Multi-task Learning for Aggregated Data using Gaussian Processes

Fariba Yousefi, Michael T. Smith, Mauricio A. Álvarez

Machine Learning Group, Department of Computer Science, University of Sheffield, UK

f.yousefi@sheffield.ac.uk

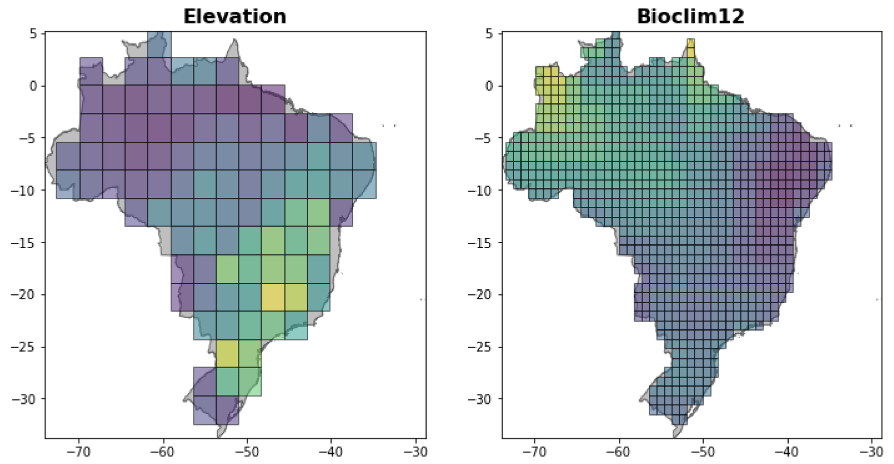
Code on GitHub

Summary

- A powerful framework for working with aggregated datasets that allows the user to combine observations from disparate datasets, with varied support is purposed.
- Both finely resolved and accurate predictions are possible by using the accuracy of low-resolution data and the fidelity of high-resolution side-information.
- Our model represents each task as the linear combination of the realizations of latent processes that are integrated at a different scale per task.
- We choose our inducing points to lie in the latent space, a distinction which allows us to estimate multiple tasks with different likelihoods.
- SVI and variational-EM with mini-batches make the framework scalable and tractable for potentially very large problems.

Introduction

- Census data are usually sampled or collected as aggregated at different administrative divisions, e.g. borough, town.
- In sensor networks, correlated variables are measured at different resolutions or scales.
- Joint modelling of the variables at different scales can improve predictions at the point or support levels.



Motivation

- We are interested in providing a general framework for multi-task learning.
- Our motivation is to use multi-task learning to jointly learn models for different tasks where each task is defined at:
  - Different support of any shape and size
  - Different nature, i.e. it is a continuous, binary, categorical or count variable.
- We appeal to the flexibility of Gaussian processes (GPs) for developing a prior over such type of datasets.
- We also provide a scalable approach for variational Bayesian inference.

Change of support using Gaussian processes

We start by defining a stochastic process over the input interval (x_a, x_b) using

$$f(x_a, x_b) = \frac{1}{\Delta_x}\int_{x_a}^{x_b} u(z)dz,$$

- $u(z)$ is a latent process that follows a Gaussian process with mean zero and covariance $k(z, z')$ and $\Delta_x = |x_b - x_a|$.
- The covariance for $f(x_a, x_b)$ follows as $\text{cov}[f(x_a, x_b), f(x'_a, x'_b)] = \frac{1}{\Delta_x\Delta_{x'}}\int_{x_a}^{x_b}\int_{x'_a}^{x'_b} k(z, z')dz'dz$
- We can now use these mean and covariance functions for representing the Gaussian process prior for $f(x_a, x_b) \sim \mathcal{GP}(0, k(x_a, x_b, x'_a, x'_b))$.
- For some forms of $k(z, z')$ it is possible to obtain an analytical expression for $k(x_a, x_b, x'_a, x'_b)$.

Multi-task learning setting

- Our inspiration for multi-task learning is the linear model of coregionalisation (LMC).

