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.

March 2022

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen
icassp 2022 Singapore
Radboud University

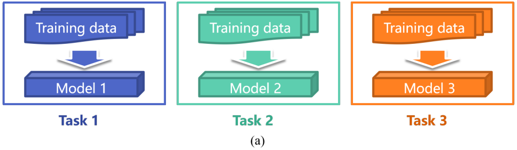
Overview

1. Introduction
    - Multitask Gaussian process (MTGP) and kernel function
    - Our motivation and contribution
2. Multitask Gaussian process with interactions
    - Manner of latent interaction
    - Function interaction between LFs
    - Coefficient interaction between LMCs
    - MTGP with hierarchical interactions
3. Experiment results
    - Performance of multitask learning
4. Summary & Future work

Outline

1. Introduction
   - Multitask Gaussian process (MTGP) and kernel function
   - Our motivation and contribution
2. Multitask Gaussian process with interactions
3. Experiment results
4. Summary & Future work

Multitask Gaussian process (MTGP)



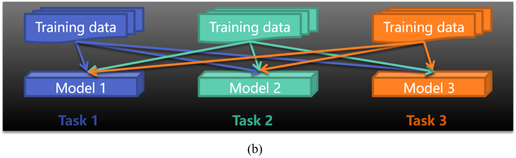
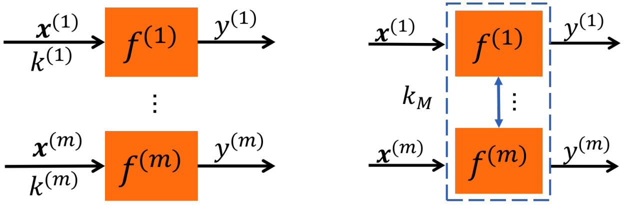


Figure 1: Single task (a) and multitask (b) learnings

Why multitask learning ?

- A complex system usually consists of multiple correlated tasks;
- Joint learning of these tasks can help us understand the system;
- To obtain higher accuracy and concurrent predictions by transferring knowledge;

Multitask Gaussian process (MTGP)



(a) Multiple single task GPs (b) MTGP with multiple single tasks

Figure 2: Conception of MTGP

Why multitask Gaussian process ?

- Interpretability, the correlation between tasks is quantifiable and explainable;
- Uncertainty representation and prediction;
- Less prone to overfitting;

Multitask Gaussian process (MTGP)

Definition of MTGP

A MTGP models m tasks as a joint Gaussian distribution,

$$\boldsymbol{f} \sim \mathcal{GP}(0, k_{\mathrm{M}}(X, X')), \tag{1}$$

where $\mathbf{y} = \boldsymbol{f} + \epsilon$, $\boldsymbol{f} = [f^{(1)}(\mathbf{x}^{(1)}), ..., f^{(m)}(\mathbf{x}^{(m)})]^\top$, $X = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(m)}]^\top$, $\mathbf{y} = [y^{(1)}, ..., y^{(m)}]^\top$, and $k_{\mathrm{M}}(X, X')$ describes both the auto-covariance of each task and the cross covariance between tasks.

- k_{M} determines the representation, smoothness, interpretability, and expressiveness of MTGP.

Multitask kernel function

The general formula of k_{M} is,

$$K_{\mathrm{M}}(X, X') = \begin{bmatrix} K^{(1,1)}, & K^{(1,m)} \\ K^{(m,1)}, & K^{(m,m)} \end{bmatrix}, \tag{2}$$

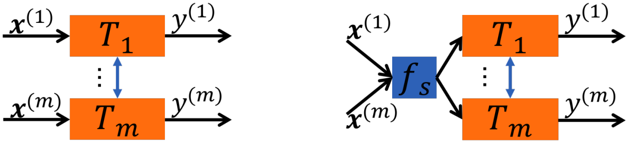
where $K^{(m,1)}$ describes the cross covariance between $f^{(m)}$ and $f^{(1)}$.

An expressive k_{M} using linear model of coregionalization (LMC) linearly combines a mixture of Q covariance components to ameliorate the representation of MTGP :

$$K_{\mathrm{M}} = \sum_{i=1}^{Q} B_i \otimes K_{s,i}, \tag{3}$$

where B_i is a LMC matrix representing task correlation and $K_{s,i}$ is a matrix constructed by arbitrary kernel of STGP.

Multitask kernel function



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

Figure 3: Frameworks of MTGPs. In subplot (a), task correlation in MTGP is described by the convolution between latent functions (LFs). In subplot (b), all tasks share a LF space f_s adumbrating their common qualities.

- The LMC has a clearer hierarchical architecture for model explanation;
- The LMC has more compact hyperparameter space due to the shared LF;

Outline

Our motivation

Limitation of existing MTGPs

- Considers linear combination of LFs as well as kernels;
- Ignores interactions between them;

Our contribution

We develop a novel LMC framework with hierarchical interactions for MTGP:

- 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 demonstrate the rich representation, interpretability, and expressiveness of a kernel framework incorporating both the function and coefficient interactions for MTGP.

