Graphical Multioutput Gaussian Process with Attention

ICLR 2024 – *Spotlight*

The Chinese University of Hong Kong

Yijue Dai, Wenzhong Yan, Feng Yin

April 18, 2024

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

1. How to measure output **dependence**? (graphical representation)

1. How to measure output **dependence**? (graphical representation)
2. How to learn **individual** knowledge while aggregating **related** information?

1. How to measure output **dependence**? (graphical representation)
2. How to learn **individual** knowledge while aggregating **related** information?
3. A **flexible/scalable** model with tractable predictions and uncertainty quantification.

1. How to measure output **dependence**? (graphical representation)
2. How to learn **individual** knowledge while aggregating **related** information?
3. A **flexible/scalable** model with tractable predictions and uncertainty quantification.
4. Distributed framework/workflow with Pareto optimal hyperparameters.

1. How to measure output **dependence**? (graphical representation)
2. How to learn **individual** knowledge while aggregating **related** information?
3. A **flexible/scalable** model with tractable predictions and uncertainty quantification.
4. Distributed framework/workflow with Pareto optimal hyperparameters.

- *At the same time!*

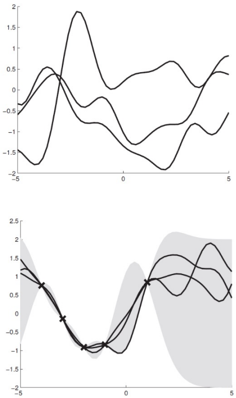
Table of Contents

Gaussian process

1 Gaussian process regression (GPR)

- Given a dataset $\mathcal{D} : \{X, \mathbf{y}\} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, the GPR model can be described as:

$$y_n = f(\mathbf{x}_n) + \epsilon_n, \ n = 1, 2, \ldots, N, \tag{1}$$

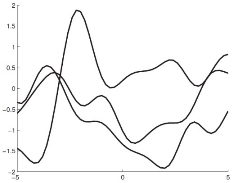


Gaussian process

1 Gaussian process regression (GPR)

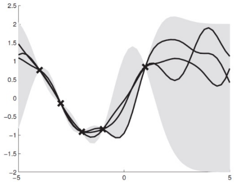
- Given a dataset $\mathcal{D}:\{X,\mathbf{y}\}=\{\mathbf{x}_n,y_n\}_{n=1}^{N}$, the GPR model can be described as:

$$y_n=f(\mathbf{x}_n)+\epsilon_n,\;\; n=1,2,\ldots,N, \tag{1}$$



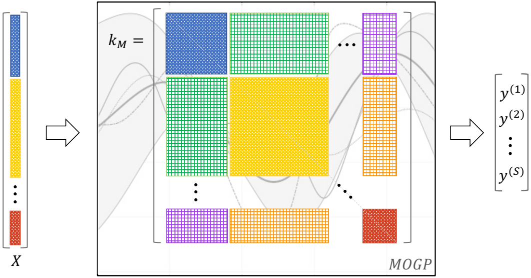
A GP characterizes a distribution over functions fully by a mean function $m(\mathbf{x})$ and a kernel function $k(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})$, i.e.,

$$f(\mathbf{x})\sim\mathcal{GP}\left(m(\mathbf{x}),k(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})\right). \tag{2}$$



For multi-output regression, an MOGP can be derived as:

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\boldsymbol{m}_M(\mathbf{x}), K_M(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}_M)), \tag{3}$$



where $\mathbf{f}(\mathbf{x}) = [f^{(1)}(\mathbf{x}), f^{(2)}(\mathbf{x}), \ldots, f^{(S)}(\mathbf{x})]$ and matrix-valued kernel $K_M(X, X) \in \mathbb{R}^{SN \times SN}$.

