Deep Mahalanobis Gaussian Process

Daniel Augusto de Souza
University College London
daniel.souza.21@ucl.ac.uk

Diego Mesquita
Getulio Vargas Foundation
diego.mesquita@fgv.br

César Lincoln Mattos
Federal University of Ceará
cesarlincoln@dc.ufc.br

João Paulo Gomes
Federal University of Ceará
jpaulo@dc.ufc.br

Abstract

We propose a class of hierarchical Gaussian process priors in which each layer controls the kernel lengthscales of the next. While this has been explored before, our proposal extends previous work on the Mahalanobis distance kernel bringing an alternative construction of non-stationary RBF-style kernels. The new approach has desirable properties that enables the analysis of input-dependent lengthscales. More specifically, we interpret our model as a GP that performs locally non-linear dimensionality reduction. We directly compare it with compositional deep Gaussian process, a popular model that uses successive latent space mappings to alleviate the burden of choosing a kernel function. Our experiments show promising results in both synthetic and real regression datasets.

1 Introduction

Gaussian processes (GP) priors are a non-parametric alternative to more traditional learning methods in various tasks in machine learning, however, the flexibility of this prior is mostly determined by the kernel function used, therefore, expressive kernels with tuneable hyperparameters are the most common choices. Deep Gaussian processes (DGP) priors [5, 1] move away from expert-designed kernels and instead learn feature spaces from data without any feature engineering, this is done through the outputs-to-inputs composition of functions sampled from simple GP priors. However, naïve composition of GPs adds extra complications not present in, e.g., stacked neural network layers. As discussed in the literature, this stems from the non-injective transformations learned by each GP unit [7, 3].

Nevertheless, as described by Dunlop et al. [2], this input-output compositional setup is not the only way to compose GPs. After all, the parameters of the GP distribution (mean and kernel function) are function themselves. In particular, we focus on the tradition of considering the lengthscales of stationary kernels $\boldsymbol{\ell}^2$ to be functions of the inputs $\boldsymbol{\ell}^2(\boldsymbol{x})$ as a way to make deep non-stationary GPs.

The model we propose follows the footsteps of Titsias and Lázaro-Gredilla [9] which presents a variational inference algorithm for marginalizing the lengthscale matrix of the Mahalanobis distance kernel, a generalization of the RBF kernel. By extending their prior to a GP prior, we obtain a hierarchical GP model which is not susceptible to classical pathologies of DGPs, as shown in Figure 1, can learn non-stationary behaviour and has a bias for dimensionality reduction. We approximate the posterior using variational inference and call deep variational Mahalanobis GP (DVMGP) the model built on this approximate posterior.

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

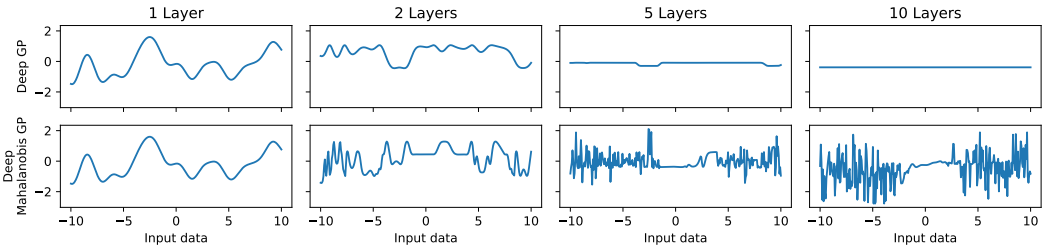

Figure 1: Samples from zero mean priors of DGP and our proposal (DMGP). Each column represents the number of layers of each model, with one layer being a regular GP.

2 Proposed model

Any stationary kernel k can be written as a scalar function of a quadratic form of the inputs, in other words, $k(\boldsymbol{a}, \boldsymbol{b}) = \phi\big((\boldsymbol{a}-\boldsymbol{b})^\intercal \boldsymbol{\Delta}^{-1}(\boldsymbol{a}-\boldsymbol{b})\big)$ where the positive-definite matrix $\boldsymbol{\Delta}$ is the lengthscales matrix. If $\phi(\tau) = \sigma_f \exp[-0.5\tau]$, then k is a Mahalanobis distance kernel and, additionally, if $\boldsymbol{\Delta}$ is diagonal, then it is usually called a RBF kernel with automatic relevancy determination (ARD).

