**MASCOT-NUM 2022**
**Clermont Ferrand, France**
**7-9 June 2022**

# Leveraging spatial regularity in distribution field estimation to accelerate statistical inference for the sciences.

Athénaïs Gautier
*IMSV, University of Bern*

**Supervisor(s):** David Ginsbourger (IMSV, University of Bern)

**PhD expected duration:** Nov. 2018 - Oct. 2022

**Address:** Alpeneggstrasse 22, CH-3012 Bern

**E-mail:** athenais.gautier@stat.unibe.ch

**Abstract:**

Precise statistical inference in spatially-dependent complex systems usually requires to either benefit from a high computational budget which allows system evaluation at a dense network of inputs variables $\mathbf{x}$, or to rely on assumptions on the model output $Y_{\mathbf{x}}$. For practical reasons, the latter is generally preferred and widely used in Uncertainty Quantification (UQ), with the resulting class of approaches being broadly called Surrogate modelling. In this setting, practitioners build upon an initial design of experiments to construct a proxy which accurately reflects users' knowledge of the system.
While techniques exist to approximate $Y_{\mathbf{x}}$ or its distribution $\mu_{\mathbf{x}}$, Gaussian Process Regression has gained in popularity as it provides a probabilistic prediction of $Y_{\mathbf{x}}$. This enables us not only to perform inference, but also to guide further data acquisition. However, this comes at the cost of a strong (Gaussian) distributional hypothesis. Other popular methods such as Artificial Neural Networks or Polynomial Regression focus on leveraging spatial regularity of the system, but do not provide a probabilistic prediction, which makes their application to design of experiment less straightforward.

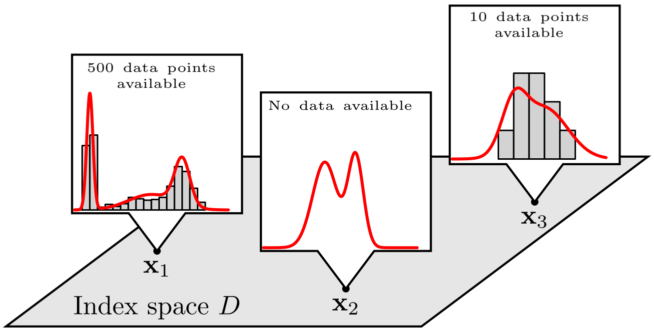



Figure 1: Typical setting: probability distribution $\mu_{\mathbf{x}}$ (red curves) versus collected data (histograms). Here, our goal is to perform sample-based inference simultaneously on the whole field $\{\mu_{\mathbf{x}}\}_{x\in D}$

In this work, we focus on a non-parametric approach, which allows to break free of distributional hypothesis on the system output. We use a class of distribution-valued fields called Spatial Logistic Gaussian Process (SLGP) to model and estimate the distribution field $\{\mu_{\mathbf{x}}\}_{x\in D}$ based on a scattered observations $(\mathbf{x_i}, t_i)_{1\leq i\leq n}$. The SLGP inherit its spatial regularity from a latent Gaussian Process [1]. This framework accommodates for heterogeneous sample sizes across space, with little to no replicates, as well as changes of shapes and modalities on the target densities. This distinguishes our approach from most techniques arising in distributional regression and conditional density estimation. Since the model provides a probabilistic prediction of $\mu_{\mathbf{x}}$, it allows for UQ, but also for adapting principled sequential approaches from the GP framework to our non-parametric setting [3].

We explore the potential of SLGPs in stochastic inverse problems, where we revisit the framework of Approximate Bayesian Computation (ABC) [2]. We demonstrate on application cases from natural sciences that by surrogating the ABC-likelihood, our model enables scientists to perform accurate inference even in the low to moderate data regime. Additionally, we can benefit from the SLGP's probabilistic nature to guide data acquisition, with approaches inspired from those presented in a Gaussian setting [4, 5].

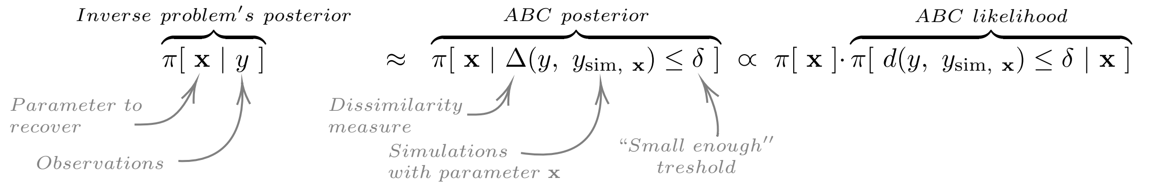
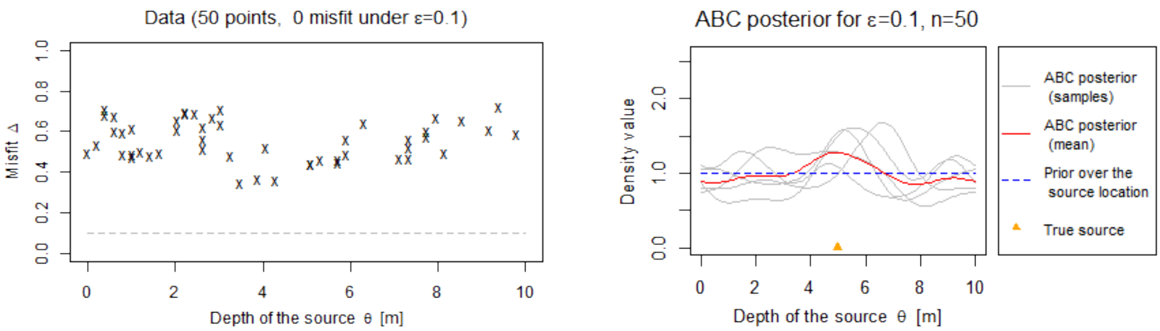



Figure 2: Informal summary of SLGP-ABC: we surrogate the ABC likelihood (i.e. the dissimilarity distribution at $\mathbf{x}$) with a SLGP, and derive the ABC posterior from it.




(a) Dissimilarities between observed and simulated data.

(b) SLGP-ABC posterior on the source depth.

Figure 3: SLGP in stochastic inverse problem: using 50 simulations to infer a contaminant source depth under uncertain geological structure (collaboration with G. Pirot, Univ. of Western Australia).

Our method's soundness is evaluated through the adequation between our predicted SLGP-ABC posteriors and histograms obtained by running classical ABC for an extended number of simulations. This highlights the sped-up achieved and the consistency between both approaches.

## References


[1] Athénaïs Gautier and David Ginsbourger. Continuous logistic gaussian random measure fields for spatial distributional modelling. *arXiv preprint arXiv:2110.02876*, 2021.

[2] 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*, 2020.

[3] Gautier, Athénaïs, Ginsbourger, David, and Pirot, Guillaume. Goal-oriented adaptive sampling under random field modelling of response probability distributions. *ESAIM: ProcS*, 71:89–100, 2021.

[4] Marko Järvenpää, Michael U Gutmann, Arijus Pleska, Aki Vehtari, and Pekka Marttinen. Efficient acquisition rules for model-based approximate bayesian computation. *Bayesian Analysis*, 14(2):595–622, 2019.

[5] Marko Jarvenpaa, Aki Vehtari, and Pekka Marttinen. Batch simulations and uncertainty quantification in gaussian process surrogate approximate bayesian computation. In *Conference on Uncertainty in Artificial Intelligence*, pages 779–788. PMLR, 2020.


**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".