MOGP Inference

1 Gaussian process regression (GPR)

Conditioning the joint Gaussian prior on the observations, the predictive distribution for a test input $\mathbf{x}_*$ turns out to be:

$$p(\mathbf{f}_* | \mathbf{x}_*, X, Y, \boldsymbol{\theta}_M) = \mathcal{N}(\bar{\mathbf{f}}_*, \mathbb{V}_*) \tag{4}$$

with (omitting kernel hyperparameters)

$$\begin{cases} \bar{\mathbf{f}}_* = K_M(\mathbf{x}_*, X)(K_M(X, X) + \boldsymbol{\Sigma})^{-1} Y & (5) \\ \mathbb{V}_* = K_M(\mathbf{x}_*, \mathbf{x}_*) - K_M(\mathbf{x}_*, X)(K_M(X, X) + \boldsymbol{\Sigma})^{-1} K_M(X, \mathbf{x}_*). & (6) \end{cases}$$

Benchmark and Chanllenges

1 Gaussian process regression (GPR)

- The popular LMC models tailor distinct coefficients to each output via Q shared independent GPs. The elements of their kernel functions are formulated as following:

$$K_{i,i'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} a_{iq} a_{i'q} k_q(\mathbf{x},\mathbf{x}'), \quad \forall i, i \in \{1, 2, \ldots, S\}. \tag{7}$$

- The popular LMC models tailor distinct coefficients to each output via Q shared independent GPs. The elements of their kernel functions are formulated as following:

$$K_{i,i'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} a_{iq}a_{i'q}k_q(\mathbf{x},\mathbf{x}'), \quad \forall i, i \in \{1, 2, \ldots, S\}. \tag{7}$$

- Different modes of the coefficients $a_{iq}a_{i'q}$ correspond to varied LMC variants.

Benchmark and Chanllenges

1 Gaussian process regression (GPR)

- The popular LMC models tailor distinct coefficients to each output via Q shared independent GPs. The elements of their kernel functions are formulated as following:

$$K_{i,i'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} a_{iq}a_{i'q}k_q(\mathbf{x},\mathbf{x}'), \quad \forall i, i \in \{1, 2, \ldots, S\}. \tag{7}$$

- Different modes of the coefficients $a_{iq}a_{i'q}$ correspond to varied LMC variants.

Common challenges

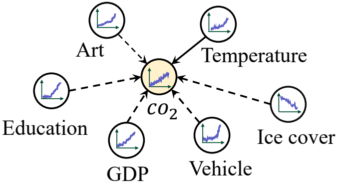
(i) High computational and storage burdens w.r.t. the SN-dimensional correlation matrix;
(ii) Model mismatch when the underlying likelihood deviates from Gaussian;
(iii) Inflexible/symmetric dependence measure (covariance).

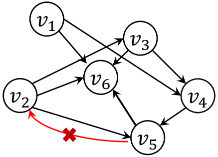
Table of Contents

Directed Graphical Model

2 Graphical MOGP

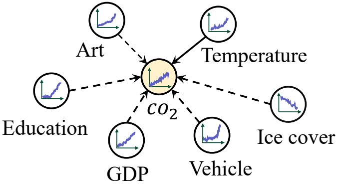
- Ubiquitously, there exists interplay among multiple outputs, such as,

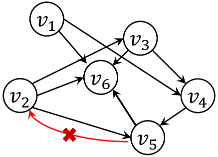




Directed Graphical Model

- Ubiquitously, there exists interplay among multiple outputs, such as,





- Given nodes $\{v_1, v_2, \dots, v_6\}$ and graph G_v shown in the RHS, the joint distribution can be decomposed by repeatedly applying the product rule of probability:

$$
\begin{aligned}
p(v_1, v_2, \dots, v_6) &\overset{\mathrm{G}_v}{=} p(v_6|v_1, v_2, v_3, v_5)p(v_5|v_2, v_4)p(v_4|v_1, v_3)p(v_3|v_2)p(v_2)p(v_1) \\
&= \prod_{i=1}^{6} p(v_i|\mathrm{pa}_i). \qquad (8)
\end{aligned}
$$

Graphical MOGP Model

- For $S > 1$ outputs, we can model each output as an SOGP, and generate multivariate Gaussian random variables evaluated at X (represented by node $f_X^{(j)},\ j \in \mathcal{I}$).

Graphical MOGP Model

- For $S > 1$ outputs, we can model each output as an SOGP, and generate multivariate Gaussian random variables evaluated at X (represented by node $f_X^{(j)},\ j \in \mathcal{I}$).