The idea of controlling the lengthscales of stationary kernels with a function can be seen as desirable at first, but just changing them to be dependent on $\boldsymbol{X}$ does not always produce semi-positive-definite functions [4]. However, there is a procedure to transforms any stationary kernel k into a valid non-stationary kernel k_{NS} [4, 6]:

$$k_{\text{NS}}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{2^d \frac{\sqrt{|\boldsymbol{\Delta}(\boldsymbol{a})|}\sqrt{|\boldsymbol{\Delta}(\boldsymbol{b})|}}{|\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b})|}}\phi\Big(2(\boldsymbol{a}-\boldsymbol{b})^\intercal(\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b}))^{-1}(\boldsymbol{a}-\boldsymbol{b})\Big)$$

However, the input-varying lengthscales of these kernels do not inherit the usual semantics of the lengthscales associated with the original stationary kernel, due to the extra term in the front [4]. Worse still, the quadratic form in this k_{NS}does not induce a inner product space because the triangle inequality is violated [6], this is due to the lengthscale $(\boldsymbol{\Delta}^2(\boldsymbol{a}) + \boldsymbol{\Delta}^2(\boldsymbol{b}))$ depending on $\boldsymbol{a}$ and $\boldsymbol{b}$ simultaneously. Therefore, two properties that contribute to the lengthscales' interpretability are lost in this family of kernels.

A different way to approach this problem is to rewrite the equation for the stationary kernel:

$$\begin{aligned} k(\boldsymbol{a}, \boldsymbol{b}) &= \phi\big((\boldsymbol{a}-\boldsymbol{b})^\intercal \boldsymbol{\Delta}^{-1}(\boldsymbol{a}-\boldsymbol{b})\big) = \phi((\boldsymbol{a}-\boldsymbol{b})^\intercal \boldsymbol{W}\boldsymbol{W}(\boldsymbol{a}-\boldsymbol{b})) \\ &= \phi((\boldsymbol{W}\boldsymbol{a} - \boldsymbol{W}\boldsymbol{b})^\intercal(\boldsymbol{W}\boldsymbol{a} - \boldsymbol{W}\boldsymbol{b})) \end{aligned}$$

At this point, we can see that the division by the lengthscales is just an arbitrary linear transformation of the inputs applied to the stationary kernel without lengthscales. Now, when replacing the lengthscales with a function of the inputs, there is no risk of the kernel being invalid:

$$k_{\text{NS}}(\boldsymbol{a}, \boldsymbol{b}) = \phi((\boldsymbol{W}(\boldsymbol{a})\,\boldsymbol{a} - \boldsymbol{W}(\boldsymbol{b})\,\boldsymbol{b})^\intercal(\boldsymbol{W}(\boldsymbol{a})\,\boldsymbol{a} - \boldsymbol{W}(\boldsymbol{b})\,\boldsymbol{b}))$$

Unlike the previous approach, the quadratic term still defines an inner product. Indeed, each point's projection only depends on that point and not on both input points, thus preserving the triangle inequality and recovering one of the lost properties.

Titsias and Lázaro-Gredilla [9] develops a variational bayesian inference method for learning $\boldsymbol{W}$, the square root of the lengthscale, in the Mahalanobis distance kernel. We will refer to this model as MGP. In an analogous manner to the development of DGP by composing outputs of Bayesian GPLVM nodes into their inputs, we extend MGP to build a model that places a function prior on $\boldsymbol{W}(\boldsymbol{x})$, implying that, instead of composing inputs into outputs, this hierarchical model is built as a composition of outputs into lengthscales. Therefore, every node of the network still has a direct dependency on the input data $\boldsymbol{x}$, in contrast with the compositional DGP models where only the first hidden nodes directly depend on the input data.

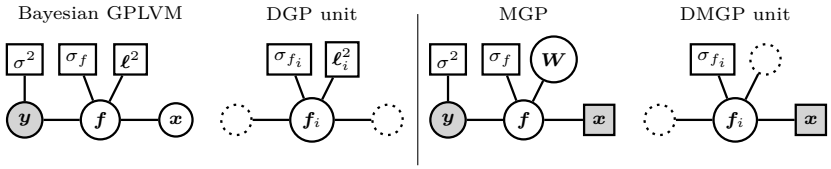

Figure 2: Graphical models for the discussed models. Dashed nodes represent the connections for composition.