Let $\{f_d(v)\}_{d=1}^{D}$ be a set of tasks where each task is defined at a different support v. We use the set of realizations $u_q^i(\mathbf{z})$ to represent each task $f_d(v)$ as

$$f_d(v) = \sum_{q=1}^{Q}\sum_{i=1}^{R_q}\frac{a_{d,q}^i}{|v|}\int_{\mathbf{z}\in v} u_q^i(\mathbf{z})d\mathbf{z},$$

where the coefficients $a_{d,q}^i$ weight the contribution of each integral term to $f_d(v)$. The cross-covariance $k_{f_d,f_{d'}}(v, v')$ between $f_d(v)$ and $f_{d'}(v')$ is then given as

$$k_{f_d,f_{d'}}(v, v') = \sum_{q=1}^{Q}\frac{b_{d,d'}^q}{|v||v'|}\int_{\mathbf{z}\in v}\int_{\mathbf{z}'\in v'} k_q(\mathbf{z}, \mathbf{z}')d\mathbf{z}'d\mathbf{z},$$

where $b_{d,d'}^q = \sum_{i=1}^{R_q} a_{d,q}^i a_{d',q}^i$. Let us define the function $\mathbf{f}(v) = [f_1(v), \cdots, f_D(v)]^\top$.
A GP prior over $\mathbf{f}(v)$ can use the kernel defined above so that

$$\mathbf{f}(v) \sim \mathcal{GP}\left(\mathbf{0}, \sum_{q=1}^{Q}\frac{1}{|v||v'|}\mathbf{B}_q\int_{\mathbf{z}\in v}\int_{\mathbf{z}'\in v'} k_q(\mathbf{z}, \mathbf{z}')d\mathbf{z}'d\mathbf{z}\right),$$

where each $\mathbf{B}_q \in \mathbb{R}^{D\times D}$ is known as a coregionalisation matrix.

- We will use ***stochastic variational inference*** to compute a posterior distribution $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f})$, by means of the the well known idea of *inducing variables.*

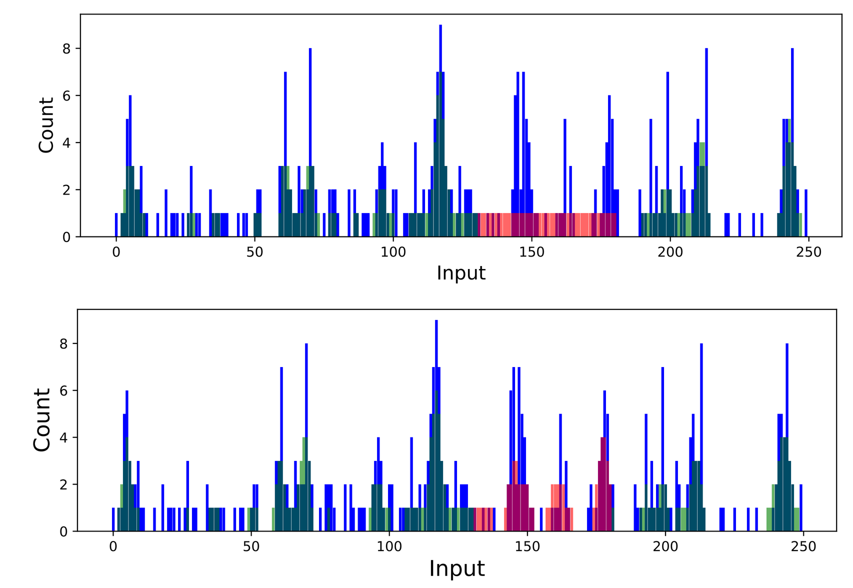
Lower-bound The lower bound for the log-marginal likelihood follows as

$$\mathcal{L} = \sum_{d=1}^{D}\sum_{j=1}^{N_d}\mathbb{E}\left[\log p(y_d(v_{d,j})|f_d(v_{d,j}))\right] - \text{KL}(q(\mathbf{u})\|p(\mathbf{u})).$$

- The optimal $q(\mathbf{u})$ is chosen by numerically maximizing $\mathcal{L}$ with respect to $\boldsymbol{\mu}$ and $\mathbf{S}$.
- The expected value is taken with respect to the $q(\mathbf{f}) = \int q(\mathbf{f}, \mathbf{u})d\mathbf{u}$ distribution.

