Probabilistic ABC with Spatial Logistic Gaussian Process modelling

Athénaïs Gautier
Institute of Mathematical Statistics and Actuarial Science
University of Bern, Switzerland
athenais.gautier@stat.unibe.ch

David Ginsbourger
Institute of Mathematical Statistics and Actuarial Science
University of Bern, Switzerland
david.ginsbourger@stat.unibe.ch

Guillaume Pirot
Centre for Exploration Targeting,
The University of Western Australia, Australia
guillaume.pirot@uwa.edu.au

Abstract

Simulation-based inference has become key in a number of application domains from physical sciences and beyond. While the Bayesian framework lends itself well to inverse problems, it often happens that involved likelihoods are intractable or very cumbersome to evaluate, so that likelihood-free methods such as Approximate Bayesian Computation (ABC) are appealed to. However, ABC methods can also be quite simulation-consuming, and efforts have recently been paid to alleviate associated costs by means of surrogating the considered dissimilarity across parameter space, yet typically under strong distributional assumptions and/or with space exploration strategies that can prematurely exclude parameter regions. In this work we propose a non-parametric approach that allows speeding-up ABC via probabilistic prediction of the dissimilarity distribution field. Our proposed spatial logistic Gaussian Process ABC approach is finally illustrated based on a test case application in contaminant localization under uncertain geology.

1 Introduction

In physical sciences, complex simulation models are extensively used as they allow accurate modelling of complex phenomena [6, 7, 15, 23]. However, such models are often difficult to use in Bayesian inference as their associated likelihood function is either too expensive to evaluate or too complicated to derive.
The framework of Likelihood Free Inference (LFI) has been developed to address this issue. Approximate Bayesian Computation (ABC) methods [1, 13] have arguably become the most popular class of approaches to perform LFI in the context of simulation models. ABC aims at identifying parameters leading to simulation results similar to observed data, by-passing in turn the need to evaluate the likelihood function.
One of the main current challenges in this field is the fact that ABC techniques generally require a large number of simulations to deliver precise inference, which can be computationally expensive.
In this work, we propose a non-parametric probabilistic model of the conditional distributions of

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

misfit between simulations and observed data, given the simulation parameters. Our approach allows modelling density-valued fields with complex dependencies of the parameter space and delivers probabilistic predictions of the distribution field. The considered approach delivers a generative model for the misfits, which in turn, yields probabilistic prediction of the ABC posterior.
We introduce the general framework and idea of ABC methods in the first section of this document. Then, we present the Spatial Logistic Gaussian Process and discuss its potential for LFI in Section 3. Finally, results on an application test case pertaining to contaminant localization under uncertain geology are presented in Section 4.

2 Bayesian inference and Approximate Bayesian Computation

2.1 The classical framework of Bayesian inference

Let us consider a parametric statistical model $\mathcal{F}_\theta$ and some observed data y_{obs} assumed to stem from this model, with a value of θ that is unknown and to be estimated. In Bayesian inference, the parameter θ is treated as random, and a prior distribution is assumed for it. Assuming further that the prior distribution possesses a density $\pi[\theta]$ (with dominating measure being typically the Lebesgue measure in finite-dimensional cases), the likelihood function can be written as $\theta \mapsto \pi[y_{obs}|\theta]$, and the posterior distribution of θ knowing y_{obs} can be expressed in virtue of Bayes theorem as

$$\pi[\theta|y_{obs}] \propto \pi[y_{obs}|\theta]\pi[\theta] \tag{1}$$

However, often the likelihood function is intractable or prohibitively costly to evaluate. Approximate Bayesian Computation is a popular framework to address this issue.