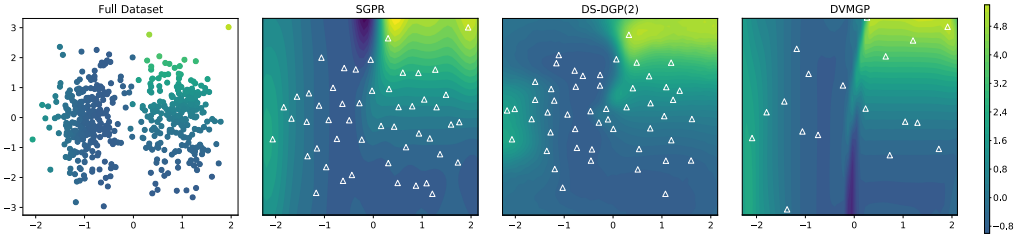

Figure 3: The mean of each model's predicted value. Empty triangles are the pseudo-inputs of the first layer of each model. As expected, DVMGP manages to recover the sharp divide between both clusters. The NLPD for each model is 0.04, -0.18, and -1.88, respectively.

The two-layer deep MGP network is characterized by joint distribution:

$$p(\boldsymbol{f}, \boldsymbol{W} \mid \boldsymbol{X}) = \mathcal{N}(\boldsymbol{s} \mid \boldsymbol{0}, \boldsymbol{K}_f) \prod_{q,d}^{Q,D} \mathcal{N}\Big(\boldsymbol{w}_{:qd} \mid \boldsymbol{0}, \boldsymbol{K}_w^{(q)}\Big)$$

$$[\boldsymbol{K}_f]_{ij} = \sigma_f \exp\Big[-0.5 \cdot \|\boldsymbol{W}_i \boldsymbol{x}_i - \boldsymbol{W}_j \boldsymbol{z}_j\|^2\Big]$$

In other words, we have added GP prior to each entry of the $\boldsymbol{W}$ $(N \times Q \times D)$-tensor, with kernels being shared for each 'row of the matrix $\boldsymbol{W}_i$. This sharing allows the determination of the relevancy of the latent dimensions by optimizing the kernel variance of each $k^{(q)}$ to the model evidence. Our variational inference scheme follows Titsias and Lázaro-Gredilla [9] very closely except that we need to introduce inducing points $\boldsymbol{V}$ on the $\boldsymbol{W}$ process and we use a mean-field approximate posterior on $q(\boldsymbol{V}) = \prod_{i,q,d} \mathcal{N}(v_{iqd} \mid \mu_{iqd}, \sigma_{iq})$.

3 Experimental results

We apply the two-layer DVMGP against a two-layer double stochastic DGP [7] and the shallow sparse GP [8] on synthetic and UCI regression datasets. These models are evaluated on average negative log predictive density and mean relative absolute error. The source code for the model is available at `https://github.com/spectraldani/DeepMahalanobisGP`.

3.1 Synthetic experiment

The input of this dataset is given by two 2D clusters $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ and the output is a polynomial that only depends on the first dimension in cluster $\boldsymbol{C}_0$ but for $\boldsymbol{C}_1$ only depends on the second. Figure 3 displays the mean of the function that each model learned and their NLPD. Because we know the mapping from each cluster it is feasible to see if DS-DGP and DVMGP recover the latent transformation.

3.2 UCI datasets

Figure 4 shows a brief overview of the datasets and the results of 5-fold cross-validation. All models are better or equal to the single-layer model and DVMGP holds comparable to DS-

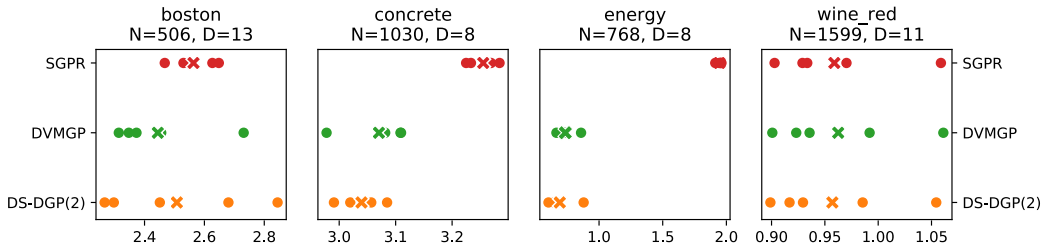

Figure 4: Test NLPD of each model on each of the datasets. Each dot represents the result of a fold, and the cross is the mean of all folds.

