Dalle Molle Institute for Artificial Intelligence
IDSIA SUPSI

Lay Summary

State space Gaussian processes for big data analytics

Project team
Prof. Dr. Marco Zaffalon
Prof. Dr. Alessio Benavoli
Dr. Dario Azzimonti
Manuel Schürch

Contact adress
Prof. Dr. Marco Zaffalon
Istituto "Dalle Molle" di Studi sull'Intelligenza Artificiale (IDSIA)
Polo universitario Lugano, Via la Santa 1
CH-6962 Lugano, Switzerland
Tel.: +41 (0)58 666 666 5
E-mail: zaffalon@idsia.ch.

23/02/2022

FNSNF Swiss National Science Foundation
75 NRP Big Data National Research Programme

1. Background

Machine learning is a field of research that creates increasingly smarter computer procedures (algorithms) for analysing Big Data. This combination of Big Data and intelligent algorithms offers an unprecedented opportunity to learn more about the world and, thus, to accelerate progress. But the sheer amount of data available also presents challenges: large amounts of complex data must be analysed in ways that are reliable and relevant, and the data must be processed efficiently.

2. Goals of the project

In this project we focused on improving Gaussian Processes (GPs), a principled machine learning method that provides both prediction and a sound quantification of the uncertainty about those predictions.

Gaussian processes are a *Bayesian, non-parametric, probabilistic* method that can be used in regression and classification tasks. GPs are a *probabilistic* model: their predictions are not just point-predictions (i.e. a single value/vector), but a probability distribution. GPs are thus particularly indicated in applications where evaluating the uncertainty on a prediction is key, such as, for example, Bayesian Optimization or forecasting. Moreover, GPs are *non-parametric* so they require very weak assumptions and provide more reliable models. Finally, the *Bayesian* nature of GPs provides a natural framework that allows for updating the model as new knowledge comes along. GPs have been developed in statistics, surrogate modelling and machine learning since the 1960s, however they are usually trained with algorithms that require $O(n^3)$ time and $O(n^2)$ space for learning, where n is the size of the training set. In practice, these computational limitations mean that GPs are restricted to applications with no more than a few thousands data points.

In the last couple of decades, many approximation methods have been proposed to overcome the complexity limitation of GPs. Such methods can be broadly classified in four groups: *numerical analysis methods, state-space methods, sparse inducing points approximations and local approximations*. Approximations in the first group study the *numerical* properties of the problem and provide approximations based on new numerical techniques and specialised hardware. Recently, some methods from this group have successfully been deployed (see,e.g. ExactGP), however they only provide numerical guarantees on the quality of the approximation which are hard to translate into statistical guarantees. *State-space* methods instead exploit the links between GPs and stochastic differential equations (SDE). In particular those methods exploit the fact that a GP can be expressed as the solution of a particular SDE. Moreover, linear SDEs can be written in state-space form, i.e. as a system of SDEs of degree one. A GP can then be associated to a state-space model which is then recursively solved by using Kalman filter and Kalman smoothing. Such methods proved to work very well for GPs defined on one-dimensional input spaces. In the case of big data and multidimensional inputs, state-space methods require approximations and there is no clear theory that provides statistical guarantees. *Sparse inducing points* approximations are a third group of GP approximations that are instead built to provide statistical guarantees. The general idea is to summarise the GP distribution by its values at m input points, called inducing points. The number of inducing points, m, is selected much smaller than the training set size, n, and determines the overall complexity of the approximations. The choice of the inducing points' positions is a key step in tuning such methods and influences the final quality of the approximation. Finally the fourth group of GP approximation consists of *local GP* methods. The idea behind this approximation is that, instead of fitting a GP to the whole training set, the data can be split into local blocks of m data points. A full GP can be trained on each block with a complexity $O(m^3)$ and then a global prediction can be issued by aggregating the local predictions. Such methods perform very well on tasks where the local structure is very important; however, in order to work efficiently, often strong independence assumptions between the local GPs are needed. Such assumptions can strongly hinder the quality of the uncertainty quantification provided by those methods.

In this project **our primary goal is to improve two types of GP approximations**: sparse inducing points and local GP approximations. Our improvements in both directions are guided by the idea of including more of the mathematical structure of GPs into the approximation itself. In particular it is well known that Gaussian process training algorithms share many similarities with state space methods to solve SDE such as Kalman and Information filter procedures. In the case of full GP this has been thoroughly studied in the literature, however those links were mostly missing in the approximation methods. By exploiting those links we were able to improve the choice of inducing points' positions and we developed a local approximation method which accounts for correlations.