2.2 Approximate Bayesian Computation

In the ABC framework, we assume that, as often in physical systems, it is possible to simulate the response associated to any given instance of θ. It is also assumed that we have access to a measure of dissimilarity Δ between responses, allowing us to compare simulated versus observed data.
Denoting by y_θ a random response with input θ and viewing θ as random with prior density π, the essence of ABC is to approximate the posterior as follows:

$$\pi[\theta|y_{obs}] \approx \pi[\theta|\Delta(y_{obs}, y_\theta) \leq \epsilon], \tag{2}$$

where $\epsilon > 0$ is a prescribed "small enough" threshold. The most basic ABC algorithm, the ABC rejection sampler [18, 20], can be summarized by the following pseudo code:

input : Prior distribution $\pi[\theta]$, simulation model $\pi[y_\theta|\theta]$, threshold ϵ, number of steps T
for $i \leftarrow 1$ **to** T **do**
  Draw θ_i from $\pi[\theta]$
  Simulate y_i from $\mathcal{F}_{\theta_i}$
  Accept θ_i if $\Delta(y_{obs}, y_i) \leq \epsilon$
end
output : Parameters θ_i that have been accepted

Algorithm 1: ABC rejection sampler

The main issue that we are tackling in this work is the computational efficiency of this approach. Indeed, if the prior is substantially broader than the posterior, as is common in practice, most simulations are rejected and the ABC rejection sampler becomes very inefficient.
To address this problem, several approaches aiming at more efficient ABC posterior sampling have been developed. Among these techniques, one can cite Markov chain Monte Carlo ABC [14], sequential and population Monte Carlo ABC [1, 2, 3, 4, 12, 19, 22]. The two latter consist in replacing draws of θ from the prior with draws from an adapted proposal density. Another class of methods consist in Synthetic Likelihood (SL) methods [17, 24], where the misfit distribution is assumed to stem from a parametric family (usually Gaussian).
The approach we present here is similar to the one used is SL methods, with the main difference being that we model the misfit distributions non-parametrically. Therefore, we do not require strong distributional hypotheses on misfits.

3 A non-parametric model for density field estimation

3.1 The Spatial Logistic Gaussian Process model

We appeal here to a flexible non-parametric Bayesian approach for density-valued field estimation that allows modelling densities $\pi(\cdot|\theta)$ which variations with respect to θ do not only concern specific quantities such as mean and variance, but also allow other distribution features to evolve over parameter space, including for instance their shape, their uni- versus multi-modal nature, etc. The considered approach builds upon a class of models, coined Spatial Logistic Gaussian Process (SLGP), that generalizes logistic Gaussian process models used in density estimation [11, 21] to the case of density field estimation.
Considering a deterministic function $\mu : (D \times \mathcal{I} \mapsto \mathbb{R})$ and a covariance function k on $(D \times \mathcal{I}) \times (D \times \mathcal{I})$ such that for $W \sim \mathcal{GP}(\mu, k)$ and all $\theta \in D$, $\int_{\mathcal{I}} e^{W(\theta,u)}\, du < \infty$ a.s., one defines a field of conditional probability densities based on a SLGP via:

$$p(t|\theta) = \frac{e^{W(\theta,t)}}{\int_{\mathcal{I}} e^{W(\theta,u)}\, du} \quad \forall(\theta, t) \in D \times \mathcal{I} \tag{3}$$

The stochastic process $\{p(\cdot|\theta), \theta \in D\}$ takes values in the space of densities $(D \to \mathcal{I})$ and is used here to induce a prior over this space.
This prior allows performing Bayesian non-parametric estimation of fields of probability density functions. The considered models deliver probabilistic predictions of distribution fields, allowing for instance to perform (approximate) posterior simulations of probability density functions as well as jointly predicting multiple moments or other functionals of target distributions.

