Modelling and predicting distribution-valued fields with applications to inversion under uncertainty

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Athénaïs Gautier
from **France**

Supervisor of the doctoral thesis:
Prof. Dr. David Ginsbourger

Institute of Mathematical Statistics and Actuarial Science
University of Bern
Switzerland

Modelling and predicting distribution-valued fields with applications to inversion under uncertainty

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Athénaïs Gautier
from **France**

Supervisor of the doctoral thesis:
Prof. Dr. David Ginsbourger

Institute of Mathematical Statistics and Actuarial Science
University of Bern
Switzerland

Accepted by the Faculty of Science.

Bern, 19 May 2023 The Dean: Prof. Dr. Marco Herwegh

Acknowledgements

Dear reader, you find these words at the opening of your journey through my Ph.D. thesis. However, to me, these words mark the final addition to this manuscript. I cannot think of a more fitting conclusion than to express my sincere thanks to those who have supported me throughout this journey.

Naturally, I will start this joyous task by expressing my deepest gratitude to my supervisor, David Ginsbourger, for offering me this incredible opportunity and for introducing me to the thrills of conducting research. From the highs of successful research venues to the frustrations of failed attempts, you were there for me, providing unwavering support and encouragement. Your constructive criticism and feedback pushed me to do better, and I learned and discovered things far beyond the scope of statistics and kernels under your guidance. Your mentorship extended beyond academic research to my personal growth, and I am grateful for the strength you lent me during difficult times.

I wish to continue with words for my co-author, Guillaume Pirot, for his help and collaboration in the results presented here, but also for his companionship in eating dessert. And although we don't have any published work together, I also believe that my "co-Ph.D. student", Cedric Travelletti deserves recognition as one of the first people to be thanked in this thesis. Starting our journey together on the same day, I am grateful for sharing this adventure with you. You brought the fresh air of your mountains to this office we shared, and always were by my side, whether it was bringing Kolmogorov formalism to this manuscript or embarking on daring journeys to reach the Danube, your presence and support have been invaluable.

Furthermore, my sincere gratitude goes to the thesis reviewers and jury members. I express my warmest thanks to Robert Gramacy for accepting the role of the external reviewer and providing thoughtful notes, encouragement, and guidance. I am also deeply grateful to Ilya Molchanov for sharing insights into the theoretical aspects of this research and for agreeing to be my jury president.

I would also like to express my gratitude to Wolfgang Polonik for his active engagement in providing feedback through meetings and reviewing earlier versions of sections in this manuscript. His constructive input has significantly enhanced the clarity and coherence of my research. Additionally, I am thankful to Yves Deville for his invaluable advice and support in coding and advancing the implementations. His expertise and guidance have been crucial in overcoming computational challenges. The contributions of these esteemed individuals to the improvement and evaluation of this thesis are greatly appreciated.

This work has been shaped by numerous meetings, discussions, and presentations over several years. The invaluable feedback received from our colleagues in our "extended research group" have in more than one way shaped the outcome of this work. Therefore, I would like to express my heartfelt appreciation to Eliane Maalouf, as well as extend my thanks to Niklas Linde, Lea Friedli, Shiran Levy, and Macarena Amaya for their support and feedback. Additionally, I am grateful to Kevin Heng and Tomasz Kacprzak for their contributions at the stellar level, as well as Philippe Schwaller, Henry Moss, and Ryan-Rhys Griffiths for their insights into the molecular aspect of work not presented here.

I am indebted to various other meetings, particularly during academic visits, which have played a significant role both in my scientific and personal development. I sincerely thank François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Olivier Roustant for welcoming me at the Institut Mathématique de Toulouse in 2021; Jo Eidsvik for welcoming me to NTNU in Trondheim during the same year; and Peter Frazier for introducing me to his research activities and his research group during my short visit to Cornell University in 2022 (with a special thought to Raul Astudillo and Poompol Buathong, who were great guides there). These visits and interactions have broadened my horizons, fostered valuable connections, and nurtured my scientific growth.

Having mentioned all these interactions, it would be wrong to forget the MASCOTNUM community. It is trully rare to be part of a community so supportive, and to benefit from the friendship and the scientific mentorship of fellow students and established researcher. With such a large research group, I am deemed to forget some names, but I still want to extend my gratitude to Amandine Marrel, Bertrand Iooss, Baptiste Kerleguer, Clément Gauchy, Clémentine Prieur, Delphine Sinoquet, Elias Fekhari, Faouzi Hakimi, Gael Poëtte, Guillaume Chennetier, Julien Bect, Julien Pelamatti, Luc Pronzato, Marouane Il Idrissi, Mickaël Binois, Naoufal Acharki, Noé Fellmann, Nora Lüthen, Paul Novello, Paul Saves, Rodolphe Le Riche, Sébastien Da Veiga, Sébastien Petit, Thierry Klein, Victor Picheny and Vincent Chabridon, among others.

Undoubtedly, the completion of this work would not have been possible without the support of the Swiss National Science Foundation (SNSF), Idiap Research Institute, and of the Institute for Mathematical Statistics and Actuarial Science (IMSV) at the University of Bern who supported me financially throughout my doctoral studies. I have been very fortunate to do my Ph.D. at the IMSV, an institute that not only fosters a stimulating scientific environment but yet remains deeply human and supportive. Amidst this remarkable alliance of professional support and personal connections, I find a shining example in Johanna Ziegel. Whether seeking advice on my academic trajectory or in need of personal support, Johanna has consistently been there for me. Her dedication exemplifies the harmonious blend of professional expertise and genuine care that characterizes the IMSV community.

I am grateful to the lecturers Michel Piot and Lutz Dümbgen (and David again) from IMSV, and Phil Garner at Idiap, for giving me the opportunity to assist in their respective courses. This experience has taught me a lot, both in terms of teaching practice and of statistics, and although it was sometimes challenging, I am glad I had this opportunity. This is not the only chance I had in learning and working for the institute, since I was also lucky to be a member of the consulting task force. For trusting me with this opportunity and guiding me through this, I am grateful to Michael Vock. Despite not having worked under his guidance, Riccardo Gatto has also been a formidable support, and I am grateful to have my office so close to his, and thankful for all our interactions, ever so kind and caring.

This kindness was shared by our office managers, who always made me feel welcome, and I want to particularly thank Andrea Fraefel and Andrea Hitz for their presence and their impressive efficiency. And of course, the life at the institute would not be the same without all the other (current and former) members that I had the chance to meet over the last four-to-five-ish years, among who: Alexander and Alexander, Alexandre (Mich-Mich), Antoine, Bellinda, Chinmoy, Christof, Clément, Fabian, Federico, Jon, Jorge, Jose, Patric, Philip, Riccardo, Sam L. (the L is Welsh), Sara, Sebastian, Tobias... I have also been very lucky to share many tea breaks / gossip sessions with some of you, the most regular participant being Anna Broccard, Laetitia Colombani, Carolin Kirsch... Though I am sure that no-one will be crossed if the honour place in this list is reserved to the sweet, loving and amazing Claire Descombes, one of the best person I have had the chance to meet. Oh, and I don't forget dogtor Mina, never.

Après avoir remercié toutes ces personnes rencontrées pendant ma thèse, je me tourne vers les mots de ma langue natale pour m'adresser à celles et ceux qui étaient là bien avant que je n'envisage de poursuivre mes études aussi longtemps, ou qui ont eu la chance de moins m'entendre répéter mes talks que les personnes sus-mentionnées. Les copains de Faerix et de l'X, en particulier TLBA, Bora, Nasta et Etienne.

Les copains de Lyon, rencontrés récemment mais qui ont sû se faire une place dans mon coeur en étant des gens bons (même les végétariens): Loïs pour ses anecdotes et sa bienveillance, Lilian pour ses conseils en amour, en culture de champignons et ses supers cocktails, Alizée et sa douceur, Redouane toujours là surtout quand on a besoin de lui à 4h du matin, Pablito dont la fulgurance aux mots croisés n'a d'égale que sa propension à lancer des punchlines, et surtout Camille en qui j'ai découvert une amie, partenaire de danse et de potins!

Les copains de prépa: Cécile, ma voisine de classe en 2013 qui est restée toute proche dans mon coeur, Alexandre (lequel ?) pour ses conseils en macaron et les puzzle banane, Froux, qui n'a jamais démissionné de son rôle d'ami, Arthur, qui a toujours donné à mes cookies le respect qui leur est dû et sait faire preuve d'une constance sans égale, Nicolas, qui même au bout du monde reste le meilleur des coachs, Sbeum!

Il ne faut pas oublier dans cette liste les amis qui sont bien plus que tout ça: Alexandre, les bébés, et Emma. Vous avez ~~toujours~~ longtemps (au moins 10 ans!) été là, et nos différences nous font grandir. J'espère partager de nombreuses autres années à vos côtés...
Tout comme j'espère le faire avec celui que j'aime tendrement: Clément.
Je suis chanceuse de rencontrer de si belles personnes, au côté de qui je me sens plus forte, plus confiante.

Et enfin, puisque je parle de belles personnes et de support indéfectible, c'est au tour de ma famille de clore cette longue et heureuse liste. Dans nos points communs comme dans nos différences, on s'aime, on se soutient, et je n'ai jamais douté un instant qu'ils étaient là pour moi. Que vous soyez cousin, cousine, oncle, tante, et plus je pense à vous tendrement.

Mais mes mots, mes pensées et mon amour vont tout particulièrement à ma Mamé, que j'adore. A ma grande soeur qui a parfois été comme une seconde maman, Marie-Alexandrine. Ses deux petits lutins, qui apportent chaque minute de la joie et de la tendresse dans ma vie, Waïl et Ambre. Mon grand frère, Sergeï, mon roc dans la tempête. Ma grande soeur Myrtille, l'une de mes plus fervente supporters. Mon papa, Régis, ma maman, Lydie... Je vous aime.

Contents

List of Figures, Algorithms, and Tables

List of Figures

List of Algorithms

List of Tables

Abstract

Capturing the dependence between a random response and predictors is a fundamental task in statistics and stochastic modelling. The focus of this work is on density regression, which entails estimating response distributions given predictor values. It enables the derivation of various statistical quantities, including the conditional mean, threshold exceedance probabilities, and quantiles.

This thesis presents a flexible approach, based upon the class of so-called Spatial Logistic Gaussian Processes (SLGPs). The SLGP framework utilizes a well-behaved latent Gaussian Process that undergoes a non-linear transformation, resulting in a class of models suitable for capturing spatially-dependent probability measures. SLGP models overcome limitations associated with strong distributional assumptions (e.g. shapes constraints, log-concavity, Gaussianity, etc.), varying sample sizes, and changes in target density shapes and modalities.

The first part of this work is dedicated to the development of SLGP models and gaining a deep understanding of the associated mathematical concepts. We introduce SLGPs from the perspective of random measures and their densities, and investigate links between properties of SLGPs and underlying processes. We show that SLGP models can be characterized by their log-increments and leverage this characterization to establish theoretical results with a main focus on spatial regularity.

We then focus on applicability of our approach, and propose an implementation relying on finite rank Gaussian Processes. We demonstrate it on synthetic examples and on temperature distributions at meteorological stations.

Finally, we address the potential of SLGPs for statistical inference, focusing on their potential in stochastic optimization and stochastic inverse problems. Notably, for inverse problems, an Approximate Bayesian Computation (ABC) framework is introduced, leveraging SLGP-surrogated likelihoods to accommodate situations with limited to moderate data. This methodology, inspired by GP-ABC methods, harnesses the probabilistic nature of SLGPs to guide data acquisition, thereby facilitating accelerated inference. We illustrate these approaches on synthetic examples as well as on a hydrogeological inverse problem in which a contaminant source is sought under uncertain geological scenario.

Chapter 1

Introduction

One of the central problems in statistics and stochastic modelling is to capture and encode the dependence of a random response on predictors in a flexible manner. Estimating some (conditional) response distributions given values of predictors $\mathbf{x} = (x_1, ..., x_d)$ is sometimes referred to as density regression and has received attention in many scientific application areas. However, this problem becomes particularly challenging when this dependence does not only concern the mean and/or the variance of the distribution, but also other features, for instance their shape, their uni-modal versus multi-modal nature, etc.

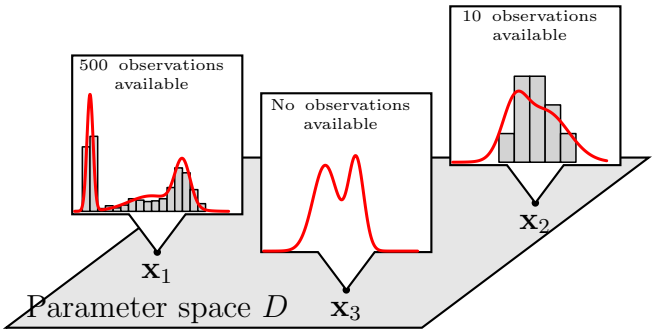


Figure 1.1: One challenging setting: Estimating the underlying probability distributions (red curve) by integrating data heterogeneously scattered across space (histogram).

The main focus of this thesis is to present a flexible and non-parametric approach for modelling spatially-dependent distributions. The approach is based on the class of distribution-valued fields called Spatial Logistic Gaussian Processes (SLGPs). The SLGP framework builds upon a *well-behaved* GP, which can be interpreted as a field of random functions, and considers the stochastic process obtained from applying *spatial logistic density* transformation to it,

hence mapping it to a field of positive random functions that integrate to one (i.e. a probability-density field).

SLGP models present a novel approach to density regression and provide a flexible solution to modelling the dependence of a random response on predictors. This non-parametric framework is designed to take advantage of the assumed spatial regularity and overcome the challenges associated with traditional methods, such as the need for strong distributional hypotheses, heterogeneous sample sizes, and changes in target density shapes and modalities.

The development of the SLGP models and a detailed understanding of the mathematical objects involved are the focus of the first part of this thesis. The second part of the thesis delves into the potential of SLGPs in statistical inference, particularly in the areas of Bayesian optimization and stochastic inverse problems.

In the case of stochastic inverse problems, the ABC framework is advanced through the use of the SLGP-surrogated likelihood. This leads to a methodology that enables accurate inference in low to moderate data regimes and can be guided by the probabilistic nature of SLGPs to drive data acquisition. The result is a powerful tool for accelerating scientifical discovery.

The remainder of this thesis is organized into four chapters, each building upon the previous one. The recommended reading order is as follows. Chapter 2 provides a background on the concepts and notions that will be used in the rest of the manuscript. Although it does not contain any original contributions, it is important to familiarize oneself with this chapter in order to fully grasp the subsequent chapters. Chapter 3 introduces the Spatial Logistic Gaussian Process (SLGP) model, and explains its construction and properties in detail. The implementation of the SLGP model is discussed in Chapter 4. Finally, in Chapter 5 the focus shifts to the application of the SLGP model in statistical inference in natural sciences. In conclusion, the four chapters in this thesis provide a comprehensive overview of the SLGP model, from its construction and implementation to its practical applications in statistical inference.

A summary of the contribution within each chapter is summarized thereafter and concludes this section.

- In Chapter 2, we introduce the essential elements of spatial statistics that are relevant to the developments in this work. Our introduction starts with definitions and properties pertaining to stochastic processes, paying particular attention to Gaussian Processes. This foundation leads us to delve into the subject of kernel methods and the related area of reproducing kernel Hilbert spaces. In a second part, we make connections to Gaussian Measure theory, which will prove useful later on as we are able to draw on the wealth of knowledge in this field to inform our own work. Finally,

we touch upon the basics of random measures, highlighting key elements from the construction outlined by Kallenberg (2017) that are applicable to this thesis.

- The central focus of Chapter 3 is to introduce and establish the proposed approach in this work. To begin, we first examine the current most common methods for estimating spatially dependent distributions, evaluating their effectiveness in addressing the challenge at hand. In this review process, we also delve into the literature on LGP models for density estimation, which served as inspiration for our work. We propose a comprehensive mathematical framework that builds upon the historical perspective of LGP models while incorporating a sound mathematical construction. After laying this foundation, we then introduce the main focus of this chapter: the Spatial Logistic Gaussian Process (SLGP). Here, we provide a detailed construction relying on random probability measure fields and discuss the mathematical characterisation of SLGPs. In the second part of the chapter, we turn our attention to the properties of SLGPs. One section focuses on revisiting notions of spatial regularity from spatial statistics and applying them to probability-distribution valued fields. The other section explores the concept of posterior consistency of the considered prior class. Throughout, we provide thorough explanations and mathematical proofs to support our claims.

- In Chapter 4, we address the practicalities of implementing SLGP models for the estimation of probability density fields. This includes a discussion of the mathematical properties of the likelihood and posterior distributions, as well as strategies for efficient estimation. Our work in this chapter is divided into three parts. Firstly, we propose models and formulations suitable for likelihood computations, with a focus on computational efficiency. Secondly, we explore implementation choices for Maximum A Posteriori (MAP) estimation and posterior inference via Markov Chain Monte Carlo (MCMC). Finally, we demonstrate the validity of our claims from the previous chapter, as well as the practical relevance of our approach through numerical illustrations. The data illustrations are particularly noteworthy as they allow us to reinforce the results presented in Chapter 3. For example, we use unconditional simulations to showcase the sharpness of the bounds derived from studying the spatial regularity of SLGPs. Additionally, we use analytical test cases to demonstrate the posterior consistency of our models in a controlled setting. Finally, we demonstrate the applicability of our models to higher dimensions using a 3D meteorological dataset.

- Finally, in Chapter 5, we explore the main motivation behind this thesis, which is to use statistical modelling to accelerate scientific discovery and guide experiments in the natural sciences. We examine the potential of both GP and SLGP modelling in Bayesian Optimization and stochastic inverse problems. For Bayesian Optimization, our focus is on defining criteria for exploring the parameter space and developing numerical methods for computing these criteria. In the context of stochastic inverse problems, we provide an overview of the framework and its key concepts, with a strong emphasis on Approximate Bayesian Computation (ABC). We review popular approaches for improving the numerical efficiency of ABC algorithms and introduce the SLGP-ABC framework, which builds upon the GP-ABC framework. Additionally, we present a suitable method for probabilistic forecasting to assess the quality of our approaches. To demonstrate the practicality of our methods, we present an illustration and benchmarking of both Bayesian Optimization and stochastic inversion on a hydrogeological application case.

Chapter 2

Background properties and methods

As discussed in the motivations, we are interested in statistical modelling of spatially dependent probability measures. This topic naturally orients us towards both the field of spatial statistics and that of random measures. We aim here at covering fundamental concepts and standard results that are deemed relevant for our work's self-containedness, with a focus on (Gaussian) Random Fields and Gaussian Measures in Sections 2.1, on kernels and native spaces thereof in 2.2; and on Random Measures in Section 2.3. The reader is expected to have a basic understanding of probability theory, and can refer to Billingsley (2008) for definitions. Indeed, for the sake of conciseness, we will primarily provide definitions and properties that are directly relevant to this work.

Also note that we introduce the concepts here in all generality. To avoid confusion with notations in Chapter 3 and after, we insist on the fact that, unless stated otherwise, our processes will be indexed by a generic set S.

2.1 Random Fields, Gaussian Random Fields and Gaussian Measures.

2.1.1 Stochastic processes.

We will begin by introducing basic concepts from spatial statistics in Subsection 2.1.1, and gradually build towards the topic of Gaussian Processes and Measures in Subsection 2.1.2.

General definitions in Stochastic Processes.

Definition 2.1.1 (Stochastic Process)**.** Let (B, Σ) be a measurable space. A B-valued stochastic process on a set S is a collection of B-valued random variables $(Z_s)_{s\in S}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$.

In the previous definition, the set S can be referred to as index set, whereas B is the state space. Whenever the index-set is a subset of $\mathbb{R}^d$, a stochastic process is also called Random Field (RF), a term that is commonly encountered in spatial statistics.

Remark. It is common to use the notation $(X_s)_{s\in S}$ for generic stochastic process. However, to make this chapter consistent with the following ones and to ease readability, we decided to borrow the notation $(Z_s)_{s\in S}$ from spatial statistics.

There are two ways to look at a stochastic process:

- For fixed $s \in S$, $\omega \in \Omega \mapsto Z_s(\omega)$ is a B-valued random variable.
- For fixed $\omega \in \Omega$, $s \in S \mapsto Z_s(\omega)$ is an element of B called a sample path or a realisation of Z.

While these two points of view are concomitant, it is important to note that the measurability of $\omega \in \Omega \mapsto Z_s(\omega)$ for any $s \in S$ is not sufficient to ensure that the whole sample path $s \in S \mapsto Z_s(\omega)$ is itself a measurable element for any $\omega \in \Omega$. This leads us to the following definition.

Definition 2.1.2 (Measurable Stochastic Process)**.** Let (B, Σ) and (S, Σ') be two measurable spaces. A B-valued stochastic process indexed by S, denoted $(Z_s)_{s\in S}$ is called measurable if all its sample paths are measurable with respect to the product σ-algebra $\Sigma' \otimes \mathcal{F}$.

Because of the duality between B-valued random processes and samples paths, defining equality between (non-necessarily measurable) stochastic processes is not straightforward, and there are several notions co-existing in the literature.

Definition 2.1.3 (Equality of stochastic processes)**.** Two stochastic processes $(Z_s^{(1)})_{s\in S}$ and $(Z_s^{(2)})_{s\in S}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$ are equal if:

$$Z_s^{(1)}(\omega) = Z_s^{(2)}(\omega) \quad \text{for all } (s, \omega) \in S \times \Omega$$

The latter is the strongest notion of equality for stochastic processes, but there exists relaxed versions of it.

Definition 2.1.4 (Indistinguishability of stochastic processes)**.** Two stochastic processes $(Z_s^{(1)})_{s\in S}$ and $(Z_s^{(2)})_{s\in S}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$ are indistinguishable if:

$$P\left[Z_s^{(1)} = Z_s^{(2)} \text{ for all } s \in S\right] = 1$$

Definition 2.1.5 (Stochastic equivalence of stochastic processes)**.** Two stochastic processes $(Z_s^{(1)})_{s\in S}$ and $(Z_s^{(2)})_{s\in S}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$ are stochastically equivalent if:

$$P\left[Z_s^{(1)} = Z_s^{(2)}\right] = 1 \text{ for all } s \in S$$

$Z^{(1)}$ and $Z^{(2)}$ are also called version (resp. modification) of one another, equal up to a version (resp. modification), or equal in distribution.

It is straightforward to deduce that equality implies indistinguishability, which in turns implies stochastic equivalence. The converse generally does not hold, except under some restricting assumptions, some of which will be stated later in this document.

As stated earlier, one of our focuses will be path properties of stochastic processes. However, not all path properties considered lead to measurable events, as pointed out in Scheuerer (2009):

Proposition 2.1.1. *For any $d \geq 1$ and an open subset $S \in \mathbb{R}^d$, let $\mathcal{C}(S) \subset \mathbb{R}^S$ denote the subset of all continuous functions $f : S \to \mathbb{R}$. Then, $S \notin \mathcal{B}(\mathbb{R})^S$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra of $\mathbb{R}$. In other word, the set of continuous functions on S is not measurable.*

In order to avoid this technical difficulty, it is advisable to restrict our interest to so-called separable random fields.

Definition 2.1.6 (Separable Random Field as defined in (Gihman and Skorohod, 1974))**.** A random field $(Z_s)_{s\in S}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and indexed by a topological set S is separable if there exists a countable dense subset $D \subset S$ and a set $N \in \mathcal{F}$ of probability 0 so that for any open set $I \subset S$ and any closed set $B \subset \mathbb{R}$, the two sets:

$$A_{B,I} = \{\omega : Z_s(\omega) \in B,\ \forall y \in I\}$$

$$A_{B,I\cap D} = \{\omega : Z_s(\omega) \in B,\ \forall y \in I \cap D\}$$

differ from each other only on a subset of N.

Due to the countability of D, $A_{B,I\cap D}$ is measurable even when $A_{B,I}$ is not. Separability means that the behaviour of the process is essentially determined by its values on a countable set. We will need this notion later to characterise the continuity of separable random fields.

Remark. Note that in all definitions but the last one, S was assumed to be either a generic set, or endowed with a σ-algebra Σ'). However, in the last definition we also required it to be equipped with a topology, so as to select a dense subset. In practice, we will often work with index sets that are subsets of $\mathbb{R}^d$ for some $d \geq 1$, and as such can be equipped with the topology and Borel σ-algebra inherited from the natural topology of the ambient Euclidean space.