The joint distribution defined by the specific graph structure, with the target node $f_X^{(i)}$ connected to the heads of arrows, can be derived as follows:

$$p(f_X^{(1)}, f_X^{(2)}, \ldots, f_X^{(S)}) = p(f_X^{(i)} | \mathrm{pa}_i) \prod_{j \in \mathrm{pa}_i} p(f_X^{(j)}). \tag{9}$$

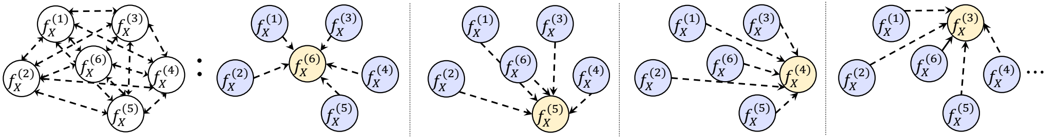
Graphical MOGP Model

2 Graphical MOGP

- For $S > 1$ outputs, we can model each output as an SOGP, and generate multivariate Gaussian random variables evaluated at X (represented by node $f_X^{(j)},\ j \in \mathcal{I}$).

The joint distribution defined by the specific graph structure, with the target node $f_X^{(i)}$ connected to the heads of arrows, can be derived as follows:

$$p(f_X^{(1)}, f_X^{(2)}, \ldots, f_X^{(S)}) = p(f_X^{(i)} | \mathrm{pa}_i) \prod_{j \in \mathrm{pa}_i} p(f_X^{(j)}). \tag{9}$$



GMOGP Prior

2 Graphical MOGP

- We model the conditional distribution as Gaussian with aggregated information, i.e.,

$$p(f_X^{(i)}|\mathrm{pa}_i) = \mathcal{N}\left(f_X^{(i)} \middle| \sum_{j\in\mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i, k_{\boldsymbol{\theta}_i}(X,X)\right), \quad i \in \mathcal{I}, \tag{10}$$

Conditioning on the states of its parents, each target node is of the form:

$$f_X^{(i)} = \sum_{j\in\mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i + \psi_i, \tag{11}$$

with $\boldsymbol{m}_i$ and $\psi_i \sim \mathcal{N}(\mathbf{0}, k_{\boldsymbol{\theta}_i}(X,X))$ characterizing the i^{th} output.

GMOGP Prior

- We model the conditional distribution as Gaussian with aggregated information, i.e.,

$$p(f_X^{(i)}|\mathrm{pa}_i) = \mathcal{N}\left(f_X^{(i)} \middle| \sum_{j\in\mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i, k_{\boldsymbol{\theta}_i}(X,X)\right), \quad i \in \mathcal{I}, \tag{10}$$

Conditioning on the states of its parents, each target node is of the form:

$$f_X^{(i)} = \sum_{j\in\mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i + \psi_i, \tag{11}$$

with $\boldsymbol{m}_i$ and $\psi_i \sim \mathcal{N}(\boldsymbol{0}, k_{\boldsymbol{\theta}_i}(X,X))$ characterizing the i^{th} output.

In the context, the GMOGP prior for each target node ($i \in \mathcal{I}$) follows:

$$p(f_X^{(i)}) = \mathcal{N}\left(\sum_{j\in\mathrm{pa}_i} \alpha_{i,j} \boldsymbol{m}_j + \boldsymbol{m}_i, \sum_{j\in\mathrm{pa}_i} \alpha_{i,j}^2 k_{\boldsymbol{\theta}_j}(X,X) + k_{\boldsymbol{\theta}_i}(X,X)\right) \tag{12}$$

Learning Parents with Attention

2 Graphical MOGP

- The attention coefficients can be learned by using an attention mechanism $\alpha_{i,j} = \exp(e_{i,j})/(1 + \sum_{j \in \mathsf{pa}_i} \exp(e_{i,j}))$ and a modified scoring function (dynamic):

$$e_{i,j} = \text{LeakyReLU}\left(\langle f_X^{(i)}, f_X^{(j)} \rangle \mathrm{w}_{ij} + c_{ij}\right), \tag{13}$$

Learning Parents with Attention

2 Graphical MOGP

- The attention coefficients can be learned by using an attention mechanism $\alpha_{i,j} = \exp(e_{i,j})/(1 + \sum_{j \in \mathrm{pa}_i} \exp(e_{i,j}))$ and a modified scoring function (dynamic):

$$e_{i,j} = \mathrm{LeakyReLU}\left(\langle f_X^{(i)}, f_X^{(j)} \rangle \mathrm{w}_{ij} + c_{ij}\right), \tag{13}$$

In practice, we substitute the observation values $\langle \mathbf{y}^{(i)}, \mathbf{y}^{(j)} \rangle$ at the initial learning phase. When $\alpha_{i,j} \approx 0$, the parents set pa_i is adjusted by unlinking the node j ($j \in \mathcal{I}, j \neq i$).