3.2 Leveraging the SLGP for likelihood free inference

We consider the standard ABC framework, where the posterior is approximated with the ABC posterior recalled in Equation 2. We assume that the available data consists in n couples of parameters and misfits, noted $\{(\theta_1, \Delta(y_{obs}, y_1)), ..., (\theta_n, \Delta(y_{obs}, y_n))\}$. Using these data, we estimate $\pi[\Delta(y_{obs}, y_\theta)|\theta]$ relying on the probability density field model defined by Equation 3.
Considering that for a given $\epsilon > 0$ and a given prior π, this ABC posterior can be written as:

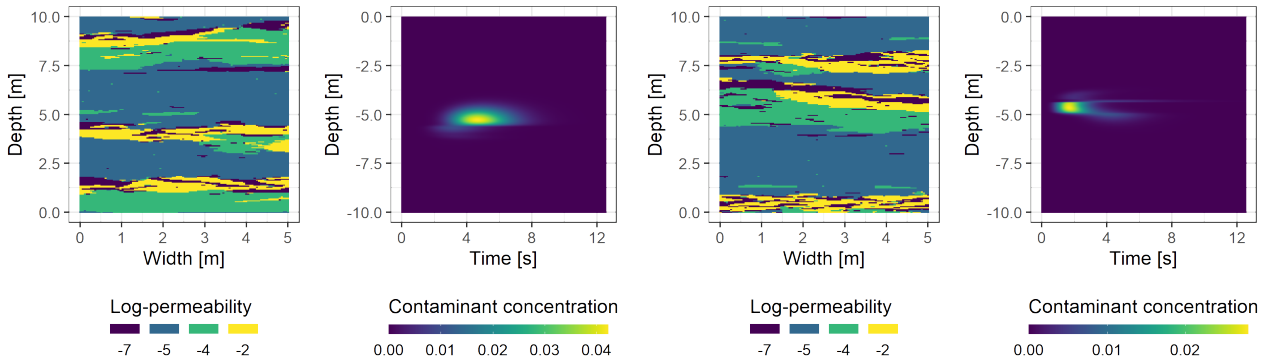
$$\pi[\theta|\Delta(y_{obs}, y_\theta) \leq \epsilon] \propto \pi[\Delta(y_{obs}, y_\theta) \leq \epsilon|\theta]\pi[\theta], \tag{4}$$

we use our estimation of the misfit distribution field to deliver a probabilistic surrogate to this ABC posterior by replacing $\pi[\Delta(y_{obs}, y_\theta) \leq \epsilon|\theta]$ in the latter equation by its estimated value.
Some of the main strengths of our approach, compared to classical ABC, is that our model can be trained on data sets where the parameters θ_i do not stem from the prior, and we incorporate information from all the simulations, not just the best ones.

4 An application in geosciences

To demonstrate applicability of this approach, we consider a one dimensional contaminant problem. We want to localize the source of a contaminant propagating into a saturated aquifer when the geological structure is unknown. Indeed, characterization of subsurface properties is very uncertain as soon as the distance to the scarce measurement locations increases. Therefore, hydro-geologists must rely on the use of analogues and expert knowledge to generate an ensemble of plausible geological realizations that can be used to quantify prediction uncertainty.

The reference observations consist of concentration breakthrough curves at different depths of the domain outlet. Simulations are obtained in two steps. First, a plausible geological realization is obtained as multiple-point statistics realizations generated with the Deesse algorithm (Mariethoz et al., 2010 [5]) that reproduce the complex features of braided-river aquifer models (Pirot et al., 2013 [16]). Then, the contaminant flow is simulated under steady-state flow and fixed boundary conditions (constant hydraulic gradient) using the Maflot Matlab code (Kunze and Lunati, 2011 [10]), hence yielding simulated concentration breakthrough curves. Examples of plausible geologies and breakthrough curves are shown in Figure 1. In this application, our reference observation is a simulation in itself. Therefore, we know the exact localization of the contaminant source.

