Infinite Mixture of Global Gaussian Processes

Melanie F. Pradier
Department of Signal Theory
and Communications
Univ. Carlos III in Madrid, Spain
melanie@tsc.uc3m.es

Fernando Perez-Cruz*
Department of Signal Theory
and Communications
Univ. Carlos III in Madrid, Spain
fernando@tsc.uc3m.es

Abstract

In this paper, we propose a simple and powerful approach to solve nonlinear regression problems using an infinite mixture of global Gaussian processes (IMoGGP). Our method is able to deal with arbitrary output distributions, non-stationary signals, heteroscedastic noise and multimodal predictive distributions straightforwardly, without the modeler needing to know these attributes a priori. The IMoGGP can be interpreted as a mixture of experts, in which the experts are not local and they cooperate in the whole input space to provide accurate regression estimates. It can also be framed as a Dependent Dirichlet Process to solve discriminative tasks. Experiments show that our method gives comparative results to state-of-the-art approaches and its simplicity makes it an attractive method for non-ML-expert practitioners. Code will be available at the authors' webpage.

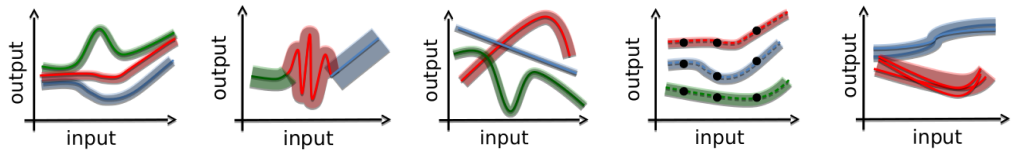
1 Introduction

Gaussian processes (GPs) are the best-known Bayesian non-parametric solution for regression [21], and in general for discriminative modeling. The standard GP assumes a Gaussian likelihood, although extensions to deal with other likelihood models exist either by using non-conjugate distributions [10] or wrappers [22, 11]. The GP can also be modified to deal with non-stationary signals by changing the kernel function [18] or even combining multiple kernels [3]. GPs have also been extended to manage heteroscedastic noise either directly [7, 13] or by relying on mixture of experts [20, 15], as well as colored noise [16, 2]. However, all these modifications imply additional design parameters which increase their complexity, and they fail to provide a unified solution to all these aspects at once.

This paper presents a simple yet powerful approach to do all these modifications seamlessly and at once. Instead of using a single GP that tracks the mean underlying function, we have several GPs that model the underlying distribution for each input vector as an infinite mixture of Gaussians. These GPs cover the whole input space, i.e., globally, allowing the predicted posterior probability to be non-Gaussian, multimodal, heteroscedastic and/or non-stationary, without the need of explicitly indicating or even knowing that those effects might be into play. Our infinite mixture of GPs can potentially capture any complex functional behavior, in a similar fashion that an infinite mixture of Gaussians can approximate any arbitrary density function.

For test input vector, our model provides an estimate of the output that is a linear combination of GPs. The proposed method is a discriminative regression algorithm, in the same way standard GPs are. Hence, we make no probabilistic assumptions over the input space. If we are strictly interested in making a probabilistic statement over the output space given the input, adding a probabilistic model over the input might be detrimental, both in terms of computational complexity and accuracy of our predictions. The mixing proportions are held constant throughout the input space, so that all GPs are active in the whole input domain. Hence, we avoid the need of relying on a gating function, which is typically used to select a particular local functions. The mixing proportions, as well as the hyper-parameters of the Gaussian processes, are inferred given the data.

*Also Machine Learning Scientist at Bell Labs.



(a) (b) (c) (d) (e)

Figure 1: **Conceptual comparison of different approaches.** Sketch comparing (a) Infinite Mixture of Global GPs, (b) Infinite Mixture of Experts, (c) Overlapping GPs for multi-tracking, (d) Spatial Dirichlet Process, and (e) time series clustering. Each color represents a different GP.