Real-valued random fields

A class of interest is that of real-valued stochastic processes (where $\mathbb{R}$ is typically equipped with its Borel σ-algebra), as it allows for studying moments of the process.

Definition 2.1.7 (Moments of stochastic processes)**.** A real-valued stochastic process $(Z_s)_{y\in S}$ over the probability space $(\Omega, \mathcal{F}, P)$ is of first order if $Z_s \in L^1(\Omega, \mathcal{F}, P)$. Then, the mapping:

$$s \mapsto \mathbb{E}[Z_s] \quad \text{for all } s \in S$$

is called the mean function.

Similarly, Z is of second order if $Z_s \in L^2(\Omega, \mathcal{F}, P)$, and the mapping:

$$(s, s') \mapsto \mathrm{Cov}(Z_s, Z_{s'}) \text{ for all } s, s' \in S$$

is called the covariance function or covariance kernel on $S \times S$.

Despite being introduced in the context of covariance functions, studying kernels is not constrained to focusing on second order fields, and we shall explore it in an upcoming section of this chapter. However, let us first introduce a class of stochastic processes called Gaussian Processes (GPs).

2.1.2 Gaussian Processes.

Gaussian processes are a type of stochastic process characterised by their Gaussian distribution and are widely used in various areas of machine learning and statistics (Williams and Rasmussen, 2006).

General settings and important properties of Gaussian Processes

Definition 2.1.8 (Gaussian Process)**.** A real-valued stochastic process $(Z_s)_{s\in S}$ over the probability space $(\Omega, \mathcal{F}, P)$ is a Gaussian Process (GP) if its finite-dimensional distributions are Gaussian, meaning that for all $n \geq 1$ $s_1, ..., s_n \in S$, the random vector $(Z_{s_1}, \ldots, Z_{s_n})$ is multivariate-Gaussian distributed.

The distribution of a Gaussian process is fully characterised by its mean function and its covariance function. The measurability structure considered here being the cylindrical σ-algebra of $\mathbb{R}^S$.

The mean and covariance affect regularity properties of the associated GP, particularly its sample path continuity. While the impact of the mean's regularity can easily be ruled out by subtracting it, the kernel's influence is more indirect. Necessary conditions on the kernel are available in the literature, with one of the most well-known one being Dudley's theorem (Dudley, 1967).

Definition 2.1.9 (Canonical semi-metric)**.** For a generic set S and a GP $(Z_s)_{s\in S}$, the canonical semi-metric associated to Z is:

$$d_Z^2(s, s') = Var[Z_s - Z_{s'}] \quad \text{for all } (s, s') \in S^2. \tag{2.1}$$

Assuming that the covariance kernel of Z is denoted by k, we have:

$$d_Z^2(s, s') = k(s, s) + k(s', s') - 2k(s, s') \quad \text{for all } (s, s') \in S^2. \tag{2.2}$$

The canonical-semi metric allows one to study sample-paths regularity.

Theorem 2.1.2 (Dudley's integral)**.** *For a GP $Z \sim \mathcal{GP}(0, k)$ over a domain S and $\epsilon > 0$:*

$$\mathbb{E}\left[\|Z\|_\infty\right] \leq 24 \int_0^\infty \sqrt{\log(N(\epsilon, S, d_Z))}\, d\epsilon. \tag{2.3}$$

where for $\epsilon > 0$, $N(\epsilon, S, d_Z)$ is the entropy number, i.e. the minimal number of (open) d_Z-balls of radius ϵ required to cover S.

If the entropy integral on the right-hand side converges, then Z has a version with uniformly continuous paths on (S, d_Z) almost-surely.

Remark. We note that, for any metric d on S, if Z admits a version with almost all sample paths uniformly continuous on (S, d_Z) and if d_Z is continuous with respect to d then Z also admits a version with almost all sample paths uniformly continuous on (S, d).

Dudley's theorem provides us with a sufficient condition to ensure existence of an almost surely continuous version of a GP. However, whenever the process at hand is separable, we obtain a stronger result, as noted in the following remark.

Remark. As pointed out in Lemma 5.2.8. of Scheuerer (2009), if a (not necessarily Gaussian) process Z is separable and admits a version $\tilde{Z}$ which is continuous almost surely, then Z and $\tilde{Z}$ are indistinguishable. As a consequence, Z itself is continuous almost surely.

A consequence of Dudley's theorem is the following:

Proposition 2.1.3. *Let us consider a GP $Z \sim \mathcal{GP}(0, k)$ defined on S, a convex and compact subset of $\mathbb{R}^d$. Assume that the following Hölder condition holds:*

$$d_Z^2(s, s') \leq K\|s - s'\|_\infty^\beta \tag{2.4}$$

for some constant $K > 0$ and $0 > \beta$. Then Z admits a version with almost surely uniformly continuous sample paths.

The full proof and derivation of this classical result is available in Appendix A.2.

Remark. We considered a centred GP in the previous proposition, as the influence of the mean is easy to rule out. Indeed, for Z a centred GP and m a function, the non-centred GP defined by $m + Z$ is continuous almost surely if m is continuous and Z is continuous almost surely.

Dudley's theorem provides us with an upper bound for the expected sup-norm of a Gaussian Process Z. However, it does not provide us with any information about the behaviour of more complex quantities like the expected value of a non-linear function of the norm of Z (i.e. $\mathbb{E}\left[f(\|Z\|_\infty)\right]$). To better understand and control these types of quantities, we need to consider the framework and powerful results of Gaussian measure theory and its connections to Banach-valued Gaussian Processes.

Banach-valued Gaussian Processes, and Gaussian Measures

A class of Gaussian processes that is particularly interesting to study is that of Banach-valued GPs. The framework of Gaussian measures on Banach spaces provides a rich set of powerful tools and results for studying these processes. By connecting the theory of Banach-valued Gaussian processes with the theory of Gaussian measures on Banach spaces, we can gain a deeper understanding of these processes and their properties.

To begin, we will focus on introducing the context that leads to Fernique's theorem. This theorem is a fundamental result in the theory of Gaussian measures on Banach spaces and is widely used in the study of Banach-valued Gaussian processes. After that, we will ensure that we can establish an equivalence between Gaussian processes and Gaussian measures on Banach spaces.

Elements from Gaussian Measure theory
The results listed in this first half of the section were adapted from (Hairer, 2009).

Definition 2.1.10 (Gaussian measure on a Banach space)**.** A probability measure μ over a Banach space $\mathfrak{B}$ is Gaussian if and only if for all $\ell \in \mathfrak{B}^*$ (the topological dual of $\mathfrak{B}$, i.e. the space of continuous linear forms on $\mathfrak{B}$), the push-forward measure $\mu \circ \ell^{-1}$ (of μ through ℓ) is a Gaussian measure over $\mathbb{R}$.

The notions of mean function and covariance kernel established in Definition 2.1.7 are respectively mirrored by mean element and covariance operator. For simplicity, we restrict ourselves to the setting where the mean can be seen an element of the Banach space itself. We will discuss two suitable set of hypotheses for this construction to be sound shortly after the definition.

Definition 2.1.11. For a probability measure μ over a suitable Banach space $\mathfrak{B}$, the mean of μ is the unique element $m_\mu \in \mathfrak{B}$ such that:

$$\int_{\mathfrak{B}} \ell(f)\, d\mu(f) = \ell(m_\mu) \text{ for all } \ell \in \mathfrak{B}^* \tag{2.5}$$

The covariance operator is the bilinear operator $C_\mu : \mathfrak{B}^* \times \mathfrak{B}^* \to \mathbb{R}$ defined by:

$$C_\mu(\ell, \ell') = \int_{\mathfrak{B}} (\ell(f) - \ell(m_\mu))(\ell'(f) - \ell'(m_\mu))\, d\mu(f) \text{ for all } \ell, \ell' \in \mathfrak{B}^* \tag{2.6}$$

Remark. The mean of a Gaussian Measure is typically an element of $\mathfrak{B}^{**}$, not of $\mathfrak{B}$. However, it is easier to work with Banach space that are reflexive (i.e. $\mathfrak{B}^{**} = \mathfrak{B}$). Another hypothesis that allows one to relax this assumption is to consider separable Banach spaces. In this setting, the mean of a Gaussian measure is always an element of $\mathfrak{B}$ itself, as established by Bogachev (1998).

One result that is of importance for us is Fernique's theorem.

Theorem 2.1.4 (Fernique 1970)**.** *Let μ be any probability measure on a separable Banach space $\mathfrak{B}$ and $R : \mathfrak{B}^2 \to \mathfrak{B}^2$ be the rotation defined by:*

$$\forall f, f' \in \mathfrak{B}^2,\ R(f, f') = \left(\frac{f + f'}{\sqrt{2}}, \frac{f - f'}{\sqrt{2}} \right).$$

If μ satisfies the invariance condition:

$$R^*(\mu \otimes \mu) = \mu \otimes \mu.$$

Then, there exists $\alpha > 0$ such that $\int_{\mathfrak{B}} \exp(\alpha \|x\|^2)\, d\mu(x) < \infty$.

Remark. Here we used the notation R^* for the push-forward of the measure $\mu \otimes \mu$ through R.

This Theorem ensures the existence about one such α, but does not provide control over its value. Thankfully, Theorem 4.1 in Ledoux (1996) does so, by providing a sharp bound on acceptable values of α. This calls for an additional definition that we state here before enunciating the proposition of interest.

Definition 2.1.12 (Dual norm)**.** Let us consider a Banach space $\mathfrak{B}$ and its dual $\mathfrak{B}^*$. The dual norm of a linear functional $\ell \in \mathfrak{B}^*$ is defined as:

$$\|\ell\| := \sup_{f \in \mathfrak{B}, \|f\| \leq 1} |\ell(f)| \tag{2.7}$$

Proposition 2.1.5. *In Theorem 2.1.4, one can take any α such that*

$$0 < \alpha \leq \frac{1}{2\|C_\mu\|} \tag{2.8}$$

where

$$\|C_\mu\| := \sup \{|C_\mu(\ell, \ell')| \ \forall \ \ell, \ell' \in \mathfrak{B}^* \ s.t. \ \|\ell\| = \|\ell'\| = 1\} \tag{2.9}$$

The theorem is stated for any probability measure, in particular it holds for a Gaussian Measure, hence the following proposition also taken from (Hairer, 2009).

Proposition 2.1.6. *There exist universal constants α, $K > 0$ with the following properties. Let μ be a Gaussian measure on a separable Banach space $\mathfrak{B}$ and let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be any measurable function such that $f(x) \leq C_f \exp(\alpha x^2)$ for every $x \geq 0$. Then, with $M = \int_B \|x\| \mu(\, dx)$ denoting the first moment of μ, one has the bound $\int_{\mathfrak{B}} f(\|x\|/M) \mu(\, dx) \leq K C_f$*

It follows from it that a Gaussian measure admits moments (in a Bochner sense) of all orders.

Finally, another proposition will also prove handy to derive the almost sure continuity results and is stated thereafter.

Proposition 2.1.7. *Let $\mathfrak{B}$ be a separable Banach space, S be a compact and convex subset of $\mathbb{R}^d$, and let $(Z_s)_{s \in S}$ be a collection of $\mathfrak{B}$-valued Gaussian random variables such that*

$$\mathbb{E}[\|Z_{\boldsymbol{s}} - Z_{\boldsymbol{s}'}\|] \leq C \|\boldsymbol{s} - \boldsymbol{s}'\|^\alpha \ \forall \boldsymbol{s}, \boldsymbol{s}' \in S \tag{2.10}$$

for some $C > 0$ and some $\alpha \in (0, 1]$. Then, there exists a unique Gaussian measure μ on $\mathcal{C}(S, \mathfrak{B})$ such that any $(\tilde{Z}_{\boldsymbol{s}})_{\boldsymbol{s} \in S}$ having law μ is a version of Z. Furthermore, $\tilde{Z}$ is almost surely β-Hölder continuous for every $\beta < \alpha$.

Can we equivalently work with Process and Measure ? In order to make use of the powerful results from the theory of Gaussian measures, we want to ensure that a Gaussian process Z with sample paths in a given Banach Space induces a Gaussian measure on it. To simplify the analysis, we will consider that Z is indexed by a compact subset of $\mathbb{R}^d$, denoted by S. While results exist in more general settings, we will not be discussing them here.

Given that our focus is on studying the spatial regularity and sup-norm of the process, it is natural to consider two choices of $\mathfrak{B}$:

1. The space of continuous functions, equipped with the sup-norm
2. The space of bounded functions, equipped with the sup-norm

In order to determine whether they are suitable for further studies, we will review the properties of the two aforementioned spaces. Fernique's theorem, stated in Theorem 2.1.4, requires the Banach space to be separable, however it is not always the case of the latter.

1. The first functional space considered is separable. This is because it is possible to approximate any continuous function on a compact subset of $\mathbb{R}^d$ using d-variate polynomials with rational coefficients.
2. The second functional space is generally not separable. This is because, for an uncountable set S, the family of indicator functions $(\delta_s(x))_{s\in S}$, which indicate whether $x = s$, are bounded and form an uncountable set. The metric open balls of radius $1/2$ around each function are pairwise disjoint, and constructing a dense subset would require picking at least one element in each of these balls. As a result, there is no countable dense subset of the space of bounded functions if S is uncountable.

Due to its better properties, we shall focus our efforts on the space of continuous functions on S.

From now on, we will consider $(\mathfrak{B}, \|\cdot\|_\infty)$ to be the Banach space of continuous functions, equipped with the sup-norm. It is known that the Gaussian process and Gaussian measure perspectives are equivalent in this setting, as established in 1D by Rajput and Cambanis (1972), and extended to higher dimensions in Travelletti and Ginsbourger (2022). Furthermore, the relationships between the mean and covariance functions of a process can be easily established as special cases of the mean element and covariance operator of the corresponding measure.

Lemma 1. *The evaluation functionals defined for any $s \in S$ by $e_s : f \in \mathfrak{B} \mapsto f(s)$ are in $\mathfrak{B}^*$.*

Proof of Lemma 1. They are indeed linear, as for any $f, f' \in \mathfrak{B}$ and $\lambda \in \mathbb{R}$, $f + \lambda f' \in \mathfrak{B}$ and $\mathrm{e}_s(f + \lambda f') = f(y) + \lambda f'(y) = \mathrm{e}_s(f) + \lambda \mathrm{e}_s(f')$.

Moreover, for any $s \in S$, $|\mathrm{e}_s(f) - \mathrm{e}_s(f)| = |f(s) - f'(s)| \leq \|f - f'\|_\infty$ by definition, hence ensuring continuity of the functionals. We deduce from this that $\mathrm{e}_s \in \mathfrak{B}^*$ for all $s \in S$. □

Proposition 2.1.8 (Relation between the covariance operator C_μ and the covariance kernel k)**.** *Consider a $\mathfrak{B}$-valued GP $Z \sim \mathcal{GP}(m, k)$, and μ the corresponding Gaussian Measure. Then:*

$$m(s) = m_\mu(e_s) \text{ for any } s \in S \tag{2.11}$$

$$k(s, s') = C_\mu(e_s, e_{s'}) \text{ for any } s, s' \in S \tag{2.12}$$

Where e_s are the evaluation functionals introduced in Lemma 1.

It follows that with the considered Banach space, not only Fernique's theorem can be applied as stochastic process and measure point of view can be interchanged, but that we can also reformulate the bound in Proposition 2.1.5 in terms of the kernel k.

Corollary 2.1.9. *Consider a $\mathfrak{B}$-valued GP $Z \sim \mathcal{GP}(m, k)$, and μ the corresponding Gaussian Measure. Then:*

$$\|C_\mu\| \geq \sup_{s \in S} k(s, s) \tag{2.13}$$

and it follows that α in Proposition 2.1.5 must satisfy:

$$0 < \alpha \leq \frac{1}{2\|C_\mu\|} \leq \frac{1}{2 \sup_{s \in S} k(s, s)} \tag{2.14}$$

Proof of Corollary 2.1.9. We already proved in Lemma 1 that the evaluation functionals are in the continuous dual. We also note that for any $s \in S$:

$$\|\mathrm{e}_s\| := \sup_{\substack{f \in \mathfrak{B} \\ \|f\|_\infty = 1}} |\mathrm{e}_s(f)| = \sup_{\substack{f \in \mathfrak{B} \\ \|f\|_\infty = 1}} |f(s)| = 1 \tag{2.15}$$

the required result then follows from Equation 2.12 and $\|C_\mu\|$'s definition. □

2.2 On kernels, Reproducing Kernel Hilbert Spaces and Gaussian Processes

In Definition 2.1.7, we informally presented covariance kernels as being bivariate functions interpretable as covariance functions of second order random fields.

Although it is an omnipresent aspect of kernels (Stein, 1999), it does not do justice to Kernel methods (Aronszajn, 1950; Kimeldorf and Wahba, 1970; Schölkopf and Smola, 2018; Saitoh and Sawano, 2016) which is a versatile framework allowing for probabilistic prediction (Williams and Rasmussen, 2006), classification (Steinwart and Christmann, 2008) and function approximation based on scattered data (Wendland, 2004). In this section, we explore other aspects of kernel methods that will be relevant later in this thesis.

2.2.1 Kernels

On general kernels: definitions and important properties

First, let us define essential properties for kernels, namely the positive definiteness and conditional positive definiteness of functions.

Definition 2.2.1 (Positive definite function)**.** For a generic space S, we call positive definite (or p.d.) on S any function $k : S \times S$ such that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(s_i, s_j) \geq 0 \tag{2.16}$$

for any $n \geq 1$, $s_1, ...s_n \in S$ and any $a_1, ..., a_n \in \mathbb{R}$.

Definition 2.2.2 (Conditionally positive definite function)**.** For a generic space S, we call conditionally positive definite (or c.p.d.) on S any function $k : S \times S$ such that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(s_i, s_j) \geq 0 \tag{2.17}$$

for any $n \geq 1$, $s_1, ...s_n \in S$ and any $a_1, ..., a_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} a_i = 0$.

This allows us to define kernels. Some authors consider conditionally positive kernels, or negative kernels (where the sum in (2) of Definition 2.2.1 has to be non-negative). In this work, unless explicitly stated otherwise, we will consider that a kernel has to be p.d., as stated in the following definition.

Definition 2.2.3 (Kernels)**.** For a generic space S, we call kernel on S any function $k : S \times S$ such that:

1. k is symmetric, i.e. $k(s, s') = k(s', s)$ for all $s, s' \in S$.
2. k is a p.d. function.

One easily checks that the covariance functions of second order fields are indeed kernels as we just defined.

Definition 2.2.4 (Stationarity)**.** A function $k : S^2 \to \mathbb{R}$ is called stationary if there exists a function $k_0 : S \to \mathbb{R}$ such that

$$\forall (s, s') \in S^2,\ k(s, s') = k_0(s - s')$$

Usually, the abuse of notation $k(s - s')$ is used to refer to a stationary function.

It is interesting to study stationary kernels, as we have a deeper understanding of their structure than of that of general kernels. In particular, Bochner's theorem (Bochner, 1933) gives more insight on the Fourier transform of stationary kernels. We focus thereafter on kernels defined on $\mathbb{R}^d$ (and subspaces thereof) and introduce a bold notation to emphasize the fact that inputs are typically expected to be vectors rather than scalars.

Theorem 2.2.1 (Bochner)**.** *A continuous stationary function $k(\boldsymbol{s}, \boldsymbol{s}') = k_0(\boldsymbol{x} - \boldsymbol{x}'),\ (\boldsymbol{s}, \boldsymbol{s}' \in \mathbb{R}^d)$ is positive definite if and only if k_0 is the Fourier transform of a finite positive measure η:*

$$k_0(\boldsymbol{y}) = \frac{1}{2\pi} \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^\top \boldsymbol{y}}\, d\eta(\boldsymbol{\omega})\ (\boldsymbol{y} \in \mathbb{R}^d) \tag{2.18}$$

If η has a density with respect to Lebesgue measure, it is called the spectral density and will be denoted thereafter by $\mathfrak{s}(\boldsymbol{\omega})$.

The so-called Fourier duality of spectral densities and covariance functions derives from it, and is also known as the Wiener-Khinchin theorem (Khinchin, 1934).

$$k(\mathbf{s} - \mathbf{s}') = \frac{1}{2\pi} \int_{R^d} \mathfrak{s}(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^\top (\mathbf{s} - \mathbf{s}')}\, d\boldsymbol{\omega} \tag{2.19}$$

$$\mathfrak{s}(\boldsymbol{\omega}) = \int_{R^d} k(\boldsymbol{y}) e^{-i\boldsymbol{\omega}^\top \boldsymbol{y}}\, d\boldsymbol{y} \tag{2.20}$$

Introducing this more formal definition of kernels enables us to discuss a family of kernel methods that will be of particular interest to us thereafter.

2.2.2 Reproducing Kernel Hilbert Spaces

Kernels are deeply linked to a class of Hilbert spaces called Reproducing Kernel Hilbert Spaces that we define now.

Definition 2.2.5 (Reproducing kernel Hilbert space: RKHS)**.** A (real) Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ of functions from some set S to $\mathbb{R}$ is said to be a reproducing kernel Hilbert space if there exists a function $k : S \times S \to \mathbb{R}$ such that:

1. $k(s, \cdot) \in \mathcal{H}$ for all $s \in S$
2. $f(s) = \langle f, k(s, \cdot) \rangle$ for all $f \in \mathcal{H}$ and all $s \in S$

The aforementioned function k is called reproducing kernel, and the property 2. is called reproducing property. Such a k is indeed a kernel as in Definition 2.2.3.

The relationship between RKHS and kernel is straightforward, indeed:

Theorem 2.2.2 (Moore-Aronszajn theorem (Aronszajn, 1950))**.** *Let k be a kernel on a generic set S, then there corresponds one and only one class of functions on S forming a Hilbert space $\mathcal{H}$ and admitting k as a reproducing kernel.*

Consider a continuous kernel k defined on a compact metric space S. It is possible to write expansions of k in suitable families of basis function. One such expansion relies on the spectral decomposition of the operator $T_k : f \in L^2(S) \mapsto \int_S f(u)k(u, \cdot)\, du \in L^2(S)$, where the integral is taken with respect to Lebesgue measure. By classical spectral theory, it is known that this self-adjoint compact operator possesses non-negative eigenvalues $\lambda_1, \lambda_2, ...$ and respectively associated normalised continuous eigenfunctions $\varphi_1, \varphi_2, ...$ forming an orthonormal basis of $L^2(S)$.

Theorem 2.2.3 (Mercer theorem)**.** *For a continuous kernel k defined on a compact metric space S, and $(\lambda_j)_{j\geq 1}$, $(\varphi_j)_{j\geq 1}$ as above, we have:*

$$k(s, s') = \sum_{j\geq 1} \lambda_j \varphi_j(s)\varphi_j(s') \quad \text{for all } s, s' \in S \tag{2.21}$$

where the series converges absolutely and uniformly on $S \times S$.

Remark. It is most common to work with continuous kernels on compact metric spaces, but the reader can refer to Steinwart and Scovel (2012) for Mercer's theorem on more general domains.

By utilizing the framework of kernel methods, it is possible to map suitable probability measures into a reproducing kernel Hilbert space, as demonstrated by Berlinet and Thomas-Agnan (2004); Smola et al. (2007); Sriperumbudur et al. (2010); Muandet et al. (2017).

Definition 2.2.6 (Kernel mean embedding)**.** Let k be a kernel on a generic space S and P be a probability measure on S such that $\mathbb{E}_{X\sim P}[k(X, X)] < \infty$ The kernel mean embedding of P into k's RKHS $\mathcal{H}$ is defined by the mapping:

$$P \mapsto \mu_P := \int_S k(u, \cdot)\, dP(u) \tag{2.22}$$

Then, $\mu_P \in \mathcal{H}$ and $\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mu_P \rangle$.

Remark. Whenever the considered kernel k is bounded, the kernel mean embedding is defined for any probability measure as $\mathbb{E}_{X\sim P}\left[k(X,X)\right] \leq \sup_{s\in S} k(s,s)$ for any P.

Note that in practice we rarely have access to the probability distribution P but rather to realisations of i.i.d. random variables, that we denote $(Y_1, ..., Y_n)$. The most common estimator of the kernel mean embedding is simply the Monte-Carlo one:

$$\hat{\mu}_P = \frac{1}{n}\sum_{i=y}^{n} k(Y_i, \cdot) \tag{2.23}$$

The framework of kernel mean embeddings can be used to define pseudo-metrics on probability measures, which can be applied in various ways in machine learning and statistics.

