Modular Gaussian Processes for Transfer Learning

Pablo Moreno-Muñoz, Antonio Artés, Mauricio A. Álvarez

Section for Cognitive Systems, Technical University of Denmark (DTU)
Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
Dept. of Computer Science, University of Sheffield, UK

to appear @ NeurIPS 2021

Introduction

Imagine a supervised learning problem, for instance *regression*, where *N* data points are processed for training a model. At a later time, new data are observed, this time corresponding to a binary *classification* task, that we know are generated by the same phenomena, e.g. using a different sensor. Having kept the observations from regression stored, a common approach would be to use them in combination with the classification dataset to generate a new model. This practice might be **inconvenient** because of:

1) the need of **centralising** the data to train the model
2) the rising data-dependent **computational cost**
3) the **obsolescence** of fitted models, whose usability is not guaranteed for new data

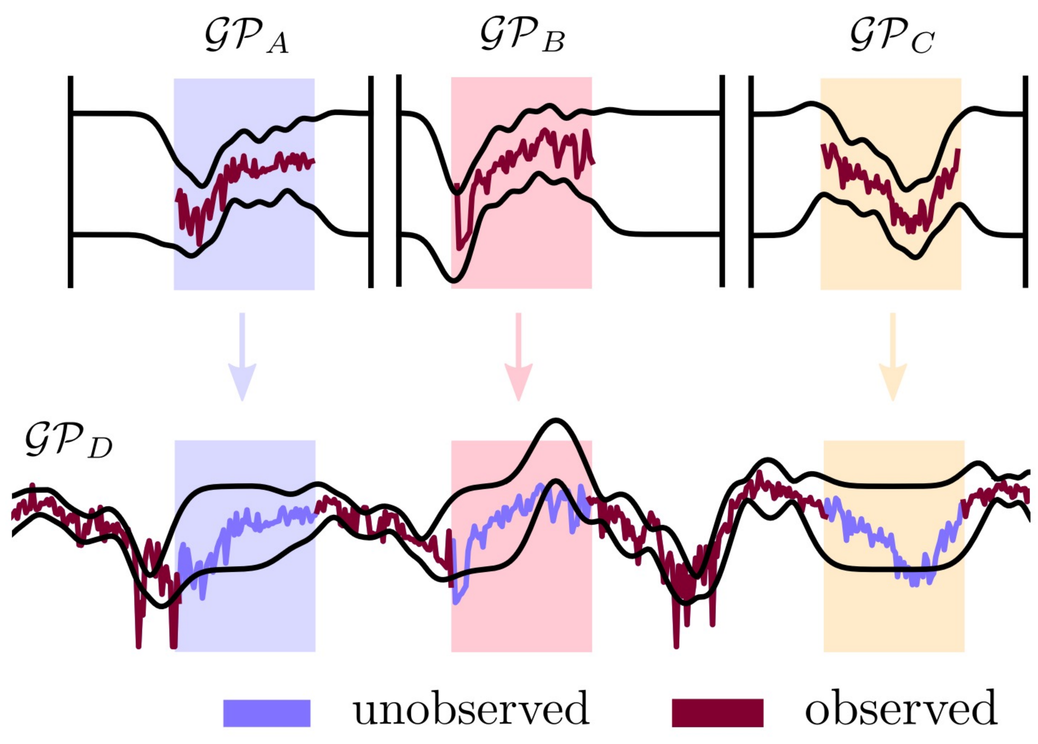

Figure 1: GP modules (A, B, C) are used for training (D) without revisiting any sample from the upper row.

