MULTITASK GAUSSIAN PROCESS WITH HIERARCHICAL LATENT INTERACTIONS

Kai Chen[†‡] *Twan van Laarhoven*[‡] *Elena Marchiori*[‡] *Feng Yin*[⋆†] (✉) *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.
[‡]Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands.

ABSTRACT

Multitask Gaussian process (MTGP) is powerful for joint learning of multiple tasks with complicated correlation patterns. However, due to the assembling of additive independent latent functions (LFs), all current MTGPs including the salient linear model of coregionalization (LMC) and convolution frameworks cannot effectively represent and learn the hierarchical latent interactions between its LFs. In this paper, we further investigate the interactions in LMC of MTGP and then propose a novel kernel representation of the hierarchical interactions, which ameliorates both the expressiveness and the interpretability of MTGP. Specifically, we express the interaction as a product of function interaction (FI) and coefficient interaction. The FI is modeled by using cross convolution of LFs. The coefficient interaction between the LMCs is described as a free-form coupling coregionalization term. We validate that considering the interactions can promote knowledge transferring in MTGP and compare our approach with some state-of-the-art MTGPs on both synthetic- and real-world datasets.

***Index Terms*—** Gaussian process, multitask learning, linear model of coregionalization, latent interaction, spectral mixture kernel

1. INTRODUCTION

Gaussian process (GP) [1] is an extraordinary Bayesian nonparametric model for representing the underlying function of a complex system due to its powerful fitting ability [2]. The inferred GP model provides a posterior distribution over the underlying function, which can be refined as evidence is accumulated. The data fitting performance as well as its natural uncertainty quantification make GP competent in various sectors, for instance 6G wireless communication system [3, 4, 5]. However, the choice of kernel is a cornerstone for GP and influences the profile of its posterior distribution. The extension of GP to multiple correlated tasks is known as multitask (MT) GP (MTGP) [6], which shows the representation power in joint MT learning. MTGPs not only model nonlinear correlation among infinitely many random variables, as single task GPs (STGPs), but also account for high level correlation across different tasks [6, 7]. For designing a MTGP, a very delicate problem is how to choose an expressive MTGP kernel, k_{M}, to jointly represent the cross covariance between tasks and auto-covariance within each single task [8, 7]. Given an expressive kernel, a MTGP can leverage knowledge from all tasks to obtain higher prediction accuracy than STGP independently leaning on each single task [9, 10].

Early approaches to MTGPs, like linear model of coregionalization (LMC) [11, 12, 6], focused on linear combination of independent latent functions (LFs) represented by STGPs. More advanced MTGPs like the multi-kernel [13, 14] and LF convolution frameworks [15] adopt convolution to construct cross covariance structure and assume that each task has its own auto-covariance structure. Comparing among the existing MTGPs, the LMC framework has favorable and useful qualities, such as compact hyperparameter space and interpretable hierarchical covariance structures. Since the introduction of the expressive and generalized spectral mixture (SM) kernels [16, 17], the learning capacities of MTGPs, including the LMC and convolution frameworks, have been ameliorated by incorporating SM kernels during the past years [18, 19, 20]. However, these MTGPs have a generalized form with a sum of additive independent LFs and dismiss the underlying interactions between the LFs. In this paper, we aim at investigating and embodying interaction particularly for LMC due to its popularity. Our work also provides an interesting cue for other MTGP frameworks, such as the convolution framework [13, 19, 20].

Contributions: We develop a novel LMC framework with hierarchical interactions for MTGP and the new features are as follows:

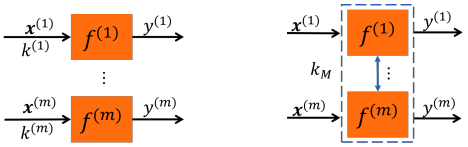
- By using convolution between LFs, we offer a kernel encoding function interactions (FIs) in MTGP for the first time;
- We derive free-form coupling coregionalization (CC) between Cholesky factors of LMCs for coefficient interactions;
- We propose a kernel framework based on both the function and coefficient interactions for MTGP, which demonstrates rich representation, interpretability, and expressiveness.
- We investigate the learning capacity of the framework and collate its performance with recent competent MTGPs.

For the rest of the paper, GPs and related MTGPs are reviewed in Section 2. In Section 3, our framework for MTGP is introduced and compared with some existing methods. Experiments on both synthetic- and real-world data are discussed in Section 4. We conclude our work and touch upon some future directions in Section 5.

