3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–9

Supplementary File for The Gaussian Process Latent Autoregressive Model

Rui Xia RUI.XIA@U.NUS.EDU
Wessel Bruinsma WPB23@CAM.AC.UK
William Tebbutt WCT23@CAM.AC.UK
Richard E. Turner RET26@CAM.AC.UK
University of Cambridge

1. Optimization of Inducing Points

Selection and optimization of the inducing inputs location are less explicit parts of the model. Recall that for the first layer corresponding to the GP mappings for first output y_1, there is only one single input, $\mathbf{x}$. Treatment of inducing inputs is standard and similar to other sparse approximation situations. The original GPAR model used fixed and regularly-spaced inducing inputs locations over time, since they are known to perform well for time-series datasets (Bui and Turner, 2014). However, for higher layers, the selections are less straightforward. As noted by Cutajar et al. (2019), since both points in the original input space (temporal space) and their **corresponding** evaluations of functions at previous output level are passed through the next layer, the inducing points of higher layers should also be intrinsically linked due to these correspondences. For example, suppose $\mathbf{z}_m$ is the inducing location of one of the inducing inputs of first layer, the following vector should be passed to layer l after previous $l-1$ propagation,

$$\begin{bmatrix}\mathbf{z}_n & f_1(\mathbf{z}_n) & f_2(\mathbf{z}_n, f_1(\mathbf{z}_n)) & \dots & f_{l-1}(\mathbf{z}_n, f_{1:l-2}(\mathbf{z}_n))\end{bmatrix}$$

Since the inducing inputs are associated across layers, free optimization of inducing inputs at each layer is no longer appropriate in contrast to the case in DGPs. The original GPAR used the posterior predictive means of previous layers evaluated at inducing locations $\mathbf{Z}$ (which are fixed and evenly-spaced over time) as "optimized" inducing points. To make the optimization of inducing locations possible, we need to relate inducing inputs over output dimensions to inducing inputs over the original input space. Inspired from the last strategy that uses posterior predictive mean of GPAR, we can use posterior mean of GPLAR. Since $q(\mathbf{u}_l)$ is taken to approximate the posterior $q(\mathbf{u}_l|\mathbf{y})$, summarizing sufficient statistics from the training observations $\mathbf{y}$, we could take mean of $p(\mathbf{u}_l)$ as inducing inputs to the next layer $l+1$. The correspondence is also clear as $q(\mathbf{u}_1)$ is the posterior distribution over inducing locations $\mathbf{Z}$, such that,

$$\begin{aligned}
q(\mathbf{u}_1)_m &= q(u_{1m}), && \text{where } u_{1m} = f_1(\mathbf{z}_m)\\
q(\mathbf{u}_2)_m &= q(u_{2m}), && \text{where } u_{2m} = f_2(\mathbf{z}_m, \mathbb{E}[q(u_{1m})])\\
&\vdots\\
q(\mathbf{u}_L)_m &= q(u_{Lm}), && \text{where } u_{Lm} = f_L(\mathbf{z}_m, \mathbb{E}[q(u_{1m})], \dots, \mathbb{E}[q(u_{(L-1)m})])
\end{aligned}$$

The resulting inducing inputs to each layer l is as follows,

$$\begin{bmatrix}\mathbf{Z} & \mathbf{m}_1 & \dots & \mathbf{m}_{l-1}\end{bmatrix}$$

where $\mathbf{m}_l$ denotes mean of each variational distribution. In this setting, inducing inputs are "automatically" optimized since they are variational parameters themselves, except for $\mathbf{Z}$ which are inducing inputs over the original input space. Experiments have shown little overhead in computation after enabling optimization over inducing locations. Optimizing the inducing locations are beneficial in high-dimensional problems (Nguyen et al., 2014).

2. Experiments Dataset Details

Suppose we have,

$$\begin{aligned} k_1(\mathbf{x}, \mathbf{x}') &= k_{SE}(\mathbf{x}, \mathbf{x}') \\ k_2((\mathbf{x}, h_1(\mathbf{x})), (\mathbf{x}', h_1(\mathbf{x}'))) &= k_{SE}(\mathbf{x}, \mathbf{x}') + k_{SE}(h_1(\mathbf{x}), h_1(\mathbf{x}')) \\ k_3((\mathbf{x}, h_{1:2}(\mathbf{x})), (\mathbf{x}', h_{1:2}(\mathbf{x}'))) &= k_{SE}(\mathbf{x}, \mathbf{x}') + k_{SE}(h_{1:2}(\mathbf{x}), h_{1:2}(\mathbf{x}')) \end{aligned}$$