Our approach is universal for solving any regression problem, but it is more effective for moderate-sized input spaces and a larger number of samples per input dimension, because if the input dimension is large the fitting with one Gaussian process will prevail (i.e., we get the standard GP fitting) and if the input dimension were low (or nonexistent), the solution would be that of a Dirichlet process for a Gaussian mixture. In short, our method transitions seamlessly between the two limiting processes. The proposed algorithm is a direct application of the Dependent Dirichlet Process (DDP) [14], but our interpretation as a non-parametric universal regression discriminative procedure is novel, as the standard interpretation of DDPs is that of an indexed collection of distributions[1].

Related works. In Figure 1a, we show a one-dimensional cartoon solution that our discriminative regression algorithm would be able to provide. Our algorithm is a mixture of experts, but not a typical one in which the input space is chopped locally, using a gating function that decides who is the expert for each input (see cartoon in Figure 1b). The proposed approach divides the available data in independent GPs, so it presents the same computational savings as the standard mixture of expert algorithms based on GPs. The data division is not based on local proximity rules, but on improving the prediction accuracy of the output.

In [12], the authors infer trajectories, i.e., time series, in which each trajectory is represented by a GP, as shown in Figure 1c, but there is no regression interpretation nor future predictions, and the number of time series has to be known beforehand. In [6], the authors want to estimate the 10-day aggregated rainfall in 39 locations in southern France and 6 test locations. They assume that the input space has a low cardinality of potential locations, and each sample had an observation in all input locations (see 1d), which can be seen as a particular case of our method. Our algorithm is also similar to the clustering of time series using Dirichlet processes (DP) [8], as illustrated in Figure 1e, but in that case samples are treated as whole sequences a priori, and the goal is to identify clusters. Our algorithm is able to solve these applications directly or with minor modifications, while those papers cannot be applied to solve the general regression problem.

2 Infinite Mixture of Global Gaussian Processes

The aim in a regression problem is to estimate $y \in \mathbb{R}$ given an input $\boldsymbol{x} \in \mathbb{R}^D$ and a database $\mathcal{D}_n = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$. From the available data it induces a general relation between the input $\boldsymbol{x}$ and the output y. In probabilistic modeling this relation is expressed by a conditional model:

$$p(y|\boldsymbol{x}, \mathcal{D}_n). \tag{1}$$

To introduce our Infinite Mixture of Global Gaussian Processes (IMoGGP) we are going to start from the standard stick-breaking construction of Dirichlet Processes (DP) for countably infinite mixture models [23]. Observations are then generated as follows:

$$\boldsymbol{\pi}|\alpha \sim \text{GEM}(\alpha) \tag{2}$$
$$z_i|\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi}) \tag{3}$$
$$\theta_m|H \sim H \tag{4}$$
$$y_i|z_i, \{\theta_m\} \sim F(\theta_{z_i}), \tag{5}$$

[1]Actually, the author in [14] proposed a simple 1-D linear regression application in which the strength of the DDP for non-parametric regression and its many desirable properties are not exploited nor hinted.

where GEM stands for the stick-breaking prior by Griffiths, Engen and McCloskey [19], α is the concentration parameter of the DP, and the mixing proportions $\boldsymbol{\pi}$ are sampled using a stick breaking procedure [9]. z_i indicates the cluster assignment of observation i, and θ_m designates the cluster parameters for cluster m, which are sampled from the base measure H. Finally, observations are sampled from $F(\cdot)$ given the cluster assignments and parameters. One standard selection for $F(\cdot)$ is a Gaussian distribution in which θ_m represents its mean and variance.