2. BACKGROUND AND RELATED WORK

Gaussian process and kernel function: A GP defines a distribution over functions, specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ [1] for given input vector $\mathbf{x} \in \mathbb{R}^P$. Mathematically, it is denoted as $f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. By placing a GP prior over functions through the choice of a kernel and hyperparameter initialization, from the training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ of finite size n, we can predict the unknown function value $\tilde{y}_*$ and its variance $\mathbb{V}[y_*]$ (that is, its uncertainty) for a test point $\mathbf{x}_*$ using Bayesian inference. The smoothness and generalization properties of a GP depend on the kernel function and its hyperparameters Θ. Consequently, various kernels, e.g. the squared exponential (SE) and kernel design methods have been introduced [1]. Based on the Bochner's theorem, the recent SM kernel have been proposed in

[16], which is derived through modeling spectral densities $\hat{k}_{\text{SM}}(\mathbf{s})$ (Fourier transform of a kernel) with a Gaussian mixture. A desired property of SM kernel is that it can generalize any stationary kernel. The SM kernel has a form as $k_{\text{SM}}(\tau) = \sum_{i=1}^{Q} k_{\text{SM},i}(\tau)$, where $k_{\text{SM},i}(\tau) = w_i \exp(-2\pi^2\tau\Sigma_i\tau^\top)\cos(2\pi\boldsymbol{\mu}_i\tau^\top)$, $\tau = \mathbf{x} - \mathbf{x}'$, Q is the number of components, w_i, $\boldsymbol{\mu}_i = [\mu_i^{(1)}, ..., \mu_i^{(P)}]$, and $\Sigma_i = \text{diag}([(\sigma_i^2)^{(1)}, ..., (\sigma_i^2)^{(P)}])$ are the weight, mean, and variance of the i-th Gaussian $\mathcal{N}_i(\mathbf{s}; \boldsymbol{\mu}_i, \Sigma_i)$ in frequency domain, respectively.

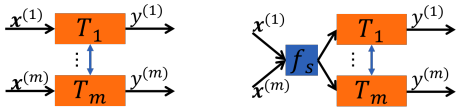


(a) m STGPs with m tasks (b) A MTGP with m tasks

Fig. 1. The frameworks of STGP (a) and MTGP (b).

Multitask Gaussian process: To clarify the mechanism of MTGP, we depict the conception of STGP and MTGP in Fig. 1. The GP model designated to the m-th task shown in subplot (a) has functional expressions: $f^{(m)} \sim \mathcal{GP}(0, k^{(m)}(\mathbf{x}^{(m)}, \mathbf{x}'^{(m)}))$ and $y^{(m)} = f^{(m)} + \epsilon$, where ϵ is the noise of the model. The black arrow denotes data flow and inference direction. The orange rectangle denotes the underlying function of each task. For m single tasks, each task modeled as an STGP and no correlation between tasks is taken into account. For MTGP shown in subplot (b), we wish to learn the m underlying functions simultaneously with $\boldsymbol{f} \sim \mathcal{GP}(0, k_{\text{M}}(X, X'))$ and $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{f} = [f^{(1)}(\mathbf{x}^{(1)}), ..., f^{(m)}(\mathbf{x}^{(m)})]^\top$, $X = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(m)}]^\top$, $\mathbf{y} = [y^{(1)}, ..., y^{(m)}]^\top$, and $k_{\text{M}}(X, X')$ describes both the auto-covariance of each task and the cross covariance between tasks. The dashed box in subplot (b) denotes the target MTGP model representing a joint Gaussian distribution of m underlying functions. Hence, k_{M} determines the learning capacity of MTGP.