where k_{SE} denotes squared-exponential kernel. With zero mean function in each layer, we randomly draw samples layer by layer. As for heterogeneous dataset experiments, we first draw data from 4 synthetic GPs similarly in the regression experiments above, where relations with previous outputs are made explicit by linear or non-linear kernels. In this experiment, the last two outputs are converted to binary outputs by first transforming samples of evaluations of the latent process to valid probability values using the sigmoid function, i.e. logistic probability, and then labels are generated from a Bernoulli distribution. The process of drawing the third output is shown as follows,

$$\begin{aligned} p(f_3|\theta_3) &= \mathcal{GP}(f_3; \mathbf{0}, k(\mathbf{x}, \mathbf{x}') + k(h_{1:2}(\mathbf{x}), h_{1:2}(\mathbf{x}'))) \\ p(\mathbf{h}_3|f_3, \mathbf{X}, \mathbf{h}_{1:2}, \sigma^2) &= \prod_n \mathcal{N}(h_{3,n}; f_3(\mathbf{x}_n, h_{1:2,n}), \sigma_3^2) \\ p(y_{3n} = 1|h_{3n}) &= \sigma(h_3 n) \;\text{ where, } \sigma(x) = 1/(1+\exp(x)) \end{aligned}$$

The training inputs are uniformly drawn ranging from $[0.0, 2.0]$, with $N = 200$. $N_{missing} = 50$ observations for the last binary output are deliberately removed from interval $[0.5, 1.0]$. The remaining points are fed to the GPLAR model and independent GPs.

3. Real-World Data Experiments

3.1. Base model comparison

In this section, we evaluate GPLAR's performance on two standard datasets commonly used to evaluate multi-output modelling power, and compare GPLAR against GPAR.

3.2. Electroencephalogram (EEG) dataset[1]

There are 256 measurements in voltage in one second from 7 electrodes mounted on a patient's scalp when the patient is presented with a certain image. We took the measurements from patient number 337, and use full 256 observations from electrodes F3-F6 and first 156 signals from electrodes, FZ, F1, and F2 as training points, and last 100 observations of FZ, F1, and F2 as the test points to predict. Fig. 1 visualize predictions for the three electrodes by only using non-linear kernels over outputs, and it is observed that predictions of GPAR over $F1$ are over-confident which leads to large HLL in Table. 1. While uncertainty over F1 from GPLAR is well-calibrated and the 95% confidence interval covers nearly every point except those in time $[0.9, 1.0]$. The SMSE for every output is also lower in results provided by GPLAR. As FZ, F1 and F2 are last three outputs fed to the model, posterior mean from GPAR already has high accuracy.

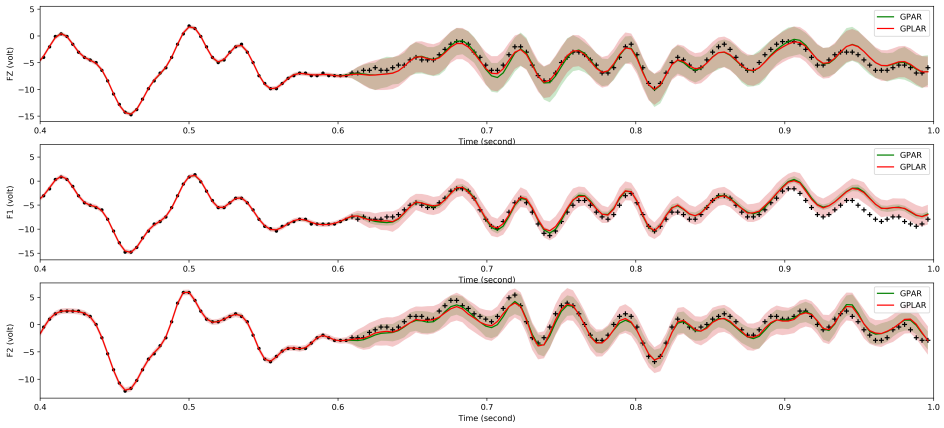


Figure 1: Predictions for electordes FZ, $F1$, and $F2$ from the EEG datasets by GPAR(green) and GPLAR(red). Dots and crosses denote training and test points.