In the regression setting, each y_i is associated with an input $\boldsymbol{x}_i$ and we can directly modify (5) as

$$y_i|z_i, \{\theta_m\}, \boldsymbol{x}_i \sim F\big(\theta_{z_i}(\boldsymbol{x}_i)\big), \tag{6}$$

$$\theta_m|H, \phi_m \sim H_{\phi_m}, \tag{7}$$

where we assume that $F\big(\theta_{z_i}(\boldsymbol{x}_i)\big)$ is Gaussian-distributed with mean $\mu_{z_i}(\boldsymbol{x}_i)$ and variance $\sigma^2_{z_i}(\boldsymbol{x}_i)$, and H_{ϕ_m} is a Gaussian process prior with hyperparameters ϕ_m. Now each cluster parameter θ_m corresponds to a latent function over the input space. This construction with a general $F(\cdot)$ is known as a dependent Dirichlet process (DDP) with constant weights, and more specifically as a single-p DDP [14], in which the parameters of each component of the infinite mixture model is indexed by the input variable $\boldsymbol{x}$.

Inference in this model is straightforward, because given $\{\theta_m\}$, sampling z_i and $\boldsymbol{\pi}$ is identical to the inference of cluster assignments in DPs, and given the cluster assignments for all pairs $(\boldsymbol{x}_i, y_i)$, $\{\theta_m\}$ can be inferred by sampling from the posterior GP distribution. We perform inference by a simple MCMC procedure based on the auxiliary variable approach from Algorithm 8 in [17] that allows to sample z_i in parallel. The procedure for each iteration is detailed in Algorithm 1, in which we have used the standard vector notation, i.e. $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ and $\boldsymbol{z} = [z_1, \ldots, z_n]^\top$.

Algorithm 1 Inference for the IMoGGP - description of one Gibbs Sampling iteration:

1: Sample extended vector of mixture proportions (propose T new GPs):

$$\boldsymbol{\pi}|\boldsymbol{z}, \alpha \sim \text{Dirichlet}\big(n_1, \ldots n_K, \underbrace{\alpha/T \ldots \alpha/T}_{T \text{ times}}\big) \tag{8}$$

2: Sample latent functions, i.e., cluster parameters θ_m, $m = 1, \ldots, M^+$:

$$p(\theta_m|\boldsymbol{\pi}, \boldsymbol{y}, \mathbf{X}, \boldsymbol{z}) \propto\ p(\theta_m|H_{\phi_m})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \theta_m) \tag{9}$$

3: Sample cluster assignments:

$$p(z_i|\boldsymbol{\pi}, y_i, \boldsymbol{x}_i, \boldsymbol{z}_{-i}, \{\theta_m\}) \propto p(z_i|\boldsymbol{\pi})\ p(y_i|\boldsymbol{x}_i, \boldsymbol{z}, \{\theta_m\}) \tag{10}$$

4: Get hyperparameters ϕ_m, $m = 1, \ldots, M^+$ by sampling (or maximizing) the evidence for each individual GP in parallel [21]. Hyperparameters for new clusters are sampled from the prior.

5: Sample concentration parameter α using an auxiliary variable η as in [5].

Finally the predicted distribution for a new input $\boldsymbol{x}^*$ is given by:

$$p(y^*|\boldsymbol{x}^*, \mathcal{D}_n) = \sum_{m=1}^{M^+} \pi_m p(y^*|\boldsymbol{x}^*, \mathcal{D}_n, \boldsymbol{z}, \{\theta_m\}) \tag{11}$$

where

$$p(y^*|\boldsymbol{x}^*, \mathcal{D}_n, \boldsymbol{z}, \{\theta_m\}) = \mathcal{N}(\mu_m(\boldsymbol{x}^*), \sigma^2_m(\boldsymbol{x}^*)) \tag{12}$$

and

$$\mu_m(\boldsymbol{x}^*) = \mathbf{k}_m^\top \mathbf{C}_m^{-1} \boldsymbol{y}_m \tag{13}$$