Definition 2.2.7 (Maximum Mean Discrepancy)**.** Consider a kernel k on S and its RKHS H. For two probability measures on S denoted P and Q that admit a kernel mean embedding, the Maximum Mean Discrepancy (MMD) between P and Q is defined by:

$$\begin{aligned}
\mathrm{MMD}(P,Q) &:= \sup_{\substack{f\in H\\ \|f\|\leq 1}} \left\{ \int_S f(u)\, dP(u) - \int_S f(v)\, dQ(v) \right\} && (2.24)\\
&= \sup_{\substack{f\in H\\ \|f\|\leq 1}} \left\{ \langle f, \mu_P - \mu_Q \rangle \right\} && (2.25)\\
&= \|\mu_P - \mu_Q\| && (2.26)
\end{aligned}$$

In light of the linearity of the operator involved and of the reproducing property, one can also formulate the MMD to emphasize its dependency on the kernel k:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{X,X'}\left[k(X,X')\right] + \mathbb{E}_{Y,Y'}\left[k(Y,Y')\right] - 2\mathbb{E}_{X,Y}\left[k(X,Y)\right] \tag{2.27}$$

where $X, X' \sim P$ and $Y, Y' \sim Q$ are independent.

Whenever the kernel mean embedding is an injective map, the kernel k is called characteristic, This property ensures that $\mathrm{MMD}(P,Q) = 0$ is equivalent to $P = Q$. This is significant as it guarantees that MMD can effectively differentiate between probability measures. It is crucial in various applications such kernel two-sample tests (Gretton et al., 2012), independence test (Gretton et al., 2005) or learning on distributional data (Sutherland, 2016; Szabó et al., 2016).

It is interesting to note that the MMD being a measure of dissimilarity can be expressed in the kernel terminology:

Proposition 2.2.4 (MMD as a Kernel)**.** *The function defined for any two probability measures (P, Q) by $-MMD^2(P, Q)$ is a conditionally positive kernel on the space of probability measures on S.*

Proof. For $n \geq 1$, consider $P_1, ..., P_n$ n probability measures on S and $a_1, ..., a_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} a_i = 0$.

$$\star := -\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{MMD}^2(P_i, P_j) \tag{2.28}$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left\|\mu_{P_i} - \mu_{P_j}\right\|^2 \tag{2.29}$$

$$= -2\sum_{i=1}^{n} a_i \left\|\mu_{P_i}\right\|^2 \underbrace{\sum_{j=1}^{n} a_j}_{=0} + 2\left\|\sum_{i=1}^{n} a_i \mu_{P_i}\right\|^2 \geq 0 \tag{2.30}$$

The symmetry of MMD^2 being an immediate consequence of the definition, we conclude that it is indeed a conditionally positive kernel. □

This property will be of significance later on, as it enables the comparison of probability measures, which is a key focus of this thesis. However, before delving into that, let us examine other aspects of RKHS in more detail.

2.2.3 Gaussian Processes in relation to kernels and RKHS

The small ball property of Gaussian Processes

There exist strong connections between Gaussian Processes and the RKHS that are induced by their covariance kernels. One such connection is known as the small-ball property of Gaussian Processes.

Proposition 2.2.5 (Small ball probabilities for a Gaussian process)**.** *If $Z \sim GP(0, k)$, is such that $\|Z\|_\infty < \infty$ a.s, then for all $\epsilon > 0$:*

$$P\left[\|Z\|_\infty < \varepsilon\right] > 0 \tag{2.31}$$

Moreover if f is an element of the reproducing kernel Hilbert space of k, then for all $\epsilon > 0$:

$$P\left[\|Z - f\|_\infty < \varepsilon\right] > 0 \tag{2.32}$$

The proof, as well as rates are available in van der Vaart and van Zanten (2008). This result allows one to relate properties of the process to the RKHS spanned by its kernel.

Transferring kernel expansions to Gaussian Processes

We also leverage connections between Gaussian Processes and the RKHS of their kernel to derive expansions of the processes. We review two of them, the first one gives raise to exact expansions of kernels and processes, whereas the other one yields a stochastic approximation of them.

Mercer theorem and Karhunen-Loève expansion Among other, Bochner's theorem allows for approximating kernels (and subsequent Gaussian Processes), which is useful for accelerating kernel methods and reducing dimensionality of the problems at hand. In particular, the Karhunen-Loève expansion (Karhunen, 1946, 1947; Loeve, 1948) transfers the deterministic expansion provided by Mercer's theorem 2.2.3 to stochastic processes.

Proposition 2.2.6 (Karhunen-Loève expansion)**.** *Consider a kernel k on a generic set S satisfying the assumptions of Theorem 2.2.3. For a mean function m on S, and $Z \sim \mathcal{GP}(m, k)$, we have:*

$$Z_s = m(s) + \sum_{j \geq 1} \sqrt{\lambda_j} \varphi_j(s) \varepsilon_j \text{ for all } s \in S \tag{2.33}$$

where $(\varepsilon_j)_{j\geq 1}$ are i.i.d standard normal and $(\lambda_j)_{j\geq 1}$, $(\varphi_j)_{j\geq 1}$ are as described in Theorem 2.2.3. The convergence is in L^2 and uniform in s.

Bochner theorem and Fourier Features Another key framework that will be considered in this thesis is that of Random Fourier Features (RFF). Random Fourier Features is a method for approximating stationary kernels. The essential element of the approach of Random Fourier Features (Rahimi and Recht, 2008, 2009) is the realisation that the Wiener-Khinchin integral (Equation 2.19) can be approximated by a Monte Carlo sum. Indeed, recall that the Wiener-Khinchin integral is given by:

$$k(\mathbf{s} - \mathbf{s}') = \frac{1}{2\pi} \int_{R^d} \boldsymbol{s}(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')} \, d\boldsymbol{\omega}$$

Then, taking $\sigma = k(0)$, and ω_i's to be draws of independent random variables that have a density equal to the spectral density associated to k, we have the Monte Carlo approximation:

$$k(\mathbf{s} - \mathbf{s}') \approx k_{RFF}(\mathbf{s}, \mathbf{s}') := \frac{\sigma}{p} \sum_{i=1}^{p} \cos(\omega_i^\top (\mathbf{s} - \mathbf{s}')) \tag{2.34}$$

A direct consequence of this approximation is that the approximate kernel has a finite basis expansions as:

$$\phi(s) := \left[\cos(\omega_1^\top s), ..., \cos(\omega_p^\top s), \sin(\omega_1^\top s), ..., \sin(\omega_p^\top s)\right] \tag{2.35}$$

It is often possible to sample from the spectral densities of the most common kernels. The Gaussian Process defined by setting:

$$Z_{RFF,s} = \frac{\sigma}{\sqrt{p}}\phi(s)^\top \boldsymbol{\varepsilon} \tag{2.36}$$

where $\boldsymbol{\varepsilon}$ is a $2p$-variate standard normal vector, is a Gaussian Process with mean zero and covariance kernel $k_{RFF}(s, s')$.

2.2.4 The Matérn family of kernels

The Matérn family of kernels is a popular family of stationary covariance kernels (Matérn, 1960; Stein, 1999), defined (in the isotropic setting) for any $d \in \mathbb{N}$ by:

$$k_\nu(s, s') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|s - s'\|_2}{\ell}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{\|s - s'\|_2}{\ell}\right) \ (s, s' \in \mathbb{R}^d)$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind. The parameter $\sigma^2 > 0$ denotes the variance of the kernel, ν is a positive parameter and ℓ is the lengthscale. We use $\|\cdot\|_2$ to denote the Euclidean norm.

These kernels admit a simpler expression for some values of ν, namely the ones that write $p + 1/2$, $p \in \mathbb{N}$:

$$k_{\nu=p+1/2}(s, s') = \exp\left(-\frac{\sqrt{2\nu}\|s - s'\|_2}{\ell}\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}\|s - s'\|_2}{\ell}\right)^{p-i}$$

The most commonly encountered examples are given below:

- for $\nu = 1/2$, the exponential covariance $k_{1/2}(s, s') = \sigma \exp\left(-\frac{\|s-s'\|_2}{\ell}\right)$,
- for $\nu = 3/2$, $k_{3/2}(s, s') = \sigma\left(1 + \frac{\sqrt{3}\|s-s'\|_2}{\ell}\right)\exp\left(-\frac{\sqrt{3}\|s-s'\|_2}{\ell}\right)$
- for $\nu = 5/2$, $k_{5/2}(s, s') = \sigma\left(1 + \frac{\sqrt{5}\|y-y'\|_2}{\ell} + \frac{5\|s-s'\|_2^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}\|s-s'\|_2}{\ell}\right)$

The family is well studied, and in particular we know that the RKHS of a Matérn kernel k_ν is norm-equivalent to the Sobolev space of order $\nu + d/2$. The reader can refer to Kanagawa et al. (2018) for proofs of the latter, and other properties.

Remark. The spectral density of a Matérn kernel is known, and given by:

$$\mathfrak{s}_\nu(\boldsymbol{w}) = \frac{\Gamma(\nu + \frac{d}{2})}{\Gamma(\nu)\prod_{i=1}^{D} \ell^2 (2\nu)^{D/2}} \left(1 + 4\pi^2 \|\boldsymbol{w}\|_2^2\right)^{-\left(\nu+\frac{d}{2}\right)}.$$

One recognizes here a multivariate student distribution with 2ν degrees of freedom, location 0 and scale matrix identity. Knowing the spectral density will be instrumental for a RFF-type approach.

Additionally, one can prove that they are characteristic kernels, no matter their parameters. Indeed, it was shown in Theorem 7 of Sriperumbudur et al. (2008) that bounded continuous translation-invariant kernels on $\mathbb{R}^d$ are characteristic if and only if their Fourier transform has support on all of $\mathbb{R}^d$. Hence, Matérn kernels being characteristic is a straightforward consequence of this result, and will come in handy later in Section 5.2.3.

2.3 A detour through Random Measures

We briefly state here some basic properties and definitions of (locally finite) random measures. We rely on the definitions from (Kallenberg, 2017). In the terminology of (Kallenberg, 2017), our *sample space* of interest is here a compact space $\mathcal{I} \subset \mathbb{R}^{d_t}$, equipped with the Euclidean metric (and hence Polish by the compactness assumption) and the corresponding Borel σ-algebra is $\mathcal{B}(\mathcal{I})$.

Definition 2.3.1 (Considered sigma-algebra on probability measures on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$)**.** We denote $\mathfrak{M}$ the collection of all probability measures on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$, and take the σ-algebra $\mathcal{M}$ on $\mathfrak{M}$ to be the smallest σ-algebra that makes all maps $M \mapsto M(B)$ from $\mathfrak{M}$ to $\mathbb{R}$ measurable for $B \in \mathcal{B}(\mathcal{I})$.

Definition 2.3.2 (Random Measures)**.** A random measure Ξ is a random element from $(\Omega, \mathcal{F}, P)$ to $(\mathfrak{M}, \mathcal{M})$ such that for any $\omega \in \Omega \setminus N$, where N is a P-null set, we have:

$$\Xi\left(\tilde{B}; \omega\right) < \infty \text{ for all (bounded) measurable sets } \tilde{B} \in \mathcal{B}(\mathcal{I}) \tag{2.37}$$

Note that here the term *bounded* is between parentheses as $\mathcal{I}$ is assumed compact and so all elements of $\mathcal{B}(\mathcal{I})$ are bounded. Among the motivations listed for the choice of this structure, we retain that the σ-field $\mathcal{M}$ is identical to the Borel σ-field for the weak topology of $\mathfrak{M}$. This structure ensures that the random elements Ξ considered are regular conditional distributions on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$:

1. For any $\omega \in \Omega$, the mapping $B \mapsto \Xi(B; \omega)$ is a measure on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$.

2. For any $B \in \mathcal{B}(\mathcal{I})$, $\omega \mapsto \Xi(B;\omega)$ is $(\Omega, \mathcal{F})$-$(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ measurable.

Another strong advantage of this choice of measurability structure lies in the fact that if the state space $\mathcal{I}$ is equipped with a localizing structure, then the class of Probability Measures is a measurable subspace. The localizing structure arises naturally in our setting, as we can simply equip $\mathcal{I}$ with the class of all (bounded) Borel sets.

Finally, we also make a simple remark:

Remark (RMs seen as scalar random fields). We can see a RM Ξ as a particular instance of a scalar-valued random field indexed by $\mathcal{B}(\mathcal{I})$, namely $(\Xi(B))_{B\in\mathcal{B}(\mathcal{I})}$. Therefore, it is natural to revisit the notions of equality in distribution and of indistinguishability for RMs. In particular, we will call two random measures Ξ and $\tilde{\Xi}$ indistinguishable from one another if and only if:

$$P\left[\Xi(B) = \tilde{\Xi}(B),\ \forall B \in \mathcal{B}(\mathcal{I})\right] = 1 \tag{2.38}$$

Chapter 3

A new framework for probability density field modelling: From Logistic to Spatial Logistic Gaussian Process models.

This chapter is based on elements presented in the preprint Gautier and Ginsbourger (2021).

3.1 Modelling (conditional) probability distributions

As already mentioned in the introduction, 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 $T_{\mathbf{x}}$. 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.

However, the problem of learning $T_{\mathbf{x}}$'s distribution becomes particularly challenging when this dependence does not only concern the mean and/or the variance of the distribution, but other features can evolve, including for instance their shape, their uni-modal versus multi-modal nature, etc.

In addition, when dealing with real-world applications, it is often the case that the data is heterogeneously scattered across space, or that there are no replicates in the dataset, making the problem even more difficult. A schematic

example of a typical application can be seen in Figure 1.1 of the introduction.

Given these challenges, it is important to find a flexible model that can accurately capture the dependencies and changes in the data, while also being able to handle the complexities that arise in real-world applications and datasets. Furthermore, it would be ideal for the chosen model to not only provide predictions of the response distribution, but also an uncertainty estimate at each predictor value.

From now-on and throughout the document, we consider a compact and convex response space $\mathcal{I} \subset \mathbb{R}^{d_t}$ with $d_t \geq 1$ and $D \subset \mathbb{R}^{d_x}$ a compact and convex index space with $d_x \geq 1$. We will call probability field estimation problem the task of learning the densities $(p_{\mathbf{x}})_{\mathbf{x}\in D}$ of a collection of $\mathcal{I}$-valued random variables from observations of i.i.d. samples at various locations. To improve readability, we will always denote by $\mathbf{x} \in D$ the predictors (also referred to as index-variables), and $t \in \mathcal{I}$ (resp. T in random form) the responses of interest.

3.1.1 Literature review

We review here the most notable approaches typically used in a frequentist framework to address the challenge of (spatially dependent) density estimation, with a brief summary of the main idea behind each approach. Note that some of these approaches slightly differ from our setting, as they aim at modelling conditional density, hence assuming that the predictors $\mathbf{x}$ are themselves realisations of a random variable.

Finite Mixture models This approach presented in Rojas et al. (2005) consists in assuming that the conditional density can be written:

$$f_{T|\mathbf{X}=\mathbf{x}}(t) = \sum_{i=1}^{K} w_i(\mathbf{x}) g_i(y; \boldsymbol{\theta}_i(\mathbf{x})) \tag{3.1}$$

where the $g_i(y; \boldsymbol{\theta}_i(x))$ are densities with a set of parameters $\theta_i(\mathbf{x})$ that depend on $\mathbf{x}$ and the w_i's are a set of mixing proportions that sum to one for each $\mathbf{x}$. The component densities $g_i(y; \boldsymbol{\theta}_i(x))$ are typically assumed to be from a known parametric family, such as the Gaussian or Poisson distribution. The most popular algorithm to estimate the parameters is the Expectation-Maximisation algorithm (Dempster et al., 1977), which gives the maximum-likelihood estimator.

Kernel density estimation Kernel methods are typically applied in settings where $\mathbf{x}$ is also considered as random (and will hence be denoted $\mathbf{X}$ thereafter).

Such methods can be used to estimate the joint density of $\mathbf{X}$ and T (Fan et al., 1996; Hall et al., 1999). The basic idea being that the joint density at a point $(\mathbf{x}, t)$ is estimated by averaging the density of its local neighbourhood, where the size of the neighbourhood is determined by a smoothing parameter, typically referred to as the bandwidth. The choice of kernel function and bandwidth are important for the performance of the method. The bandwidth determines the smoothness of the density estimate and the kernel function determines the shape of the density estimate. Estimating the bandwidth is commonly done with cross-validation (Fan and Yim, 2004), bootstrap (Hall et al., 1999) or other methods. Once the joint distribution is estimated, the conditional probabilities are a by-product obtained by marginalizing over $\mathbf{X}$.

Semi-parametric modelling with Generalized lambda distributions Generalized lambda distributions have recently been used in Zhu and Sudret (2020) for flexible semi-parametric modelling of unimodal distributions depending on covariables. Indeed, the generalized lambda family can yield good approximations of a wide class of unimodal distributions. Modelling the parameters as spatially dependant (e.g. as a polynomial of $\mathbf{x}$) allows for spatially dependant probability density estimation.

Distributional Kriging Another branch of study pertaining to geostatistics and that does not rely on Gaussian or specific distributional assumptions is the so-called distributional Kriging (Aitchison, 1982; Egozcue et al., 2006; Talská et al., 2018). However, such approaches are ill-suited in the case of moderate sample size heterogeneously scattered across space, as they rely on interpolating (partially) observed cumulative distribution functions using Aitchison geometry.

Shape restricted distributional regression Under the assumption of a strong relationship between covariates and a response variable, such as monotonicity or convexity, estimating conditional distributions is called Shape-constrained regression. The reader can refer for instance to the survey by Guntuboyina and Sen (2018). Due to the strong shape requirements, such problems yield a shape-constrained non-parametric estimator, which does not involve any tuning parameters.

Within a Bayesian context, it is natural to put a prior on probability density functions and derive posterior distributions of such probability density functions given observed data. We next list a couple of Bayesian approaches to spatial density estimation.

Infinite mixture models This approach differs from finite Mixture models by allowing the number of mixture components to be infinite. The Bayesian approach provides a natural way to incorporate uncertainty about the number components into the density estimation process. Its popularity is partly due to the wide literature on algorithms for posterior sampling within a Markov Chain Monte Carlo framework (Jain and Neal, 2004; Walker, 2007; Papaspiliopoulos and Roberts, 2008) or fast approximation (Minka, 2001).

Generalized stick breaking processes These Bayesian methods represent a probability density function (pdf) as a mixture of simpler components, where the mixture weights are generated via a sequence of random "stick breaking" events. (Dunson and Park, 2008; Dunson et al., 2007; Chung and Dunson, 2009; Griffin and Steel, 2006)

Multivariate transformation of a Beta distribution Multivariate transformations of a Beta distribution can be useful in settings where the dependencies between variables are important and need to be explicitly modelled. These approaches involve transforming a univariate Beta distribution into a multivariate distribution, which can better capture the dependencies between the variables (Trippa et al., 2011). Two popular examples of a multivariate transformation of a Beta distribution are the Dirichlet distribution (commonly used as a prior distribution in Bayesian models and can be used for density estimation in a multivariate setting) and copulae-based approaches.

Conditional density estimation with neural network models A neural network model can be used for conditional density estimation (Papamakarios et al., 2017; Rothfuss et al., 2019; Papamakarios, 2019). One advantage of using neural networks for conditional density estimation is their ability to handle non-linear relationships between variables. Neural networks can also model high-dimensional covariates, making them well-suited for complex data sets. Another advantage of using neural networks for conditional density estimation is their ability to handle multi-modality, where the density of the target variable has multiple modes. This is achieved by using flexible activation functions, such as the sigmoid function or the softplus function, which can capture multiple modes in the density of the target variable. However, neural network models for conditional density estimation can be computationally expensive and may require large amounts of data to train effectively. They can also be sensitive to the choice of network architecture and hyperparameters, which must be carefully selected to ensure good performance. Finally, they are rarely interpretable and performing UQ on it is not straightforward.

Transformed Gaussian Process (GP) GP provide a flexible and powerful tool for modelling continuous functions, and can be used to model the relationship between the covariates and the target variable. In order to use a GP for conditional density estimation, we first need to transform the GP into a form that can be used for density estimation. This is typically done by nonlinear transformation of the GP, (Jara and Hanson, 2011; Donner and Opper, 2018; Tokdar et al., 2010; Gautier et al., 2021). We will explore in depth one approach of this class of methods in the coming subsections.

All the aforementioned approaches come with their own specificities. However, to the extent of our understanding of the field, only a few of them address the challenges we are facing.

Indeed, most approaches listed here struggle with the low to moderate data regime, and even more so with heterogeneously scattered data. When combining this observation with our need for uncertainty quantification and flexibility in the model, we noted that GP-based methods were the most natural candidates for our purposes.

A particular case of transformed GPs: (S)LGP models The Spatial Logistic Gaussian Process (SLGP) model, related to Tokdar et al. (2010) and being at the centre of the present contribution following up on (Gautier et al., 2021; Gautier and Ginsbourger, 2021), is itself a spatial generalization of the Logistic Gaussian Process (LGP) model.

The LGP for density estimation was established and studied in (Lenk, 1988, 1991; Leonard, 1978) and is commonly introduced as a random probability density function obtained by applying a non-linear *transformation* (or "mapping") ψ to a *sufficiently well-behaved* GP $Z = (Z_t)_{t\in\mathcal{I}}$, resulting in

$$\psi[Z](t) = \frac{e^{Z_t}}{\int_{\mathcal{I}} e^{Z_u}\, d\lambda(u)} \text{ for all } t \in \mathcal{I} \tag{3.2}$$

Here and throughout the document, we consider a compact and convex response space $\mathcal{I} \subset \mathbb{R}^{d_t}$ with $d_t \geq 1$ and we denote by λ the Lebesgue measure on $\mathbb{R}^{d_t}$. We further assume that $\lambda(\mathcal{I}) > 0$.

For the Spatial Logistic Gaussian Process (SLGP), we will similarly build upon a *well-behaved* GP $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ (now indexed by a product set) and study the stochastic process obtained from applying the *spatial logistic density* transformation to Z as follows:

$$\Psi[Z](\mathbf{x},t) = \frac{e^{Z_{\mathbf{x},t}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \text{ for all } (\mathbf{x},t) \in D \times \mathcal{I} \tag{3.3}$$

where $D \subset \mathbb{R}^{d_x}$ is a compact and convex index space with $d_x \geq 1$.

At any fixed $\mathbf{x}$, $\Psi[Z](\mathbf{x}, \cdot)$ hence returns an LGP, so that an SLGP can be seen as a field of LGPs. What the mathematical objects involved precisely are (in terms of random measures or densities, and fields thereof) calls for some careful analysis. In this section, we will start by giving an historical perspective on the LGP models that inspired this work. We will then focus on our first two questions: questioning the stochastic nature of (S)LGPs and characterising their distributions.

Throughout the rest of this chapter, we will denote by $(\Omega, \mathcal{F}, P)$ the ambient probability space. We will also denote by $\mathcal{B}(\mathcal{I})$ the Borel σ-algebra induced by the Euclidean metric on $\mathcal{I}$. For a set S (here $\mathcal{I}$ or $D \times \mathcal{I}$), we denote by $\mathcal{C}^0(S)$, $\mathcal{A}(S)$, $\mathcal{A}^+(S)$ the sets of continuous real functions, Probability Density Functions (PDFs) and positive PDFs on S, respectively.

$$\mathcal{A}(S) := \{p : p \text{ is a pdf on } S\} \tag{3.4}$$
$$\mathcal{A}^+(S) := \{p : p \text{ is a positive pdf on } S\} \tag{3.5}$$

Finally, we denote by $\mathcal{A}_d(D; \mathcal{I})$ the set of fields of PDFs on $\mathcal{I}$ indexed by D, and by $\mathcal{A}_d^+(D; \mathcal{I})$ its counterpart featuring positive PDFs.

$$\mathcal{A}_d(D; \mathcal{I}) := \{(p(x, \cdot))_{x \in D} : p(x, \cdot) \text{ is a pdf on } \mathcal{I} \text{ for all } x \in D\} \tag{3.6}$$
$$\mathcal{A}_d^+(D; \mathcal{I}) := \{(p(x, \cdot))_{x \in D} : p(x, \cdot) \text{ is a positive pdf on } \mathcal{I} \text{ for all } x \in D\} \tag{3.7}$$

It is also important to note that to alleviate technical difficulties, we will always assume that the Random Fields (RF) considered are measurable, as well as separable whenever almost sure continuity is mentioned.