If we fit standard DGPs to this dataset, only one small modification is needed to deal with missing data, which is separating the calculations of the variational expectation terms in the final layer. Missing values are identified and skipped during the calculations. Since every output in such a multi-output task is correlated to input to some extend, skip connection is important which propagates the input to every intermediate layer and the final layer. As an alternative, salimbeni2017doubly introduced an identity mean function at each intermediate layer, however, explicit propagation of input is validated to perform better although still worse than autoregressive models. Despite that DGPs can discover non-Gaussian dependencies between inputs and outputs since all multi-output layers use independent outputs with shared covariance functions, such input and output wrappings

1. The EEG datset is available at https://archive.ics.uci.edu/ml/datasets/eeg+database.

Output	SMSE		HLL	
	GPAR	GPLAR	GPAR	GPLAR
FZ	0.1340	0.1273	-135.7	-141.3
F1	0.3285	0.3130	**-663.1**	-183.1
F2	0.1536	0.1317	-132.4	-136.6

Table 1: SMSE and HLL for every output: GPAR vs GPLAR for the EEG datasets

Output	SMSE		HLL	
	GPAR	GPLAR	GPAR	GPLAR
USD/CAD	0.0215	0.0439	148.60	153.95
USD/JPY	0.0170	0.0234	843.18	860.95
USD/AUD	**0.2089**	**0.0685**	523.97	464.58

Table 2: SMSE and HLL for every output: GPAR vs GPLAR for the Exchange datasets

are implicit and independent outputs prevent DGPs from exploiting dependencies between outputs, limiting their predictive performance on highly correlated data.

3.3. Exchange Rates Dataset.[2]

The Pacific Exchange Rates Service keeps records of exchange rates of all currencies against US dollars every day. We extract exchange rates of ten international currencies and three metals in the year 2007, and take $50-100$th days for "USD/CAD", $50-150$th days for "'USD/JPY' and $50-200$th days for "'USD/AUD" as missing values to be predicted, and take information on all other days and full-year observations for all other currencies as training points. By using both linear and non-linear kernels over outputs, Fig. 2 presents predictions of GPAR and GPLAR for the three currencies with missing values. Although as shown in Table. 2, only SMSE of GPLAR over "USD/AUD" is significantly lower than that of GPAR, it is also observed that GPLAR give predictions with more uncertainty even outside the missing area, while GPAR would have high confidence and model with more wigglings.

3.4. Real-data: Heterogeneous Output

To compare with the large-scale experiments in hensman2013gaussian and moreno2018heterogeneous, we test GPLAR on the complete records of house properties sold in the Greater London

2. The exchange rates dataset is available at http://fx.sauder.ubc.ca.

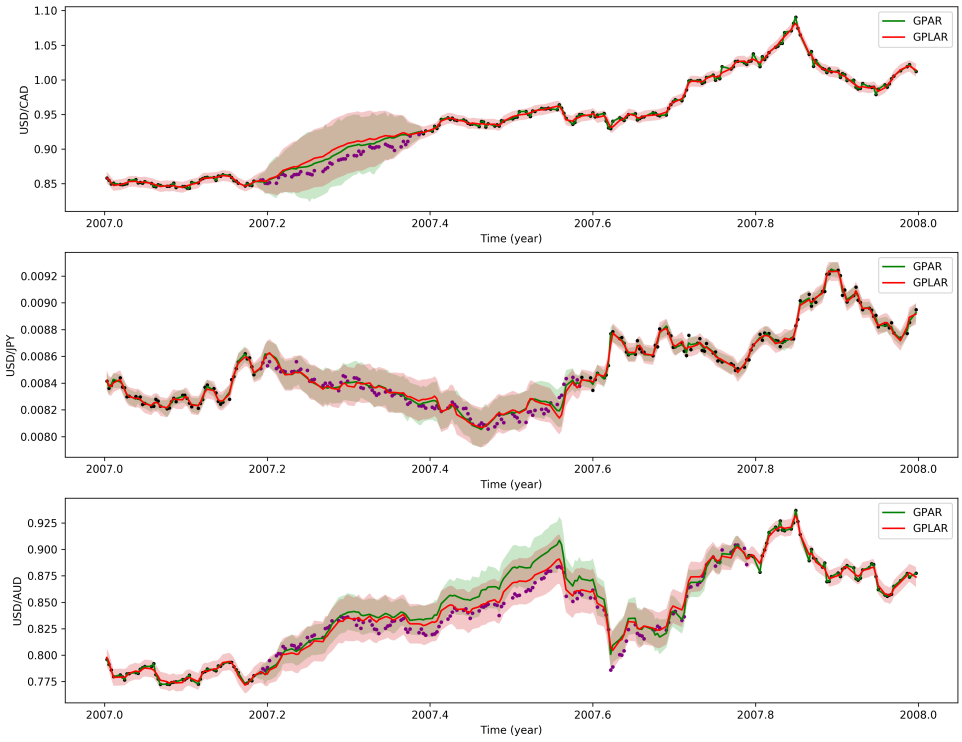


Figure 2: Predictions for "USD/CAD", "USD/JPY" and "USD/AUD" from the exchange rates datasets by GPAR(green) and GPLAR(red). Black dots are training observations, purple dots are test points.