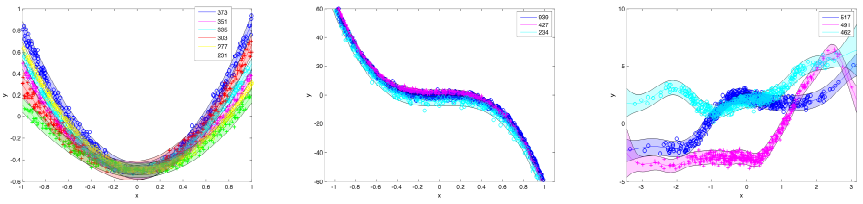
$$\sigma^2_m(\boldsymbol{x}^*) = k_m(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_m^\top \mathbf{C}_m^{-1} \mathbf{k}_m \tag{14}$$

where $\mathbf{k}_m = [k_m(\boldsymbol{x}_1^m, \boldsymbol{x}^*), k_m(\boldsymbol{x}_2^m, \boldsymbol{x}^*), \ldots, k_m(\boldsymbol{x}_{n_m}^m, \boldsymbol{x}^*)]^\top$ and $\mathbf{C}_m = \mathbf{K}_m + \sigma_m^2 \mathbf{I}$. The set $\{\boldsymbol{x}_j^m\}_{j=1}^{n_m}$ are the $\boldsymbol{x}_i$ for which z_i is equal to m and $(\mathbf{K}_m)_{rs} = k_m(\boldsymbol{x}_r^m, \boldsymbol{x}_s^m)$. Finally $k_m(\boldsymbol{x}, \boldsymbol{x}')$ is the kernel or covariance function for each GP.

		(a)	(b)	(c)	(d)	(e)	(f)
PLLH	sGP	-0.0217	-3.4920	-3.3030	-0.5855	-1.6373	0.2033
	IMoE	0.7017	-2.1248	-2.1604	1.9452	-1.6308	**0.9943**
	IMoGGP	**0.9008**	**-2.1237**	**-1.2575**	**2.3587**	**-1.5723**	0.9846
MSE	sGP	0.0288	**4.8115**	4.2815	93.4640	0.7877	86.6815
	IMoE	0.0331	4.8394	5.2263	93.4640	0.7780	82.8929
	IMoGGP	**0.0287**	4.8500	**4.2703**	**43.6710**	**0.7754**	**82.4264**

Table 1: **Comparison of the single GP (sGP), the Infinite Mixture of Experts (IMoE) and the proposed Infinite Mixture of Global Gaussian Processes (IMoGGP).** The three first columns correspond to the toy examples plotted in Section 2: (a) Heteroscedasticity, (b) Non-Gaussianity, (c) Multimodality. The last three columns correspond to real databases available online: (d) Concrete, (e) Marathon, (f) RSSI.

Figure 2 shows the posterior GPs inferred by the IMoGGP model. The method is able to deal with heteroscedastic data, heavy tailed and asymmetric noise, and multimodal distributions with the same regression model, by adding global GPs over the whole input space to capture these behaviors.



(a) Heteroscedasticity. (b) Non-Gaussianity. (c) Multimodality.

Figure 2: **Properties that can be captured by the IMoGGP model**: (a) non-stationary, heteroscedastic noise; (b) non-Gaussian likelihoods, specifically a Student's t with Gamma distributed noise; and, (c) multimodal predictive distributions.

3 Results

This Section compares the performance of the IMoGGP against a single GP (sGP) and the Infinite Mixture of Experts (IMoE) from [20]. The three algorithms are compared in exactly the same conditions, with the same hyperparameters and input splits into training and test data. Each simulation is run for 1000 iterations, and averaging is done for the last 500 iterations. All results are computed on an independent test set corresponding to 20% of the total input data, and we perform 10 different splits at each time. For all our experiments, we use the popular Noisy Squared Exponential (NSE) kernel [21].