Multitask kernel function: The construction of a MTGP mainly involves the designation of $k_{\text{M}}(X, X')$. Some state-of-the-art works on MT kernel design include: free form LMC first appeared in [6, 21], multi-kernel method [13], GP regression network (GPRN) [10, 22], asymmetric focused MTGP [23], and multitask SM kernels [18, 19, 20]. The general formula of k_{M} can be written as matrix $K_{\text{M}}(X, X') = \begin{bmatrix} K^{(1,1)}, & K^{(1,m)} \\ K^{(m,1)}, & K^{(m,m)} \end{bmatrix}$, where the submatrix $K^{(m,1)}$ describes the cross covariance between $f^{(m)}$ and $f^{(1)}$. An expressive k_{M} using LMC linearly combines a mixture of Q covariance components to ameliorate the representation of MTGP with multiple LFs, which has an eventual form: $K_{\text{M}} = \sum_{i=1}^{Q} B_i \otimes K_{s,i}$, where B_i is a LMC matrix representing task correlation and $K_{s,i}$ is a matrix constructed by arbitrary kernel of STGP. Since the introduction of SM kernels [16, 17], various MTGPs have been developed [22, 18, 19, 20] by substituting k_i with k_{SM} or by building $k^{(m,1)}$ as a convolution form. Due to the neat generalization and interpretability of SM, here we mainly review MTGPs incorporating SM. Firstly, the cross SM (CSM) kernel [18] improved the expressiveness of GPRN: it contains cross phase spectrum and is also defined in a LMC form with $K_{\text{M}} = \sum_{i=1}^{Q} B_i \otimes k_{\text{SG},i}(\tau; \Theta_i)$, where $k_{\text{SG},i}(\tau; \Theta_i)$ is phase notation of a spectral Gaussian kernel. The multi-output SM (MOSM) kernel [19] provides a principled matrix factorization way to construct multivariate covariance functions with a better interpretation of the correlation between tasks. Even if MOSM extends existing MTGPs in expressiveness and interpretation, it still considers linear combination of LFs and ignores interactions between them. By considering the imperfection of compatibility in MOSM when $m = 1$, the more recent multi-output convolution SM (MOCSM) [20] employs both SM kernel and convolution mechanism and gain a more compatible MTGP.



(a) MTGP using convolution (b) MTGP using LMC

Fig. 2. Convolution (a) and LMC (b) frameworks of MTGP.

Comparison between MTGPs: In Fig. 2, we present and compare the two main frameworks of MTGPs: MTGPs using (a) convolution and (b) LMC. Here, T_m labeled in the orange rectangle denotes the m-th task. In subplot (a), task correlation in MTGP is described by the convolution between LFs of tasks and therefore each task has its own kernel function. The bidirectional blue arrow denotes task correlation based on the convolution. In subplot (b), task correlation in LMC is hard-coded in a symmetrical hyperparameter matrix and the $\{m, m'\}$-th entry of the matrix denotes the cross covariance between a pair of tasks m, m'. Diagonal elements in the matrix encode the auto-covariances and signal magnitudes of tasks. The differences between tasks are accounted by their different signal magnitudes. In particular, all tasks share a LF space f_s (denoted by blue rectange) adumbrating their common qualities. The shared f_s means these tasks employ the same shared kernel for knowledge transferring and sharing. From Fig. 2, the LMC has a clearer hierarchical architecture for model explanation. In particular, the LMC has more compact hyperparameter space than the convolution framework due to the shared LF. The architecture of LMC is simple and popularly used for many applications [9, 24]. That's one of the reason we mainly investigate the interaction of LMCs.

3. MULTITASK SPECTRAL MIXTURE KERNEL WITH INTERACTIONS

For our attention on the LMC, a more general form of underlying function in MTGP can be amply expressed as linear combination of independent LFs [7, 8],

$$f^{(m)}(\mathbf{x}^{(m)}) = \sum_{i=1}^{Q} \alpha_i^{(m)} f_{s,i}(\mathbf{x}^{(m)}), \tag{1}$$

where $\alpha_i^{(m)}$ is a scalar coefficient and the LF $f_{s,i}(\mathbf{x}^{(m)})$ is a zero mean GP with $\text{cov}[f_{s,i}(\mathbf{x}^{(m)}), f_{s,i}(\mathbf{x}'^{(m)})] = k_{s,i}(\mathbf{x}^{(m)}, \mathbf{x}'^{(m)})$. The linear combination form of such separate functions assumes that $f_{s,i}(\mathbf{x}^{(m)}) \perp f_{s,j}(\mathbf{x}^{(m)})$ and $\text{cov}[f_{s,i}(\mathbf{x}^{(m)}), f_{s,j}(\mathbf{x}^{(m)})] = 0$ when $i \neq j$. As mentioned before, kernel $k_{s,i}$ ensures the generalization of $f_{s,i}(\mathbf{x}^{(m)})$. Due to the expressiveness and generalization of SM kernel, we substitute $k_{s,i}$ with $k_{\text{SM},i}$ and therefore obtain a more expressive $f^{(m)}(\mathbf{x}^{(m)}) = \sum_{i=1}^{Q} \alpha_i^{(m)} f_{\text{SM},i}(\mathbf{x}^{(m)})$, where $f_{\text{SM},i} \sim \mathcal{GP}(0, k_{\text{SM},i})$. On the other hand, for STGP using SM kernel, the dependency between LFs $\{f_{\text{SM},i}, f_{\text{SM},j}\}$ has been investigated and proved by [25], which is somewhat significant. However, the interaction between LFs of aforementioned MTGPs is unclear and unexplained. Hence, in this paper, we consider two hierarchical interactions in LMC, FI between LFs $f_{\text{SM},i}$ and $f_{\text{SM},j}$ and coefficient interaction between $\alpha_i^{(m)}$ and $\alpha_j^{(m)}$.

