Sample-based estimation of probability density fields: a spatial extension of the logistic Gaussian process

A. Gautier
Idiap Research Institute and University of Bern

Supervisor(s): Prof. Ginsbourger (Idiap Research Institute and University of Bern)

Ph.D. expected duration: Nov. 2018 - Oct. 2022

Adress: Rue Marconi 19, CH-1920 Martigny

Email: athenais.gautier@idiap.ch

Abstract:

We consider natural or artificial systems for which for each instance of input variables $\mathbf{x}$, the corresponding output is random and follows a probability distribution $\mu_{\mathbf{x}}$ that depends on $\mathbf{x}$. We denote the input set by D, typically assumed to be a compact set in Euclidean space, and use the letter t (resp. T, in random form) to denote outputs, which range of values is denoted $\mathcal{I} \subset \mathbb{R}$. Our aim here is to estimate the field $\{\mu_{\mathbf{x}},\ \mathbf{x} \in D\}$ based only on a finite number of observations $(\mathbf{x_i}, t_i)_{1 \leq i \leq n} \in D \times \mathcal{I}$ where the t_i's were independently sampled from the $\mu_{\mathbf{x}_i}$'s, respectively. Such settings are notably inspired by stochastic optimization and optimization problems, for which estimates of $\{\mu_{\mathbf{x}},\ \mathbf{x} \in D\}$ and associated quantifications of uncertainty could be highly beneficial.

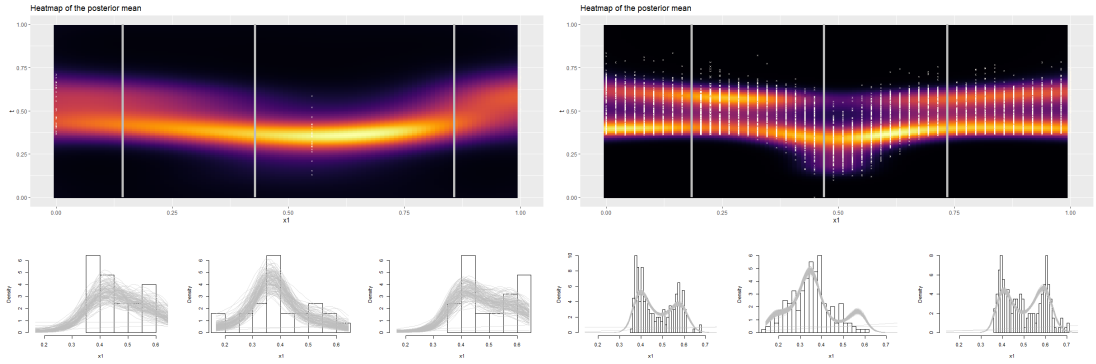
Related problems have been tackled in geostatistics within distributional extensions of kriging, motivated in particular by compositional data analysis. Yet, the considered framework of heterogeneous sample sizes accross space does not easily fit into standard distributional kriging frameworks where probability densities are either given or estimated by differentiating smooth cumulative distribution function estimates. On the other hand, some conditional density estimation approaches allow handling heterogeneous sample sizes but can be constraining in terms of which distributional features are allowed to vary and/or how they may vary over space. In contrast, the approach that we introduce here generalizes to spatial contexts a class of non-parametric Bayesian density models based on logistic Gaussian processes [1, 3], and allows modelling density-valued fields with complex dependences of $\mu_{\mathbf{x}}$ on $\mathbf{x}$ while accomodating heterogeneous sample sizes.

Spatial extension of the logistic Gaussian process : We focus here on cases where the $\mu_{\mathbf{x}}$'s are absolutely continus with respect to a common reference measure on $\mathcal{I}$, and we further denote the associated density field by $\{p_{\mathbf{x}},\ \mathbf{x} \in D\}$. Our proposed approach crucially relies on the definition of a Gaussian Process (GP) indexed by the cartesian product $D \times \mathcal{I}$. Considering indeed a deterministic trend $m : D \times \mathcal{I} \to \mathbb{R}$ and a covariance kernel $k : (D \times \mathcal{I})^2 \to \mathbb{R}$, we define $\{p_{\mathbf{x}}, \mathbf{x} \in D\}$ by

$$p_{\mathbf{x}}(t) = \frac{e^{Z(\mathbf{x},t)}}{\int_{\mathcal{I}} e^{Z(\mathbf{x},u)}\, du} \qquad ((\mathbf{x}, t) \in D \times \mathcal{I}), \tag{1}$$

where Z is a GP indexed by $D \times \mathcal{I}$, with mean function m and covariance kernel k such that for all $\mathbf{x} \in D$, $\int_{\mathcal{I}} e^{Z(\mathbf{x},u)}\, du < \infty$ a.s.. Under such assumptions, the random field $\{p_{\mathbf{x}},\ \mathbf{x} \in D\}$ takes values in the space of probability densities $(D \to \mathcal{I})$ and therefore induces a prior over this space.

Our contributions build upon bayesian non parametric inferences on fields of probability density functions under this prior. The considered models allow for instance performing (approximate) posterior simulations of probability density functions as well as jointly predicting multiple moments or other functionals of target distributions.



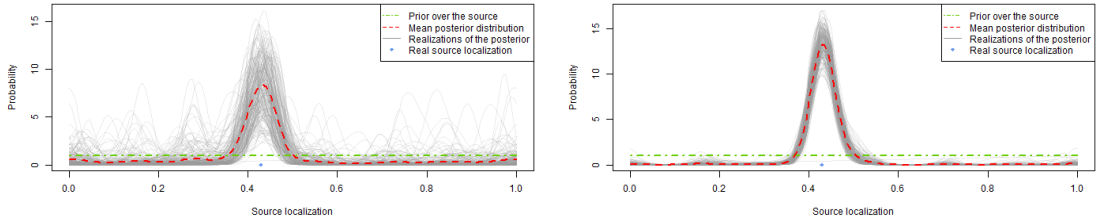
(a) Using 150 data points (b) Using 10000 data points

Figure 1: Estimating a field of probability density functions based on spatially spread samples : colormap of the mean field (top) and posterior pdfs vs histogram of data on slices (bottom)

In particular, we investigate ways of using the proposed class of model to speed up Approximate Bayesian Computing (ABC) methods [2] and further iterative algorithms involving decision making on where in input space to run stochastic simulations, be it for optimization for inversion goals.

Ongoing work directions include the design and estimation of look-ahead aquisition criteria leveraging the model's probabilistic nature towards efficient uncertainty reduction sampling schemes.

We illustrate the applicability of this approach to a problem of contaminant source localization under uncertain geological structure (collab. with G. Pirot, Univ. of Western Australia).



(a) Using 500 samples (b) Using 5000 samples

Figure 2: Realizations of the random posterior distribution of the source localization

References

[1] Peter J. Lenk. Towards a practicable bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.

[2] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6), November 2012.

[3] Surya Tokdar and Jayanta K. Ghosh. Posterior consistency of logistic gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137:34–42, 01 2007.

Short biography – Athénaïs Gautier graduated with an engineering degree from Mines de Saint Etienne (2015-2018) as well as a MSc in applied mathematics from University Paris Dauphine (2018). Her PhD takes place within the framework of the Swiss National Science Foundation project number 178858 on "Uncertainty quantification and efficient design of experiments for data and simulation-driven inverse problem solving".