.3) to show the importance of modeling dependency between the transition outputs. More concretely, in this example, the underlying car dynamic is characterized by the linear Gaussian state-space model (LGSSM),

$$\mathbf{x}_t = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{I}_2 \\ \mathbf{0} & \boldsymbol{I}_2 \end{bmatrix} \mathbf{x}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}),$$

$$\mathbf{y}_t = [\boldsymbol{I}_2, \mathbf{0}] \, \mathbf{x}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}),$$

where the partially observable state $\mathbf{x}_t \in \mathbb{R}^4$ consists of 2-D car positions and 2-D velocities; $\mathbf{y}_t$ is a noisy observation of the car positions. For more details about this model, one can refer to [1]. Note that the entries of the state vector, $\{\mathbf{x}_{t,d}\}_{d=1}^4$, are correlated due to the linearity and Gaussianity of the state transition in the LGSSM. Hence, by exploiting the correlations with the observed states $\{\mathbf{x}_{t,d}\}_{d=1}^2$, it is possible to infer the unobserved ones, $\{\mathbf{x}_{t,d}\}_{d=3}^4$.

We use the underlying LGSSM to generate T=120 observations $\vec{\mathbf{y}} = \{\mathbf{y}_t\}_{t=1}^T$ for training the *PRSSM* [10] and the newly proposed *ODGPSSM*. For both PRSSM and ODGPSSM, we adopt the following initialization: 1) The initial state, $\mathbf{x}_0 = [0, 0, 1, -1]^\top$, is assumed to be known, hence there is no need to train the recognition network; 2) The dimension of the latent state, d_x, is set to be 4, and the number of the GP inducing points m is set to be 20; 3) The GP transition models are pretrained/initialized using 20 true latent state pairs with the same training epochs. The number of the latent GP functions for ODGPSSM remains the same as the state dimension, thus only slightly increasing the overall computational complexity (see Remark 2). Fig. 3 depicts the learning results of ODGPSSM and PRSSM. It can be observed that both PRSSM and ODGPSSM infer the first two dimensions of the latent states well. However, the PRSSM fails to capture the underlying dynamics in the 3rd and 4th unobserved dimensions. This is due to the fact that the independent modeling in PRSSM ignores the correlations between the states, resulting in the fluctuations of the GP transition posterior around the zero-mean prior. In contrast, ODGPSSM establishes the dependencies through the LMC framework, making it capable of correctly learning the GP transition posterior and inferring the unobserved states by exploiting the knowledge from the shared correlations and the first two dimensions that are fully observed.

4.2. Real Datasets

Since the superiority of PRSSM compared to classic time-series modeling approaches has been shown in [10], we will skip similar comparisons due to space limitations. In this subsection, we only compare ODGPSSM with its two competitors, PRSSM and DSSM, on five real system identification datasets introduced in [10] (see [10] for more details). For each dataset, the first half of a sequence is
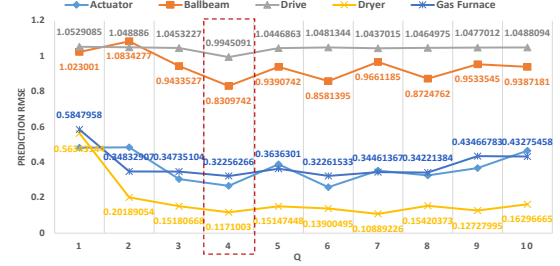

Fig. 4: Test performance of ODGPSSM under different values of Q

Table 1: Prediction RMSE comparison (on standardized test sets) between the proposed ODGPSSM (with $Q = 4$) and its competitors.