Function interaction between LFs: In this paper, we follow the investigation in [25] and view the dependency as a general sense of FI between LFs. Furthermore, for $k_{\text{SM},i}$, we borrow the parameterization of time and phase delayed SM kernel in [26, 20] to enrich the features of FI. Then, we meliorate the FI in LMC by describing it related to time and phase delay, as modeled in [25, 26] through the so-called generalized convolution spectral mixture (GCSM) kernel, defined as follows:

$$\begin{aligned} k^{i\times j}_{\text{GCSM}}(\tau) =& \mathcal{F}^{-1}_{s\to\tau}\big[\hat{k}^{i\times j}_{\text{GCSM}}(\mathbf{s}) + \hat{k}^{i\times j}_{\text{GCSM}}(-\mathbf{s})\big](\tau) \\ =& c_{ij}\exp\big(-\frac{1}{2}\tau_{\boldsymbol{\theta}}^{\top}\Sigma_{ij}\tau_{\boldsymbol{\theta}}\big)\cos\big(\tau_{\boldsymbol{\theta}}^{\top}\boldsymbol{\mu}_{ij} - \boldsymbol{\phi}_{ij}\pi\big), \end{aligned} \tag{2}$$

where $\mathbf{s}$ is a spectral vector, $c_{ij} = w_{ij}a_{ij}$ is the cross constant incorporating cross weight and amplitude, and $\tau_{\boldsymbol{\theta}} = 2\pi(\tau - \frac{\boldsymbol{\theta}_{ij}}{2})$ is the Euclidean distance with time delay. Other parameters related to $\{f_{\text{SM},i}, f_{\text{SM},j}\}$ are defined as: cross weight, $w_{ij} = \sqrt{w_i w_j}$; cross amplitude, $a_{ij} = \big|4\pi^2\Sigma_i\Sigma_j\big|^{\frac{1}{4}}\mathcal{N}(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \frac{\Sigma_i+\Sigma_j}{2})$; cross mean, $\boldsymbol{\mu}_{ij} = \frac{\Sigma_i\boldsymbol{\mu}_j+\Sigma_j\boldsymbol{\mu}_i}{\Sigma_i+\Sigma_j}$; cross length scale, $\Sigma_{ij} = \frac{2\Sigma_i\Sigma_j}{\Sigma_i+\Sigma_j}$; cross time delay, $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j$, and cross phase delay, $\boldsymbol{\phi}_{ij} = \boldsymbol{\phi}_i - \boldsymbol{\phi}_j$, where $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$ respectively denote latent time and phase delays in $f_{\text{SM},i}$. Here, $\hat{k}^{i\times j}_{\text{GCSM}}(\mathbf{s})$ is a cross spectral density (SD) between the $k_{\text{SM},i}$ and $k_{\text{SM},j}$,

$$\hat{k}^{i\times j}_{\text{GCSM}}(\mathbf{s}) = c_{ij}\mathcal{N}\big(\mathbf{s}; \boldsymbol{\mu}_{ij}, \Sigma_{ij}\big)\exp\big(-\pi\imath(\boldsymbol{\theta}_{ij}\mathbf{s} + \boldsymbol{\phi}_{ij})\big), \tag{3}$$

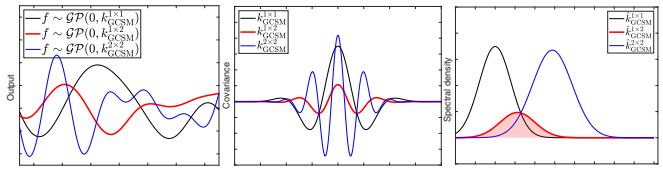
where the hat symbol denotes SD. Note that $k^{i\times j}_{\text{GCSM}}(\tau) = k_{\text{SM},i}$ and $\hat{k}^{i\times j}_{\text{GCSM}}(\tau) = \hat{k}_{\text{SM},i}$ for $i = j$, where $\hat{k}_{\text{SM},i}$ is the SD of $k_{\text{SM},i}$. The SD and covariance of the interaction (in red) between $f_{\text{SM},i}$ and $f_{\text{SM},j}$ are illustrated in Fig. 3. Through the GCSM in Eq. (3), we can model the FI between $\{f_{\text{SM},i}, f_{\text{SM},j}\}$.