Contribution

We propose a framework based on *modules* of Gaussian processes (GP). Given the previous example, we would consider the *regression* model (or module) intact. Once new data arrives, one fits a *meta-GP* using the module, but **without revisiting any sample.**

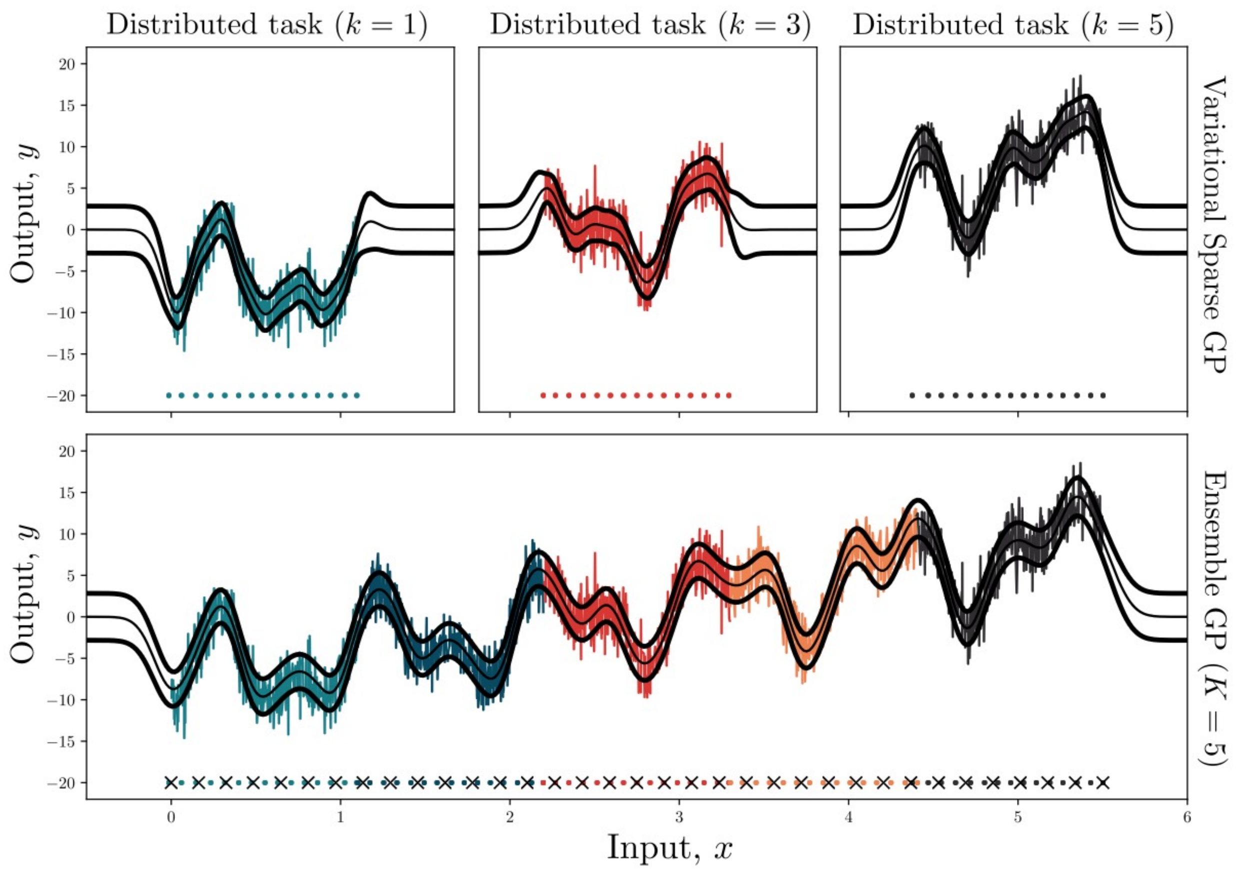

Figure 2: Modular Gaussian process regression models. We show three of five tasks united in a new GP model. Tasks are GP modules fitted independently with 500 data points and we consider 15 inducing variables per module.

What is a GP module?

$$\mathcal{M}_k = \{\boldsymbol{\phi}_k, \boldsymbol{\psi}_k, \boldsymbol{Z}_k\}$$

$\boldsymbol{\phi}_k$ -- variational parameters
$\boldsymbol{\psi}_k$ -- *kernel* hyperparameters
$\boldsymbol{u}_k, \boldsymbol{Z}_k$ -- inducing points

We learn the GP modules *independently* using the standard sparse variational GP framework.