Methods	Actuator	Ballbeam	Drive	Dryer	Gas Furnace
DSSM	3.0752	2.6311	1.8417	2.7055	1.7746
PRSSM	0.3592	0.9082	1.0459	0.1334	0.3356
ODGPSSM	**0.2668**	**0.8309**	**0.9945**	**0.1171**	**0.3225**

used for training and the rest for testing. All datasets are standardized using training sequence and the latent state dimension is set to be $d_x = 4$. More detailed model settings can be found in the accompanying code online available at: *https://github.com/zhidilin/ODGPSSM*. For DSSM, we refined the code from a public implementation at *http://pyro.ai/examples/dmm.html*, and set the emission function to be the same as the two GPSSMs. The test results are reported in Table 1, where the root-mean-square error (RMSE) is averaged over 100-step ahead forecasting.

From Table 1 we can observe that ODGPSSM consistently outperforms PRSSM in terms of the prediction RMSE across all the datasets, which convincingly illustrates the benefits of output dependency modeling. It can also be observed that both ODGPSSM and PRSSM outperform DSSM in terms of the prediction RMSE. The reason is that both the transition function and the variational distributions in DSSM are modeled by deep neural networks that require big data to tune the large number of parameters. However, in our case, the datasets are relatively small, e.g. the training set of the *Gas Furnace* dataset is merely of length 148, which is insufficient to support the DSSM learning. In contrast, GPSSMs inherit the merits of GP, showing unique superiority in small dataset regimes.

Finally, we investigate the impact of Q in ODGPSSM, since the number of latent GPs, Q, determines the model flexibility. The results in Fig. 4 show that the prediction RMSE across almost all the datasets reach the lowest points when $Q = d_x = 4$, even though the models with $Q > 4$ is more flexible than the ones with $Q = 4$. This is probably because the additional parameters (additional coregionalization coefficients and variational parameters) makes the model unidentifiable and the learning more difficult. Future work will attempt to remedy this problem by introducing sparse constraints on the coregionalization coefficients and verify it on real data provided by China Unicom.

5. CONCLUSION

In this paper, we propose an output-dependent GPSSM by explicitly modeling the output dependency of GP transition using the well-known, simple yet practical LMC framework. We also propose a variational learning algorithm that only gently increases the computational complexity to learn the output-dependent GPSSM. Experimental results show that modeling the output dependency in GPSSM not only facilitates latent state inference when the latent state is partially observed, but also makes the GPSSM more competent than its competitors in terms of prediction.

6. REFERENCES

[1] Simo Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.

[2] Marc Deisenroth and Carl E Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, United states, July 2011, pp. 465–472.

[3] Rahul Krishnan, Uri Shalit, and David Sontag, "Structured inference networks for nonlinear state space models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, San Francisco, CA, United states, Feb. 2017, pp. 2101–2109.

[4] Ahmed M Alaa and Mihaela van der Schaar, "Attentive state-space modeling of disease progression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 11338–11348.

[5] Haitao Liu, Changjun Liu, Xiaomo Jiang, Xudong Chen, Shuhua Yang, and Xiaofang Wang, "Deep probabilistic time series forecasting using augmented recurrent input for dynamic systems," *Mech. Syst. Signal Process.*, vol. 177, pp. 109212, 2022.

[6] Andrew James McHutchon, *Nonlinear modelling and control using Gaussian processes*, Ph.D. thesis, University of Cambridge, 2014.

[7] Roger Frigola, Fredrik Lindsten, Thomas B Schön, and Carl E Rasmussen, "Bayesian inference and learning in Gaussian process state-space models with particle MCMC," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Lake Tahoe, NV, United states, Dec. 2013, pp. 3156–3164.

[8] Roger Frigola, Yutian Chen, and Carl E Rasmussen, "Variational Gaussian process state-space models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, Dec. 2014, pp. 3680–3688.

[9] Roger Frigola, *Bayesian time series learning with Gaussian processes*, Ph.D. thesis, University of Cambridge, 2015.

[10] Andreas Doerr, Christian Daniel, Martin Schiegg, Nguyen-Tuong Duy, Stefan Schaal, Marc Toussaint, and Trimpe Sebastian, "Probabilistic recurrent state-space models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, July 2018, pp. 1280–1289.