Coefficient interaction between LMCs: For the coefficient interaction between $\alpha^{(m)}_i$ and $\alpha^{(m)}_j$, we reformulate $\alpha^{(m)}_i$ in a LMC matrix B_i as $B_i(m, m') = \alpha^{(m)}_i\alpha^{(m')}_i$, where $B_i(m, m')$ is the element of B_i located at the m-th column and m'-th row. In addition, we consider the free-form parameterization of B_i in [6, 8, 7] and decompose B_i using Cholesky factorization as: $B_i = B_{L,i}B_{L,i}^{\top}$ with the Cholesky factor, low triangle $B_{L,i} = \begin{bmatrix} \ell^i_{1,1}, & 0 \\ \ell^i_{m,1}, & \ell^i_{m,m} \end{bmatrix}$, where $\ell^i_{m,m'}$ can be seen as the correlation between tasks m and m' [6]. Note that there are in total $m(m+1)/2$ hyperparameters for the free-form parameterization of B_i in LMC. Furthermore, for any two arbitrary LMC terms B_i and B_j, we construct $B_{ij} = B_{L,i}B_{L,j}^{\top}$ to encode the coefficient interaction between LMCs. Here we interpretate $B_{i,j}$ as a free-form CC that different from [12], which is capable of capturing complicated interaction between coefficients. There is no extra hyperparameter introduced for the coefficient interaction. In particular, a compatibility is guaranteed such that B_{ij} becomes B_i when $i = j$. Thus, for MTGP with Q LFs constructed by using LMC and SM kernel, we have a kernel with hierarchical interactions as follows:

$$K_{\text{GCSM-CC}}(\tau) = \sum_{i=1}^{Q}\sum_{j=1}^{Q} B_{ij} \otimes k^{i\times j}_{\text{GCSM}}(\tau). \tag{4}$$

Due to the use of GCSM and CC, we call the proposed kernel as GCSM-CC. The positive semi-definite (PSD) of GCSM-CC is guaranteed by the LMC framework and the PSD of both B_i [6] and k_{GCSM} [25]. Note that GCSM-CC is the first work of non-separable LMC using SM.

The interpretation of interaction: In addition to existing MTGPs, the GCSM-CC has following desired properties: (1) it does not treat LFs separately; (2) it allows to model FI including time and phase delay; (3) it uses CC to represent coefficient interaction across all LMCs; (4) it includes more interaction terms without the increasing of hyperparameter space (if remove time and phase delays). As shown in Fig. 3, there are two LFs, $f \sim \mathcal{GP}(0, k^{1\times1}_{\text{GCSM}})$ (in black) and $f \sim \mathcal{GP}(0, k^{2\times2}_{\text{GCSM}})$ (in blue). From subplot (c), the SD $\hat{k}^{1\times2}_{\text{GCSM}}$ of the interaction in the frequency domain can be seen as an intersection between $\hat{k}^{1\times1}_{\text{GCSM}}$ and $\hat{k}^{2\times2}_{\text{GCSM}}$, where $\hat{k}^{1\times1}_{\text{GCSM}} = \hat{k}_{\text{SM},1}$. The covariance $k^{1\times2}_{\text{GCSM}}$ (subplot (b)) of the interaction in the time domain has a period smaller than $k_{\text{SM},1}$ and bigger than $k_{\text{SM},2}$. For the LF $f \sim \mathcal{GP}(0, k^{1\times2}_{\text{GCSM}})$, covariance $k^{1\times2}_{\text{GCSM}}$, and SD $\hat{k}^{1\times2}_{\text{GCSM}}$ of the interaction, their magnitudes are smaller than the original LFs. Due to the distributivity and commutativity of interaction, there are Q^2 interaction terms with $Q^2 - Q$ cross interaction terms plus Q auto interaction terms in GCSM-CC. The cross interactions allow extensive communication between LFs and hence bring potent fitting capacity.



(a) The LFs (b) The covariances (c) The SDs

Fig. 3. The illustrations of LFs (a), covariances (b), and SDs (c) of their interaction.