What is a meta-GP model?

$$\mathcal{M}_* = \{\boldsymbol{\phi}_*, \boldsymbol{\psi}_*, \boldsymbol{Z}_*\}$$

$\boldsymbol{\phi}_*$ -- **new** variational parameters
$\boldsymbol{\psi}_*$ -- **new** *kernel* hyperparameters
$\boldsymbol{u}_*, \boldsymbol{Z}_*$ -- **new** inducing points

Equations

Log-marginal likelihood factorization and initial lower bound

$$\log p(\boldsymbol{y}) = \log \iint q(\boldsymbol{u}_*) p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{y}|f_+) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{+\neq \boldsymbol{u}_*} d\boldsymbol{u}_* \geq \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(f_{+\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)] + \log\frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)}\right]$$

Log-likelihood approximation from modular variational densities

$$\log p(\boldsymbol{y}|f_+) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K|f_+) = \log \prod_{k=1}^{K} p(\boldsymbol{y}_k|f_+) \approx \sum_{k=1}^{K} \log Z_k \frac{q_k(f_+)}{p_k(f_+)}$$

Module-based lower bound for learning meta-GPs.

$$\mathcal{L}_{\mathcal{E}} = \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)}[\log q_k(\boldsymbol{u}_k)] - \log p_k(\boldsymbol{u}_k)] - \mathrm{KL}[q(\boldsymbol{u}_*)||p(\boldsymbol{u}_*)]$$

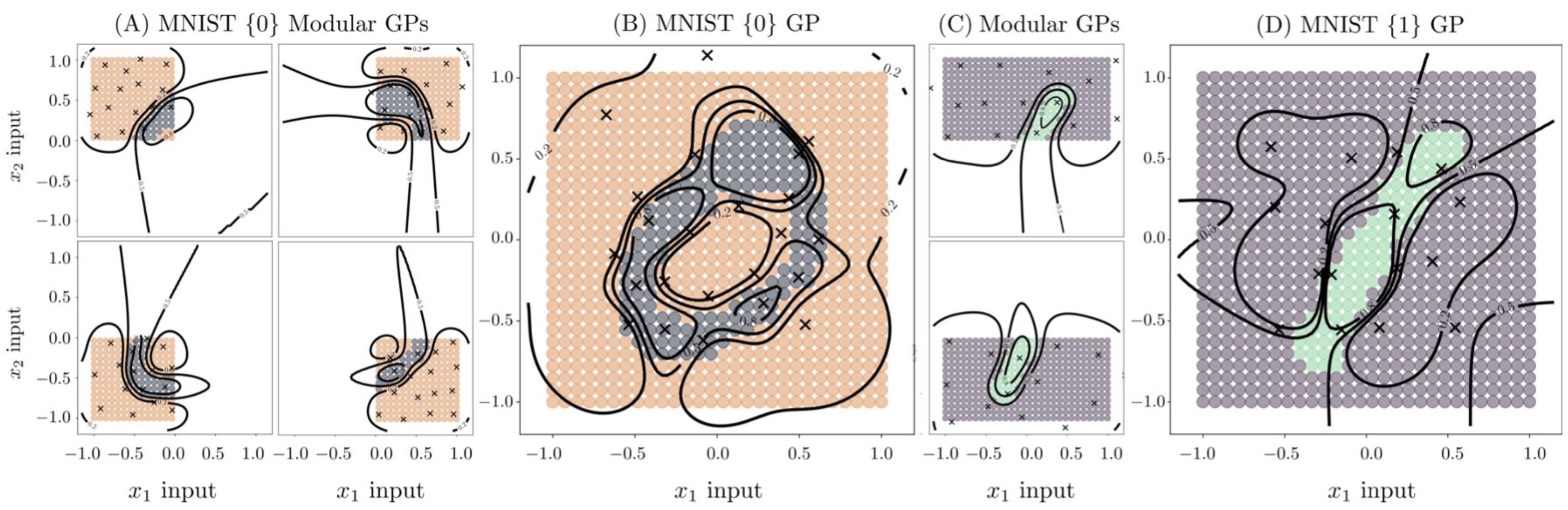

Figure 4: Modular GPs for {0,1} MNIST data samples. The meta-GPs (B--D) are built from fitted modules and do not revisit samples, only parameters and variables stored in each *k*th module.