Learning Parents with Attention

2 Graphical MOGP

- The attention coefficients can be learned by using an attention mechanism $\alpha_{i,j} = \exp(e_{i,j})/(1 + \sum_{j \in \mathrm{pa}_i} \exp(e_{i,j}))$ and a modified scoring function (dynamic):

$$e_{i,j} = \text{LeakyReLU}\left(\langle f_X^{(i)}, f_X^{(j)} \rangle \mathrm{w}_{ij} + c_{ij}\right), \tag{13}$$

In practice, we substitute the observation values $\langle \mathbf{y}^{(i)}, \mathbf{y}^{(j)} \rangle$ at the initial learning phase. When $\alpha_{i,j} \approx 0$, the parents set pa_i is adjusted by unlinking the node j $(j \in \mathcal{I}, j \neq i)$.

- Asymmetric dependence measure and capture covariance from $\mathrm{cov}(f_X^{(j)}, f_X^{(i)}) = \sum_{j' \in \mathrm{pa}_i} \alpha_{i,j'} \mathrm{cov}(f_X^{(j)}, f_X^{(j')}) + I_{ij} k_{\theta_i}(X, X)$.

Learning Parents with Attention

- The attention coefficients can be learned by using an attention mechanism $\alpha_{i,j} = \exp(e_{i,j})/(1 + \sum_{j \in \mathrm{pa}_i} \exp(e_{i,j}))$ and a modified scoring function (dynamic):

$$e_{i,j} = \mathrm{LeakyReLU}\left(\langle f_X^{(i)}, f_X^{(j)} \rangle \mathrm{w}_{ij} + c_{ij}\right), \tag{13}$$

In practice, we substitute the observation values $\langle \mathbf{y}^{(i)}, \mathbf{y}^{(j)} \rangle$ at the initial learning phase. When $\alpha_{i,j} \approx 0$, the parents set pa_i is adjusted by unlinking the node j $(j \in \mathcal{I}, j \neq i)$.

- Asymmetric dependence measure and capture covariance from $\mathrm{cov}(f_X^{(j)}, f_X^{(i)}) = \sum_{j' \in \mathrm{pa}_i} \alpha_{i,j'} \mathrm{cov}(f_X^{(j)}, f_X^{(j')}) + I_{ij} k_{\theta_i}(X, X)$.
- Heterotopic data and new comings.

Model Learning and Inference

2 Graphical MOGP

- Model parameters w.r.t. each target, $\gamma^{(i)} := \{\Theta, \alpha_i, \sigma_i\}$, can be updated via minimizing:

$$\mathcal{L}_{\gamma^{(i)}}^{(i)} \propto \left\{ (\tilde{\mathbf{y}}^{(i)})^T \left(k_G^{(i)}(X,X) + \sigma_i^2 I_N \right)^{-1} \tilde{\mathbf{y}}^{(i)} + \log \left| k_G^{(i)}(X,X) + \sigma_i^2 I_N \right| \right\}, \tag{14}$$

where $\tilde{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)} - (\sum_{j \in \mathrm{pa}_i} \alpha_{i,j} \boldsymbol{m}_j + \boldsymbol{m}_i)$, $k_G^{(i)}(X,X) = \sum_{j \in \mathrm{pa}_i} \alpha_{i,j}^2 k_{\boldsymbol{\theta}_j}(X,X) + k_{\boldsymbol{\theta}_i}(X,X)$.

Model Learning and Inference

2 Graphical MOGP

- Model parameters w.r.t. each target, $\gamma^{(i)} := \{\Theta, \alpha_i, \sigma_i\}$, can be updated via minimizing:

$$\mathcal{L}_{\gamma^{(i)}}^{(i)} \propto \left\{ (\tilde{\mathbf{y}}^{(i)})^T \left(k_G^{(i)}(X,X) + \sigma_i^2 I_N \right)^{-1} \tilde{\mathbf{y}}^{(i)} + \log \left| k_G^{(i)}(X,X) + \sigma_i^2 I_N \right| \right\}, \tag{14}$$

where $\tilde{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)} - (\sum_{j \in \mathrm{pa}_i} \alpha_{i,j} \boldsymbol{m}_j + \boldsymbol{m}_i)$, $k_G^{(i)}(X,X) = \sum_{j \in \mathrm{pa}_i} \alpha_{i,j}^2 k_{\boldsymbol{\theta}_j}(X,X) + k_{\boldsymbol{\theta}_i}(X,X)$.