Synthetic data

- A synthetic example for two tasks that follow a Poisson likelihood each.
- We sample from the latent vector-valued GP process and use those samples to modulate the Poisson likelihoods using $\exp(f_1(\cdot))$ and $\exp(f_2(\cdot))$ as the respective rates.
- The first task is generated using intervals of $v_1 = 1$ units, whereas the second task is generated using intervals of $v_2 = 2$ units.
- In this experiment, we evaluated our model's capability in predicting one task, sampled more frequently, using the training information from a second task with a larger support.

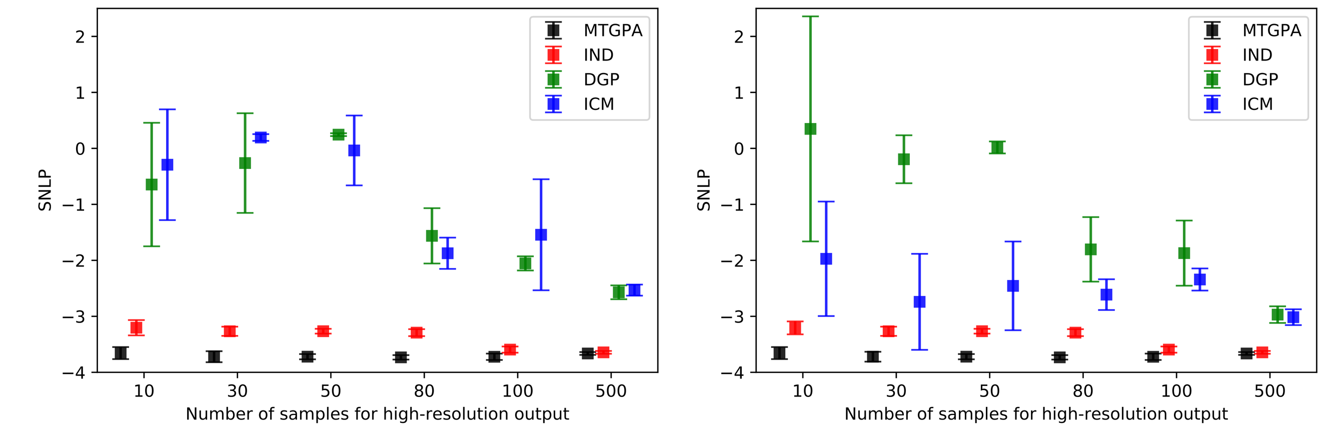


Counts for the Poisson likelihoods and predictions using the single-task vs multi-task models. Predictions are shown only for the first task (the one with support of $v_1 = 1$). The blue bars are the original one-unit support data, the green bars are the predicted training count data and the red bars are the predicted test results in the gap $[130, 180]$.