3. Methods

In this project we worked on improving sparse inducing points GP approximations and local GP approximations. Here we provide more details on those methods.

Sparse inducing points approximations are based on the assumption that it is possible to represent the GP distribution by the GP values at a small number of inducing points. Those values are considered as a sufficient statistic for the GP distribution and, by exploiting the Gaussian properties, the posterior and predictive distribution can be derived. In regression tasks, with a Gaussian likelihood, the posterior can be derived analytically. In the past 15 years many inducing points methods were proposed, such methods differ mainly in the definition of the joint prior over the latent functions and test values. The Variational Free Energy method (Titsias 2009) is a method that fixes the joint prior in a way that guarantees convergence to full GP as the number of inducing points increases. The research in this field has mainly focussed on the full batch case, i.e. the whole dataset is loaded and used at the same time. This case already expands the use of full GP to large datasets, however in order to use GPs on big data (datasets with millions of points), we can only use approaches that split the data in mini-batches, i.e. small subsets of data, and proceed with training on those mini-batches. This setup is called mini-batch learning or online learning. In our works we focused on increasing the performance of sparse inducing points algorithms in this setup.

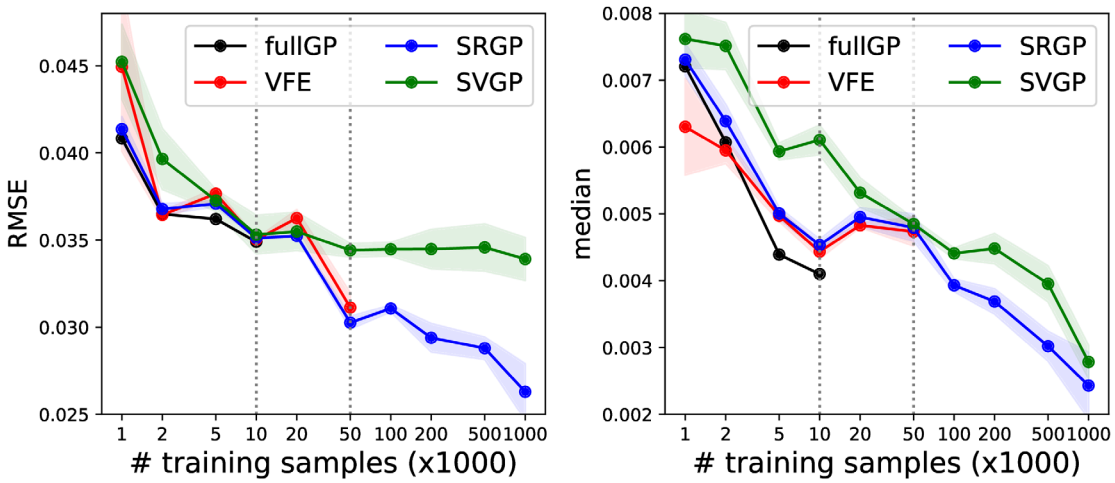
Local GP approximations have been used for many decades to approximate GPs with large data. They take their roots in the spatial statistics literature, but they were recently developed also for machine learning applications. The basic idea is that small (computationally cheap) local models can be fitted on subsets of data in place of fitting a large model on the whole dataset. In order to provide predictions everywhere then the local models need to be aggregated. The previous works on this part have thus focused on how to create local GP approximations and how to aggregate their predictions. In particular most methods start with an assumption of independence between the local GPs approximations. This assumption makes training local GPs feasible, however it hinders the quality of the final predictions, especially it reduces the quality of the uncertainty quantification. In our works we proposed a method to address this shortcoming of local GPs.

4. Results

In order to achieve our primary goal we developed three different methods that improve GP approximations. Each method is outlined in the papers listed below, either published or under submission.

