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

A. Gautier
University of Bern

Supervisor(s): Prof. Ginsbourger (University of Bern)

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

Adress: Alpeneggstrasse 22, CH-3012 Bern

Email: athenais.gautier@stat.unibe.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 $\mathcal{T} \subset \mathbb{R}$. Our aim here is to estimate the field $\{\mu_{\mathbf{x}},\ \mathbf{x} \in D\}$ based only on a finite number of observations $(\mathbf{x_i}, t_i)_{1 \leq i \leq n} \in D \times \mathcal{T}$ 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 inversion problems, for which estimates of $\{\mu_{\mathbf{x}},\ \mathbf{x} \in D\}$ and associated uncertainty quantification could be instrumental.

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.

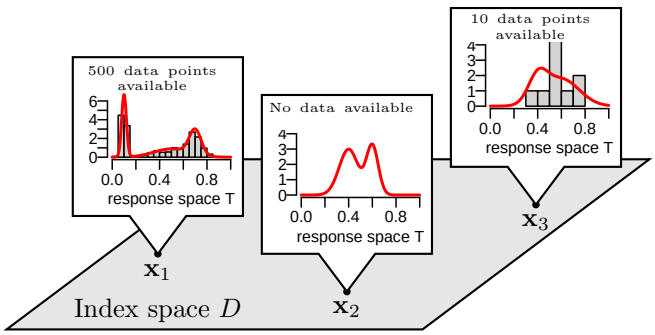


Figure 1: Typical setting: probability density of $\mu_{\mathbf{x}}$ (red curve) versus collected data (histogram)

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 investigate here generalizes to spatial contexts a class of non-parametric Bayesian density models based on logistic Gaussian processes [2, 3], and allows modelling density-valued fields with complex dependences of $\mu_{\mathbf{x}}$ on $\mathbf{x}$ while accommodating heterogeneous sample sizes.

Spatial extension of the logistic Gaussian process : We focus here on cases where the $\mu_{\mathbf{x}}$'s are absolutely continuous 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

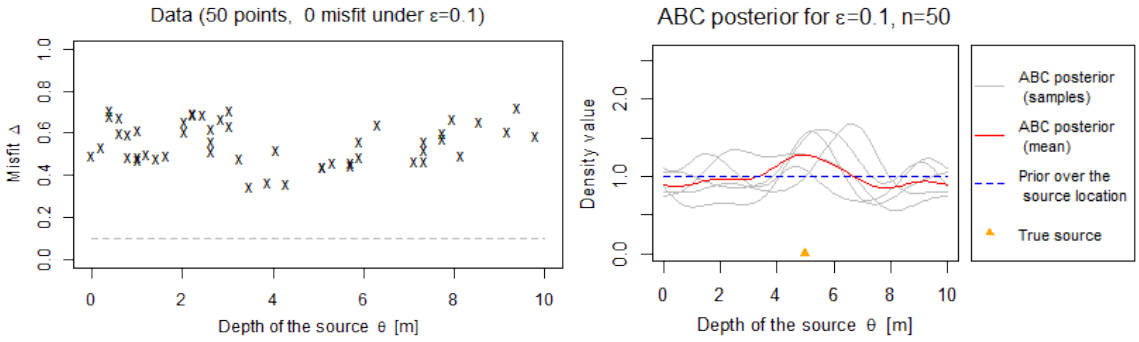
definition of a Gaussian Process (GP) indexed by the Cartesian product $D \times \mathcal{T}$. Indeed, we define $\{p_{\mathbf{x}}, \mathbf{x} \in D\}$ by:

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

where Z is a GP indexed by $D \times \mathcal{T}$ such that for all $\mathbf{x} \in D$, $\int_{\mathcal{T}} e^{Z(\mathbf{x},u)}\, du < \infty$ a.s.. Under such assumptions, the random field $\{p_{\mathbf{x}},\ \mathbf{x} \in D\}$ takes values in the space of probability densities $(D \to \mathcal{T})$ 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.

We present some sufficient conditions on covariance kernels underlying SLGPs for associated models to enjoy spatial regularity properties. And propose an implementation of SLGP. Finally, we investigate ways of using the proposed class of model to speed up Approximate Bayesian Computing (ABC) methods and further iterative algorithms involving decisions on points in D where to run new stochastic simulations, be it for optimization or for inversion goals [1] .



(a) Misfits between observed and simulated data. (b) Estimated ABC-posterior of the source depth

Figure 2: SLGP in stochastic inverse problem: using 50 simulations to infer a contaminant source depth under uncertain geological structure (collab. with G. Pirot, Univ. of Western Australia)

References

[1] Athénaïs Gautier, David Ginsbourger, and Guillaume Pirot. Probabilistic ABC with Spatial Logistic Gaussian Process modelling. In *Third Workshop on Machine Learning and the Physical Sciences. NeurIPS 2020.*

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

[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".