3.1.2 A historical perspective: the LGP

Recall that we informally introduced the LGP in Equation 3.2 as being obtained through exponentiation and normalisation of a *well-behaved* GP Z:

$$\psi[Z](t) = \frac{e^{Z_t}}{\int_{\mathcal{I}} e^{Z_u} \, d\lambda(u)} \text{ for all } t \in \mathcal{I}$$

These models intend to provide a flexible prior over positive density functions, where the smoothness of the generated densities is directly inherited from the GP's smoothness.

In the literature, various assumptions and theoretical settings have been proposed that (often, implicitly) specify what *well-behaved* refers to and in what sense the colloquial definition above is meant. We present a concise review of a few papers among the ones we deem to be most representative on the topic.

What we find notable is that working assumptions fluctuate between different contributions, and there is not yet a consensus on the most appropriate set of hypotheses. In particular, the choice between having Z enjoy properties almost-surely (continuity, being exponentially integrable) or surely (separability of the process, or taking values in a suitable function space) is far from straightforward.

- In the seminal paper Leonard (1978), the LGP was studied in a uni-dimensional setting, with $\mathcal{I}$ being a compact interval. The authors considered a.s. surely continuous GPs with exponential covariance kernel.

- Later, the authors of Lenk (1988) claimed that LGPs should be seen as positive-valued random functions integrating to 1 but fail to provide an explicit construction of the corresponding measure space. They extended their construction to derive a generalized logistic Gaussian process (gLGP) and elegant formulations of the posterior distribution of the gLGP conditioned on observations were derived. Numerical approaches for calculating the Bayes estimate were proposed, constituting the starting point of the follow-up paper Lenk (1991).

- In Tokdar and Ghosh (2007), the LGP was introduced from a hierarchical Bayesian modelling perspective, allowing in turn to handle the estimation of GP hyper-parameters. This paper considered a separable GP Z that is exponentially integrable almost surely, stating that the LGP thus takes values in $\mathcal{A}(\mathcal{I})$. The main result in the paper is that the considered hierarchical model achieves weak and strong consistency for density estimation at functions that are piece-wise continuous. It is completed by another paper, Tokdar (2007), where the authors propose a tractable implementation of the density estimation with such a model. Let us note that the GP's separability alleviates some technicalities regarding the measurability assumptions to consider, and having $\int_{\mathcal{I}} e^{Z_u}\, d\lambda(u) < \infty$ a.s. allows to state that LGP realisations are PDFs almost surely.

- Meanwhile, the authors of van der Vaart and van Zanten (2008) work with a bounded-functions-valued GP, which allowed it to be viewed as a Borel measurable map in the space of bounded functions of $\mathcal{I}$ equipped with the sup-norm. This paper derived concentration rates for the posterior of the LGP. With these assumptions, the LGP can be considered as a Borel measurable map in the same space as Z and is guaranteed to have sample paths that are bounded probability density functions.

This short review emphasizes the lack of consensus regarding the LGP's definition including underlying structures and assumptions. It is interesting to

note that in van der Vaart and van Zanten (2008), the authors require Z to be bounded surely, whereas the authors of the three other papers worked with almost sure properties of Z (mostly, the almost surely continuity of the process).

We claim thereafter that working with sure properties allows us to draw links between the LGP and the fertile framework of random measures. We will revisit the definition of LGP in order to build up our subsequent analyses and generalizations on transparent mathematical foundations. Note that this section can be considered as a particular case of the following one, as we will later introduce indexed versions of the LGP and transpose all the upcoming properties to the spatialized version.

Here, rather than viewing LGPs as random functions satisfying constraints (namely: non-negativity, and integrating to 1), we propose viewing them through the scope of Random-Measures (RM). We rely on the definitions from Kallenberg (2017), that are recalled in Section 2.3 and that allow us to work with Random Probability Measures (RPM) (i.e. RM that are surely probability measures). With this in mind, we can establish a connection between LGPs and RPMs, and enjoy the measurability structure of the latter.

To lighten notations, here we call a random process $(Z_t)_{t\in\mathcal{I}}$ exponentially integrable when, for any $\omega \in \Omega$, we have $\int_{\mathcal{I}} e^{Z_u(\omega)}\, d\lambda(u) < \infty$.

Proposition 3.1.1 (RPM induced by a RF, or L-RPM)**.** *For $Z = (Z_t)_{t\in\mathcal{I}}$ an exponentially integrable RF, then:*

$$\Xi(B) := \frac{\int_B e^{Z_u}\, d\lambda(u)}{\int_{\mathcal{I}} e^{Z_u}\, d\lambda(u)} \quad (B \in \mathcal{B}(\mathcal{I})) \tag{3.8}$$

defines a random probability measure that we call random probability measure induced by Z. For notational conciseness, we will denote it L-RPM(Z).

Proof of proposition 3.1.1. Since Z is a measurable RF, e^Z and its integrals are measurable as well. Therefore, for any $B \in \mathcal{B}(\mathcal{I})$. the mapping $\omega \mapsto \Xi(B;\omega)$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Furthermore, for any $\omega \in \Omega$, $B \mapsto \Xi(B;\omega)$ is a probability measure on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$, so *a fortiori* locally finite. □

Remark. We consider the condition of sure-exponential-integrability made in Definition 3.1.1 not to be overly restrictive. Indeed, let us consider a RF Z that is a.s. exponentially integrable (meaning that $e^{Z(\omega)}$ is integrable for all $\omega \in \Omega$ except some P-null set noted N). Then, $\mathcal{I}$ being compact, we can always construct a surely exponentially integrable RF $\tilde{Z}$ indistinguishable from Z via

$$\tilde{Z}(\omega) = \begin{cases} Z(\omega) & \text{if } \omega \in \Omega \setminus N \\ 0 & \text{else} \end{cases} \tag{3.9}$$

Following this construction, we can formally define a class of processes slightly more general than the LGP's one.

Definition 3.1.1. For a RF Z that is exponentially integrable,

$$\psi[Z](t) = \frac{e^{Z_t}}{\int_{\mathcal{I}} e^{Z_u}\, d\lambda(u)} \text{ for all } t \in \mathcal{I} \tag{3.10}$$

is a representer of the Radon–Nikodym derivative of L-RPM(Z), denoted $\psi[Z]$.

While it is tempting to characterise a L-RPM by the underlying transformed RF, it is insufficient as highlighted by the next remark.

Remark. Let us consider an exponentially integrable RF $(Z_t)_{t\in\mathcal{I}}$ and a random variable R defined on the same probability space. Then, $(Z_t)_{t\in\mathcal{I}}$ and $(Z_t + R)_{t\in\mathcal{I}}$ induce the same L-RPM, since:

$$\frac{\int_B e^{Z_u(\omega)}\, d\lambda(u)}{\int_{\mathcal{I}} e^{Z_u(\omega)}\, d\lambda(u)} = \frac{\int_B e^{[Z_u+R](\omega)}\, d\lambda(u)}{\int_{\mathcal{I}} e^{[Z_u+R](\omega)}\, d\lambda(u)} \quad \forall B \in \mathcal{B}(\mathcal{I}),\ \forall \omega \in \Omega \tag{3.11}$$

Due to the normalisation constant in Equation 3.8, there is no one-to-one correspondence between RFs and associated random measures.

To address this caveat, we derive and prove a characterisation of the L-RPM in terms of its underlying increment field.

Proposition 3.1.2 (Condition for the indistinguishability of L-RPM)**.** *Let $Z := (Z_t)_{t\in\mathcal{I}}$ and $\tilde{Z} := (\tilde{Z}_t)_{t\in\mathcal{I}}$ be two RFs that are exponentially integrable, and for all $(t, t') \in \mathcal{I}^2$, let $\Delta Z_{t,t'} := Z_t - Z_{t'}$ (respectively $\Delta \tilde{Z}_{t,t'} := \tilde{Z}_t - \tilde{Z}_{t'}$) be associated increment processes. Then,*

(1) *$(\Delta Z_{t,t'})_{(t,t')\in\mathcal{I}^2}$ is indistinguishable from $(\Delta \tilde{Z}_{t,t'})_{(t,t')\in\mathcal{I}^2}$.*

$\Leftrightarrow$ (2) *$\psi[Z]$ is indistinguishable from $\psi[\tilde{Z}]$.*

$\Rightarrow$ (3) *L-RPM(Z) is indistinguishable from L-RPM($\tilde{Z}$).*

Additionally, if Z and $\tilde{Z}$ are almost surely continuous, (3) $\Rightarrow$ (2).

Proof of proposition 3.1.2. Let us consider two such RFs Z and $\tilde{Z}$. Assuming the two increment fields are indistinguishable, for an arbitrary $t_0 \in \mathcal{I}$, both $\Delta Z_{\cdot t_0}$ and $\Delta \tilde{Z}_{\cdot t_0}$ are indistinguishable RFs that are exponentially integrable. Note that $\psi[\Delta \tilde{Z}_{\cdot t_0}] = \psi[\tilde{Z}]$, and therefore:

$$\begin{aligned} 1 &= P\left[\psi[\Delta \tilde{Z}_{\cdot t_0}](t) = \psi[\Delta Z_{\cdot t_0}](t)\ \forall t \in \mathcal{I}\right] \\ &= P\left[\Psi[\tilde{Z}](t) = \Psi[Z](t)\ \forall t \in \mathcal{I}\right] \end{aligned}$$

It follows that (1) $\Rightarrow$ (2). Moreover, it also follows that:

$$\begin{aligned} 1 &= P\left[\psi[\tilde{Z}](t) = \psi[Z](t)\ \forall t \in \mathcal{I}\right] \\ &= P\left[\int_B \psi[\tilde{Z}](u)\, d\lambda(u) = \int_B \Psi[Z](u)\, d\lambda(u),\ \ \forall B \in \mathcal{B}(\mathcal{I})\right] \end{aligned}$$

Which proves that (2) $\Rightarrow$ (3).

Conversely, for (2) $\Rightarrow$ (1), let us assume that $Y = \psi[Z]$ is indistinguishable from $\tilde{Y} = \psi[\tilde{Z}]$. By SLP's construction, we can consider $\log Y_t = Z_t - \log \int_{\mathcal{I}} e^{Z_u}\, d\lambda(u)$ (resp. $\log \tilde{Y}_t = \tilde{Z}_t - \log \int_{\mathcal{I}} e^{\tilde{Z}_u}\, d\lambda(u)$), and:

$$\begin{aligned} 1 &= P\left[\log Y_t = \log \tilde{Y}_t\ \forall t \in \mathcal{I}\right] \\ &= P\left[\log Y_t - \log Y_{t'} = \log \tilde{Y}_t - \log \tilde{Y}_{t'}\ \forall t \in \mathcal{I}\right] \\ &= P\left[Z_t - Z_{t'} = \tilde{Z}_t - \tilde{Z}_{t'}\ \forall (t,t') \in \mathcal{I}^2\right] \end{aligned}$$

which is, indeed, proving (2) $\Rightarrow$ (1).

Finally, let us assume that (3) holds and that $\Xi = \text{L-RPM}(Z)$ is indistinguishable from $\tilde{\Xi} = \text{L-RPM}(\tilde{Z})$. By indistinguishability:

$$\begin{aligned} & P\left[\Xi(B) = \tilde{\Xi}(B),\ \forall B \in \mathcal{B}(\mathcal{I})\right] = 1 \\ \Leftrightarrow\ & P\left[Y_t = \tilde{Y}_t,\ \ \text{for } \lambda\text{-almost every } t \in \mathcal{I}\right] = 1 \end{aligned}$$

Under the general setting, this is not enough to prove that (3) $\Rightarrow$ (2).

However, assuming that both Z and $\tilde{Z}$ are a.s. continuous, we deduce that so are Y and $\tilde{Y}$. This allows for going from almost sure equality λ-almost everywhere to almost sure equality everywhere, and therefore:

$$P\left[Y_t = \tilde{Y}_t,\ \ \text{for all } t \in \mathcal{I}\right] = 1$$

□

Noticeably, we made no Gaussianity assumption on the transformed RF Z. Indeed, this hypothesis is not required to properly define L-RPMs in a general setting. Gaussianity is mostly instrumental, as it allows for parametrization of GPs through their mean and covariance functions. Moreover, one can rely on the flourishing literature on the topic to derive properties of the processes at hand. We now turn to L-RPM induces by GPs, with a special focus on LGPs.

Definition 3.1.2 (Logistic Gaussian Process)**.** A LGP is a process $(Y_t)_{t\in\mathcal{I}}$ such that there exists a measurable GP Z that is exponentially integrable and:

$$Y_t = \psi[Z](t) = \frac{e^{Z_t}}{\int_{\mathcal{I}} e^{Z_u}\, d\lambda(u)} \text{ for all } t \in \mathcal{I} \tag{3.12}$$

We call Y the LGP induced by Z.

We can connect previous work on LGPs with our RPM framework. As we mentioned earlier, several authors worked under continuity assumptions to alleviate technical difficulties. In our framework, working under almost-surely continuity of the transformed GP allows us to link the LGPs as they are defined in 3.1.2 and the L-RPMs in Proposition 3.1.1.

Proposition 3.1.3 (LGP induced by an a.s. continuous GP)**.** *Let us consider a GP $Z = (Z_t)_{t\in\mathcal{I}}$ that is exponentially integrable and almost-surely continuous. Then, $\psi[Z]$ is a.s. the continuous representer of $\frac{d\Xi}{d\lambda}$, where $\Xi = L\text{-}RPM(Z)$.*

With these basic ideas in mind, let us now introduce the SLGP with more formalism. Note that all the properties we will prove in the coming section applies to LGP as well.

3.2 The SLGP model and its characterisation

In this section, we build upon the work of (Pati et al., 2013) and present the considered spatial extension of the Logistic Gaussian Process. We point out that rather than focusing on the posterior consistency of the model as the authors of the aforementioned paper did, we will study its spatial regularity.

From now-on, we will call a measurable RF $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ exponentially integrable alongside $\mathcal{I}$ if $\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}(\omega)}\, d\lambda(u) < \infty$ for any $(\mathbf{x},\omega) \in D \times \Omega$.

To introduce a spatial extension to L-RPMs, we first need to introduce a spatial extension of the logistic density transformation:

Definition 3.2.1 (Spatial logistic density transformation)**.** The spatial logistic density transformation Ψ is defined over the set of measurable $w : D \times \mathcal{I} \to \mathbb{R}$ such that for all $\mathbf{x} \in D$, $\int_{\mathcal{I}} e^{w(\mathbf{x},u)}\, d\lambda(u) < \infty$:

$$\Psi[w](\mathbf{x},t) := \frac{e^{w(\mathbf{x},t)}}{\int_{\mathcal{I}} e^{w(\mathbf{x},u)}\, d\lambda(u)} \quad \text{for all } (\mathbf{x},t) \in D \times \mathcal{I} \tag{3.13}$$

hence being a mapping between functions that are exponentially integrable alongside $\mathcal{I}$ and $\mathcal{A}_d^+(D;\mathcal{I})$.

Recall that we informally introduced the SLGP in Equation 3.3 as being an indexed version of the LGP:

$$\Psi[Z](\mathbf{x},t) = \frac{e^{Z_{\mathbf{x},t}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \text{ for all } (\mathbf{x},t) \in D \times \mathcal{I}$$

We start working with few assumptions on the transformed RF. This naturally leads us to working with fields of random measures (i.e. collections of RPMs defined on the same probability space). For notational conciseness, we refer to such fields as RPMFs.

Random Probability Measure Fields induced by a RF: definition and characterisation

It is natural to present a spatialised version of the L-RPMs introduced in Definition 3.1.1.

Definition 3.2.2 (L-RPMF)**.** Let us consider $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$, a RF that is exponentially integrable alongside $\mathcal{I}$, then:

$$\Xi_{\mathbf{x}}(B) = \int_B \Psi[Z](\mathbf{x},u)\, d\lambda(u) = \frac{\int_B e^{Z_{\mathbf{x},u}}\, d\lambda(u)}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \quad (\mathbf{x}\in D,\ B\in\mathcal{B}(\mathcal{I})) \tag{3.14}$$

defines a RPMF that we call Logistic Random Probability Measure Field induced by Z. We also use the notation $\Xi = \text{L-RPMF}(Z)$.

Definition 3.2.3. Let us consider $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$, a RF that is exponentially integrable alongside $\mathcal{I}$, then:

$$\Psi[Z](\mathbf{x},t) = \frac{e^{Z_{\mathbf{x},t}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \text{ for all } (\mathbf{x},t) \in D \times \mathcal{I} \tag{3.15}$$

defines a process such that, for any $\mathbf{x}\in D$, $\Psi[Z](\mathbf{x},\cdot)$ is a representer of $\dfrac{d\Xi_{\mathbf{x}}}{d\lambda}$, the Radon–Nikodym derivative of $\Xi_{\mathbf{x}}$, where $\Xi = \text{L-RPMF}(Z)$.
We denote this process by $\Psi[Z]$ and refer to it as Spatial Logistic Process (SLP).

While it is tempting to characterise a L-RPMF by its underlying RF, it is hopeless. In fact, different RFs may yield the same L-RPMF.

Remark. Let us consider two RFs $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ and $(R_{\mathbf{x}})_{\mathbf{x}\in D}$ defined on the same probability space, and assume that Z is exponentially integrable alongside $\mathcal{I}$. Then, $\Psi[Z]$ and $\Psi[Z+R]$ are equal, indeed:

$$\frac{e^{Z_{\mathbf{x},t}(\omega)}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}(\omega)}\, d\lambda(u)} = \frac{e^{[Z_{\mathbf{x},t}+R_{\mathbf{x}}](\omega)}}{\int_{\mathcal{I}} e^{[Z_{\mathbf{x},u}+R_{\mathbf{x}}](\omega)}\, d\lambda(u)} \quad (\mathbf{x},t,\omega)\in D\times\mathcal{I}\times\Omega \tag{3.16}$$

It follows that L-RPMF(Z) and L-RPMF$(Z+R)$ are also equal.

The arising questions that we will try to address through the rest of this section is: how to characterise the random measure fields that can be obtained through Equation 3.14, and can we give sufficient conditions on measurable and exponentially integrable RFs for them to yield the same L-RPMF?

This calls for a proper definition of what "the same" encapsulates, as there are several notions of coincidence between RFs (and a fortiori RPMFs). Here, we will mostly focus on indistinguishability.

Remark. Let $(\Xi_{\mathbf{x}})_{\mathbf{x}\in D}$ be a RPMF. It is a collection of probability measure-valued random variables indexed by D. As such it is natural to call two RPMFs $\Xi_{\mathbf{x}}$ and $\tilde{\Xi}_{\mathbf{x}}$ indistinguishable if:

$$P\left[\Xi_{\mathbf{x}} = \tilde{\Xi}_{\mathbf{x}}, \forall \mathbf{x} \in D\right] = 1 \tag{3.17}$$

By definition of equality between measures, we can reformulate the latter as:

$$P\left[\Xi_{\mathbf{x}}(B) = \tilde{\Xi}_{\mathbf{x}}(B),\ \forall \mathbf{x} \in D, \forall B \in \mathcal{B}(\mathcal{I})\right] = 1 \tag{3.18}$$

Coincidentally, that last equation corresponds to the indistinguishability of the scalar-valued RFs $\Xi_{\mathbf{x}}(B)$ and $\tilde{\Xi}_{\mathbf{x}}(B)$, indexed by $(\mathbf{x}, B) \in D \times \mathcal{B}(\mathcal{I})$. This equivalence results from the construction of RPMs in (Kallenberg, 2017) that ensures that all RPMs are regular conditional distributions, as recalled in Section 2.3.

Proposition 3.2.1 (Condition for the indistinguishability of L-RPMF and SLPs)**.** *Let $Z := (Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ and $\tilde{Z} := (\tilde{Z}_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ be two RFs that are exponentially integrable alongside $\mathcal{I}$, and for all $(\mathbf{x}, t, t') \in D \times \mathcal{I}^2$, let $\Delta Z_{\mathbf{x},t,t'} := Z_{\mathbf{x},t} - Z_{\mathbf{x},t'}$ (respectively $\Delta\tilde{Z}_{\mathbf{x},t,t'} := \tilde{Z}_{\mathbf{x},t} - \tilde{Z}_{\mathbf{x},t'}$) be associated increment processes. Then,*

(1) *$(\Delta Z_{\mathbf{x},t,t'})_{(\mathbf{x},t,t')\in D\times\mathcal{I}^2}$ is indistinguishable from $(\Delta\tilde{Z}_{\mathbf{x},t,t'})_{(\mathbf{x},t,t')\in D\times\mathcal{I}^2}$.*

$\Leftrightarrow$ (2) *$\Psi[Z]$ is indistinguishable from $\Psi[\tilde{Z}]$.*

$\Rightarrow$ (3) *L-RPMF(Z) is indistinguishable from L-RPMF($\tilde{Z}$).*

Additionally, if Z and $\tilde{Z}$ are almost surely continuous, (3) $\Rightarrow$ (2).

Proof. Let us consider two such RFs Z and $\tilde{Z}$. Assuming the two increment fields are indistinguishable, for an arbitrary $t_0 \in \mathcal{I}$, both $\Delta Z_{\cdot\cdot t_0}$ and $\Delta\tilde{Z}_{\cdot\cdot t_0}$ are indistinguishable RFs that are exponentially integrable alongside $\mathcal{I}$. Note that $\Psi[\Delta\tilde{Z}_{\cdot\cdot t_0}] = \Psi[\tilde{Z}]$, and therefore:

$$\begin{aligned} 1 &= P\left[\Psi[\Delta\tilde{Z}_{\cdot\cdot t_0}](\mathbf{x}, t) = \Psi[\Delta Z_{\cdot\cdot t_0}](\mathbf{x}, t)\ \forall (\mathbf{x}, t) \in D \times \mathcal{I}\right] \\ &= P\left[\Psi[\tilde{Z}](\mathbf{x}, t) = \Psi[Z](\mathbf{x}, t)\ \forall (\mathbf{x}, t) \in D \times \mathcal{I}\right] \end{aligned}$$

It follows that (1) ⇒ (2). Moreover, it also follows that:

$$\begin{aligned}1 &= P\left[\Psi[\tilde{Z}](\mathbf{x},t) = \Psi[Z](\mathbf{x},t)\ \forall(\mathbf{x},t)\in D\times\mathcal{I}\right]\\ &= P\left[\int_B \Psi[\tilde{Z}](\mathbf{x},u)\,d\lambda(u) = \int_B \Psi[Z](\mathbf{x},u)\,d\lambda(u),\ \ \forall(\mathbf{x},B)\in D\times\mathcal{B}(\mathcal{I})\right]\end{aligned}$$

Which proves that (2) ⇒ (3).

Conversely, for (2) ⇒ (1), let us assume that $Y = \Psi[Z]$ is indistinguishable from $\tilde{Y} = \Psi[\tilde{Z}]$. By SLP's construction, we can consider $\log Y_{\mathbf{x},t} = Z_{\mathbf{x},t} - \log\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\,d\lambda(u)$ (resp. $\log\tilde{Y}_{\mathbf{x},t} = \tilde{Z}_{\mathbf{x},t} - \log\int_{\mathcal{I}} e^{\tilde{Z}_{\mathbf{x},u}}\,d\lambda(u)$), and:

$$\begin{aligned}1 &= P\left[\log Y_{\mathbf{x},t} = \log\tilde{Y}_{\mathbf{x},t}\ \forall(\mathbf{x},t)\in D\times\mathcal{I}\right]\\ &= P\left[\log Y_{\mathbf{x},t} - \log Y_{\mathbf{x},t'} = \log\tilde{Y}_{\mathbf{x},t} - \log\tilde{Y}_{\mathbf{x},t'}\ \forall(\mathbf{x},t,t')\in D\times\mathcal{I}^2\right]\\ &= P\left[Z_{\mathbf{x},t} - Z_{\mathbf{x},t'} = \tilde{Z}_{\mathbf{x},t} - \tilde{Z}_{\mathbf{x},t'}\ \forall(\mathbf{x},t,t')\in D\times\mathcal{I}^2\right]\end{aligned}$$

which is, indeed, proving (2) ⇒ (1).