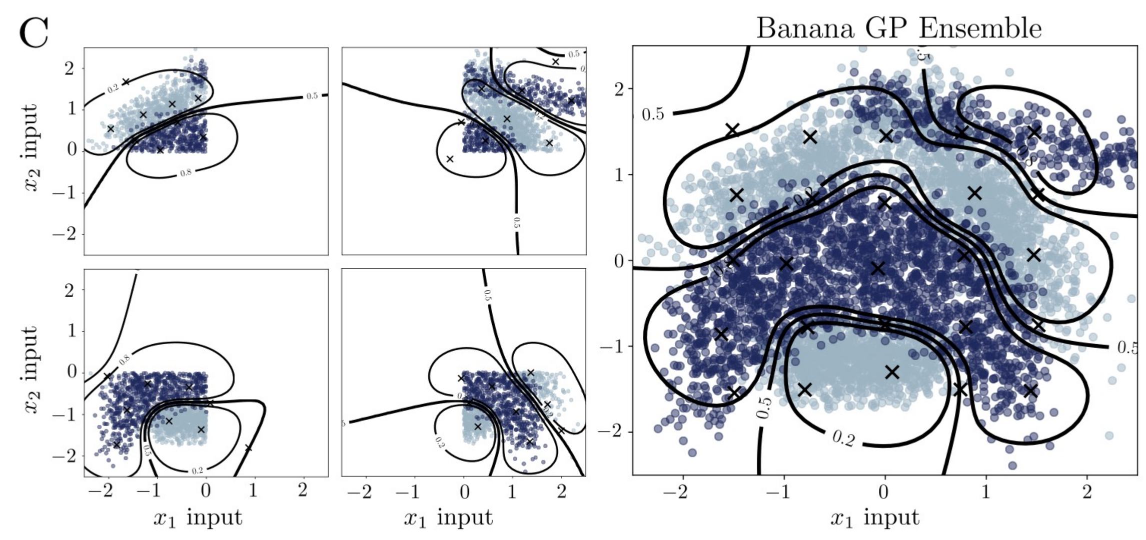

Figure 3: Modular GPs are used for training the classification model with *banana* dataset. The final *meta-GP* (right) is an sparse GP that predicts accurately without having observed any data point, only using the uncertainty metrics provided by the four GP modules (left).

Experiments

We tested the performance of *meta-GP* models built without revisiting data, only using other sparse GP models in *regression, classification* and *heterogeneous* multi-output tasks.
Results show that the *meta* model still predicts well, is competitive with other distributed approaches and scalable in the sense of number of modules and building *meta-GPs* also from *meta-GPs.*

We particularly want to show the idea of building models from models.

Conclusions

We introduced a new framework for building meta-models from independently trained GP modules. Our main contribution is to keep modules intact based on their parameters, avoid their obsolescence and mix them to form new usable tools without revisiting any data.

Key References

- Bui et al. *"Streaming sparse GP approximations". NIPS 2017.*
- *Gal et al. "Distributed variational inference in sparse GP regression and latent variable models". NIPS 2014.*
- *Moreno-Muñoz et al. "Heterogeneous multi-output GP prediction". NeurIPS 2018.*
- *Matthews et al. "On sparse variational methods and the KL divergence between stochastic processes". AISTATS 2016.*

Contact: pabmo@dtu.dk

GenU Workshop 2021 @Copenhagen