area in 2017[3]. Each record contains the postcode of the property and is transformed into a latitude-longitude 2-dimensional spatial point as input. We take two observations, one multi-class and one continuous. Unlike the experiments done in moreno2018heterogeneous where the first output only distinguish flat or non-flat properties (binary), the first observation in our case is multi-class indicating whether the property is flat, terraced, or semi-detached. The second output is the logarithm transformed sale price of the house. It is possible that multiple records exist with the same postcode and property type, for example, flats in one building, or properties sold multiple times in one year. Hence, prices of these records are averaged, making observations of each spatial point distinct. The complete datasets containing distinct records are shown in Fig. 3. A training set of randomly selected $20,000$ points is used with 200 inducing points, and the remaining $5,286$ are for

3. The London House Price data is available at https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads.

	Input Kernel		Output Linear	Output Nonlinear	
Output	Variance	Lengthscales	Variance	Variance	Lengthscales
Flat	0.1884	[0.021, 0.026]			
Terraced	0.3489	[0.026, 0.030]	[4.388]	0.3186	[0.921]
Semi-detached	**0.0001**	**[1000., 1000.]**	[0.890 1.547]	0.0001	**[10000, 9714.]**

Table 3: Hyperparameter values of kernels learnt by GPAR on London House Price datasets

test predictions.

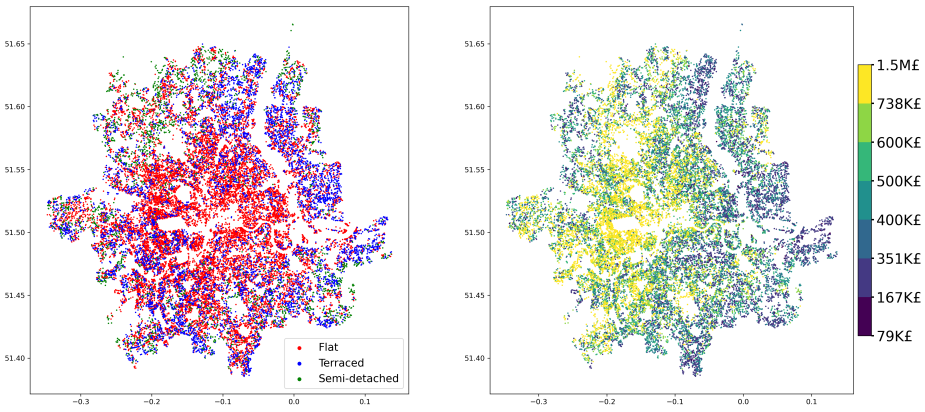
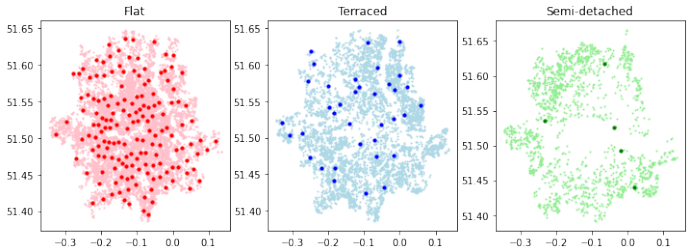


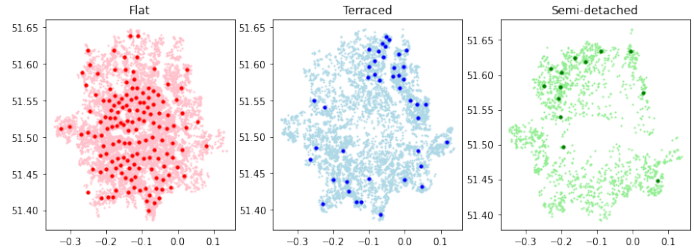
Figure 3: London House Price dataset: property type (left) and sale price (right), presented on longitude-latitude map.