Finally, let us assume that (3) holds and that $\Xi_{\mathbf{x}} = \text{L-RPMF}(Z)$ is indistinguishable from $\tilde{\Xi}_{\mathbf{x}} = \text{L-RPMF}(\tilde{Z})$. By indistinguishability:

$$\begin{aligned}&P\left[\Xi_{\mathbf{x}}(B) = \tilde{\Xi}_{\mathbf{x}}(B),\ \forall\mathbf{x}\in D,\forall B\in\mathcal{B}(\mathcal{I})\right] = 1\\ &\Leftrightarrow P\left[Y_{\mathbf{x},t} = \tilde{Y}_{\mathbf{x},t},\ \ \text{for } \lambda\text{-almost every } t\in\mathcal{I},\ \text{for all } \mathbf{x}\in D\right] = 1\end{aligned}$$

Under the general setting, this is not enough to prove that (3) ⇒ (2).

However, assuming that both Z and $\tilde{Z}$ are a.s. continuous, we deduce that so are Y and $\tilde{Y}$. This allows for going from almost sure equality λ-almost everywhere to almost sure equality everywhere, and therefore:

$$P\left[Y_{\mathbf{x},t} = \tilde{Y}_{\mathbf{x},t},\ \ \text{for all } t\in\mathcal{I},\mathbf{x}\in D\right] = 1$$

□

Remark (Indistinguishability compared to others notions of coincidence between RPMF). In Proposition 3.2.1, we worked with the indistinguishability of random measure fields, as defined in Equation 3.18. Although one could consider other types of equality between RPMF, such as the equality up to a modification:

$$P\left[\Xi_{\mathbf{x}} = \tilde{\Xi}_{\mathbf{x}}\right] = 1 \quad \forall\ \mathbf{x}\in D \tag{3.19}$$

we found out that indistinguishability seems to be the best fit, as it naturally relates indistinguishability of L-RPMFs to that of underlying fields of increments.

Remark (Working with increments). In the previous proposition and its proof, we decided to work with increments of Z rather than with Z directly. This makes it easier relate SLPs and the RF inducing it. Indeed, for $Y = \Psi[Z]$, we have $\log Y_{\mathbf{x},t} = Z_{\mathbf{x},t} - \log \int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)$ and therefore:

$$\log Y_{\mathbf{x},t} - \log Y_{\mathbf{x},t'} = Z_{\mathbf{x},t} - Z_{\mathbf{x},t'} \quad \text{for all } (\mathbf{x}, t, t') \in D \times \mathcal{I}^2 \tag{3.20}$$

From this characterisation, it appears that indistinguishability of SLPs or L-RPMFs is driven by the increments of the transformed RF. It also appears that almost sure continuity is a practical assumption to alleviate technical difficulties. However, it also highlights how general our construction is, indeed:

Lemma 2. *Let us consider a RPMF $(\Xi_{\mathbf{x}})_{\mathbf{x}\in D}$, if there exist a measurable process $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ with:*

$$P\left[Y_{\mathbf{x},\cdot} \text{ is a positive representer of } \frac{d\Xi_{\mathbf{x}}}{d\lambda} \text{ for all } \mathbf{x} \in D\right] = 1 \tag{3.21}$$

then there exist a RF Z exponentially integrable alongside $\mathcal{I}$ such that Y is indistinguishable from $\Psi[Z]$, and Ξ is indistinguishable from L-RPMF(Z).

Proof. Let us consider such a $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ and denote by N the P-null set where Y is not the positive representer of the Radon-Nikodym derivative of $\Xi_{\mathbf{x}}$ for all $\mathbf{x} \in D$. We define $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ by:

$$Z_{\mathbf{x},t}(\omega) := \begin{cases} \log Y_{\mathbf{x},t}(\omega) & \text{if } \omega \in \Omega \setminus N \\ 0 & \text{else} \end{cases} \tag{3.22}$$

By construction, Z is RF that is exponentially integrable alongside $\mathcal{I}$, and $\Psi[Z]$ is indistinguishable from Y. It follows from proposition 3.2.1 that L-RPMF(Z) is indistinguishable from Ξ. □

Therefore, L-RPMFs are quite a general object, and can model a wide class of RPMFs. However, in practice we will generally construct our models by specifying a Z, and transforming it to obtain the corresponding SLPs/L-RPMFs. Next, we will focus on L-RPMFs induced by GPs.

Logistic Random Probability Measure Fields induced by a GP, and their Radon-Nikodym derivative

We now characterise which SLPs are obtained by transforming a GP.

Proposition 3.2.2 (Characterizing SLPs obtained by transforming a GP)**.** *For a measurable RF $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ that is exponentially integrable alongside $\mathcal{I}$, the following are equivalent:*

(1) There exist a measurable GP $(\tilde{Z}_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ *that is exponentially integrable such that* $\Psi[Z] = \Psi[\tilde{Z}]$

(2) $(Z_{\mathbf{x},t} - Z_{\mathbf{x},t'})_{(\mathbf{x},t,t')\in D\times\mathcal{I}^2}$*, is a GP.*

Proof. If (1) holds, then $Z_{\mathbf{x},t} - Z_{\mathbf{x},t'} = \tilde{Z}_{\mathbf{x},t} - \tilde{Z}_{\mathbf{x},t'}$ for all $(\mathbf{x},t) \in D\times\mathcal{I}$. Since $\tilde{Z}$ is a GP, so is its increment field, and so is Z's one.

Conversely, consider Z as in (2). For an arbitrary $t_0 \in \mathcal{I}$, let us define $\tilde{Z} := Z_{\mathbf{x},t} - Z_{\mathbf{x},t_0}$. The process $\tilde{Z}$ is a GP on $D\times\mathcal{I}$, and for any $(\mathbf{x},t) \in \mathcal{D}\times\mathcal{I}$:

$$\Psi[\tilde{Z}](\mathbf{x},t) = \frac{e^{Z_{\mathbf{x},t}-Z_{\mathbf{x},t_0}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}-Z_{\mathbf{x},t_0}}\, d\lambda(u)} = \frac{e^{Z_{\mathbf{x},t}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} = \Psi[Z](\mathbf{x},t)$$

□

From this, it appears that SLPs obtained by transforming a GP are not characterised by a GP on $D\times\mathcal{I}$ but rather by an increment (Gaussian) process on $D\times\mathcal{I}^2$ with sufficient measurability.

To shorten notations and connect our work to previous contributions from other authors, from now-on we will refer to SLPs obtained by transforming a GP as Spatial Logistic Gaussian Processes (SLGPs).

SLGPs benefit from continuity assumptions, as it allows for easier characterisation and parametrisations, and highlighted in the following remark.

Remark. In practice, GPs are often defined up to stochastic equivalence, by specifying their mean and covariance kernel (and therefore their finite-dimensional distributions). However, since the definition of the SLGPs induced by some Z involves the sample path of Z over all $\mathcal{I}$, having two GPs $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ and $(\tilde{Z}_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ exponentially integrable alongside $\mathcal{I}$ with:

$$P\left[\tilde{Z}_{\mathbf{x},t} = \tilde{Z}_{\mathbf{x},t}\right] = 1 \text{ for all } (\mathbf{x},t) \in D\times\mathcal{I} \text{ (i.e. } Z \text{ and } \tilde{Z} \text{ are equivalent)} \quad (3.23)$$

is not sufficient to ensure that $\Xi = \text{L-RPMF}(Z)$ and $\tilde{\Xi} = \text{L-RPMF}(\tilde{Z})$ satisfy:

$$P\left[\Xi_{\mathbf{x}} = \tilde{\Xi}_{\mathbf{x}}\right] = 1 \text{ for all } \mathbf{x} \in D \text{ (i.e. } \Xi \text{ and } \tilde{\Xi} \text{ are equivalent)} \quad (3.24)$$

In other terms, two GPs with the same mean function and covariance function do not necessarily yield L-RPMFs that are equivalent, nor indistinguishable. One well-known exception to this arises when $\tilde{Z}$ is a.s. continuous and is a version of Z. Then, both GPs are separable and indistinguishable, and so are the L-RPMFs they induce. We refer to (Azaïs and Wschebor, 2009) Ch. 1, Sec. 4, Prop. 1.9 for a proof in dimension 1, and to (Scheuerer, 2009) Ch. 5 Sec. 2 Lemma 5.2.8. for a generalisation of it.

This enables us to derive yet another characterisation of L-RPMFs obtained by transforming continuous GPs.

Proposition 3.2.3 (Underlying increment mean and covariance)**.** *For every L-RPMF $(\Xi_{\mathbf{x}})_{\mathbf{x}\in D}$ (resp. SLGP $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$) induced by an almost surely continuous GP Z, there exist a unique mean function m_{inc} and a unique covariance kernel k_{inc}:*

$$m_{inc} : D \times \mathcal{I}^2 \to \mathbb{R} \tag{3.25}$$

$$k_{inc} : \big((D \times \mathcal{I}^2) \times (D \times \mathcal{I}^2)\big) \to \mathbb{R} \tag{3.26}$$

that characterise all the L-RPMFs indistinguishable from Ξ (resp. the SLPs indistinguishable from Y).
We call them the mean and covariance underlying the L-RPMF (resp. the SLGP).

Proof. Combining Propositions 3.2.1 and 3.2.2 emphasizes that the process that drives Ξ and Y's behaviour is $(Z_{\mathbf{x},t} - Z_{\mathbf{x},t'})_{(\mathbf{x},t,t')\in D\times\mathcal{I}^2}$. It is a continuous GP, with m_{inc} and k_{inc} being its mean function and covariance function. As noted in remark 3.2, indistinguishability of continuous GPs is driven by these functions, which ensures that m_{inc} and k_{inc} characterise Ξ (resp. Y) □

Proposition 3.2.4 (SLGP induced by an a.s. continuous GP)**.** *Let us consider a GP $Z = (Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ that is exponentially integrable alongside $\mathcal{I}$ and almost-surely continuous in t. Then, for any $\mathbf{x} \in D$, $\Psi[Z](\mathbf{x}, \cdot)$ is almost surely the continuous representer of $\dfrac{d\Xi_{\mathbf{x}}}{d\lambda}$, where $\Xi = \text{L-RPMF}(Z)$.*

A direct consequence of Proposition 3.2.4, is that whenever these assumptions on Z are fulfilled, we can refer to the corresponding SLGP as being almost surely a probability density functions field.

As mentioned in Remark 3.2, combining Gaussianity assumptions and (a.s.) continuity assumptions allows for simpler characterisation. Indeed, being equal up to a version coincide with being indistinguishable. Therefore, it is possible to characterise a.s. continuous GPs that yield indistinguishable SLGPs directly through their kernels and means.

Proposition 3.2.5 (Increment mean and covariance of GPs underlying a SLGP)**.** *Let $(Z_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ be a GP that is exponentially integrable alongside $\mathcal{I}$, and generates a SLGP (resp. a L-RPMF) with underlying mean and covariance m_{inc} and k_{inc}. Z's mean m and covariance k satisfy for all $(\mathbf{x}, \mathbf{x}') \in D^2$, $(t_1, t_2, t_1', t_2') \in \mathcal{I}^4$:*

$$m_{inc}(\mathbf{x}, t_1, t_2) = m(\mathbf{x}, t_1) - m(\mathbf{x}, t_2) \tag{3.27}$$

$$k_{inc}([\mathbf{x}, t_1, t_2], [\mathbf{x}', t_1', t_2']) = \begin{array}{c} k([\mathbf{x}, t_1], [\mathbf{x}', t_1']) + k([\mathbf{x}, t_2], [\mathbf{x}', t_2']) \\ -k([\mathbf{x}, t_1], [\mathbf{x}', t_2']) - k([\mathbf{x}, t_2], [\mathbf{x}', t_1']) \end{array} \tag{3.28}$$

This last property is central, as we already mentioned that in practice it is easier to define a SLGP by specifying Z. This generally involves choosing a suitable kernel k on $D \times \mathcal{I}$ and then deducing the corresponding SLGPs and L-RPMFs and their underlying means and kernel m_{inc} and k_{inc} from Equations 3.27 and 3.28.

In the rest of the chapter, we will study the spatial regularity of the SLGP in Section 3.3.1 and touch upon the posterior consistency of this model in Section 3.3.2.

3.3 Properties of the SLGP

3.3.1 Continuity modes for (logistic Gaussian) random measure fields

Our object of interest in this document is a random measure field. A natural question, when working with spatial objects, is to quantify the impact of a prior on the regularity of the delivered predictions. Quantifying the spatial regularity of such an object boils down to quantifying how similar two conditional measures $\Xi_{\mathbf{x}}$, $\Xi_{\mathbf{x}'}$ are when their respective predictors $\mathbf{x}, \mathbf{x}'$ become close.

This investigation requires distances (or dissimilarity measures) between both measures and locations, and we will consider different ones. To compare two measures, we will consider Hellinger distance, Kullback-Leibler divergence and Total Variation distance. For locations, we will consider the sup norm over D as well as the canonical distance associated to the covariance kernel of the Gaussian random increment field.

In this particular case, we are focusing on a two notions of regularity. The first one being the almost sure continuity of the SLGP, the second one being inspired by the Mean-Squared continuity on the scalar valued case. We will prove statements of the following form: for a given dissimilarity between measures ρ and for a SLGP $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$, under sufficient regularity of the covariance kernel k_{inc} underlying Y:

$$\lim_{\mathbf{x}'\to\mathbf{x}} \mathbb{E}\left[\rho\left(\Xi_{\mathbf{x}}, \Xi_{\mathbf{x}'}\right)\right] = 0. \tag{3.29}$$

We will also provide bounds on the convergence rate. In this work, we shall focus on ρ being either d_H the Hellinger distance, d_{TV} the Total variation distance or KL the Kullback-Leibler divergence. The choice of these three dissimilarity measures is motivated by the following Lemma.

Lemma 3 (Bounds on distances between measures)**.** *There exists two constants $C_{KL}, C_{TV} > 0$ such that for f_1 and f_2 two positive probability density functions*

on $\mathcal{I}$:

$$d_H(f_1, f_2) \leq he^{h/2} \tag{3.30}$$

$$KL(f_1, f_2) \leq C_{KL} h^2 e^h (1+h) \tag{3.31}$$

$$d_{TV}(f_1, f_2) \leq C_{TV} h^2 e^h (1+h)^2 \tag{3.32}$$

where $h := \| \log(f_1) - \log(f_2) \|_\infty$.

This is lemma 3.1 of (van der Vaart and van Zanten, 2008).

As is standard in spatial statistics, we shall derive the SLGP Y's regularity through that of it underlying kernel and mean (or equivalently, through that of the kernel and mean of a GP Z inducing Y). More precisely, we will prove that inherits its regularity from the canonical semi-distance associated to a kernel (reminded in definition 2.1.9).

Condition 1 (Condition on kernels on k_{inc} on $D \times \mathcal{I}^2$)**.** There exist $C, \alpha_1, \alpha_2 > 0$ such that for all $(\mathbf{x}, \mathbf{x}') \in D^2, (t_1, t_1', t_2, t_2') \in \mathcal{I}^4$:

$$d^2_{k_{\text{inc}}}([\mathbf{x}, t_1, t_2], [\mathbf{x}', t_1', t_2']) \leq C \cdot \max(\|\mathbf{x} - \mathbf{x}'\|_\infty^{\alpha_1}, \|t_1 - t_1'\|_\infty^{\alpha_2}, \|t_2 - t_2'\|_\infty^{\alpha_2}) \tag{3.33}$$

As we noted in Proposition 3.2.5, we are mostly interested in kernels k_{inc} on $D \times \mathcal{I}^2$ that can be interpreted as increments of kernels k on $D \times \mathcal{I}$. Therefore, we also introduce a natural counterpart to Condition 1 for kernels on $D \times \mathcal{I}$:

Condition 2 (Condition on kernels on $D \times \mathcal{I}$)**.** There exist $C, \alpha_1, \alpha_2 > 0$ such that for all $\mathbf{x}, \mathbf{x}' \in D, t, t' \in \mathcal{I}$:

$$d^2_k([\mathbf{x}, t], [\mathbf{x}', t']) \leq C \cdot \max(\|\mathbf{x} - \mathbf{x}'\|_\infty^{\alpha_1}, \|t - t'\|_\infty^{\alpha_2}) \tag{3.34}$$

Remark. In our setting, D and $\mathcal{I}$ being compact, if a covariance kernel k_u on $D \times \mathcal{I}^2$ satisfies Condition 1, it is also true that there exists C' such that for all $(\mathbf{y}, \mathbf{y}') \in (D \times \mathcal{I}^2)^2$:

$$d^2_{k_{\text{inc}}}(\mathbf{y}, \mathbf{y}') \leq C' \cdot \|\mathbf{y} - \mathbf{y}'\|_\infty^{\min(\alpha_1, \alpha_2)} \tag{3.35}$$

An analogous result is true for a covariance kernel k satisfying Condition 2. Hence, Conditions 1 and 2 can be referred to as Hölder-type conditions. Although Equation 3.35 would allow for deriving most results in the coming subsection, when deriving rates in Section 3.3.1 it is interesting to distinguish the regularity over D from the regularity over $\mathcal{I}$ as there is a strong asymmetry between both spaces.

First, we claim that in our setting, it is equivalent to be working with Condition 1 or Condition 2.

Proposition 3.3.1. *Let k be a kernel on $D \times \mathcal{I}$ and k_{inc} be a kernel on $D \times \mathcal{I}^2$. The two following statements stand:*

1. *If k satisfies Condition 2, and k_{inc} derives from it through Equation 3.28, then k_{inc} satisfies Condition 1.*

2. *Conversely, if k_{inc} is the underlying increment kernel of a SLGP Y and if it satisfies Condition 1, then it is possible to choose a kernel k on $D \times \mathcal{I}$ such that k satisfies Condition 2 and through Equation 3.28 is satisfied.*

Moreover, the constants α_1, α_2 will be the same in both conditions.

This proves to be practical in the upcoming sections, as it allows to slightly shorten notations by using k rather than k_{inc}. Moreover, as mentioned earlier, it is often easier to define a SLGP by transforming a GP.

From here, we will conduct our analysis in the setting considered in Proposition 3.2.5 and assume that the considered SLGPs are almost surely positive. With this in mind, we are now ready to introduce one of the main contributions of this chapter. We first show that Condition 2 is sufficient for the almost surely continuity (in sup norm) of the SLGP as well as mean Hölder continuity of the SLGP.

Almost sure continuity of the Spatial Logistic Gaussian Process

First, let us remark that if a covariance kernel k on $D \times \mathcal{I}$ satisfies Condition 2, then the associated centred GP admits a version that is almost surely continuous and therefore almost surely bounded. Proposition 2.1.3 proven in appendix for self-containedness constitutes a classical result in stochastic processes literature, but is essential as it ensures the objects we will work with are well-defined. It then allows us to derive a bound for the expected value of the sup-norm of our increment field, and to leverage it for our main contribution for this section.

Proposition 3.3.2. *If a covariance kernel k on $D \times \mathcal{I}$ satisfies Condition 2, then for any $0 < \delta < \dfrac{\alpha_1}{2}$, there exists a constant K_δ such that for $Z \sim \mathcal{GP}(0, k)$:*

$$M(\mathbf{x}, \mathbf{x}') := \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty\right] \leq K_\delta \|\mathbf{x} - \mathbf{x}'\|_\infty^{\alpha_1/2-\delta} \ , \ \forall (\mathbf{x}, \mathbf{x}') \in D^2 \tag{3.36}$$

Despite its reliance on standard results for spatial statistics (namely Dudley's theorem), the full proof of Proposition 3.3.2 requires precision, to ensure that the provided bounds are tight. As such, we decided to give the main idea here, but to refer the reader to proofs in the Appendix A.2 for full derivation.

Main elements for proving Proposition 3.3.2. For any $(\mathbf{x}, \mathbf{x}') \in D^2$, the process $Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}$, is a Gaussian Process whose covariance kernel can be expressed as linear combination of k. As such, the canonical semi-distance associated to it (here denoted $d^2_{\mathbf{x},\mathbf{x}'}$) inherits its regularity from Condition 2, which yields:

$$d^2_{\mathbf{x},\mathbf{x}'}(t,t') \leq 3C\|\mathbf{x}-\mathbf{x}'\|^{\alpha_1}_{\infty} \forall (t,t') \in \mathcal{I}^2 \tag{3.37}$$

$$d^2_{\mathbf{x},\mathbf{x}'}(t,t') \leq 4C\|t-t'\|^{\alpha_2}_{\infty} \forall (t,t') \in \mathcal{I}^2 \tag{3.38}$$

combining these bounds with Dudley's theorem and careful numerical development yields the required result. □

This bound on the expected value of the increments of a GP allows us to make a much stronger statement than the one in proposition 2.1.3.

Corollary 3.3.3. *If k satisfies Condition 2 and $Z \sim \mathcal{GP}(0,k)$, then for any $(\mathbf{x},\mathbf{x}') \in D^2$, the process $(Z_{\mathbf{x},t} - Z_{\mathbf{x}',t})_{t\in\mathcal{I}}$ is almost surely β-Hölder continuous for any $\beta < \frac{\alpha_1}{2}$*

Proof. To prove this result, we just need to combine the bound provided by Proposition 3.3.2 with Proposition 2.1.7 in Appendix. This induces the existence of a version $\tilde{Z}$ almost surely β-Hölder continuous. Then, $D \times \mathcal{I}$ being compact, it follows that Z and $\tilde{Z}$ are indistinguishable. □

From thereon, we will always work with assumptions ensuring the a.s. continuity of the GPs we work with. We will consider that our GPs of interest are also $\mathfrak{B}$-valued. Indeed, as stated in Remark 3.1.2, given an a.s. continuous GP Z it is always possible to construct a surely continuous GP $\tilde{Z}$ (and therefore $\mathfrak{B}$-valued) indistinguishable from it.

Theorem 3.3.4. *Let us consider a centred GP Z on $D \times \mathcal{I}$ whose covariance kernel k satisfies Condition 2. The SLGP induced by Z denoted here Y is almost surely in $\mathcal{A}^+_d(\mathcal{I}; D)$ and it is almost surely β-Hölder continuous for $\|\cdot\|_\infty$ and any $\beta < \frac{\alpha_1}{2}$.*

Proof. First, note that under Condition 2, Z is almost surely continuous (and hence a.s. bounded). This standard result of spatial statistics is recalled in Appendix, Proposition 2.1.3. It follows from it that Y is almost surely in$\mathcal{A}^+_d(\mathcal{I}; D)$ (and that we are in the setting of 3.2.4).
Now observe that we always have:

$$Z_{\mathbf{x},\cdot} - \|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty \leq Z_{\mathbf{x}',\cdot} \leq Z_{\mathbf{x},\cdot} + \|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty \tag{3.39}$$

with $\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty$ being possibly infinite on the null-set where Z is not continuous. As such:

$$\left| \frac{e^{Z_{\mathbf{x}',\cdot}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x}',u}}\, d\lambda(u)} - \frac{e^{Z_{\mathbf{x},\cdot}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \right| \leq \frac{e^{Z_{\mathbf{x},\cdot}}}{\int_{\mathcal{I}} e^{Z_{\mathbf{x},u}}\, d\lambda(u)} \left[e^{2\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty} - 1 \right] \tag{3.40}$$

By convexity of the exponential, we find that:

$$\|Y_{\mathbf{x},\cdot} - Y_{\mathbf{x}',\cdot}\|_\infty \leq \frac{e^{2\|Z\|_\infty}}{\lambda(\mathcal{I})} \left[2\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty + O(\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty^2) \right] \tag{3.41}$$

Combining the almost-sure boundedness of Z with Corollary 3.3.3 and the domain's compactness concludes the proofs. □

Remark. For simplicity of notation, we focused on the case where Z is a centred GP, but all properties are easily extended to the case where the mean of Z is β-Hölder continuous for any $\beta < \frac{\alpha_1}{2}$.

From the Proposition 3.3.2, we are also able to derive an analogue to scalar's mean square continuity, presented in the following section.

Mean power continuity of the Spatial Logistic Gaussian Process

We also leverage the bound on the expected value of the sup-norm of our increment field in our second contribution: we show that the Hölder conditions on k and k_{inc} are sufficient conditions for the mean power continuity of the SLGP.

Theorem 3.3.5 (Sufficient condition for mean power continuity of the SLGP)**.** *Consider the SLGP Y induced by a centred GP Z with covariance kernel k and assume that k satisfies Condition 2.*

Then, for all $\gamma > 0$ and $0 < \delta < \gamma\alpha_1/2$ (for Equation 3.42, resp. $0 < \delta < \gamma\alpha_1$ for Equations 3.43 and 3.44), there exists $K_{\gamma,\delta} > 0$ such that for all $\mathbf{x}, \mathbf{x}' \in D^2$:

$$\mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x} - \mathbf{x}'\|_\infty^{\gamma\alpha_1/2-\delta} \tag{3.42}$$

$$\mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x} - \mathbf{x}'\|_\infty^{\gamma\alpha_1-\delta} \tag{3.43}$$

$$\mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x} - \mathbf{x}'\|_\infty^{\gamma\alpha_1-\delta} \tag{3.44}$$

The main addition of this theorem, compared to the Proposition 3.3.4 is that it provides some control on the modulus of continuity. In theorem 3.3.5, we add to the almost-sure Hölder continuity by also providing rates on the dissimilarity between SLGPs considered at different $\mathbf{x}$'s. We give here the sketch of proof and refer the reader to appendix A.2 for detailed derivations.