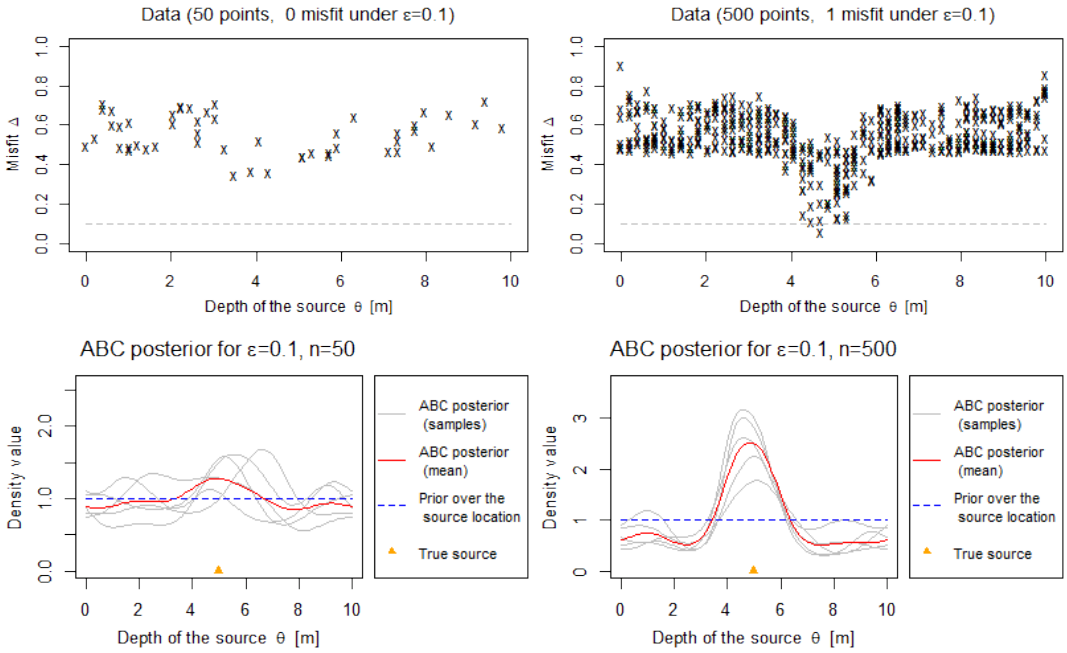


(a) First geological structure

(b) Second geological structure

Figure 1: Two geological structures and associated simulated response for a source of depth 5m.

We ran 50 (resp. 500) simulations using the approach mentioned above, computed the misfits (here, L^2 distances) between our reference observation and the simulations and used them to train our SLGP model. In this application, the considered SLGP is constructed by transforming a centered GP with a Matérn 5/2 covariance kernel, and inference of kernel hyperparameters is performed following a Bayesian approach. From an implementation perspective, the joint posterior distribution of kernel hyper-parameters and inducing values underlying the SLGP was approximated by MCMC. In turn, we use resulting approximate posterior SLGP samples to estimate $\pi[\Delta(y_{obs}, y_\theta)|\theta]$ and to derive the corresponding ABC posterior $\pi[\Delta(y_{obs}, y_\theta) \leq \epsilon|\theta]$. The results are available in Figure 2.



(a) Using 50 simulations

(b) Using 500 simulations

Figure 2: Misfit between observation and simulations and plausible ABC-posteriors for two different sample sizes.

It is worth noting that even though the number of simulations would be highly insufficient to estimate the posterior for a classical ABC approach, as there is here at best one simulation that yielded a misfit value under the threshold, our approach was still able to capture information from all the simulations and locate the true source, with a confidence increasing with the number of simulations.

