Supplementary: 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

1 Literature Review

In this section, we review the literature that is related to our approach and indicate the main differences with our proposal. We have carried out a thorough literature search and, for brevity, we only reference those that are directly related to our approach, so many papers that are not directly comparable have been left out. We know this is a widely-research field and if any relevant body of work has not been referenced, it has been unintentional.

1.1 Modeling with Mixture of GP Experts

Mixture of expert models in which each expert is a GP has been proposed in the literature in several occasions. One of the most popular methods in Bayesian regression is the Infinite Mixture of Experts (IMoE) to capture local properties of the signal in different areas of the input space [7]. A gating function is used to determine which GP is active in each area, making it useful locally. Extensions of the IMoE include modeling the input and output jointly, i.e., modeling $p(\boldsymbol{x}, y)$ [13], or allowing the use of the same experts in different input areas through hierarchical DPs [19]. Our work addresses GP components at a global scale, which is a more natural way to capture heteroscedastic noise and other gradual behaviors. Global GP priors were previously proposed in [12], but only considering a fixed number of mixtures. Also the work in [18] proposes a mixture of global Gaussian Processes for traffic flow prediction, but their approach is generative and models the input space too.

Multiresolution Gaussian Processes [5] relies on a hierarchy of GPs to partition the whole space in order to capture long-range, non-Markovian dependencies while allowing for abrupt changes. Our method does not partition the input space and allows for multiple functions at a same input location instead. Additive GPs (aGP) in [4, 14] divide the output function into low-dimensional components of varying degrees. Our method is different as we allow for multiple output functions instead of a partitioning of the output dimensions, i.e.

$$g(y_i) = \sum_{k=1}^{\infty} f_k(x_{i1}, \ldots, x_{iD}) \qquad \text{(IMoGGP)}$$

$$g(y_i) = f(x_{i1}) + f(x_{i2}) + f(x_{i3}, x_{i4}) + \ldots \qquad \text{(aGP)}$$

The aGP seems more suitable for high-dimensional input spaces, it fails to capture heteroscedasticity or multimodality, as our proposal does.

1.2 Density Regression with Dependent Dirichlet Processes

DDPs are useful to model collections of distributions that vary in time, space or experimental settings. In the literature, it is often the case to use a semiparametric model, i.e., a parametric function for the signal, and a non-parametric prior for the noise, in order to capture heteroscedasticity or

*Also Machine Learning Scientist at Bell Labs.

non-Gaussianity [3, 15, 6, 2, 8]. Our model considers a more general formulation by assuming a completely non-parametric model for both signal and noise.

In [15], a DDP prior is used to model the joint distribution $p(\boldsymbol{x}, y)$, given different experimental conditions. With such a generative approach, modeling $\boldsymbol{x}$ might dominate over y, resulting in an under-fitting of y. Our approach directly focus on the conditional distribution $p(y|\boldsymbol{x})$, and applies the DDPs in a different way, directly over the input space $\boldsymbol{x}$. Such discriminative perspective typically gives better accuracy and has the advantage of estimating less parameters than in the generative approach. The work in [10] uses single-p DDPs, i.e., DDPs with constant weights over $\boldsymbol{x}$, to cluster the behavior of multiple ANOVA models under different experimental conditions. Here again, their approach is generative, whereas we use the DDP to directly model the conditional distribution $p(y|\boldsymbol{x})$.

In [3], DDPs are used for Bayesian density regression with kernel-varying weights, assuming a linear relationship between y and $\boldsymbol{x}$. Our approach generalizes this work by replacing the linear basis functions by arbitrary non-linear functions with a GP prior each. The authors in [6, 2] introduce a DDP-based model for spatial modeling applications called the Spatial DP prior, which is a probability weighted collection of random surfaces. They use a linear process for the signal and a mixture of GPs to capture the noise. Because each atom in the DP corresponds to a realization of a random field over the input space, their algorithm needs the assumption that multiple points are available at each location $\boldsymbol{x}$ and in particular, that the number of points assigned to each GP is always uniform. Our approach removes those restrictions and assumes no particular functional form of the signal.