[11] Alessandro Davide Ialongo, Mark van der Wilk, James Hensman, and Carl Edward Rasmussen, "Overcoming mean-field approximations in recurrent Gaussian process models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, United states, June 2019, pp. 2931–2940.

[12] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[13] Richard Cornelius Suwandi, Zhidi Lin, Yiyong Sun, Zhiguo Wang, Lei Cheng, and Feng Yin, "Gaussian process regression with grid spectral mixture kernel: Distributed learning for multidimensional data," in *Proc. Int. Conf. Inf. Fusion (FUSION)*, Linkoping, Sweden, July 2022, pp. 1–8.

[14] Feng Yin, Lishuo Pan, Tianshi Chen, Sergios Theodoridis, Zhi-Quan Luo, and Abdelhak M Zoubir, "Linear multiple low-rank kernel based stationary Gaussian processes regression for time series," *IEEE Trans. Signal Process.*, vol. 68, pp. 5260–5275, Sept. 2020.

[15] Sergios Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, Academic Press, 2020.

[16] Lei Cheng, Feng Yin, Sergios Theodoridis, Sotirios Chatzis, and Tsung-Hui Chang, "Rethinking Bayesian learning for data analysis: The art of prior and inference in sparsity-aware modeling," *IEEE Signal Process. Mag.*, vol. 39, no. 6, pp. 18–52, Oct. 2022.

[17] Jack M Wang, David J Fleet, and Aaron Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Dec. 2007.

[18] Ang Xie, Feng Yin, Bo Ai, Sha Zhang, and Shuguang Cui, "Learning while tracking: A practical system based on variational Gaussian process state-space model and smartphone sensory data," in *Proc. Int. Conf. Inf. Fusion (FUSION)*, Rustenburg, South Africa, July 2020, pp. 1–7.

[19] Feng Yin, Zhidi Lin, Qinglei Kong, Yue Xu, Deshi Li, Sergios Theodoridis, and Shuguang Cui, "Fedloc: Federated learning framework for data-driven cooperative localization and location data processing," *IEEE Open J. Signal Process.*, vol. 1, pp. 187–215, Nov. 2020.

[20] Yuxin Zhao, Carsten Fritsche, Gustaf Hendeby, Feng Yin, Tianshi Chen, and Fredrik Gunnarsson, "Cramér–Rao bounds for filtering based on Gaussian process state-space models," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5936–5951, Dec. 2019.

[21] Jakob Lindinger, Barbara Rakitsch, and Christoph Lippert, "Laplace approximated Gaussian process state-space models," in *Proc. Conf. Uncertain. Artif. Intell. (UAI)*, Eindhoven, Netherlands, Aug. 2022.

[22] Stefanos Eleftheriadis, Tom Nicholson, Marc Peter Deisenroth, and James Hensman, "Identification of Gaussian process state space models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, United states, Dec. 2017, pp. 5309–5319.

[23] Yuhao Liu and Petar M Djurić, "Gaussian process state-space models with time-varying parameters and inducing points," in *Proc. European Signal Proces. Conf. (EUSIPCO)*, Amsterdam, Netherlands, Jan. 2021, pp. 1462–1466.

[24] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al., "Kernels for vector-valued functions: A review," *Found. Trends Mach. Learn.*, vol. 4, no. 3, pp. 195–266, 2012.

[26] Kai Chen, Twan van Laarhoven, Elena Marchiori, Feng Yin, and Shuguang Cui, "Multitask gaussian process with hierarchical latent interactions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process Proc. (ICASSP)*, Singapore, May 2022, pp. 4148–4152.

[27] James Hensman, Nicolò Fusi, and Neil D Lawrence, "Gaussian processes for big data," in *Proc. Conf. Uncertain. Artif. Intell. (UAI)*, Bellevue, WA, United states, July 2013, pp. 282–290.

[28] Diederik P Kingma and Max Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.