Outline

Manner of latent interaction

$k_{s,i}$ ensures the generalization of $f_{s,i}(\mathbf{x}^{(m)})$ and we therefore replace it with $k_{\mathrm{SM},i}$

$$f^{(m)}(\mathbf{x}^{(m)}) = \sum_{i=1}^{Q} \alpha_i^{(m)} f_{\mathrm{SM},i}(\mathbf{x}^{(m)}), \tag{4}$$

where $f_{\mathrm{SM},i} \sim \mathcal{GP}(0, k_{\mathrm{SM},i})$ and $k_{\mathrm{SM},i}(\tau) = \mathcal{F}^{-1}_{s \to \tau}\left[w_i \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_i, \Sigma_i)\right](\tau)$.

Due to $f_{\mathrm{SM},i}(\mathbf{x}^{(m)}) \not\perp f_{\mathrm{SM},j}(\mathbf{x}^{(m)})$ and $\mathrm{cov}[f_{\mathrm{SM},i}(\mathbf{x}^{(m)}), f_{\mathrm{SM},j}(\mathbf{x}^{(m)})] \neq 0$ when $i \neq j$, we consider two hierarchical interactions in LMC:

- FI between LFs $f_{\mathrm{SM},i}$ and $f_{\mathrm{SM},j}$;
- coefficient interaction between $\alpha_i^{(m)}$ and $\alpha_j^{(m)}$;

Outline

Function interaction between LFs

By following the so-called generalized convolution spectral mixture (GCSM) in[1] and view it as a general sense of FI between LFs.

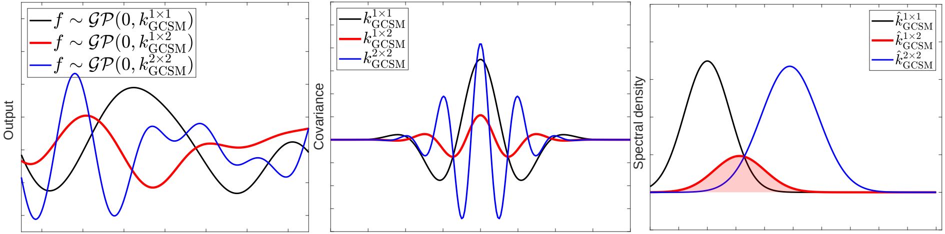
$$k_{\text{GCSM}}^{i\times j}(\tau) = c_{ij} \exp\Big(-\frac{1}{2}\tau_{\boldsymbol{\theta}}^{\top}\Sigma_{ij}\tau_{\boldsymbol{\theta}}\Big)\cos\Big(\tau_{\boldsymbol{\theta}}^{\top}\boldsymbol{\mu}_{ij} - \boldsymbol{\phi}_{ij}\pi\Big), \tag{5}$$

where c_{ij} is the cross constant, $\tau_{\boldsymbol{\theta}} = 2\pi(\tau - \frac{\boldsymbol{\theta}_{ij}}{2})$ is the Euclidean distance with time delay, $\boldsymbol{\mu}_{ij}$ is cross inverse period, Σ_{ij} is cross inverse length scale, $\boldsymbol{\theta}_{ij}$ is cross time delay, and $\boldsymbol{\phi}_{ij}$ is cross phase delay between $f_{\text{SM},i}$ and $f_{\text{SM},j}$, respectively.

Note that $k_{\text{GCSM}}^{i\times j}(\tau) = k_{\text{SM},i}$ for $i = j$

[1] Chen et al. 2019.

Function interaction between LFs



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

Figure 4: The illustrations of LFs (a), covariances (b), and SDs (c) of their interaction.

Includes more interaction terms without the increasing of hyperparameter space (if remove time and phase delays).

Outline

Coefficient interaction between LMCs

For the coefficient interaction between $\alpha_i^{(m)}$ and $\alpha_j^{(m)}$, we reformulate $\alpha_i^{(m)}$ in a free-form parameterized[2] matrix B_i as $B_i(m, m') = \alpha_i^{(m)} \alpha_i^{(m')}$.

- Decompose B_i as: $B_i = B_{L,i} B_{L,i}^\top$ with the Cholesky factor $B_{L,i} = \begin{bmatrix} \ell^i_{1,1}, & 0 \\ \ell^i_{m,1}, & \ell^i_{m,m} \end{bmatrix}$, where $\ell^i_{m,m'}$ can be seen as the correlation between tasks m and m';
- There are $m(m+1)/2$ hyperparameters for B_i;

We construct $B_{ij} = B_{L,i} B_{L,j}^\top$ to encode the coefficient interaction and interpretate $B_{i,j}$ as a free-form coupling coregionalization (CC).