Main elements for proving Theorem 3.3.5. The core idea of this proof is to leverage Lemma 3 and to apply Fernique's theorem. Careful analysis and further derivations enable us to prove that we have the following (tight) upper-bounds:

$$\begin{aligned}\mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty\right]^\gamma \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty\right]^{2\gamma} \\ \mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty\right]^{2\gamma}\end{aligned} \tag{3.45}$$

We combine this inequality with Proposition 3.3.2 to conclude the proof. □

Remark. The proof of this theorem consists in getting to the Inequalities in 3.45 and then leveraging Proposition 3.3.2. It is noteworthy to observe that the exact same proof structure can be applied, for instance, to prove that for a SLGP $Y' = \Psi[Z']$ and a density field f obtained by spatial logistic density transformation of a function g, $f = \Psi[g]$, if $\mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - g(\mathbf{x},\cdot)\|_\infty\right]^\gamma \to 0$ then for all $\mathbf{x}$:

$$\begin{aligned}\mathbb{E}\left[d_H(f(\mathbf{x},\cdot), Y_{\mathbf{x},\cdot})^\gamma\right] &\to 0 \\ \mathbb{E}\left[KL(f(\mathbf{x},\cdot), Y_{\mathbf{x},\cdot})^\gamma\right] &\to 0 \\ \mathbb{E}\left[d_{TV}(f(\mathbf{x},\cdot), Y_{\mathbf{x},\cdot})^\gamma\right] &\to 0\end{aligned} \tag{3.46}$$

Hence making these bounds applicable in the context of uniform approximation by a GP.

3.3.2 Posterior consistency for (logistic Gaussian) random measure fields

From now-on, we will consider that we have observations obtained by independent sampling of a reference field p_0 of pdfs on $\mathcal{I}$ indexed by D. By that, we mean that our dataset consist in n couples of locations and observations $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$, where the $\mathbf{x}_i$ are in D. Moreover, we assume the t_i's are obtained by independent sampling of random variables T_i with density $p_0(\mathbf{x}_i,\cdot)$. The (random) vectors of observations are denoted by $\mathbf{T} = (T_i)_{1\leq i\leq n}$ and $\mathbf{t} = (t_i)_{1\leq i\leq n}$ respectively, similarly the vector of concatenated sampled location will be denoted by $\mathbf{X} = (\mathbf{x}_i)_{1\leq i\leq n}$.

In this section, we follow the approach applied in Pati et al. (2013) and consider that the $\mathbf{x}_i$'s are i.i.d. realisations of some random variables $\mathbf{X}_i$ and we note Q their distribution. We assume that Q admits a density q with respect to the Lebesgue measure.

The distribution Q will be mostly instrumental, as it allows us to study joint densities rather than conditional density, and therefore apply Schwartz's

theorem (Schwartz, 1965). This theorem gives a general method for establishing consistency in non-parametric and semi-parametric problems, and is briefly recalled in Appendix A.1.

Definition 3.3.1 (Mapping conditional densities to joint densities)**.** We note Λ_q the map defined for any $f \in \mathcal{A}_d(D;\mathcal{I})$ by the relationship:

$$\forall \mathbf{x} \in D,\ \forall t \in \mathcal{I},\ \Lambda_q[f](\mathbf{x},t) = q(\mathbf{x})f(t|\mathbf{x}) \tag{3.47}$$

Whenever the linear map Λ_q is applied to a function $f \in \mathcal{A}_d(D;\mathcal{I})$ (i.e. a density on $\mathcal{I}$ indexed by D), it returns an element of $\mathcal{A}(D \times \mathcal{I})$ (i.e. a joint density on $D \times \mathcal{I}$).

We establish weak posterior consistency of the priors induced by the SLGP for a given class of function, meaning that this posterior consistency is achieved for the weak topology. The definition of this topology is recalled in Appendix A.1.

Proposition 3.3.6 (Weak consistency of the joint-density)**.** *Let Y be a SLGP on $\mathcal{I}$ indexed by D, and denote by Z one of the GPs inducing Y. Further assume that $Z \sim \mathcal{GP}(0,k)$ and that $\|Z\|_\infty < \infty$ a.s. For f_0 an element of the Reproducing Kernel Hilbert Space (RKHS) of k, the prior Π induced by $\Lambda_q[Y]$ achieves weak posterior consistency at $h_0 = \Lambda_q \circ \Psi[f_0]$.*

In order to prove this result, we need an intermediate result about logistic transforms, namely:

Lemma 4. *For any two functions $f_1, f_2 : (D \times \mathcal{I}) \to \mathbb{R}$ exponentially integrable alongside $\mathcal{I}$ and any $\epsilon > 0$*

$$\|f_1 - f_2\|_\infty \leq \epsilon \Rightarrow \left|\log\left(\frac{\Lambda_q \circ \Psi[f_1]}{\Lambda_q \circ \Psi[f_2]}\right)\right| = \left|\log\left(\frac{\Psi[f_1]}{\Psi[f_2]}\right)\right| \leq 2\epsilon \tag{3.48}$$

Proof of Lemma 4. By definitions of the spatial logistic density transform and of Λ:

$$\left|\log\left(\frac{\Lambda_q \circ \Psi[f_1]}{\Lambda_q \circ \Psi[f_2]}\right)\right| = \left|\log\left(\frac{q(\cdot)\int_{\mathcal{I}} e^{f_2(\cdot,u)-f_2(\cdot,\cdot)}\,du}{q(\cdot)\int_{\mathcal{I}} e^{f_1(\cdot,v)-f_1(\cdot,\cdot)}\,dv}\right)\right| \tag{3.49}$$

$$\leq \left|\log\left(\frac{\int_{\mathcal{I}} e^{\|f_1-f_2\|_\infty}\,du}{\int_{\mathcal{I}} e^{-\|f_1-f_2\|_\infty}\,dv}\right)\right| \tag{3.50}$$

$$\leq 2\|f_1 - f_2\|_\infty \leq 2\epsilon \tag{3.51}$$

□

Proof of proposition 3.3.6. Since we endowed $\mathcal{A}(D \times \mathcal{I})$ with the weak convergence topology, we can apply Schwartz's theorem, as long as we prove that for all $\epsilon > 0$:

$$\Pi\left[h \in \mathcal{A}(D \times \mathcal{I}), KL(h_0, h) < \epsilon\right] > 0 \tag{3.52}$$

In our case, it will be sufficient to only consider joint densities that are strictly positive on $D \times \mathcal{I}$, as they can be written $h = \Lambda \circ \Psi[f]$ for some $f : D \times \mathcal{I} \to \mathbb{R}$. Applying lemma 4 allows for rewriting the quantity of interest as:

$$P\left[\|Z - w_0\|_\infty < \epsilon\right] > 0 \tag{3.53}$$

This corresponds to the small ball probabilities for Gaussian processes and the property holds, as recalled in Proposition 2.2.5. □

Corollary 3.3.7 (Weak consistency of the probability density field)**.** *Let Y be a SLGP on $\mathcal{I}$ indexed by D, and denote by Z one of the GPs inducing Y. Further assume that $Z \sim \mathcal{GP}(0, k)$ and that $\|Z\|_\infty < \infty$ a.s. For f_0 an element of the Reproducing Kernel Hilbert Space (RKHS) of k, the prior Π_d induced by Y achieves weak posterior consistency at $\Psi[f_0]$.*

This ensures that asymptotically the SLGP models will be able to recover a large class of probability density fields.

Chapter 4

SLGP fitting under finite-rank kernels: parametrization and inference

In this Chapter, we discuss implementation choices and important properties motivating them. Note that all the codes used for running experiments in this Chapter and the next one are available on the GitHub repository Gautier, Athénaïs (2023).

We will consider for simplicity that our SLGP is obtained by transforming almost-surely continuous GPs on $[0,1]^{d_x+d_t}$. As such, we can relax the tone adopted in Chapter 3 in favour of less mathematical yet sound phrasing. We will see throughout this Section that our choices of GPs ensure that the a.s. continuity assumption is satisfied.

4.1 General considerations

4.1.1 Data integration

From now on, we will consider that our observations are obtained by independent sampling of a reference probability density field p_0. By that, we mean that our dataset consist in n couples of locations and observations $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$, where the $\mathbf{x}_i$ are in $[0,1]^{d_x}$. Moreover, we assume the t_i's are obtained by independent sampling of random variables T_i with respective densities $p_0(\mathbf{x}_i, \cdot)$. The random vector of observations is denoted by $\mathbf{T} = (T_i)_{1\leq i\leq n}$, while we use $\mathbf{t} = (t_i)_{1\leq i\leq n}$ for a realisation of $\mathbf{T}$.

For a suitable covariance kernel k, consider the following hierarchical model:

$$\begin{cases} Z \sim \mathcal{GP}(0, k) \\ \pi\left(T_i = t | Z\right) = \dfrac{e^{Z(\mathbf{x}_i, t)}}{\int_{[0,1]^{d_t}} e^{Z(\mathbf{x}_i, u)}\, du} \quad (t \in [0,1]^{d_t}, 1 \leq i \leq n) \end{cases} \tag{4.1}$$

Assuming that the observations stem from the model, and leveraging their independence, we obtain density of observations knowing the underlying GP:

$$\pi\left(\mathbf{T} = t | Z\right) = \prod_{i=1}^{n} \frac{e^{Z_{\mathbf{x}_i, t_i}}}{\int_{[0,1]^{d_t}} e^{Z_{\mathbf{x}_i, u}}\, du} \tag{4.2}$$

Implementation of the density field estimation can be done through MCMC sampling but causes two main issues that we will address now.

4.1.2 Hyperparameters estimation

The first problem posed is that in almost all realistic cases, GPs depend on some unknown hyperparameters that need to be estimated. This issue can be addressed in a Bayesian way, by specifying a prior on the hyperparameters. Typically, k admits a variance parameter σ^2 and other parameters, that we will denote $\boldsymbol{\theta}$. To highlight this dependency, we use the notation $k = \sigma^2 k_\theta$. We introduce an augmented Bayesian model that allows for hyperparameters estimation. It requires practitioners to specify prior beliefs on both σ^2 and θ through prior distributions.

$$\begin{cases} \sigma \sim \pi_\sigma \text{ and } \boldsymbol{\theta} \sim \pi_{\boldsymbol{\theta}} \\ Z | \sigma, \boldsymbol{\theta} \sim \mathcal{GP}(0, \sigma^2 k_\theta) \\ \pi\left(T_i = t | Z\right) = \dfrac{e^{Z(\mathbf{x}_i, t)}}{\int_{[0,1]^{d_t}} e^{Z(\mathbf{x}_i, u)}\, du} \; t \in [0,1]^{d_t}, 1 \leq i \leq n \end{cases} \tag{4.3}$$

The conditional density of observations here has a similar expression to the previous one:

$$\pi\left(\mathbf{T} = t | Z, \sigma, \theta\right) = \prod_{i=1}^{n} \frac{e^{Z_{\mathbf{x}_i, t_i}}}{\int_{[0,1]^{d_t}} e^{Z_{\mathbf{x}_i, u}}\, du} \tag{4.4}$$

In the context of Gaussian Processes, suitable choices of prior distributions have been previously discussed in the literature. Researchers have proposed various methods for selecting priors, such as the principled approach outlined in Berger et al. (2001) and more recent work on extending the Penalized Complexity framework to 3D GPs as in Fuglstad et al. (2019). We discuss the choices we made in Section 4.3.

4.1.3 Dimensionality of the problem

The remaining issue with this hierarchical model lies on the fact that the integrals in Equations 4.2 and 4.4 involve values of Z over the whole response domain. This infinite dimensional object makes likelihood-based computations cumbersome.

We propose a way to reduce the dimensionality by considering only finite rank Gaussian Processes.

Definition 4.1.1 (Finite-rank Gaussian Processes)**.** A Gaussian Process Z on a generic domain S is called a finite rank Gaussian Process if there exists $p \in \mathbb{N}$, and a family of functions $(f_i)_{1\leq i\leq p}$ on S such that

$$Z(s) = \sum_{j=1}^{p} f_j(s)\varepsilon_j,\ \forall s \in S \tag{4.5}$$

where $\boldsymbol{\varepsilon}$ is a random vector of p i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Remark. Note that from now on, we should use the notation ε (resp. $\boldsymbol{\varepsilon}$) to denote the random variable (resp. random vector), and ϵ (resp. $\boldsymbol{\epsilon}$) for fixed values and realisations thereof.

In particular, we will be interested on finite-rank GPs on $D \times \mathcal{I} = [0,1]^{d_x+d_t}$, and as such consider function $(f_{i,\theta})_{1\leq i\leq p}$ on $[0,1]^{d_x+d_t}$. This yields finite-rank GPs that we can write as follows:

$$Z(\mathbf{x}, t) = \sum_{j=1}^{p} f_{j,\theta}(\mathbf{x}, t)\varepsilon_j = \boldsymbol{\varepsilon}^\top F_\theta(\mathbf{x}, t),\ \forall \mathbf{x} \in D,\ t \in \mathcal{I}, \tag{4.6}$$

where $F_\theta(\mathbf{x}, t) := (f_{j,\theta}(\mathbf{x}, t))_{1\leq j\leq p}$ is the vector of basis functions evaluated at $(\mathbf{x}, t)$ and $\boldsymbol{\varepsilon}$ is a random vector of p i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Note that we assumed that the GP's dependency on the hyperparameters θ is only expressed through the deterministic basis functions.

Remark. When the $f_{i,\theta}$ are L^2 orthonormal, this coincides with the Karhunen-Loève expansion of the GP, as introduced in Proposition 2.2.6. However, for most kernels, this expansion is not analytically known.

Lemma 5. *A finite rank GP defined as in Equation 4.6 has the following covariance kernel:*

$$Cov(Z(\mathbf{x}, t), Z(\mathbf{x}', t')) = \sigma^2 \sum_{j=1}^{p} f_{j,\theta}(\mathbf{x}, t) f_{j,\theta}(\mathbf{x}', t') \tag{4.7}$$

The posterior distribution of $\boldsymbol{\varepsilon}, \sigma, \theta$, given data $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$ can be obtained by replacing Z's expression in Equation 4.4 and denoting by ϕ_p the pdf of the p-variate standard normal distribution:

$$\pi[\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}|\mathbf{T} = \mathbf{t}] \propto \pi_\sigma(\sigma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\phi_p\left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)\prod_{i=1}^{n}\frac{e^{\sum_{j=1}^p \epsilon_j f_{j,\theta}(\mathbf{x}_i,t_i)}}{\int_{\mathcal{I}} e^{\sum_{j=1}^p \epsilon_j f_{j,\theta}(\mathbf{x}_i,u)}\,du} \tag{4.8}$$

4.2 Basis functions considered

Although theoretically, any basis function could be used in the implementation, practice is different. Indeed, due to the exponential transformation, practitioners need to be careful with the numerical stability of their model. We identified some families of functions that seem to display sufficiently nice behaviours and make implementation less prone to numerical overflow. From here on, we review such approaches.

4.2.1 Leveraging inducing points

In (Tokdar, 2007) a finite rank approach leveraging inducing points is used. A moderate number of inducing points is introduced to reduce the dimensionality of the problem for logistic Gaussian Processes (with no more than a hundred points). We claim that this approach can be considered as a particular, adaptive choice of equation 4.6. Indeed, let us consider a Gaussian Process $Z \sim \mathcal{GP}(0, k)$ over a general domain T. For arbitrary indices $y_1, ...y_p \in T$, we can introduce $\mathbf{Z}_p := (Z_{y_1}, ..., Z_{y_p})\top$ and informally say that Z can be approximated by $W := \mathbb{E}[Z|\mathbf{Z}_p]$.

The conditioning formula in the Gaussian setting are well known and give us a closed-form formula for W: $W_y = k_p(y)^\top K^{-1}\mathbf{Z}_p$ with $K = (k(y_i, y_j))_{1\leq i,j\leq p}$ the covariance matrix of the chosen design and $k_p(y) := (k(y_i, y))_{1\leq i\leq p}$ for $y \in T$. We can rewrite this equation by introducing $\mathbf{X}_p = K^{-1/2}\mathbf{Z}_p$ a multivariate standard normal. In this case, we get $W_y = k_p(y)^\top K^{-1/2}X_p$. Therefore, setting $f_i(y)$ to be the i-th coordinate of the vector $k_p(y)^\top K^{-1/2}$ yields that $W_y = \sum_{i=1}^p X_i f_i(y)$ where the X_i's are i.i.d. $\mathcal{N}(0, 1)$.

4.2.2 Fourier functions

We propose using Fourier-type basis, i.e. collections of: $(\cos(\boldsymbol{\omega}^\top[\mathbf{x}, t]), \sin(\boldsymbol{\omega}^\top[\mathbf{x}, t]))$, for varying $\boldsymbol{\omega}$'s in $\mathbb{R}^{d_x+1}$. We implement three different ways to select these angular frequencies. The first one being inspired from the discrete Fourier basis,

while the other two are variations within the Random Fourier Feature framework. Finally, to simplify notations we will mostly focus on the setting where $\mathcal{I} = [0, 1]$.

Multivariate adaptation of the discrete Fourier basis For this approach, that is a multivariate extension of the unidimensional discrete Fourier basis, we consider that the domain D is $[0, 1]^{d_x}$. We specify two parameters, noted $p_{\mathbf{x}}$ and p_t that determine the frequencies considered in $\mathbf{x}$ and t. Then, we consider all the $\boldsymbol{\omega} \in \left\{-\frac{2\pi(p_{\mathbf{x}}-1)}{p_{\mathbf{x}}}, ..., \frac{2\pi(p_{\mathbf{x}}-1)}{p_{\mathbf{x}}}\right\}^{d_x} \otimes \left\{2\pi, ..., \frac{2\pi(p_t-1)}{p_t}\right\}$. This approach yields $(2p_{\mathbf{x}} - 1)^d(p_t - 1)$ frequencies and therefore twice as many basis functions. This family of functions excludes non-positive frequencies in t for two reasons. First, excluding negative values allows for avoiding redundancy by ensuring that one can not have both $\boldsymbol{\omega}$ and$-\boldsymbol{\omega}$ as frequency. Second, we also avoid functions independent of t, as they would be cancelled-out within the normalisation step.

When using a finite rank Gaussian Process relying on these basis functions, we recommand to weight each function depending on its frequency, but to estimate only one hyperparameter: a common variance parameter. Indeed, one could imagine having one variance parameter per function, but it would lead to estimating $(2p_{\mathbf{x}} - 1)^d(p_t - 1)$ hyper-parameters, which would prove numerically costly.

Random Fourier Features

As mentioned in Subsection 2.2.3, the framework of Random Fourier Features yields one way to construct finite rank GPs that "resemble" GPs with a prescribed covariance kernel. We propose constructing such a RFF inspired GP. For a given kernel k, we denote ω_i's draws of independent random variables that have a density equal to the spectral density of k. We consider the basis functions given for any $\mathbf{x} \in D$ and $t \in \mathcal{I}$ by:

$$\boldsymbol{\varphi}([\mathbf{x}, t]) = \left[\cos(\omega_1^\top[\mathbf{x}, t]), ..., \cos(\omega_p^\top[\mathbf{x}, t]), \sin(\omega_1^\top[\mathbf{x}, t]), ..., \sin(\omega_p^\top[\mathbf{x}, t])\right] \quad (4.9)$$

Then, let us define the process:

$$Z_{RFF,\mathbf{x},t} = \frac{\sigma}{\sqrt{p}}\boldsymbol{\varphi}([\mathbf{x}, t])^\top\boldsymbol{\varepsilon} \quad (4.10)$$

where $\boldsymbol{\varepsilon}$ is a 2p-variate standard normal vector. Z_{RFF} is a Gaussian Process with mean zero and covariance kernel $k_{RFF}([\mathbf{x}, t], [\mathbf{x}', t'])$.

One can easily prove that:

$$k_{RFF}([\mathbf{x}, t], [\mathbf{x}', t']) = \frac{\sigma^2}{p}\sum_{i=1}^{p}\cos(\omega_i^\top[\mathbf{x} - \mathbf{x}', t - t']) \quad (4.11)$$

This is a Monte-Carlo approximation of the Bochner integral of k, as recalled in Equation 2.34 and in subsection 2.2.3 of Chapter 2

Space filling random Fourier Features

Rather than randomly sampling from the spectral density to get to Equation 2.34, one can instead aim at selecting a set of points as diverse as possible that adequately reflect the (spectral) density at hand. One framework that tackles such challenges is that of Space-filling designs (with respect to a prescribed measure). Most commonly, it is studied and implemented with the Lebesgue measure (i.e. uniform distribution) being the target measure. However, we will mostly be interested in Matérn kernel in this document, and as such need to work with respect to multivariate Student distribution (as pointed out in Section 2.2.4). Thankfully, we can fully leverage existing codes and adapt them to our setting, thanks to the following observation:

Lemma 6. *Let $\mathbb{X} := (X_1, ..., X_p)^\top$ follow a d-variate Student distribution, with 2ν degrees of freedom, location 0 and scale matrix identity.*

- *Any marginal is a univariate Student distribution. In particular, X_1 follows a univariate Student with 2ν degrees of freedom.*

- *The conditionals are multivariate Student distributions. In particular, for $i \geq 2$, the re-scaled variable $\sqrt{\frac{2\nu+i-1}{2\nu+\sum_{j=1}^{i-1} X_j^2}} X_i \Big| X_1, ..., X_{i-1}$ follows a univariate student distribution with $2\nu + i - 1$ degrees of freedom*

Using our knowledge of the marginals and conditionals of the multivariate Student distribution, and applying the Rosenblatt transformation (Rosenblatt, 1952) allows for transforming any Space-filling design with respect to the uniform distribution into one with respect to the multivariate Student. Differences between frequencies obtained through random sampling and those with a space-filling design is displayed in Figure 4.1. In this work, we use maximin-LHS designs as implemented in the R package DiceDesign (Dupuy et al., 2015)

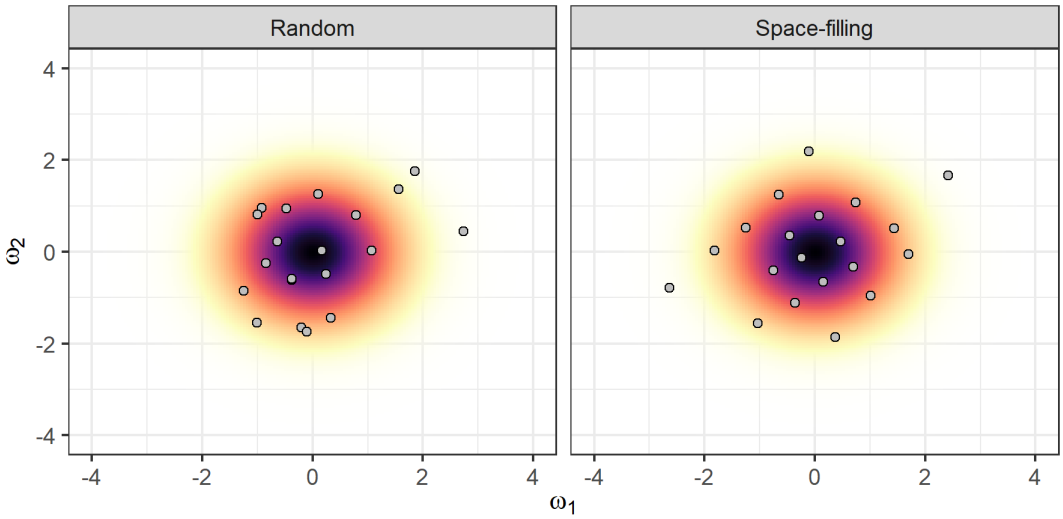


Figure 4.1: Illustrating Random versus Space-filling Fourier Features with 20 sampled frequencies in a 2D Matérn 5/2 kernel. The coloured background corresponds to the spectral density at hand.