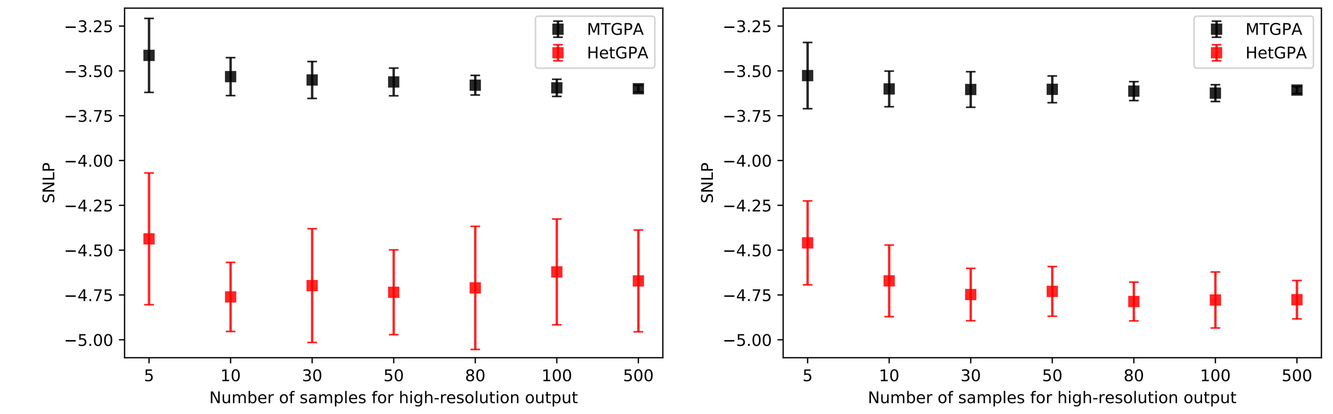
Fertility dataset

Canadian fertility dataset is used from the Human Fertility Database (humanfertility.org).

- A two-dimensional input dataset of fertility rates aggregated by year of conception and ages in Canada.
- The dataset consists of live births' statistics by year, age of mother and birth order.
- The ages of the mother are between $[15, 54]$ and the years are between $[1944, 2009]$.



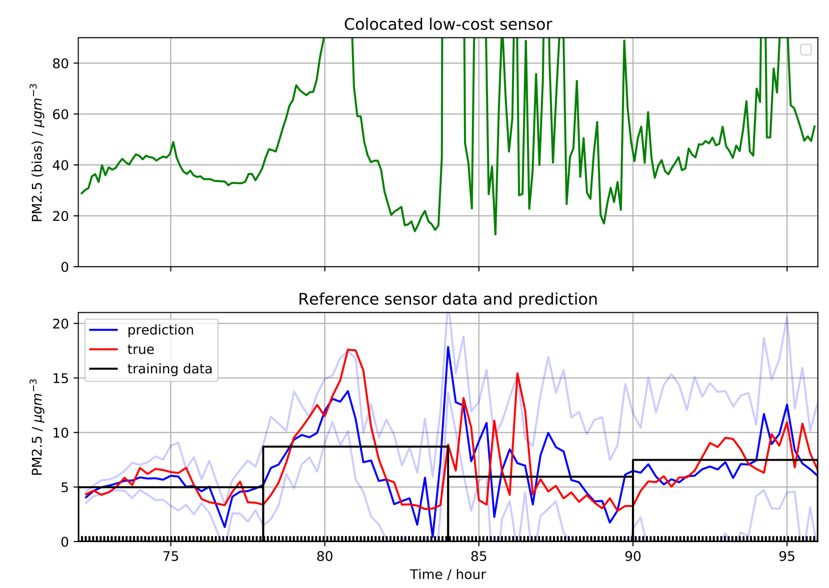
SMSE plots for different baselines for 5×5 and 2×2 aggregated data: MTGPA , Independent GPs (IND), Dependent GPs (DGP) and Intrinsic Co-regionalisation Model or Multi-task GPs (ICM).



SNLP plots for four outputs (two fertility rates) for 5×5 (left panel) and 2×2 (right panel) aggregated data. All outputs are considered as Gaussian (MTGPA) and all outputs are considered as heteroscedastic Gaussian (HetGPA).

Air pollution monitoring network

- We use data from two fine particulate (PM2.5) sensors from March, 2019 in Kampala.
- The data is taken between 2019-03-13 and 2019-03-22.
- In our model, one task represents the high accuracy low-resolution samples and the second task represents the lowaccuracy high-resolution samples.



Upper plot: a (biased) OPC low-accuracy high-frequency measurement of PM2.5 air pollution.
Lower plot: the high-precision low-frequency training data (black rectangles) the test data from the same instrument (red) and the posterior prediction for this output variable, predicting over the same 15-minute periods as the test data (blue, with pale blue indicating 95% confidence intervals).

Acknowledgement

MTS and MAA have been financed by the Engineering and Physical Research Council (EPSRC) Research Project EP/N014162/1. MAA has also been financed by the EPSRC Research Project EP/R034303/1.

Thirty-third Conference on Neural Information Processing Systems (NeurIPS), Vancouver, CANADA, 2019