1.3 Clustering of Time Series

Finally, hierarchical mixtures of Gaussian processes have often been used in the literature to cluster time series, as in [17], [9] and [16]. All these works assume prior knowledge of which points belong together to the same temporal sequence, and assign the points of a temporal sequence jointly to the same GP. All these models are generative and seek interpretable results. In our case, cluster assignments are purely auxiliary variables, we only care about predictive accuracy. Slightly different and also closely related to our approach is the work in [11], which uses a parametric mixture of GPs for the data association problem. The objective there is to find the appropriate cluster assignments for each point and recover multiple trajectories, which is useful in multi-tracking scenarios. In our case, mixture assignments are just auxiliary variables, and we assume a potentially infinite number of mixtures to represent the data.

2 Database Descriptions

The synthetic databases correspond to the examples in Section 2 of the main manuscript, and include $n = 2000$ observations for each case. The Heteroscedasticity example corresponds to a quadratic function $y = x^2 - 0.5 + \epsilon$ with added input-dependent Gaussian noise $\epsilon \sim \mathcal{N}\big(0, (0.01 + \sin(2\pi x/10)^2)^2\big)$. The Non-Gaussianity case corresponds to a cubic function $y = 4x^3 - 1 + \frac{1}{2}\epsilon + 3\gamma$ with added Student's t and gamma noise, $\epsilon \sim \mathrm{Students'\ t}(10)$ and $\gamma \sim \mathrm{Gamma}(2, 0.5)$[1]. The last toy example called Multimodality consists of 3 GPs generated using Mattern covariance functions plus Gaussian noise. In all these examples the underlying GP covariance function was a squared exponential and the likelihood model was Gaussian. Our IMoGGP model is able to deal with heteroscedastic data, heavy tailed and asymmetric noise, and parallel Gaussian processes with the same regression model, by adding global GPs over the whole input space to capture these behaviors.

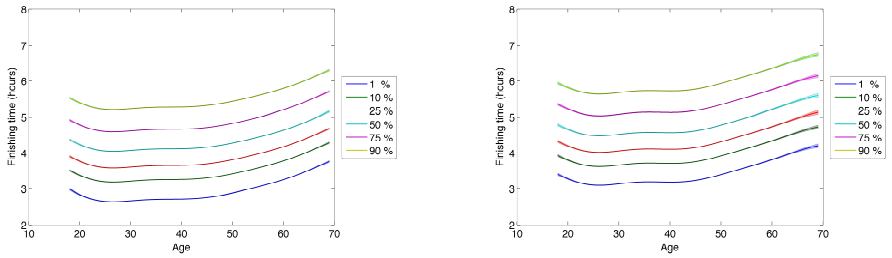
We also consider three different real databases, all of them publicly available. The concrete database from [20] consists of 1030 observations and the input dimension is 8. The objective is to predict the compressive strength (MPa) of concrete, which is one of the most important materials in civil engineering. This is a highly nonlinear function of age and ingredients that include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate. The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory.

[1] We define the Gamma distribution in terms of shape and rate parameters.

For the New York City marathon data[2], the objective is to predict the arrival time of a runner given his gender and age. The output y designates the arrival time, and $\boldsymbol{x}$ is a two dimensional vector with the age and gender for each runner. We take a subset of $4,800$ runners in total, keeping the same age/gender relative distribution. Finally, the RSSI database consists of 4799 measurements of the Received Signal Strength Indicator (RSSI), which captures the power of different wireless networks at different locations along a large corridor[3]. Modeling RSSI correctly is very important, as different signal strengths can have a strong impact on functionality in wireless planning and localization [1].