4.3 Implementation

4.3.1 Maximum a posteriori estimation

Here, we propose to find the values of σ^2, θ and ε that maximize the (unnormalised) posterior density. It allows us to perform reasonably fast density field estimation, or to initialize our MCMC with this value of $\boldsymbol{\epsilon}$ rather than with an arbitrary value.

Maximizing the (unnormalised) posterior density is equivalent to minimizing its negative log. We will favour this approach, both because it yields simpler computations, but also because it is more numerically stable. Note that the right-hand term of Equation 4.4 can be split in two parts, one relative to the prior, the other to the likelihood term. We will focus mostly on the likelihood aspect, which is at the core of this contribution.

Negative log likelihood Under model 4.3 assumptions, the likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t}) = \prod_{i=1}^{n} \frac{e^{\sum_{j=1}^{p} \epsilon_j f_{j,\theta}(\mathbf{x}_i, t_i)}}{\int_{\mathcal{I}} e^{\sum_{j=1}^{p} \epsilon_j f_{j,\theta}(\mathbf{x}_i, u)} \, du} \tag{4.12}$$

As is standard in computational statistics, instead of studying the likelihood as it is written in Equation 4.12, we will rather consider the negative log-likelihood. We simplify notation by introducing $F_\theta(\mathbf{x}, t)$, the vector with coordinates $f_{j,\theta}(\mathbf{x}, t)$, and the average observed value of it: $\bar{F}_\theta = \left(\bar{F}_{j,\theta}\right)_{1\leq j\leq p} := \frac{1}{n}\sum_{i=1}^{n} F_\theta(\mathbf{x}_i, t_i)$. Then, the negative log-likelihood can be written as:

$$\ell(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t}) \;\; = -\boldsymbol{\epsilon}^\top \bar{F}_\theta + \sum_{i=1}^{n} \log\left(\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^\top F_\theta(\mathbf{x}_i, u)}\, du\right) \tag{4.13}$$

Let us also consider that among $\mathbf{x}_1, ..., \mathbf{x}_n$ there are only K distinct values noted $\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_K$. Furthermore, consider that at $\tilde{\mathbf{x}}_k$, there are n_k distinct observations. This allows us to rewrite equation 4.13 as:

$$\ell(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t}) \;\; = -\boldsymbol{\epsilon}^\top \bar{F}_\theta + \sum_{k=1}^{K} n_k \log\left(\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, u)}\, du\right) \tag{4.14}$$

This term does not display dependencies on σ^2, but possesses a property that will be crucial to ensure the good behaviour of our estimations.

Theorem 4.3.1. *For fixed data* $\mathbf{t}$ *and hyperparameters* σ *and* $\boldsymbol{\theta}$*, the negative log-likelihood function*

$$\boldsymbol{\epsilon} \mapsto \ell(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t}) \tag{4.15}$$

is convex.

Equivalently, the likelihood negative log-likelihood function at fixed hyperparameters

$$\boldsymbol{\epsilon} \mapsto \mathcal{L}(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t}) \tag{4.16}$$

is log-concave.

Proof. To prove this statement, we compute the gradient and Hessian of the negative log likelihood.

Its gradients write:

$$\frac{\partial \ell(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t})}{\partial \epsilon_i} = -\bar{F}_{i,\theta} + \sum_{k=1}^{K} n_k \int_{\mathcal{I}} f_{i,\theta}(\tilde{\mathbf{x}}_k, u) \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, v)}\, dv}\, du \tag{4.17}$$

The second order derivatives are:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\epsilon}, \sigma, \boldsymbol{\theta}; \mathbf{t})}{\partial \epsilon_i \partial \epsilon_{i'}} &= \sum_{k=1}^{K} n_k \int_{\mathcal{I}} f_{i,\theta}(\tilde{\mathbf{x}}_k, u) f_{i',\theta}(\tilde{\mathbf{x}}_k, u) \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, v)}\, dv}\, du \\ - \sum_{k=1}^{K} n_k &\left(\int_{\mathcal{I}} f_{i,\theta}(\tilde{\mathbf{x}}_k, u) \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, v)}\, dv}\, du\right)\left(\int_{\mathcal{I}} f_{i',\theta}(\tilde{\mathbf{x}}_k, u) \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k, v)}\, dv}\, du\right)\end{aligned} \tag{4.18}$$

For a given $\boldsymbol{\epsilon}$, let us introduce $Y_1, ...Y_k$, k random variables with respective probability density $\frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}$. Then:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\epsilon},\sigma,\boldsymbol{\theta};\mathbf{t})}{\partial\epsilon_i\partial\epsilon_{i'}} &= \sum_{k=1}^{K} n_k \left(\mathbb{E}\left[f_{i,\theta}(\tilde{\mathbf{x}}_k,Y_i)f_{i',\theta}(\tilde{\mathbf{x}}_k,Y_{i'})\right] - \mathbb{E}\left[f_{i,\theta}(\tilde{\mathbf{x}}_k,Y_i)\right]\mathbb{E}\left[f_{i',\theta}(\tilde{\mathbf{x}}_k,Y_{i'})\right]\right)\\ &= \sum_{k=1}^{K} n_k \mathrm{Cov}\left(f_{i,\theta}(\tilde{\mathbf{x}}_k,Y_i), f_{i',\theta}(\tilde{\mathbf{x}}_k,Y_{i'})\right)\end{aligned} \tag{4.19}$$

Since the Hessian matrix is a (sum of) covariance matrices, it inherits their symmetry and p.d-ness. This suffices to prove that the negative log-likelihood is convex. □

Remark. Other partial derivatives are not necessary to prove Theorem 4.3.1, but still useful for implementation. We give them here.

$$\frac{\partial \ell(\boldsymbol{\epsilon},\sigma,\boldsymbol{\theta};\mathbf{t})}{\partial\theta_i} = -\boldsymbol{\epsilon}^\top \frac{\partial \bar{F}_\theta}{\partial\theta_i} + \sum_{k=1}^{K} n_k \int_{\mathcal{I}} \boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_i} \frac{e^{\boldsymbol{\epsilon}^T f(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T f(\tilde{\mathbf{x}}_k,v)}\,dv}\,du \tag{4.20}$$

where we use the informal notation $\frac{\partial F_\theta(\mathbf{x},t)}{\partial\theta_i}$ to denote the vector with coordinates $\frac{\partial f_{j,\theta}(\mathbf{x},t)}{\partial\theta_i}$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\epsilon},\sigma,\boldsymbol{\theta};\mathbf{t})}{\partial\theta_i\partial\theta_{i'}} &= \sum_{k=1}^{K} n_k \int_{\mathcal{I}} \left[\boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_i}\boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_{i'}} + \frac{\partial^2 F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_i\partial\theta_{i'}}\right] \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\\ &- \sum_{k=1}^{K} n_k \left(\int_{\mathcal{I}} \boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_i} \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\right)\left(\int_{\mathcal{I}} \boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_{i'}} \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\right)\\ &- \boldsymbol{\epsilon}^\top \frac{\partial^2 \bar{F}_\theta}{\partial\theta_i\partial\theta_{i'}}\end{aligned} \tag{4.21}$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\epsilon},\sigma,\boldsymbol{\theta};\mathbf{t})}{\partial\epsilon_i\partial\theta_{i'}} &= \sum_{k=1}^{K} n_k \int_{\mathcal{I}} \left[f_{i,\theta}(\tilde{\mathbf{x}}_k,u)\boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_{i'}} + \frac{\partial f_{i,\theta}(\tilde{\mathbf{x}}_k,u)}{\partial\theta_{i'}}\right] \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\\ &- \sum_{k=1}^{K} n_k \left(\int_{\mathcal{I}} f_{i,\theta}(\tilde{\mathbf{x}}_k,u) \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\right)\left(\int_{\mathcal{I}} \boldsymbol{\epsilon}^\top \frac{\partial F_\theta(\tilde{\mathbf{x}}_k,u)}{\partial\theta_{i'}} \frac{e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,u)}}{\int_{\mathcal{I}} e^{\boldsymbol{\epsilon}^T F_\theta(\tilde{\mathbf{x}}_k,v)}\,dv}\,du\right)\\ &- \frac{\partial \bar{F}_{i,\theta}}{\partial\theta_{i'}}\end{aligned} \tag{4.22}$$

Knowing the partial derivatives allow us to perform Maximum a posteriori estimation of the parameters. This gives us a point-estimate of the field.

Our implementation of the MAP estimation In our implementation, we relied on results from van der Vaart and van Zanten (2009) showing that the inference asymptotically behaves well when performing LGP-based density estimation with an inverse gamma distribution on the lengthscale hyperparameter. Additionally, we have experimented with different priors for the variance, including those available in the R packages INLA (Lindgren and Rue, 2015) and RSTAN (Stan Development Team, 2022). However, although the negative-log likelihood is convex in $\boldsymbol{\epsilon}$, it is not in $\boldsymbol{\theta}$, making optimisation much slower when considering the lengthscale. We also found that the numerical stability of the SLGP was greatly impacted by the choice of the variance, due to the exponentiation. Therefore, we have decided to fix the variance at a relatively low value to ensure numerical stability, rather than performing Bayesian inference on it. After performing preliminary experiences, it appears that selecting a variance such that $\text{Median}\left[\max_{\mathbf{x}\in D, t\in\mathcal{I}} \boldsymbol{\varepsilon}^\top F_\theta(\mathbf{x},t) - \min_{\mathbf{x}\in D, t\in\mathcal{I}} \boldsymbol{\varepsilon}^\top F_\theta(\mathbf{x},t)\right] \approx 5$ already gives a lot of flexibility to SLGPs models (recall that $e^5 \approx 148$) while preventing numerical instability. This value seems like a reasonable heuristic in our current setting, and severely alleviate technical issues when computing the normalising term at the denominators of SLGPs.

Also note that this selection of the variance is not directly data-dependent, and is mostly here to ensure the numerical stability of the prior. In our current setting, σ^2 depends only on the choice of basis functions and lengthscales.

Hence, our current optimisation relies on grid search and heuristics and can be summarised as follows:

Algorithm 1: SLGP-based maximum a posteriori estimation of the underlying density, with grid search for θ and heuristic for σ^2.

input : Grid Θ of values for θ, dataset $\{(\mathbf{x}_i, t_i)\}_{1 \leq i \leq n}$, basis functions $(f_{i,\theta})_{1 \leq i \leq p}$

for $\theta \in \Theta$ **do**

- Draw n_{sim} i.i.d. realisations of $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(0, 1)$.
- Evaluate the basis function on a coarse grid covering the whole domain $D \times \mathcal{I}$
- Use the last two steps to estimate:
  $\mathfrak{m}_\theta := \text{Median}\left[\max_{\mathbf{x} \in D, t \in \mathcal{I}} \boldsymbol{\varepsilon}^\top F_\theta(\mathbf{x}, t) - \min_{\mathbf{x} \in D, t \in \mathcal{I}} \boldsymbol{\varepsilon}^\top F_\theta(\mathbf{x}, t)\right]$.
- Set $\sigma_\theta = \frac{5}{\mathfrak{m}_\theta}$, to enforce numerical stability of the prior
- Perform (gradient-based) optimisation to get $\boldsymbol{\epsilon}^*_\theta$ at θ and σ^2_θ
- Store θ, σ^2_θ, $\boldsymbol{\epsilon}^*_\theta$ and the value $\pi[\boldsymbol{\epsilon}^*_\theta, \sigma_\theta, \theta | \mathbf{T} = \mathbf{t}]$

output: θ^*, $\sigma^2_{\theta^*}$ and $\boldsymbol{\epsilon}^*_{\theta^*}$ that give the highest value of $\pi[\boldsymbol{\epsilon}^*_{\theta^*}, \sigma_{\theta^*}, \theta^* | \mathbf{T} = \mathbf{t}]$

4.3.2 Markov Chain Monte Carlo

An alternative to MAP estimation is the so-called Markov Chain Monte Carlo Sampling. Although computationally more expensive than a gradient-based estimation, it yields a probabilistic prediction of the probability distribution field.

A brief overview of Markov Chain Monte Carlo algorithms

Markov Chain Monte Carlo (MCMC) is a method for estimating probability distributions by generating random samples from those distributions. It works by constructing a Markov Chain, where each step in the chain is a random transition to a new state. The chain is then run for a sufficient number of iterations to converge to a stationary distribution, which is an approximation of the target probability distribution. This stationary distribution can be used to estimate quantities of interest, such as means and variances. MCMC is widely used in Bayesian statistical modelling and other fields where the direct calculation of a probability distribution is intractable.

We identify several algorithms that could be relevant to our setting and present various degrees of computational efficiency and computation cost.

Metropolis-Hastings The simplest algorithm for MCMC is the so-called Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). We

present here the essence of MH for sampling some parameters $\boldsymbol{\epsilon}$ based on data $\mathcal{D}$ (here, we have $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}_{1 \leq i \leq n}$. Note that we kept the notation relevant to this thesis but that the algorithm is presented in the general setting, and is not dependent on a particular form of the posterior.

Algorithm 2: Metropolis-Hastings algorithm

input : Unnormalised posterior $\pi\left[\boldsymbol{\epsilon}|\mathcal{D}\right]$, number of iterations T, proposal density $q(\boldsymbol{\epsilon}|\boldsymbol{\epsilon}')$, initial value $\boldsymbol{\epsilon}^{(0)}$

for $1 \leq i \leq T$ **do**

Simulate $\boldsymbol{\epsilon}' \sim q(\cdot|\boldsymbol{\epsilon}^{(i-1)})$ and $u \sim \mathcal{U}([0,1])$

if $u \leq \frac{\pi[\boldsymbol{\epsilon}'|\mathcal{D}]q(\boldsymbol{\epsilon}^{(i-1)}|\boldsymbol{\epsilon}')}{\pi\left[\boldsymbol{\epsilon}^{(i-1)}|\mathcal{D}\right]q(\boldsymbol{\epsilon}'|\boldsymbol{\epsilon}^{(i-1)})}$ **then**

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}'$

else

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}^{(i-1)}$

output: Samples $\boldsymbol{\epsilon}^{(i)}$ from the posterior

The performance of the MH algorithm is sensitive to the choice of the proposal distribution $q(\boldsymbol{\epsilon}|\boldsymbol{\epsilon}')$. It is common to use $q(\boldsymbol{\epsilon}|\boldsymbol{\epsilon}') = \mathcal{N}(\boldsymbol{\epsilon}', \Sigma)$ with carefully tuned Σ.

A class of algorithms called adaptive Metropolis algorithms (Haario et al., 2001, 2006; Andrieu and Thoms, 2008) modify the proposal distribution during the simulation in order to improve its efficiency. The idea behind adaptive MCMC is to adjust the proposal distribution in a way that reduces the correlation between consecutive samples, speeds up convergence, and increases the acceptance rate of candidate states.

The MH's efficiency can be improved upon whenever using specific priors or posteriors. In particular, whenever the target posterior has density with respect to a Gaussian process or Gaussian random field reference measure, one can use algorithms that come with faster convergence.

Preconditioned Crank Nicholson The goal of preconditioned Crank Nicholson (pCN) for MCMC is to improve the convergence and efficiency of the MCMC algorithm by transforming the target distribution in a way that makes it easier to explore (Neal, 1998; Beskos et al., 2008; Cotter et al., 2013). The transformed distribution leads to a more efficient exploration of the state space, which can result in faster convergence and improved mixing of the Markov Chain.

Algorithm 3: Preconditioned Crank Nicholson algorithm

input : Unnormalised posterior $\pi[\boldsymbol{\epsilon}|\mathcal{D}]$, number of iterations T, tuning-parameter $\beta > 0$, proposal covariance Σ, initial value $\boldsymbol{\epsilon}^{(0)}$

for $1 \leq i \leq T$ **do**

Simulate $\boldsymbol{\epsilon}' \sim \mathcal{N}(\sqrt{1-\beta^2}\boldsymbol{\epsilon}^{(i-1)}, \beta^2\Sigma)$ and $u \sim \mathcal{U}([0,1])$

if $u \leq \frac{\pi[\boldsymbol{\epsilon}'|\mathcal{D}]\exp\left\{\boldsymbol{\epsilon}^{(i-1)\top}\Sigma\boldsymbol{\epsilon}^{(i-1)}/2\right\}}{\pi\left[\boldsymbol{\epsilon}^{(i-1)}|\mathcal{D}\right]\exp\left\{\boldsymbol{\epsilon}'^{\top}\Sigma\boldsymbol{\epsilon}'/2\right\}}$ **then**

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}'$

else

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}^{(i-1)}$

output: Samples $\boldsymbol{\epsilon}^{(i)}$ from the posterior

The pCN differs only slightly from the MH algorithm, with the main difference being that the proposal is not centred anymore. Although one may still have to tune the step size parameter β and design Σ to achieve a desired level of statistical efficiency, the performance of the pCN method is robust to the dimension of the sampling problem being considered. As for MH, generalisations (Rudolf and Sprungk, 2018) and adaptive versions (Chen et al., 2016) of the algorithm have been proposed to improve efficiency.

Yet another way to leverage properties of the posterior is to used gradient-informed methods

Metropolis Adjusted Langevin Algorithm The main idea behind MALA (Roberts and Stramer, 2002) is to use the gradient information of the target distribution to construct a proposal distribution that is closer to the target distribution. This leads to a higher acceptance rate of candidate states and faster convergence of the MCMC algorithm.

In MALA, a candidate state is generated by proposing a small step in the direction of the gradient of the target distribution, followed by a random perturbation. The acceptance-rejection rule is based on the Metropolis algorithm, and the candidate state is accepted with a probability proportional to the ratio of the target distribution evaluated at the candidate state and the current state.

MALA has several advantages over traditional MCMC algorithms, such as the Metropolis algorithm or the Gibbs sampler. It is especially effective for problems with highly correlated variables, where traditional MCMC algorithms can be slow and inefficient. However, MALA requires the gradient information of the target distribution, which can be computationally expensive to calculate, especially for high-dimensional problems.

Algorithm 4: Metropolis Adjusted Langevin algorithm

input : Unnormalised posterior $\pi\left[\boldsymbol{\epsilon}|\mathcal{D}\right]$, gradient $\nabla\pi\left[\boldsymbol{\epsilon}|\mathcal{D}\right]$, number of iterations T, step-size $\beta > 0$, proposal covariance Σ, initial value $\boldsymbol{\epsilon}^{(0)}$

for $1 \leq i \leq T$ **do**

Compute the gradient $G_{i-1} := \nabla\pi\left[\boldsymbol{\epsilon}^{(i-1)}|\mathcal{D}\right]$

Simulate $\boldsymbol{\epsilon}' \sim \mathcal{N}(\boldsymbol{\epsilon}^{(i-1)} + \beta G_{i-1}, 2\beta I_p)$ and $u \sim \mathcal{U}([0,1])$

if $u \leq \frac{\pi[\boldsymbol{\epsilon}'|\mathcal{D}]}{\pi\left[\boldsymbol{\epsilon}^{(i-1)}|\mathcal{D}\right]}$ **then**

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}'$

else

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}^{(i-1)}$

output: Samples $\boldsymbol{\epsilon}^{(i)}$ from the posterior

While MALA uses a simple diffusion process as a proposal, there exist other, more complex, gradient-informed methods that use the analogy between estimation problem and a physical system to perform inference.

Hamiltonian Monte Carlo Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) is based on the idea of simulating the dynamics of a physical system, called the Hamiltonian system, that has the target distribution as its equilibrium distribution.

In HMC, the target distribution is transformed into a new distribution defined by the energy of a physical system. The physical system is then simulated using Hamilton's equations of motion, which describe the evolution of the system over time. The trajectory of the physical system is used to generate a candidate state, which is accepted or rejected based on the acceptance-rejection rule of the Metropolis algorithm.

The key advantage of HMC is that it can efficiently explore high-dimensional target distributions that have complex covariance structures, where traditional MCMC algorithms can be slow and inefficient. This is because HMC uses gradient information to construct a proposal distribution that is closer to the target distribution, and it also uses a technique called momentum resampling to reduce the correlation between successive samples.

However, as MALA, HMC relies on gradient information of the target distribution, which can be computationally expensive to obtain. In addition, HMC requires careful tuning of the simulation parameters, such as the step size ΔT and the number of leapfrog steps L, to ensure that the samples are generated efficiently and accurately.

Algorithm 5: Hamiltonian Monte Carlo algorithm

input : Unnormalised posterior $\pi[\boldsymbol{\epsilon}|\mathcal{D}]$, gradient $\nabla\pi[\boldsymbol{\epsilon}|\mathcal{D}]$, number of iterations T, number of leapfrog steps L, time-step $\Delta t > 0$, proposal covariance Σ, initial value $\boldsymbol{\epsilon}^{(0)}$

for $1 \leq i \leq T$ **do**

Initialize the momentum $\boldsymbol{p}_i(0) \sim \mathcal{N}_p(0, I_p)$

Set $\boldsymbol{\epsilon}_i(0) = \boldsymbol{\epsilon}^{(i-1)}$

for $1 \leq j \leq L$ **do**

Update the momentum:
$\boldsymbol{p}_i((j-\frac{1}{2})\Delta t) \leftarrow \boldsymbol{p}_i((j-1)\Delta t) - \frac{\Delta t}{2}\nabla\pi[\boldsymbol{\epsilon}_i((j-1)\Delta t)|\mathcal{D}]$

Update the position: $\boldsymbol{\epsilon}_i(j\Delta t) \leftarrow \boldsymbol{\epsilon}_i((j-1)\Delta t) + \Delta t\boldsymbol{p}_i((j-\frac{1}{2})\Delta t)$

Update the momentum:
$\boldsymbol{p}_i(j\Delta t) \leftarrow \boldsymbol{p}_i((j-\frac{1}{2})\Delta t) - \frac{\Delta t}{2}\nabla\pi[\boldsymbol{\epsilon}_i(j\Delta t)|\mathcal{D}]$

Set $\boldsymbol{\epsilon}' \leftarrow \boldsymbol{\epsilon}_i(L\Delta t)$ and simulate $u \sim \mathcal{U}([0,1])$

if $u \leq \frac{\pi[\boldsymbol{\epsilon}'|\mathcal{D}]\exp\{\boldsymbol{p}_i(0)^\top\boldsymbol{p}_i(0)/2\}}{\pi[\boldsymbol{\epsilon}^{(i-1)}|\mathcal{D}]\exp\{\boldsymbol{p}_i(L\Delta t)^\top\boldsymbol{p}_i(L\Delta t)/2\}}$ **then**

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}'$

else

$\boldsymbol{\epsilon}^{(i)} \leftarrow \boldsymbol{\epsilon}^{(i-1)}$

output: Samples $\boldsymbol{\epsilon}^{(i)}$ from the posterior

MCMC implementation choices

We have to decide between the various algorithms available to us. First, we summarise the main requirements of all presented algorithms in Table 4.1.

	MH	pCN	MALA	HMC
No normality requirements	✓		✓	✓
Efficient in higher dimension		✓	✓	✓
Requires gradient evaluations			✓	✓

Table 4.1: Main requirements of the presented algorithms

The joint sampling of weights $\boldsymbol{\epsilon}$ and hyperparameters using preconditioned Crank-Nicholson is not feasible because the priors on the hyperparameters are not suitable. The gradient evaluations are available, but they come at a high computational cost, making the use of Metropolis-Hastings Adjusted Langevin Algorithm and Hamiltonian Monte Carlo impractical.

Out of the algorithms presented, Metropolis-Hastings is the only suitable option for joint estimation, despite being inefficient in higher dimensions. However,

if the hyperparameters are known and estimation is not necessary, pCN can be utilized to accelerate the inference process. Based on these considerations, we propose the following methodology.

Implementation used

1. Perform the MAP estimation using Algorithm 1. This yields the MAP estimates θ, σ_θ^2 and $\boldsymbol{\varepsilon}_\theta^*$.

2. Proceed to the MCMC sampling of $\boldsymbol{\varepsilon}$ at σ_θ^2 and θ fixed using a pCN algorithm. To improve numerical efficiency and ensure we are starting in an interesting region of the parameter space, we initialize the algorithm with $\boldsymbol{\varepsilon}_\theta^*$ (or a slightly perturbed version of it when running parallel chains).

Performing the MCMC estimation with fixed hyperparameters allows for theoretical guarantees on the convergence of the chain. Indeed, as noted earlier in this chapter, the posterior is a log-concave function of $\boldsymbol{\varepsilon}$ which ensures a fast convergence of the considered algorithm (Dwivedi et al., 2018).