multi-objective optimization (MOO)

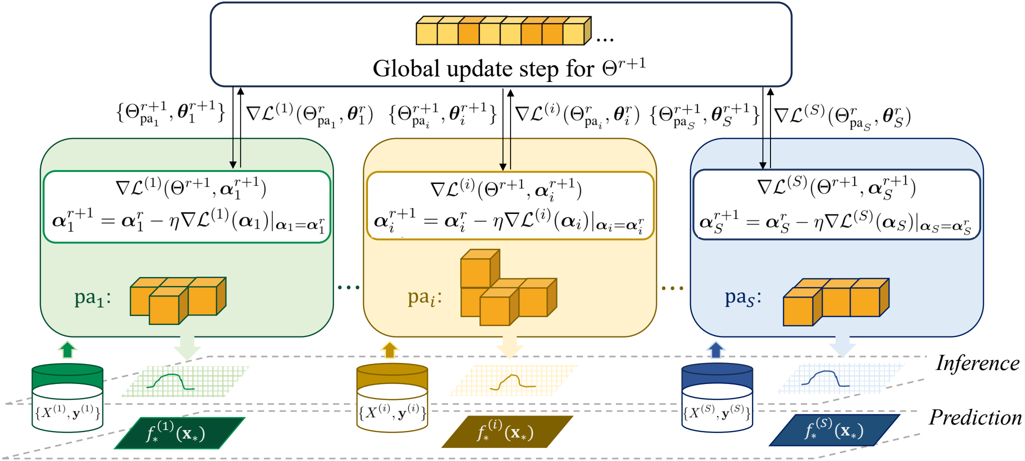
The objectives with shared kernel hyperparameters can be modeled as a MOO problem,

$$\boldsymbol{F}(\Theta) = [\mathcal{L}^{(1)}(\Theta), \mathcal{L}^{(2)}(\Theta), \ldots, \mathcal{L}^{(S)}(\Theta)]^T. \tag{15}$$

- Applying weighted sum method with loss $\sum_{i=1}^{S} w_i \mathcal{L}^{(i)}(\Theta)$, $w_i > 0$, to solve the problem provides a sufficient condition for Pareto optimality of the kernel hyperparameters.

Distributed Framework and Workflow

2 Graphical MOGP



Non-Gaussian Prior

- The target node can further aggregate non-linear relations, i.e.,

$$\mathbb{G}_{\phi_k^{(i)}}(f_X^{(i)}) = \mathbb{G}_{\phi_k^{(i)}}\Big(\sum_{j \in \mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i + \psi_i\Big). \tag{16}$$

- The target node can further aggregate non-linear relations, i.e.,

$$\mathbb{G}_{\phi_k^{(i)}}(f_X^{(i)}) = \mathbb{G}_{\phi_k^{(i)}}\left(\sum_{j\in \mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i + \psi_i\right). \tag{16}$$

- In the sequel, the transformed GMOGP prior is ($\forall k \in \{0, 1, \ldots, K-1\}$):

$$p_{\gamma^{(i)},\Phi_i}\left(f_{K_X}^{(i)} | \mathbb{G}, X\right) = p_{\gamma^{(i)}}\left(f_{0_X}^{(i)}\right) \prod_{k=0}^{K-1} \left| \det \frac{\partial \mathbb{G}_{\phi_k^{(i)}}\left(f_{k_X}^{(i)}\right)}{\partial f_{k_X}^{(i)}} \right|^{-1}, \tag{17}$$

Non-Gaussian Prior

- The target node can further aggregate non-linear relations, i.e.,

$$\mathbb{G}_{\phi_k^{(i)}}(f_X^{(i)}) = \mathbb{G}_{\phi_k^{(i)}}\left(\sum_{j\in \mathrm{pa}_i} \alpha_{i,j} f_X^{(j)} + \boldsymbol{m}_i + \psi_i\right). \tag{16}$$