We initialize inducing points of GPLAR from the posterior predictive mean of GPAR acting as if the multi-class labels are a set of three continuous outputs ranging from 0 to 1. GPLAR then imposes a robust maximum likelihood with these three latent processes. As shown in Fig. 4, the inducing points after optimization has a better representation or summary of the overall observations (The meaning of color is explained in the captions of Fig. 4). For properties of type flat, more inducing points are located in the centre of London. For properties of type terraced, more inducing points are moved to the northeast, or spread out in the southern part. As for semi-detached properties, more inducing points are located in the northwest. All the inducing points after optimization gain a more reasonable spatial meaning reflecting the true distribution of houses. This suggests that our optimization strategy of inducing points has corrected the error brought by treatments of

multi-class labels as continuous values in GPAR. An interesting and unexpected finding of GPAR is that when the first three latent processes are treated as continuous values from 0 to 1, GPAR still finds the particular relationship between the three outputs such that property type can only be one of them. If one looks at the hyperparameter values of kernels of the third output in Table. 3, variance of temporal kernel is pushed to zero and lengthscales along both longitude and latitude are pushed to large numbers. The same phenomenon can be observed with the nonlinear kernel between the third and first two outputs, indicating the third output is learnt to completely depend linearly on the first two outputs.



(a) GPAR



(b) GPLAR

Figure 4: The x-axis is longitude and the y-axis is latitude. The inducing points locations are labeled as red dots if the inducing values of first three latent process has largest value corresponding to flat, as blue dots if it is terraced, and as green dots if it is semi-detached. Figure (a) shows inducing points in GPAR and Figure (b) shows the optimized inducing points learnt by GPLAR. Both figures have the lighter and smaller scatter points denoting true observations in the background.

During test predictions, GPLAR would produce $S = 100$ samples for each test point. For each sample, we take the corresponding class with the maximum latent process value. Finally, the modal class over all samples will give the predicted class of that test point. Accuracy of multi-class property type, SMSE of the real-valued house sale price, and log-density of both outputs are presented in Table. 4. It is observed that GPLAR has better performance than independent GPs evaluated by all metrics, indicating improvement of performance after adding kernels between outputs in large-scale datasets and heterogeneous real datasets, such that property type of a house has information for predicting the sale price of the house, and vise versa. Fig. 5 shows that GPLAR has successfully recovered the distribution of type and sales-price and with well-calibrated uncertainty. For example, the

Output	Binary		Continuous	
	Accuracy	HLL	SMSE	HLL
IGP	0.6416	-2.655	0.6122	-0.8820
GPLAR	**0.6672**	**-2.444**	**0.5949**	**-0.8566**

Table 4: SMSE/accuracy and HLL for heterogeneous output: IGP vs GPLAR for the London House Price data sets

middle area has a lighter color (indicating high uncertainty) compared to darker colors on the periphery. Because the type is more mixed-up in the centre part, while more separated away from centre as observed from Fig. 3. Similarly, the predicted price distribution also matches with true observations.

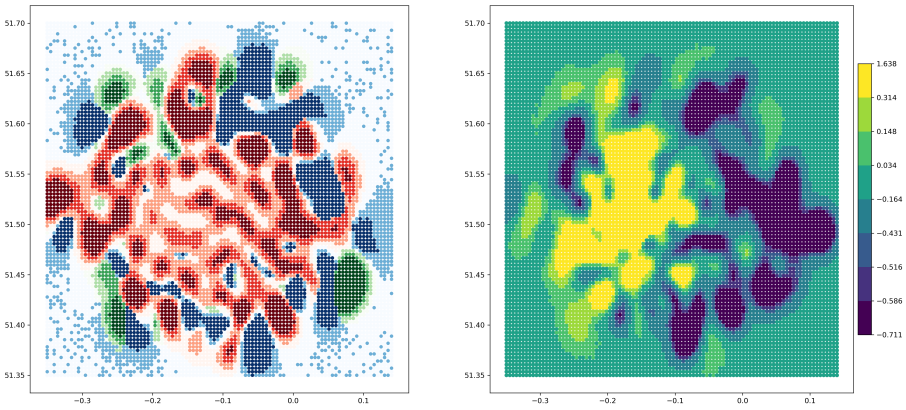


Figure 5: GPLAR predictions of London House property type (left) and easy-normalized log-sale price (right) evaluated at 100×100 grid points over longitude-latitude space. The left figure shows the color corresponding to the latent process having the largest value (Flat, terraced or semi-detached), and the darkness of color denotes uncertainty, the darker the more certain.

References

Thang D Bui and Richard E Turner. Tree-structured gaussian process approximations. In *Advances in Neural Information Processing Systems*, pages 2213–2221, 2014.

Kurt Cutajar, Mark Pullin, Andreas Damianou, Neil Lawrence, and Javier González. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*, 2019.

Trung V Nguyen, Edwin V Bonilla, et al. Collaborative multi-output gaussian processes. In *UAI*, pages 643–652, 2014.