4. EXPERIMENTS

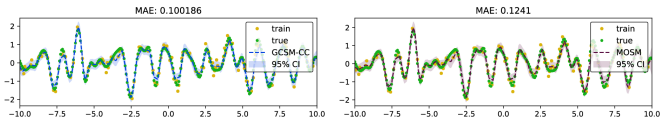
In this section, we compare GCSM-CC with some existing MTGPs [6, 22, 18, 19]. First we show the ability of GCSM-CC to simultaneously model multiple incomplete signals sampled from a GP, its integral and derivative. Then we use GCSM-CC for prediction tasks on a real-world problem with sensor signals [1] related to air pollution monitoring: Nitrogen oxide (NO) concentration. For both the synthetic and real-world datasets, we follow the experimental setting in [15, 19] to generate incomplete signal. However, we observed that in the considered experimental setting it is not difficult to capture correlation between tasks because the training data in one signal has a consecutive intersection with the training data of another signal. In particular, we consider a more challenging task that does not involve consecutive intersection of training data between tasks and consider also a signal recovery. We implemented our models using GPflow [27] to improve scalability and facilitate gradient computation. The open source codes will be made available soon. We use the mean absolute error, $\text{MAE} = \sum_{i=1}^{n}\big|y_i - \tilde{y}_i\big|/n$ as performance metric.

4.1. Synthetic experiment for symmetric predictions of all tasks

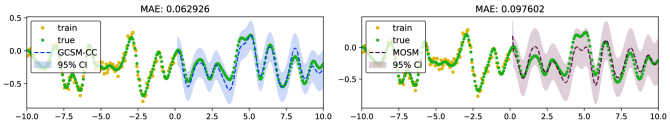
We conduct a synthetic experiment learning multiple incomplete signals. In this context, we validate the interpolation, extrapolation, and signal recovery ability of GCSM-CC and compare its pattern recognition performance with that of other MTGPs. We consider three tasks corresponding to three signals: a mixed signal, its integral, and its derivative, respectively. Specifically, we generate a Gaussian signal with length 300 in the interval [-10, 10] and numerically compute its first integral and derivative. In this experiment, MTGPs

[1] http://slb.nu/slbanalys/historiska-data-luft/

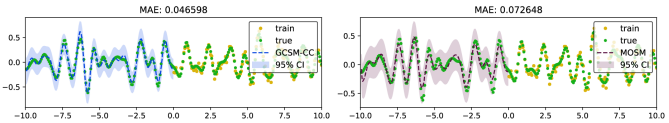
allow bi-directional knowledge transfer between tasks to improve the symmetric predictions of all tasks.



(a) The signal of task 1 sampled from $f \sim \mathcal{GP}(0, K_{SM})$



(b) Task 2 with integral of the signal



(c) Task 3 with derivative of the signal

Fig. 4. Performance of GCSM-CC (in blue dashed line) and MOSM (in plum dashed line) on synthetic MT.

For the 1st task with a signal sampled from $\mathcal{GP}(0, K_{\text{SM}})$, we randomly choose half as the training data, and the rest as testing data. For the 2nd task, the integral of the signal in the interval [-10, 0] are used for training (in dark yellow), while the remaining signal points in the interval [0, 10] are used for testing (in green). For the 3rd task, the derivative of the signal in the interval [0, 10] is used for training and the rest for testing. The performance of GCSM-CC on the generated signal is shown in Fig. 4 (a). According to Table 1, all considered MTGPs have comparable performance: they interpolate well the missing values. The second task, i.e., the integral of the signal, is shown in Fig. 4 (b). In this case its inherent patterns are more difficult to recognize and extrapolate. Here, GCSM-CC is superior to all other methods: it achieves the lowest MAE as well as the smallest confidence interval (CI). Besides, GCSM-CC excels on the last task, i.e., the derivative signal (see Fig. 4 (c)): it shows the best pattern learning and extrapolation capability while using the same number of base component ($Q = 10$). Overall, these results indicate that the learning capability of GCSM-CC in capturing integration and differentiation patterns of the generated signal simultaneously.

4.2. Nitrogen oxides concentration for asymmetric extrapolations of primary tasks

On the contrary, for this real-world experiment, MTGPs focus on improving the prediction performance of primary tasks by transferring valuable knowledge from other correlated tasks. The sensor network dataset recorded from Stockholm city monitor air pollution parameters in order to provide air quality surveillance for the regional environment. In particular, NO is an important parameter, since long term exposure at high concentration can cause inflammation of the human airways. Extrapolation and forecasting models allow to monitor NO concentration in order to control and prevent negative effects on health and environment. As the first real world dataset, we use NO concentration from 5 January, 2017 to 25 January, 2017, in one-hour interval, collected at three stations (Essingeleden, Hornsgatan, Sveavägen) in Stockholm.