- In the sequel, the transformed GMOGP prior is ($\forall k \in \{0, 1, \ldots, K-1\}$):

$$p_{\gamma^{(i)},\boldsymbol{\Phi}_i}(f_{K_X}^{(i)} | \mathbb{G}, X) = p_{\gamma^{(i)}}(f_{0_X}^{(i)}) \prod_{k=0}^{K-1} \left| \det \frac{\partial \mathbb{G}_{\phi_k^{(i)}}(f_{k_X}^{(i)})}{\partial f_{k_X}^{(i)}} \right|^{-1}, \tag{17}$$

- Variational inference and negative evidence lower bound (NELBO) are applied:

$$\min_{\substack{\{\gamma^{(i)},\boldsymbol{\Phi}_i,\boldsymbol{u}_0^{(i)},\boldsymbol{m}_u^{(i)},K_u^{(i)}\};\\ i=1,2,\ldots,S}} -\left(\sum_{i=1}^{S} \mathbb{E}_{q(f_{0_X}^{(i)})}\left[\log p(\mathbf{y}^{(i)}|\mathbb{G}_{\boldsymbol{\Phi}_i}(f_{0_X}^{(i)}))\right] + \mathbb{E}_{q(\boldsymbol{u}_0^{(i)})}\left[\log \frac{p(\boldsymbol{u}_0^{(i)})}{q(\boldsymbol{u}_0^{(i)})}\right]\right) \tag{18}$$

Table of Contents

A multi-output regression task with non-Gaussian noise and different function compositions is evaluated. Five outputs ($S = 5$) are generated by the following functions specialized at $X \in \mathbb{R}^{1800\times 2}$:

$$\mathbf{y}^{(1)} = f_1(X) + \epsilon_1, \tag{19}$$

$$\mathbf{y}^{(2)} = f_1(X) + f_2(X) + \epsilon_2, \tag{20}$$

$$\mathbf{y}^{(3)} = \sinh(2\,\text{arcsinh}(f_1(X) + f_2(X)) + \epsilon_3), \tag{21}$$

$$\mathbf{y}^{(4)} = 3\tanh(f_3(X)f_4(X) + f_1(X) + \epsilon_4), \tag{22}$$

$$\mathbf{y}^{(5)} = 5f_3(X)f_4(X) + \epsilon_5, \tag{23}$$

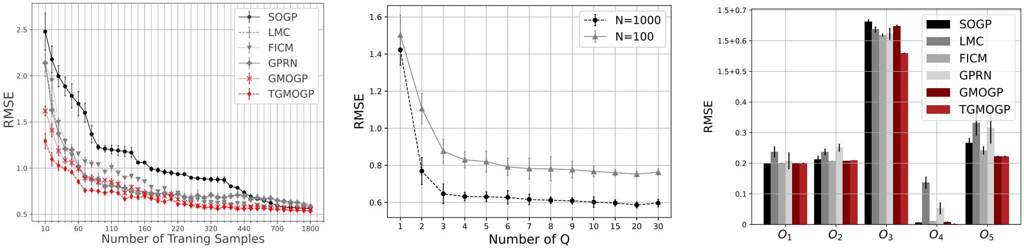
where
$f_1(\mathbf{x}) = 2\cos(x_1 + x_2), f_2(\mathbf{x}) = (x_1 + x_2)^2, f_3(\mathbf{x}) = \exp(|x_1 x_2| + 1), f_4(\mathbf{x}) = \log(x_1 + 3)$,
and $\epsilon_1, \epsilon_2, \ldots, \epsilon_5$ are i.i.d. Gaussian noise with a common standard deviation 0.2.

Table: The average test RMSE of the synthetic experiments. All metrics are compared against the baselines: [1] Isolated SOGPs, [2] LMC, [3] free-form task similarity model (FICM), [4] Gaussian process regression network (GPRN), and [5] convolution process (CMOGP). (l_{NF} : The flow parameters, V_m : The variational parameters.)