Recursive estimation for sparse Gaussian process regression

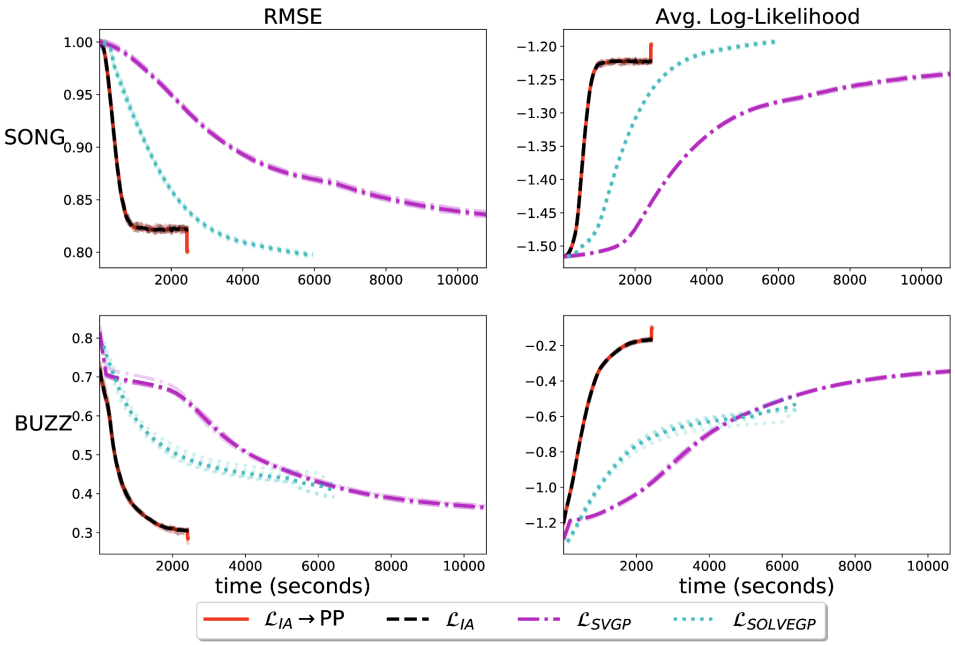
In this work we investigate a connection between a general class of sparse inducing point GP regression methods and Bayesian recursive estimation which enables Kalman Filter-like updating for online learning. The majority of previous work has focused on the batch setting for learning the model parameters and the position of the inducing points. Here instead we focus on training with mini-batches. By exploiting the Kalman filter formulation, we propose a novel approach that estimates such parameters by recursively propagating the analytical gradients of the posterior over mini-batches of the data. Compared to state of the art methods, our method keeps analytic updates for the mean and covariance of the posterior, thus reducing drastically the size of the optimization problem. As an example, we report below an application of our method to a control problem for a non-linear plant. In this problem we observe the control variables and the past values of the physical quantity we want to control. The data comes from measuring every 0.2s the physical values and the control variables so, in just 3 days, we collect around 1.2 million data points. A GP is ideal for this forecasting problem, however it can only be used with very few data points thus reducing its accuracy. In the figure below (from the refereed paper) we can see the performance of a full GP trained on a subset of data (black line, stops at 10k), a full-batch sparse method (VFE, red, stops at 50k) trained on the maximum number allowed, a competitor state-of-the-art mini-batch method (SVGP, green, uses the whole dataset) and our proposed algorithm (SRGP, blue, uses the whole dataset).



All methods increase their performances as the number of training samples increase, however our proposed method (SRGP) is much more data efficient and for the same number of data points performs better than the direct competitor (SVGP). We observed this behaviour also in all other benchmarks reported in the paper.

Sparse Information Filter for Fast Gaussian Process Regression

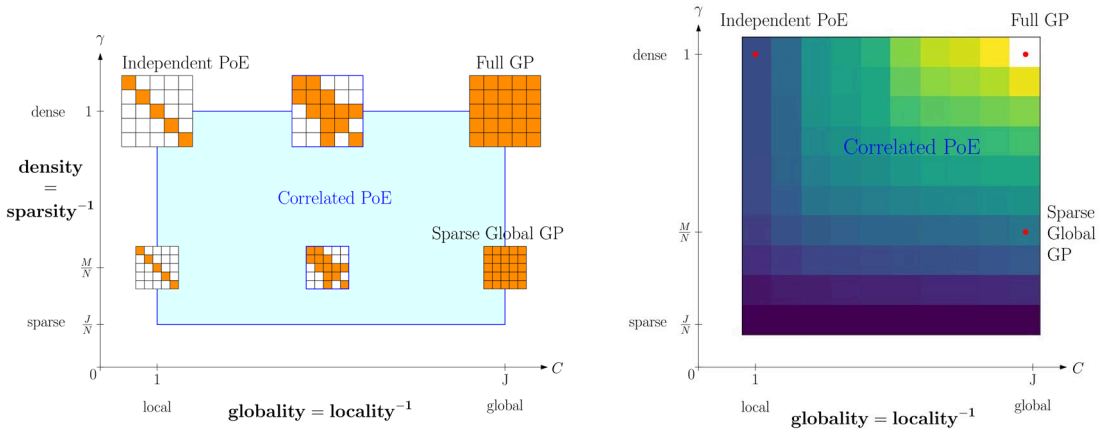
In this work, we focus on GP regression tasks and propose a new algorithm to train variational sparse GP models. In the mini-batch setting, we focus on how to choose the location of the inducing point and the model hyperparameters. We derive an objective function that we can optimise with stochastic gradient descent. This objective function is based on an independence assumption on the mini-batches (L_{IA}) with an analytical posterior propagation ($L_{IA} \rightarrow PP$) that exploits the connections between Information filter and variational sparse GP models. By using this new objective function, we can train the variational sparse GP models much more efficiently than the state-of-the-art. We benchmark our method on several real datasets with millions of data points against the state-of-the-art Stochastic Variational GP (SVGP) and sparse orthogonal variational inference for Gaussian Processes (SOLVEGP). Our method achieves comparable performances to SVGP and SOLVEGP while providing considerable speed-ups. Specifically, $L_{IA} \rightarrow PP$ is consistently 4 times faster than SVGP and, on average, 2.5 times faster than SOLVEGP. For example, the plot below shows the performance versus the training time, in seconds, for two datasets with 0.5 million data points. As shown in the graph, the performance of $L_{IA} \rightarrow PP$ (red line) is always equal or better than direct competitors (SVGP, blue, SOLVEGP, purple), but it takes substantially less time to achieve it. All methods were run on the same hardware and Tensorflow environment in order to have comparable computational times.