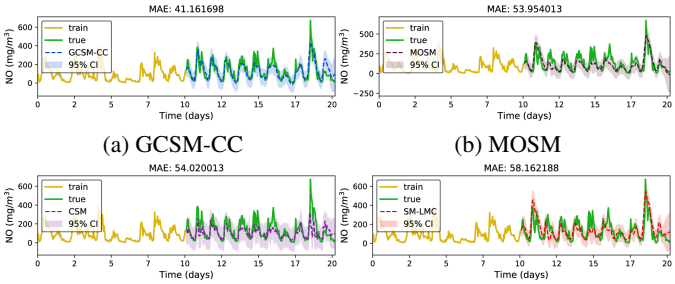
Each station corresponds to a single task: Essingeleden as task 1, Hornsgatan as task 2 and Sveavägen as task 3. NO evolution shows time and phase related patterns and their variability over the period of recording. Different stations have different local patterns

Table 1. Performance (MAE) of GCSM-CC and other MTGPs.

Signal	SE-LMC	Matérn-LMC	GPRN	CSM	MOSM	GCSM-CC
Mixed signal	0.16 ±0.01	0.11 ±0.01	0.12 ±0.01	0.12 ±0.01	0.13 ±0.01	**0.10** ±0.003
Integral	0.26 ±0.01	0.25 ±0.02	0.33 ±0.05	0.19 ±0.06	0.09 ±0.004	**0.06** ±0.003
Derivative	0.18 ±0.01	0.19 ±0.01	0.09 ±0.01	0.17 ±0.02	0.08 ±0.01	**0.04** ±0.01
NO^H	130.96 ±0.41	132.89 ±0.37	58.16 ±1.17	52.02 ±4.28	53.95 ±1.04	**41.16** ±0.95
NO^S	85.06 ±0.38	85.19 ±0.36	45.98 ±2.61	35.48 ±1.17	60.81 ±1.60	**33.39** ±1.54

impacted by the station's surroundings. For instance: Essingeleden's measurement are recorded at open path, Hornsgatan's measurement and Sveavägen's measurement at street. Still, these tasks have shared global trends because of the global seasonal change and periodic characteristics of human and industry activities. The evolution of NO concentration in each task is a result of nonlinear interaction of time- and phase dependent local and global patterns. Therefore, knowledge from correlated tasks should help each other when used to model long range trends.

We aim to assess comparatively the long range extrapolation ability of GCSM-CC for future forecasting and signal recovery simultaneously. Therefore, we perform extrapolation for primary tasks 2 and 3. For task 1, we randomly chose half of the Essingeleden time series as the training data. For task 2, the first half of the Hornsgatan time series is used for training and the remaining data for testing. For task 3, the last half of the Sveavägen time series is used for training and the rest for testing. Table 1 reports the performance of the considered MTGPs on each task of the two experiments. Significantly, the GCSM-CC achieves the lowest MAE.



(a) GCSM-CC (b) MOSM

(c) CSM (d) GPRN

Fig. 5. NO concentration extrapolations.

5. CONCLUSION

We have proposed the GCSM-CC kernel for MTGP with hierarchical interactions. Hierarchical interactions include FI modeled by using the dependency between LFs specified by SM kernel, coefficient interaction constructed by using free-form CC. In this way, GCSM-CC advances the learning capacity and interpretability of MTGPs beyond non-interactive frameworks. Interesting future research involves the development of sparse and efficient inference methods for this interactive MTGP [28].

Acknowledgement

The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Natural Science Foundation of China (NSFC) with grants No. 92067202 and No. 62106212, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, and by China Postdoctoral Science Foundation with grant No. 2020M671899. The corresponding author is Feng Yin.

6. REFERENCES

[1] Carl Edward Rasmussen, *Gaussian processes for machine learning*, Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2006.

[2] Yue Xu, Feng Yin, Wenjun Xu, Jiaru Lin, and Shuguang Cui, "Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1291–1306, 2019.

[3] Yue Xu, Feng Yin, Wenjun Xu, Chia-Han Lee, Jiaru Lin, and Shuguang Cui, "Scalable learning paradigms for data-driven wireless communication," *IEEE Communications Magazine*, vol. 58, no. 10, pp. 81–87, 2020.

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

[5] Kai Chen, Qinglei Kong, Yijue Dai, Yue Xu, Feng Yin, Lexi Xu, and Shuguang Cui, "Recent advances in data-driven wireless communication using Gaussian processes: A comprehensive survey," *China Communications*, vol. 19, pp. 218–237, 2022.