Table 3 shows quantitative results for both synthetic and real databases. We report the mean Predictive Log Likelihood (PLLH) and Mean Squared Error (MSE) for each method and input database. Our method gives the highest PLLH in five out of six databases. The highest gains are achieved for the Multimodality and Concrete databases. Indeed, the IMoGGP is the only method able to use multiple functions at a single input location. On the other hand, the Concrete database is the example with highest dimension ($D = 8$), and the curse of dimensionality makes it harder for the IMoE to learn local functions. The IMoGGP is less affected as it uses global GPs over the input space and is able to share more information across all dimensions.

Conclusion. In this paper, we have presented the IMoGGP, a simple, yet powerful approach to solve general regression problems. We have shown its connection to DDPs and have compared its performance to state-of-the-art approaches. Our method is rather suitable for very large databases of moderate dimension. The computational cost is not enormous as the data points are partitioned in several GPs, and sampling of cluster assignments can be run in parallel. As future work, we would like to extend the model to deal with high input dimensions. In such scenario, it might be desirable to have varying mixture weights across the input space like in [1] and a selection of relevant input dimensions and orders, such as in [4].

References

[1] Yeonseung Chung and David B. Dunson. Nonparametric Bayes Conditional Distribution Modeling With Variable Selection. *Journal of the American Statistical Association*, 104(488):1646–1660, December 2009.

[2] G. Cottone, M.D. Paola, and R. Santoro. A novel exact representation of stationary colored Gaussian processes (fractional differential approach). *Journal of Physics A: Mathematical and Theoretical*, 43(8):085002, February 2010.

[3] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. *arXiv:1302.4922 [cs, stat]*, February 2013. arXiv: 1302.4922.

[4] David K Duvenaud, Hannes Nickisch, and Carl E. Rasmussen. Additive Gaussian Processes. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 226–234. Curran Associates, Inc., 2011.

[5] M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.

[6] A. E. Gelfand, A. Kottas, and S. N. MacEachern. Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, September 2005.

[7] P. W. Goldberg, C. K. I. Williams, and C. M. Bishop. Regression with input-dependent noise: A Gaussian process treatment. *Advances in neural information processing systems*, 10:493–499, 1997.

[8] J. Hensman, M. Rattray, and N. D. Lawrence. Fast nonparametric clustering of structured time-series. *arXiv:1401.1605 [cs, stat]*, January 2014. arXiv: 1401.1605.

[9] H. Ishwaran and L. F. James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, March 2001.

[10] E. Khan, S. Mohamed, and K. P. Murphy. Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3140–3148. Curran Associates, Inc., 2012.

[11] M. Lazaro-Gredilla. Bayesian Warped Gaussian Processes. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1619–1627. Curran Associates, Inc., 2012.

[12] M. Lazaro-Gredilla, S. Van Vaerenbergh, and N. D. Lawrence. Overlapping Mixtures of Gaussian Processes for the data association problem. *Pattern Recognition*, 45(4):1386–1395, April 2012.

[13] Q. V. Le, A. J. Smola, and S. Canu. Heteroscedastic Gaussian Process Regression. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 489–496, New York, NY, USA, 2005. ACM.

[14] S. N. MacEachern. Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

[15] E. Meeds and S. Osindero. An alternative infinite mixture of gaussian process experts. In *In Advances In Neural Information Processing Systems*, page 2006.

[16] R. Murray-Smith and A. Girard. Gaussian Process priors with ARMA noise models. In *Irish Signals and Systems Conference, Maynooth*, pages 147–152, 2001.

[17] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[18] C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. *Advances in neural information processing systems*, 16:273–280, 2004.

[19] Jim Pitman. PoissonDirichlet and GEM Invariant Distributions for Split-and-Merge Transformations of an Interval Partition. *Comb. Probab. Comput.*, 11(5):501–514, September 2002.

[20] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. *Advances in neural information processing systems*, 2:881–888, 2002.

[21] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[22] E. Snelson and C. E. Rasmussen. Warped Gaussian Processes. *Advances in Neural Information Processing Systems 16, 337-344 (2004)*, 2004.

[23] Y. W. Teh. Dirichlet Process. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer US, 2011.