Correlated Product of Experts for Sparse Gaussian Process Regression

In this work, we shifted our focus to local approximation methods in the regression setting. We propose a new approach based on aggregating predictions from several local and correlated experts. Compared to the state-of-the-art, our method allows for a degree of correlation between the experts that can vary between independent up to fully correlated experts. The individual predictions of the experts are aggregated by taking into account their correlation resulting in consistent uncertainty estimates. Our method recovers the state-of-the-art independent Product of Experts, sparse GP and full GP in the limiting cases, however it allows for all intermediate cases in a simple and efficient way. The presented framework can deal with a general kernel function and multiple variables, and has a time and space complexity which is linear in the number of experts and data samples. This makes our approach highly scalable. As shown in the plot below, left-hand side, Correlated Product of Experts (PoE) allows for continuously adjusting

between local (independent PoE) and global (full GP) methods and between sparse and full methods. The figure on the right-hand side shows the distance (in Kullback-Leibler divergence, blue is far, yellow is close) between our approximation and the full GP model. As the globality and the density of the model increases the distance between the approximation and full GP decreases. Correlated PoE introduces two axes along which the approximation can be tuned. If we are in the case where full GP is not possible, then we can choose to have a more global approximation by increasing the sparsity or a more local but denser approximation. This is a key property in applied work because, in some problems, a better local approximation is preferable at the cost of having a worse global view on the problem. As opposed to other state-of-the-art approximation methods, Correlated PoE allows the user to choose which type of approximation is needed.



We demonstrate superior performance, in a time vs. accuracy sense, of our proposed method against state-of-the-art GP approximation methods for synthetic as well as several real-world datasets with deterministic and stochastic optimization.

Our methods are implemented in python code and are available in the following on-line repositories

- Recursive estimation for sparse Gaussian process regression
  https://github.com/manuelIDSIA/SRGP
- Sparse Information Filter for Fast Gaussian Process Regression
  https://github.com/lkania/Sparse-IF-for-Fast-GP

Refereed publications:

Manuel Schürch, Dario Azzimonti, Alessio Benavoli, Marco Zaffalon (2020) Recursive estimation for sparse Gaussian process regression. Automatica, Volume 120, 2020,109-127, https://doi.org/10.1016/j.automatica.2020.109127.

Kania L., Schürch M., Azzimonti D., Benavoli A. (2021) Sparse Information Filter for Fast Gaussian Process Regression. In: Oliver N., Pérez-Cruz F., Kramer S., Read J., Lozano J.A. (eds) Machine Learning and

Knowledge Discovery in Databases. Research Track. ECML PKDD 2021. Lecture Notes in Computer Science, vol 12977. Springer, Cham. https://doi.org/10.1007/978-3-030-86523-8_32

Manuel Schürch, Dario Azzimonti, Alessio Benavoli, Marco Zaffalon (2022) Correlated Product of Experts for Sparse Gaussian Process Regression. arXiv:2112.09519. Submitted.

Azzimonti, D. and Ginsbourger, D. (2018). Estimating orthant probabilities of high-dimensional Gaussian vectors with an application to set estimation. J. Comput. Graph. Statist., 27(2):255–267.