[6] Edwin V Bonilla, Kian M Chai, and Christopher Williams, "Multi-task Gaussian process prediction," in *Advances in neural information processing systems*, 2008, pp. 153–160.

[7] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al., "Kernels for vector-valued functions: A review," *Foundations and Trends® in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.

[8] Haitao Liu, Jianfei Cai, and Yew-Soon Ong, "Remarks on multi-output Gaussian process regression," *Knowledge-Based Systems*, vol. 144, pp. 102–121, 2018.

[9] Joseph Futoma, Sanjay Hariharan, and Katherine Heller, "Learning to detect sepsis with a multitask Gaussian process rnn classifier," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1174–1182.

[10] Ahmed M Alaa and Mihaela van der Schaar, "Deep multi-task Gaussian processes for survival analysis with competing risks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2326–2334.

[11] P Goovaerts, "Geostatistics for natural resources evaluation. oxford univ. press, new york.," *Geostatistics for natural resources evaluation. Oxford Univ. Press, New York.*, 1997.

[12] JA Vargas-Guzmán, AW Warrick, and DE Myers, "Coregionalization by linear combination of nonorthogonal components," *Mathematical Geology*, vol. 34, no. 4, pp. 405–419, 2002.

[13] Arman Melkumyan and Fabio Ramos, "Multi-kernel Gaussian processes," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, vol. 22, p. 1408.

[14] Kai Chen, Twan van Laarhoven, Perry Groot, Jinsong Chen, and Elena Marchiori, "Generalized convolution spectral mixture for multitask Gaussian processes," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 12, pp. 5613–5623, 2020.

[15] Mauricio A Álvarez and Neil D Lawrence, "Computationally efficient convolved multiple output Gaussian processes," *Journal of Machine Learning Research*, vol. 12, pp. 1459–1500, 2011.

[16] Andrew Wilson and Ryan Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1067–1075.

[17] Kai Chen, Twan van Laarhoven, and Elena Marchiori, "Gaussian processes with skewed Laplace spectral mixture kernels for long-term forecasting," *Machine Learning*, vol. 110, no. 8, pp. 2213–2238, 2021.

[18] Kyle R Ulrich, David E Carlson, Kafui Dzirasa, and Lawrence Carin, "GP kernels for cross-spectrum analysis," in *Advances in neural information processing systems*, 2015, pp. 1999–2007.

[19] Gabriel Parra and Felipe Tobar, "Spectral mixture kernels for multi-output Gaussian processes," in *Advances in Neural Information Processing Systems*, 2017, pp. 6684–6693.

[20] Kai Chen, Twan van Laarhoven, Perry Groot, Jinsong Chen, and Elena Marchiori, "Multioutput convolution spectral mixture for Gaussian processes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 2255–2266, 2019.

[21] Robert Dürichen, Marco AF Pimentel, Lei Clifton, Achim Schweikard, and David A Clifton, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 314–322, 2015.

[22] Andrew Gordon Wilson, David A Knowles, and Zoubin Ghahramani, "Gaussian process regression networks," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1139–1146.

[23] Gayle Leen, Jaakko Peltonen, and Samuel Kaski, "Focused multi-task learning in a Gaussian process framework," *Machine Learning*, vol. 89, no. 1-2, pp. 157–182, 2012.

[24] Yan Wen, Guoli Li, Qunjing Wang, Xiwen Guo, and Wenping Cao, "Modeling and analysis of permanent magnet spherical motors by a multi-task Gaussian process method and finite element method for output torque," *IEEE Transactions on Industrial Electronics*, 2020.

[25] Kai Chen, Twan van Laarhoven, Jinsong Chen, and Elena Marchiori, "Incorporating dependencies in spectral kernels for Gaussian processes," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Wrzburg, Germany*, 2019.

[26] Kai Chen, Yijue Dai, Feng Yin, Elena Marchiori, and Sergios Theodoridis, "Novel compressible adaptive spectral mixture kernels for Gaussian processes with sparse time and phase delay structures," *arXiv preprint, arXiv:1808.00560*, 2018.

[27] Alexander G de G Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman, "GPflow: A Gaussian process library using tensorflow," *Journal of Machine Learning Research*, vol. 18, no. 40, pp. 1–6, 2017.

[28] Ang Xie, Feng Yin, Yue Xu, Bo Ai, Tianshi Chen, and Shuguang Cui, "Distributed Gaussian processes hyperparameter optimization for big data using proximal admm," *IEEE Signal Processing Letters*, vol. 26, no. 8, pp. 1197–1201, 2019.