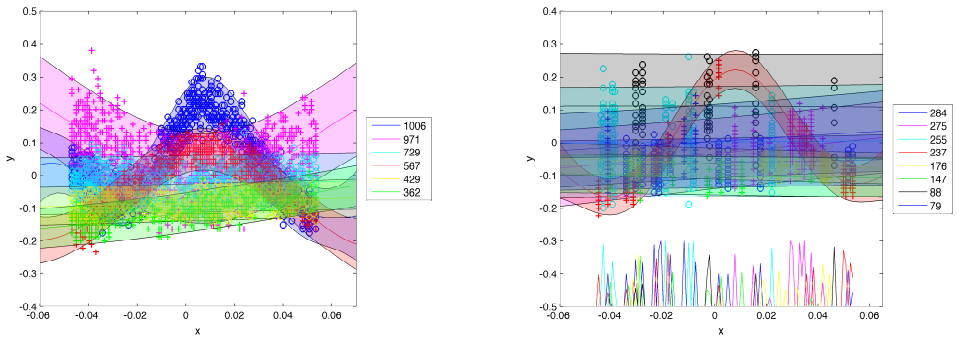
3 Further Results

Figure 1 illustrates the capacity of our IMoGGP to estimate percentiles in the marathon database for the arrival time of male and female runners. For each age, the finishing time can be modeled as a countably infinite mixture of Gaussians, and percentiles can be easily computed by integrating over this linear combination of Gaussians.



(a) Male runners. (b) Female runners.

Figure 1: **Percentile Estimation using the IMoGGP.** Application to the arrival time of male and female runners in the Marathon database.



(a) IMoGGP. (b) IMoE.

Figure 2: **Qualitative behavior of the IMoGGP and IMoE for the RSSI database.** The legend shows the number of points assigned to each individual GP. The underlying functions in the IMoE case represents the gating function for each GP. Point assignments to different GPs are represented with different colors and shapes.

Figure 2 compares the qualitative behavior of the IMoGGP and IMoE dealing with a real database, in particular, we show the estimated global Gaussian Processes in the case of the RSSI database. The gating function for the IMoE is also represented below the curves. The IMoGGP fits the data using a smaller number of GPs that cover the whole input space, whereas the IMoE tends to use several local functions.

[2]Data for the NYC marathon is available at `http://www.tcsnycmarathon.org/about-the-race/results`

[3]The authors will release this database in their webpage.

References

[1] P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *Proceedings of IEEE INFOCOM 2000*, pages 775–784. 2000.

[2] J. A. Duan, M. Guindani, and A. E. Gelfand. Generalized Spatial Dirichlet Process Models. *Biometrika*, 94(4):809–825, December 2007.

[3] D. B. Dunson, N. Pillai, and J. Park. Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183, April 2007.

[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] Emily B. Fox and David B. Dunson. Multiresolution Gaussian Processes. *arXiv:1209.0833 [stat]*, September 2012. arXiv: 1209.0833.

[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] Z. Ghahramani and C. E. Rasmussen. Bayesian monte carlo. In *Advances in neural information processing systems*, pages 489–496, 2002.

[8] J. E. Griffin and M. F. J. Steel. Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statistica Sinica*, 20(4):1507, 2010.

[9] 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.

[10] M. D. Iorio, P. Mller, G. L. Rosner, and S. N. MacEachern. An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, 99(465):205–215, March 2004.

[11] 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.

[12] J. C. Lemm. Mixtures of Gaussian process priors. *arXiv:physics/9911077*, November 1999. arXiv: physics/9911077.

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

[14] Shaan Qamar and Surya T. Tokdar. Additive Gaussian Process Regression. *arXiv:1411.7009 [stat]*, November 2014. arXiv: 1411.7009.

[15] A. Rodriguez, D. B. Dunson, and A. E. Gelfand. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96(1):149–162, March 2009.

[16] J. Ross and J. Dy. Nonparametric Mixture of Gaussian Processes with Constraints. In *International Conference on Machine Learning*, pages 1346–1354, 2013.

[17] J. Q. Shi, R. Murray-Smith, and D. M. Titterington. Hierarchical Gaussian Process Mixtures for Regression. *Statistics and Computing*, 15(1):31–41, January 2005.

[18] Shiliang Sun and Xin Xu. Variational Inference for Infinite Mixtures of Gaussian Processes With Applications to Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):466–475, June 2011.

[19] A. Tayal, P. Poupart, and Y. Li. Hierarchical Double Dirichlet Process Mixture of Gaussian Processes. In *AAAI*, 2012.

[20] I. C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, December 1998.