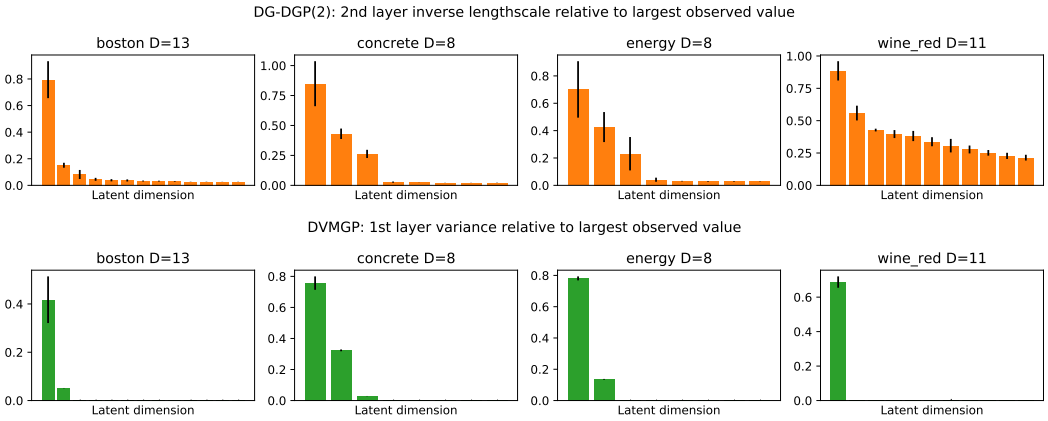

Figure 5: A plot of the mean and 1σ interval of the values that correspond to the relevance of each latent dimension.

DGP. However, as seen in the previous section, ease to project the input space into smaller dimensions is one of the main features of modelling behind DVMGP. Figure 5 displays the set of hyperparameters that correlate with the relevance of the dimensions of the latent space. Despite the similar performance, a sharper division between dimensions can be seen in DVMGP's results.

4 Discussion

We presented an alternative deep Gaussian process based on composing a GP's output into the lengthscales of the other's kernel, therefore, it is not susceptible to the pathological behaviour present in DGPs [3]. This alternative builds upon a previous variational model by Titsias and Lázaro-Gredilla [9], thus, distinguishing itself from other models based on standard compositional functions. Moreover, the use of Monte Carlo methods for inference enables extensions to big data datasets. Therefore, this approach can be extended to big data datasets using scalable inference.

By evaluating the model in synthetic and empirical datasets, we see that this new model is either comparable or better to ordinary Deep Gaussian Processes, especially in tasks that require learning projections of the input data. It also has the advantage that the quadratic form inside the kernel still defines an inner product space, contrasting with models based on Paciorek [6].

However, DVMGP may suffer from some drawbacks related to optimization and expressivity. When using the Mahalanobis kernel for dimensionality reduction, each hidden layer has $Q \times D$ GPs, then, the number of variational parameters may become very large, which can slow down optimization. It is also currently limited to RBF-style kernels in the output layer and further research is required to explore alternatives under this inference methodology.

References

[1] Andreas Damianou and Neil Lawrence. "Deep Gaussian processes". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Carlos M. Carvalho and Pradeep Ravikumar. Vol. 31. Proceedings of Machine Learning Research. Scottsdale, AZ, USA: PMLR, 2013, pp. 207–215. URL: http://proceedings.mlr.press/v31/damianou13a.html.

[2] Matthew M. Dunlop et al. "How deep are deep Gaussian processes?" In: *Journal of Machine Learning Research* 19.54 (2018), pp. 1–46. URL: http://jmlr.org/papers/v19/18-015.html.

[3] David Duvenaud et al. "Avoiding pathologies in very deep networks". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, 2014, pp. 202–210. URL: http://proceedings.mlr.press/v33/duvenaud14.html.

[4] Mark N. Gibbs. "Bayesian Gaussian processes for regression and classification". PhD thesis. University of Cambridge, 1997.

[5] Neil D. Lawrence and Andrew J. Moore. "Hierarchical Gaussian process latent variable models". In: *Proceedings of the 24th international conference on Machine learning - ICML '07*. ICML '07. Corvalis, Oregon, USA: ACM Press, 2007, pp. 481–488. ISBN: 9781595937933. DOI: 10.1145/1273496.1273557. URL: https://doi.org/10.1145/1273496.1273557.

[6] Christopher Joseph Paciorek. "Nonstationary Gaussian Processes for Regression and Spatial Modelling". PhD thesis. Carnegie Mellon University, May 2003.

[7] Hugh Salimbeni and Marc Deisenroth. "Doubly Stochastic Variational Inference for Deep Gaussian Processes". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4588–4599. URL: http://papers.nips.cc/paper/7045-doubly-stochastic-variational-inference-for-deep-gaussian-processes.pdf.

[8] Michalis K. Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Ed. by David van Dyk and Max Welling. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 2009, pp. 567–574. URL: http://proceedings.mlr.press/v5/titsias09a.html.

[9] Michalis K. Titsias and Miguel Lázaro-Gredilla. "Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 279–287. URL: http://papers.nips.cc/paper/5088-variational-inference-for-mahalanobis-distance-metrics-in-gaussian-process-regression.pdf.