Azzimonti, D., Ginsbourger, D., Rohmer, J., and Idier, D. (2019). Profile extrema for visualizing and quantifying uncertainties on excursion regions. Application to coastal flooding. Technometrics, 61(4):474–493.

Azzimonti, D. (2019). Two types of Bayesian excursion set estimates based on Gaussian process models. In 21st European Young Statisticians Meeting.

Azzimonti, D., Rottondi, C., and Tornatore, M. (2019). Using Active Learning to Decrease Probes for QoT Estimation in Optical Networks. In Optical Fiber Communication Conference (OFC) 2019, page Th1H.1. Optical Society of America.

Azzimonti, D., Rottondi, C., and Tornatore, M. (2020). Reducing probes for quality of transmission estimation in optical networks with active learning. J. Opt. Commun. Netw., 12(1):A38–A48.

Corani, G., Augusto, J. P. S. C., Azzimonti, D., and Zaffalon, M. (2020). Reconciling Hierarchical Forecasts via Bayes' Rule. In Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Lecture Notes in Computer Science. Springer.

Benavoli, A., Azzimonti, D., and Piga, D. (2020). Skew gaussian processes for classification. Machine Learning, 109, 1877–1902.

Azzimonti, D., Ginsbourger, D., Chevalier, C., Bect, J., and Richet, Y. (2021). Adaptive Design of Experiments for Conservative Estimation of Excursion Sets. Technometrics, 63(1):13–26.

Azzimonti, D., Rottondi, C., Giusti, A., Tornatore, M.,and Bianco, A. (2021). Comparison of domain adaptation and active learning techniques for quality of transmission estimation with small-sized training datasets [invited]. IEEE/OSA Journal of Optical Communications and Net-working, 13(1):A56–A66.

Benavoli, A., Azzimonti, D., and Piga, D. (2021). A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with Skew Gaussian Processes. Machine Learning, 110:3095-3133.

Benavoli, A., Azzimonti, D., and Piga, D. (2021). Preferential Bayesian optimisation with Skew Gaussian Processes. In 2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21Companion), July 10–14, 2021, Lille, France, New York, NY, USA. ACM.

Talks and Posters at Conferences:

Manuel Schürch. Correlated Product of Experts for Sparse Gaussian Process Regression. 2021 World Meeting of the International Society for Bayesian Analysis (ISBA 2021). Online. June 29, 2021.

Lucas Kania. Sparse Information Filter for Fast Gaussian Process Regression. ECML-PKDD 2021. Online. September 16, 2021.

Manuel Schürch. Correlated Product of Experts for Sparse Gaussian Process Regression. Lifting Inference with Kernel Embeddings (LIKE22). Online. January 11, 2022.

Ph.D. Thesis:

This project supported the thesis of Manuel Schürch which is currently at its final stage.

5. Significance of the results for science and practice

In this project we focused on making GPs, a powerful machine learning method, viable for large datasets. The main objective of the project was to work on algorithms to train GP approximations that handle big data and provide performances comparable to full GP algorithms. The practical implications of our work are:

- The training of sparse inducing points methods for large datasets can be very costly from a computational point of view. By exploiting the connections between sparse GPs and Kalman filter we provided a method that can outperform state-of-the-art algorithms for sparse GPs on big data. Moreover by using the Information filter point of view we provided a method to train sparse GPs that achieves comparable performance to state-of-the-art much faster.
- Local GP methods have long suffered from strong independence assumptions or from costly inclusion of the correlations. Our correlated product of expert framework lets the user decide how much locality and sparsity is required from the approximation. This is a shift from state-of-the-art methods where this adjustment was either not possible (independent PoE) or hidden from the final user.

The main scientific implications for science are:

- More viable training of GP approximations opens the door to better time series forecasting on large datasets. For example, in electricity demand forecasting, large datasets are common because collecting data is automated by the electricity provider. Standard time series methods only work on subsets of data and do not provide reliable uncertainty quantification. In order to make better decisions, a GP can be used in this application; however, a big data GP approximation is necessary in order to exploit all information available. In our correlated PoE paper we show that this GP approximation is particularly well suited to this type of time series.
- Both sparse inducing points and correlated PoE methods could be used in weather forecast applications. In particular the task of providing good spatial interpolations between different sources of information if well suited to correlated PoE, where locality in a geographic sense could be enforced.
- Finally, as we already showed in our papers, control theory applications are naturally suited for GP models. The sound uncertainty quantification provided by (approximate) GPs, allows for better decisions in a risk-based framework.