[2] Bonilla, Chai and Williams 2008.

Outline

MTGP with hierarchical interactions

For MTGP with Q LFs constructed by using LMC and SM kernel, we have a kernel with hierarchical interactions as follows:

$$K_{\text{GCSM-CC}}(\tau) = \sum_{i=1}^{Q} \sum_{j=1}^{Q} B_{ij} \otimes k_{\text{GCSM}}^{i \times j}(\tau). \tag{6}$$

- MTGP with $k_{\text{GCSM-CC}}$ has richer representation capacity than existing LMC and convolution frameworks because there are $Q^2 - Q$ cross interaction terms when $i \neq j$.

Experiment setting

Performance metric: $\mathrm{MAE} = \sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right| / n$.

We compare GCSM-CC with some advanced baselines:

- Traditional MTGP[3]
- GPRN[4]
- CSM[5]
- MOSM[6]
- ......

[3] Bonilla, Chai and Williams 2008.

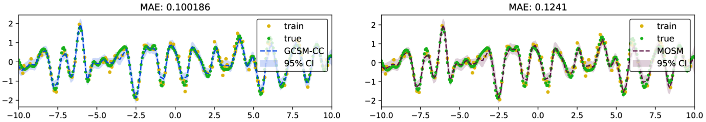
[4] Wilson, Knowles and Ghahramani 2012.

[5] Ulrich et al. 2015.

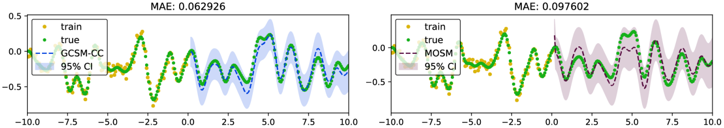
[6] Parra and Tobar 2017.

Outline

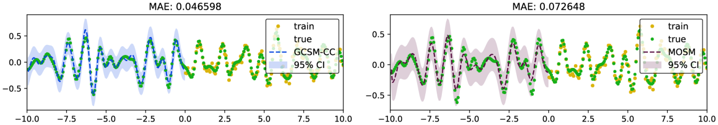
Performance on synthetic multitask



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



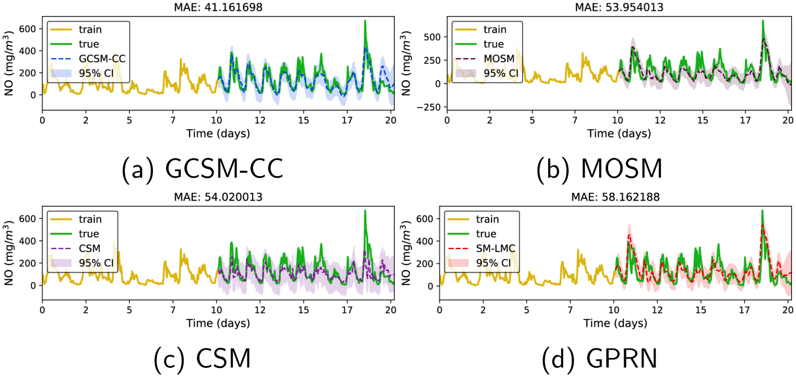
(b) Task 2 with integral of the signal



(c) Task 3 with derivative of the signal

Figure 5: Performance of GCSM-CC (in blue dashed line) and MOSM (in plum dashed line) on synthetic MT.

Asymmetric extrapolations of primary tasks



(a) GCSM-CC (b) MOSM

(c) CSM (d) GPRN

Figure 6: Nitrogen oxides concentration extrapolations.

Experiment performance

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

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

Summary & Future work

Summary

- GCSM-CC with hierarchical interactions includes FI modeled by convolution of SM kernels and coefficient interaction constructed by using free-form coupling coregionalization.
- GCSM-CC advances the learning capacity and interpretability of MTGPs beyond non-interactive frameworks.

Future work

- Interesting future research involves sparse and efficient inference methods for current MTGPs.

References I

Alvarez, Mauricio A, Lorenzo Rosasco, Neil D Lawrence et al. (2012). "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3, pp. 195–266.

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

Chen, Kai et al. (2019). "Incorporating Dependencies in Spectral Kernels for Gaussian Processes". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany*.

Parra, Gabriel and Felipe Tobar (2017). "Spectral Mixture Kernels for Multi-Output Gaussian Processes". In: *Advances in Neural Information Processing Systems*, pp. 6684–6693.

Rasmussen, Carl Edward (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press. xviii, 248 Seiten. ISBN: 9780262182539.

Ulrich, Kyle R et al. (2015). "GP kernels for cross-spectrum analysis". In: *Advances in neural information processing systems*, pp. 1999–2007.

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

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

Feedback welcome!

The End