	Average RMSE	Test NLL	K_{dim}	Number of Parameters
[1] SOGP	0.5653±0.0023	0.4891±0.0043	N	$4S$
[2] LMC	0.5917±0.0096	0.5543±0.0506	$S \times N$	$(2+S)Q+2S+1$
[3] FICM	0.5544±0.0046	0.4798±0.0176	$S \times N$	$(S(S+5)+4)/2$
[4] GPRN	0.5819±0.0207	0.5787±0.0445	$S \times N$	$2(S+Q)+3$
[5] CMOGP	0.5539±0.0089	0.4689±0.0143	$S \times N$	$(2+S)Q+2S+1$
[6] GMOGP	0.5541±0.0054	0.1636±0.0143	N	$S(S+5)+1$
[7] TGMOGP	0.5343±0.0023	-0.6354±0.0023	N	$S(S+5)+1+3l_{NF}+V_m$

Synthetic Data Experiment Results

3 Experiments and Summary



(e) (f) (g)

Figure: The sub-figures show (e) the average RMSE changes with the number of training samples, (f) the RMSE versus the number of latent independent GPs, and (g) the test error of each output.

Real-World Data Experiments

Table: Comparison of test RMSE on real datasets, where SGPRN and variational LMC (V-LMC) are tested. The shadowed results are learned with two distributed computing units.

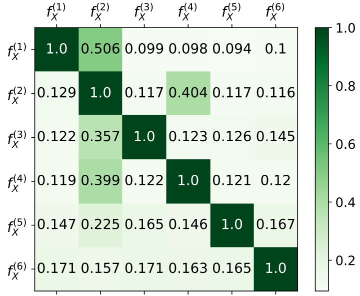
Datasets	SOGP	V-LMC$_{100}$	FICM	SGPRN	GMOGP	TGMOGP$_{100}$
JURA	0.605±0.01	0.443±0.01	0.394±0.05	0.438±0.02	**0.376±0.01**	0.382±0.01
ECG	0.245±0.01	0.229±0.01	0.222±0.01	0.232±0.02	0.219±0.00	**0.217±0.00**
EEG	0.343±0.05	0.207±0.03	0.147±0.03	0.261±0.03	**0.082±0.01**	0.117±0.00
SARCOS$_1$	1.139±0.01	1.063±0.01	0.792±0.04	0.844±0.04	0.643±0.03	**0.558±0.00**
KUKA	0.05±0.01	0.14±0.01	0.03±0.01	0.12±0.02	**0.02/ 0.02 ±0.00**	0.04/ 0.04 ±0.01
Test NLL	-0.25±0.01	-0.51±0.01	-0.65 ±0.02	-0.55±0.01	-1.81/ -1.76 ±0.02	**-3.49/ -3.48 ±0.01**
SARCOS$_2$	0.26±0.05	0.29±0.04	0.33±0.03	-	0.21/ 0.22 ±0.02	**0.16/ 0.16 ±0.01**
Time/Iter	20.75(s)	370.2(s)	419.3(s)	>2400(s)	32.41/ 21.07 (s)	**4.65/ 3.43 (s)**

Test Attention Coefficients

3 Experiments and Summary

Coefficient	$\alpha_{1,2}$	$\alpha_{1,3}$	$\alpha_{1,4}$	$\alpha_{2,1}$	$\alpha_{2,3}$	$\alpha_{2,4}$	$\alpha_{3,1}$	$\alpha_{3,2}$	$\alpha_{3,4}$	$\alpha_{4,1}$	$\alpha_{4,2}$	$\alpha_{4,3}$
SARCOS_2	1.9e-4	0.998	2.9e-4	5.4e-4	6.9e-4	0.898	7.3e-5	5.1e-5	0.989	1.3e-3	0.966	6.7e-4

Figure: The attention coefficients for real traffic data.



Summary

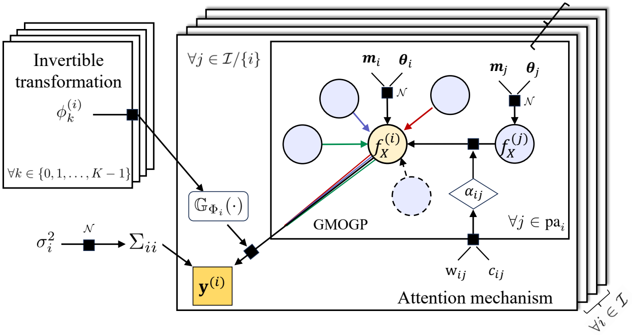


Figure: Visualizing variable dependencies in the GMOGP. Each target output has their own parents, transformations, and samples with knowledge exchanged by kernel hyperparameters $\boldsymbol{\theta}_{j,\, j \in \mathrm{pa}_i}$.

Graphical Multioutput Gaussian Process with Attention

Thank you for listening!

Any questions?