5 Conclusion and discussions

We presented a methodology allowing to enhance ABC-posterior estimation by leveraging the regularity of dissimilarity distributions across parameter space. This methodology appeared to be particularly promising for small to moderate sample sizes, as illustrated on a simple contaminant localization example under uncertain geology. Some particularly appealing features of the proposed approach is the use of the whole data set as opposed to keeping only those couples leading to small dissimilarities to the observations, as well as the probabilistic nature of the employed model. Having a probabilistic model enables delivering uncertainty assessments along with the estimates of interest, a property that can be especially useful when working with data sets of moderate size. Furthermore, disposing of a generative model could help exploring the effect of tuning ϵ on the overall approach, potentially leading to more integrated and automated strategies. Of course, this is all relying on model adequacy in the first place, and it would be valuable to further develop diagnostics and model adjustment methods in order to best fit SLGPs to the data at hand as well as signal potential fitting issues. In turn, consistency results in terms of the ABC posterior ought to be investigated. Finally, we started to explore potentialities of SLGP models in order to create acquisition functions, and we are looking forward to novel approaches pertaining to sequential design of experiments dedicated to ABC (following up on recent works by Järvenpää et al. that rely on GPs [8, 9]) with a specific focus on the potential benefits and challenges brought by SLGP models.

Broader impact

In physical sciences, acquiring new data is an expensive process, especially for the large scale or complex phenomena. Therefore, being able to deliver consistent estimations of the quantity of interests while accounting for the uncertainty due to the moderate sample size is crucial.
Upcoming work also aims at studying the potential of our model when it comes to guiding data acquisition.

Acknowledgements

The authors wish to thank the reviewers for their remarks that helped to improve the paper. AG's and DG's contributions have taken place within the Swiss National Science Foundation project number 178858. AG and DG would like to warmly thank Dr. Tomasz Kacprzak (ETH Zürich) for early discussions having motivated part of this work.

References

[1] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[2] Fernando V Bonassi and Mike West. Sequential Monte Carlo with Adaptive weights for Approximate Bayesian Computation. *Bayesian Anal.*, 10(1):171–187, 03 2015.

[3] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[4] Christopher C Drovandi and Anthony N Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

[5] Philippe Renard Grégoire Mariethoz and Julien Straubhaar. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46(11), 2010.

[6] Jörg Herbel, Tomasz Kacprzak, Adam Amara, Alexandre Refregier, Claudio Bruderer, and Andrina Nicola. The redshift distribution of cosmological samples: a forward modeling approach. *Journal of Cosmology and Astroparticle Physics*, 2017(08):035, 2017.

[7] Philip B Holden, Neil R Edwards, James Hensman, and Richard D Wilkinson. ABC for climate: dealing with expensive simulators. *Handbook of approximate Bayesian computation*, pages 569–95, 2018.

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

[9] Marko Järvenpää, Aki Vehtari, and Pekka Marttinen. Batch simulations and uncertainty quantification in Gaussian process surrogate-based approximate Bayesian computation. *arXiv preprint arXiv:1910.06121*, 2019.

[10] Rouven Künze and Ivan Lunati. A matlab toolbox to simulate flow through porous media. Technical report, University of Lausanne, Switzerland, 2011.

[11] Peter J. Lenk. Towards a Practicable Bayesian Nonparametric Density Estimator. *Biometrika*, 78(3):531–543, 1991.

[12] Maxime Lenormand, Franck Jabot, and Guillaume Deffuant. Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013.

[13] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[14] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[15] Trevelyan J McKinley, Ian Vernon, Ioannis Andrianakis, Nicky McCreesh, Jeremy E Oakley, Rebecca N Nsubuga, Michael Goldstein, Richard G White, et al. Approximate Bayesian Computation and simulation-based inference for complex stochastic epidemic models. *Statistical science*, 33(1):4–18, 2018.

[16] Guillaume Pirot, Julien Straubhaar, and Philippe Renard. A pseudo genetic model of coarse braided-river deposits. *Water Resources Research*, 51(12):9595–9611, 2015.

[17] Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

[18] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[19] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[20] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[21] 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.

[22] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[23] Anja Weyant, Chad Schafer, and W Michael Wood-Vasey. Likelihood-free cosmological inference with type Ia supernovae: approximate Bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, 764(2):116, 2013.

[24] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.