4.4 Applications on analytical test cases and a meteorological application

We shall evaluate and illustrate the methods considered in this thesis on various datasets, either artificial test-cases, or datasets from natural sciences.

4.4.1 Assessing the expected power continuity with unconditional realisations

We consider some of the most popular covariance kernels, and visualize how their continuity modulus affects their expected power continuity. For the sake of simplicity in deriving the Hölder exponents α_1 and α_2 in Equation 3.34, we focus on the setting where $D = \mathcal{I} = [0, 1]$. For two commonly-used kernels, we derived their Hölder exponents in $\mathbf{x}$. The considered functions are summarised in Table 4.2.

Kernel	Associated α_1 in Hölder condition
Exponential: $k(y, y') := e^{-\|y-y'\|_2}$	$0 \leq \alpha_1 \leq 1$
Gaussian: $k(y, y') := e^{-\|y-y'\|_2^2/2}$	$0 \leq \alpha_1 \leq 2$

Table 4.2: Kernels used and their Hölder exponents

By drawing 1000 unconditional realisations of SLGPs induced by centred GPs with the corresponding kernels, we can represent a re-scaled version of $\mathbb{E}\left[\Delta\left(\Xi_0, \Xi_{\mathbf{x}'}\right)^\gamma\right]$ for the three dissimilarities Δ considered in Section 3.3.1 and varying γ. We also represent the corresponding theoretical rate. Re-scaling is used solely to allow all curves to appear on the same plots.

The results, represented in Figure 4.2 support our claim that the bounds obtained in previous derivations (Theorem 3.3.5) are sharp. We now continue working on synthetic fields, and check that our SLGP models allow for learning the underlying fields.

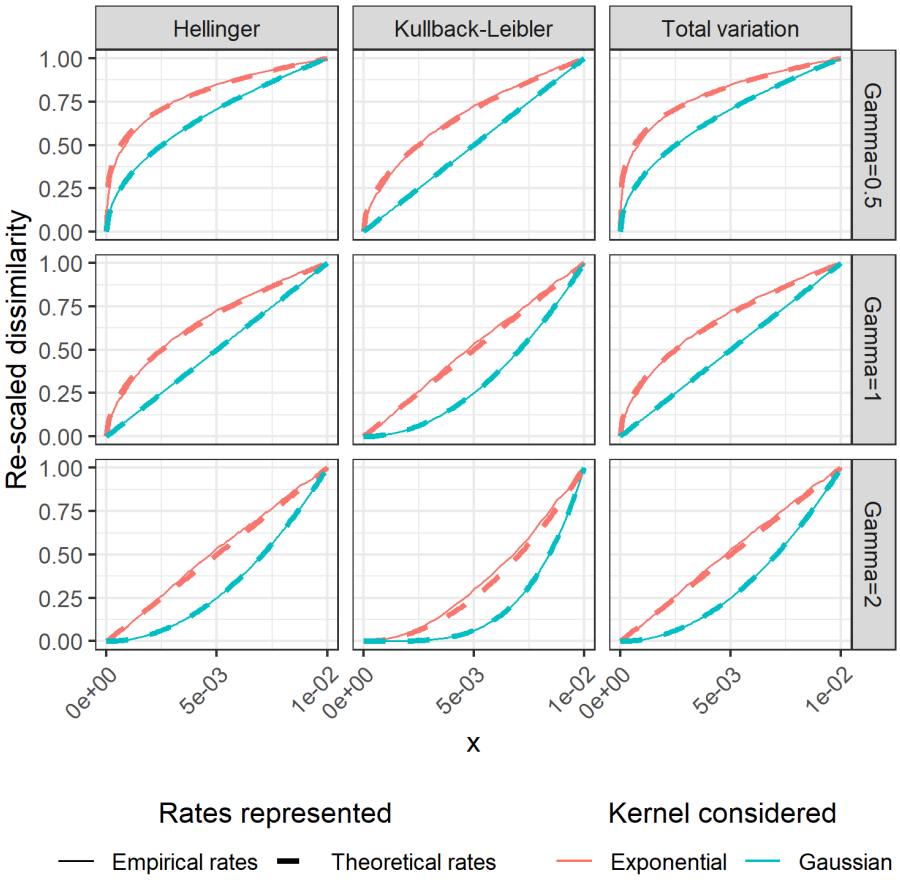


Figure 4.2: Visualising $\mathbb{E}\left[\Delta\left(\Xi_0, \Xi_{\mathbf{x}'}\right)^\gamma\right]$ (plain lines) and the theoretical bound (dotted lines) for both kernels, all three dissimilarities and $\gamma \in \{0.5, 1, 2\}$.

4.4.2 Illustrating the Posterior Consistency with an artificial dataset

We consider two density valued-fields, perfectly known, represented in Figure 4.3. We obtain them by applying the spatial logistic density transformation to realisations of GPs with Matérn 5/2 kernels. The index spaces we consider here are $D = [0, 1]$ and $\mathcal{I} = [0, 1]$.

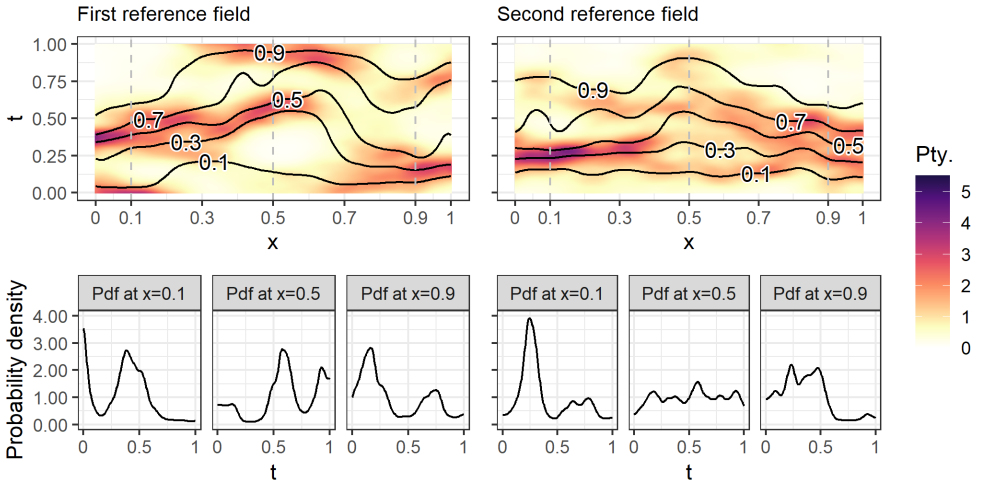


Figure 4.3: Representation of the two density fields used as reference: heat-map of the probability density field $f_1(\mathbf{x}, t)$ and $f_2(\mathbf{x}, t)$ with main quantiles of the field (top) and probability density functions over slices at $\mathbf{x} \in \{0.1, 0.5, 0.9\}$ (bottom).

We run the density field estimation, without hyperparameters estimation. Figures displaying the mean posterior field are available in Figure 4.4 for the first reference field and in Figure 4.5 for the second. We observe that higher sample size seems to yield a better estimation as the models manages to capture the shape and modalities of the true density field.

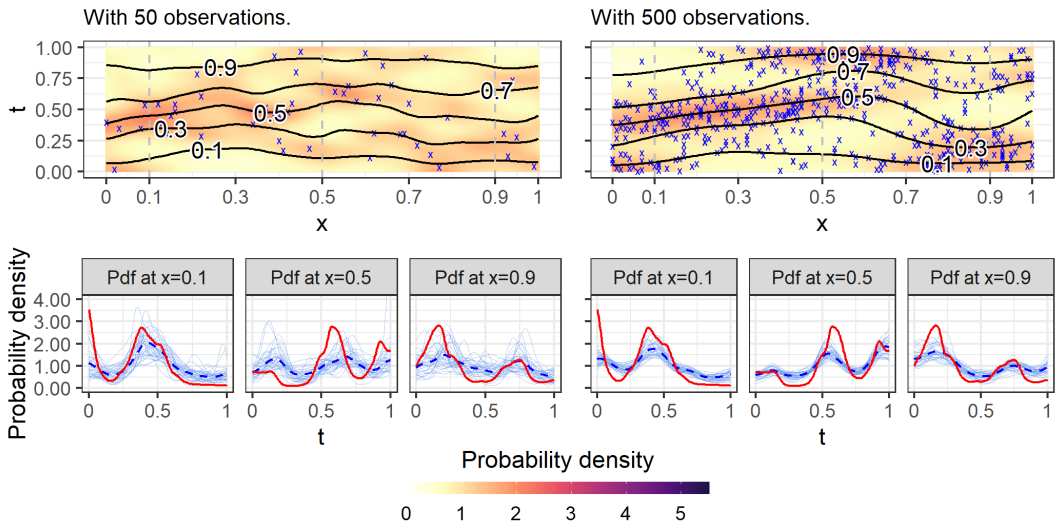


Figure 4.4: Results for the first reference field. [Top figures:] Heat-map of the mean posterior probability density field with main quantiles and sample used. [Bottom figures:] 100 realisations of the posterior pdf (thin blue lines), posterior mean (blue dotted line) and true pdf (red) at $\mathbf{x} \in \{0.1, 0.5, 0.9\}$.

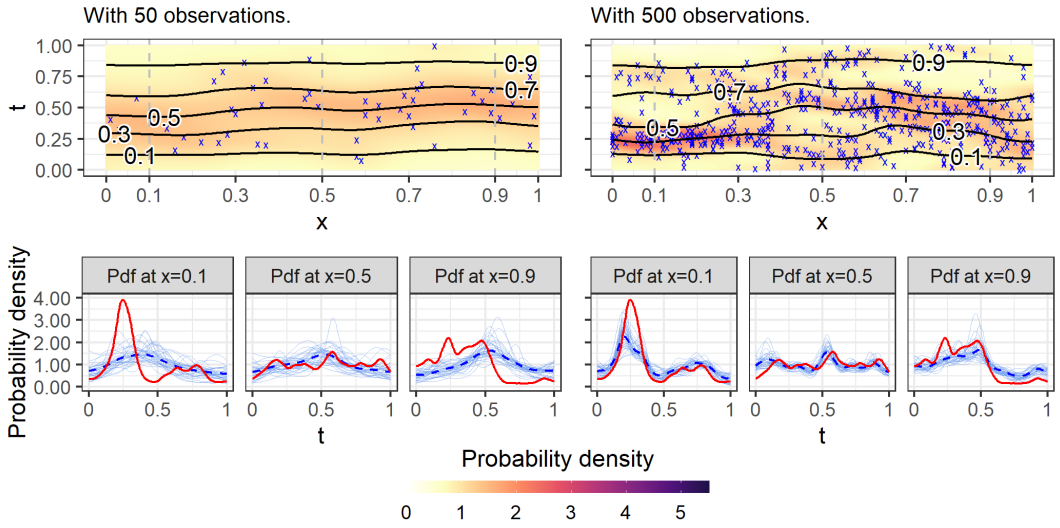


Figure 4.5: Results for the second reference field. [Top figures:] Heat-map of the mean posterior probability density field with main quantiles and sample used. [Bottom figures:] 100 realisations of the posterior pdf (thin blue lines), posterior mean (blue dotted line) and true pdf (red) at $\mathbf{x} \in \{0.1, 0.5, 0.9\}$.

We expect the goodness of fit of our density estimation procedure to increase

with the number of available observations. Since we only consider finite rank GP, the order p (number of Fourier components used) may also determine how precise our estimation can be. In order to quantify the prediction error for different sample sizes and GP's order, we define an Integrated Hellinger distance to measure dissimilarity between two probability density valued fields $f(\mathbf{x}, \cdot)$ and $f'(\mathbf{x}, \cdot)$:

$$d_{IH}^2(f(\mathbf{x}, \cdot), f'(\mathbf{x}, \cdot)) = \frac{1}{2} \int_D \int_{\mathcal{I}} \left(\sqrt{f(\mathbf{v}, u)} - \sqrt{f'(\mathbf{v}, u)} \right)^2 du \, d\mathbf{v} \tag{4.23}$$

In Fig. 4.6, we display the distribution of d_{IH} between true and estimated fields for various sample sizes and SLGP orders, obtained by running 50 replication of the experiment with varying seed. We see that the errors are comparable for small sample sizes. The order becomes limiting when more observations are available, as those of the considered SLGPs relying on the smallest numbers of basis functions appear to struggle to capture small scale variations.

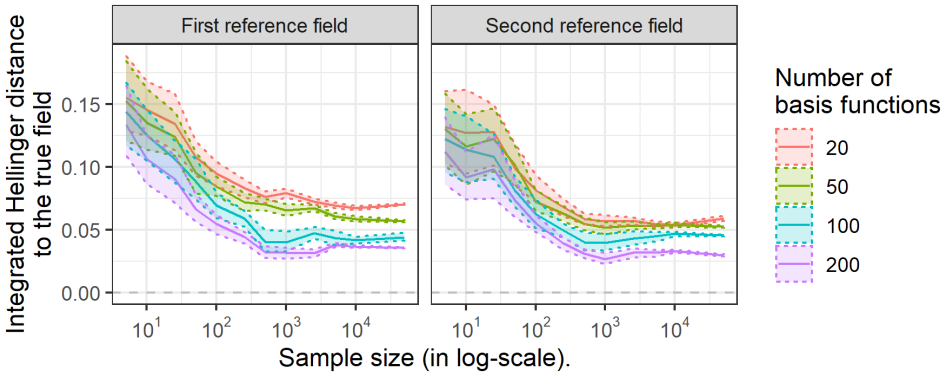


Figure 4.6: Integrated Hellinger distance distribution for different sample sizes and process orders.

Although the goodness of fit are often comparable when few observations are available, when numerous data points are used, the order becomes limiting. We attribute this threshold phenomenon to the SLGP being unable to model small scale variations.

4.4.3 Demonstrating applicability in higher dimensions with a meteorological dataset

We present an application on a data-set of temperatures in Switzerland. This application is by no mean a real forecasting application, and its only aim is to

illustrate the applicability of the SLGP density estimation on real data. The temperature data-set is provided by of Meteorology and MeteoSwiss (2019) and the topographical data is provided by of Topography swisstopo (2019).

Our data consist in daily average temperatures in 2019, measured at 29 stations in Switzerland and marginalised over time. The stations considered are represented in Figure 4.7.

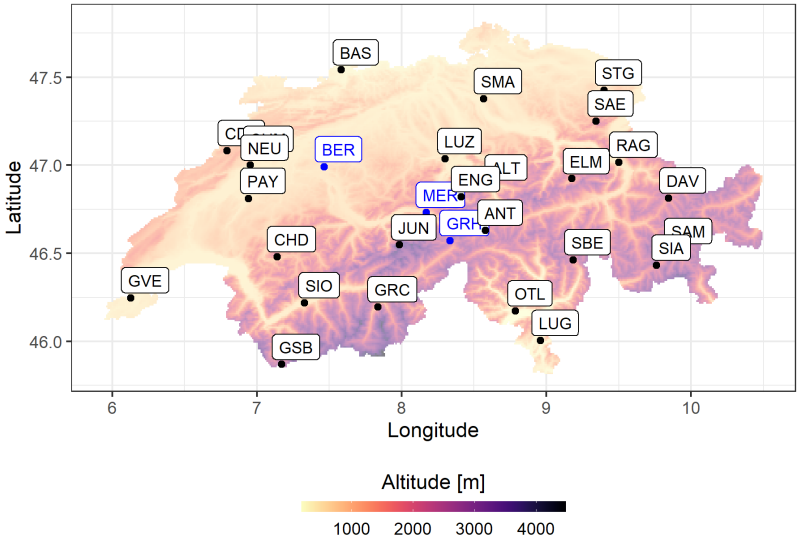


Figure 4.7: Map of Switzerland showing the 29 Stations present in the data-set, the stations located in the canton of Bern are in blue.

We consider that the distribution of these temperatures depends on the latitude, longitude, and altitude of the stations, and we fit the SLGP model on all the stations but three (for the purpose of this illustration, we arbitrarily excluded the stations located in the Canton of Bern). Since we are not taking the measurement date into account, we are actually working with marginal distributions. An example of the available data is displayed in Table 4.3, and panels representing histograms and pointwise kernel density estimator of the temperatures at each station are visible in Figure 4.9.

Station	Date	Daily avg. T. [°C]	Altitude	Longitude	Latitude
Altdorf	2019-01-01	1.5	438	46.88707	8.621894
Altdorf	2019-01-02	0.0	438	46.88707	8.621894
		⋮			
St. Gallen	2019-12-31	-4.1	776	47.42547	9.398528

Table 4.3: First and last rows of the temperature data-set in Switzerland.

We observe that this dataset presents changes of shape, by displaying both uni and multi modality, various degrees of skewness, as well as shifts in temperatures. It appears, as one can expect, that stations located at a high altitude tend to have colder temperatures than the ones located in the Swiss plateau.

Prior knowledge on the Swiss climate hints toward the altitude being the most relevant coordinate, with latitude and longitude having far less impact on predictions, in most application cases. We leverage this insight to slightly simplify the model, by assuming that the rescaled latitude and longitude share a common lengthscale.

We specify a finite-rank GP with 250 random Fourier features (i.e. 500 basis functions) drawn from the spectral density of a Matérn 5/2 kernel. We follow the methodology from Section 4.3, by first performing MAP estimation, with the hyperparameters being determined with a grid-search. One can refer to Figure 4.8 for the negative-log-posterior profiles. With this approach, we identify promising values of the lengthscale for latitude and longitude to be at 40% of the range, while it is at 15% for the altitude and 7.5% for the temperature value.

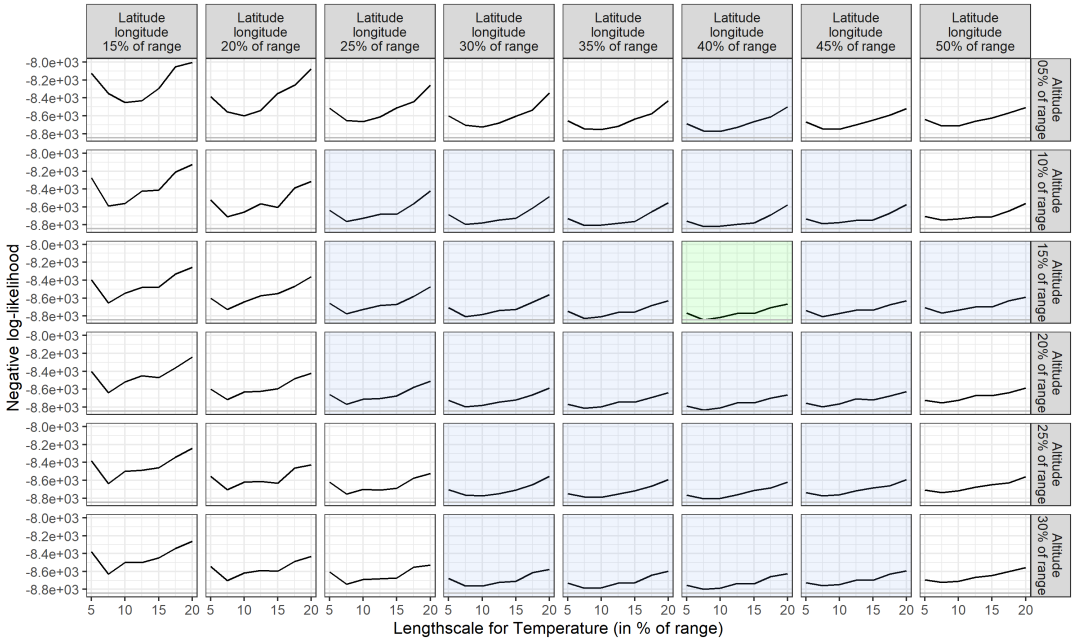


Figure 4.8: Showing the values of the negative log-posterior when varying the lengthscale parameters for Latitude/Longitude, Altitude and Temperature. The panel achieving the minimum is highlighted in green, the panels whose smallest values are less than 1% away from the minimum are in light blue.

We noted during this step of our work that the lengthscale for the temperature is generally highly identifiable, while the posterior is much flatter when varying the other lengthscales. We experimented on the choice of prior for these hyperparameters, but concluded that given the relatively large volume of data available, prior parametrization had little impact on this behaviour.

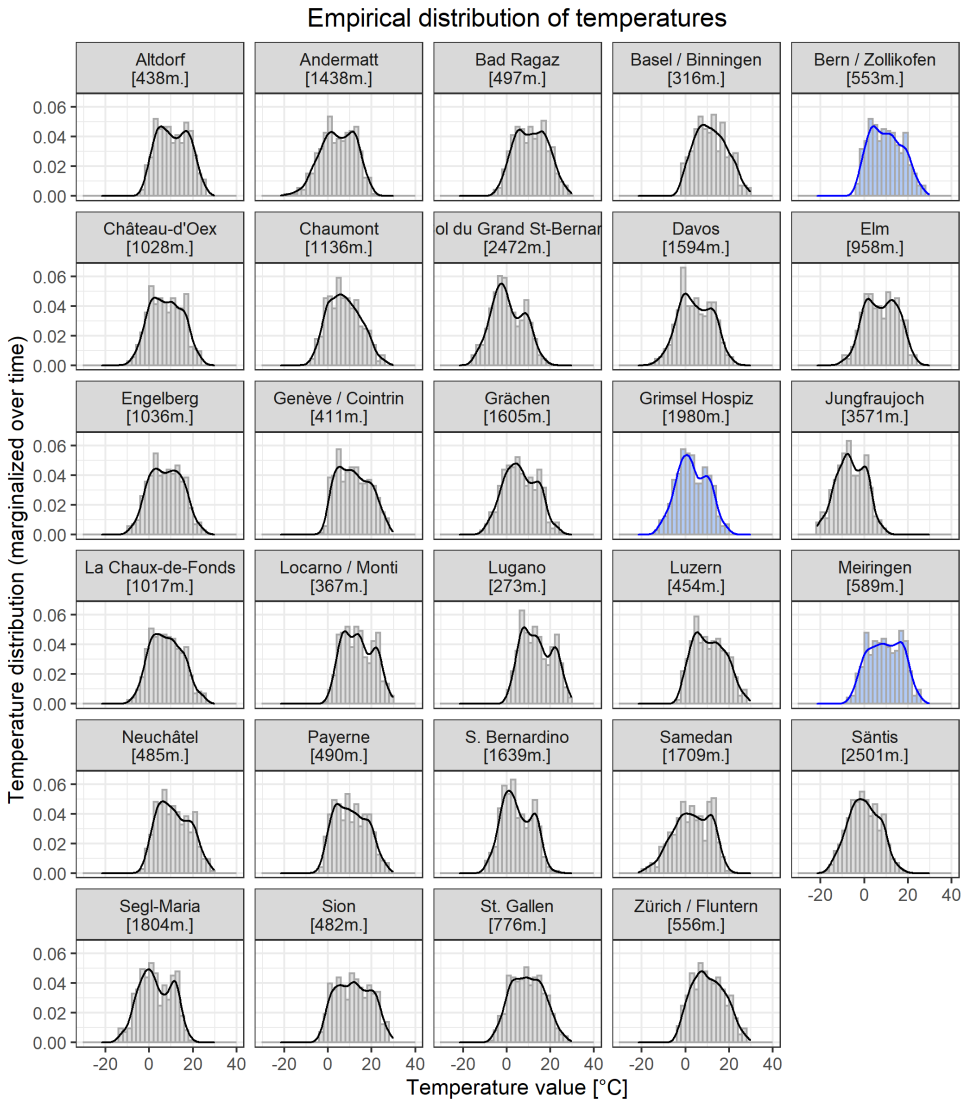


Figure 4.9: Histograms and pointwise kernel density estimator of the data at each of the 29 Stations present in the data-set, the stations located in the canton of Bern are in blue.

Once the hyperparameters were estimated, we performed a MCMC-based

estimation to draw 100 realisations from the SLGP's posterior. We display the estimation results on stations present in the training set in Figure 4.10, as well as out of the training set in Figure 4.10. A collection of similar plots for all stations in the dataset is also available in Appendix A.3.

Let us start with the model displayed in Figure 4.10. For all the station, the MAP estimates follow the available histograms quite closely. However, it appears that the MCMC draws have more variability and sometimes struggle to reproduce all the modes. In particular, it still puts non negligible probability mass on every region of the domain. In addition to this artefact, it appears that at Col du grand St-Bernard, the model fails to reproduce the mode around 12 °C. This motivated us to pay particular attention to study stations of interest close to the Col du grand St-Bernard to see if we could partly explain this discrepancy between data and model predictions. Namely, we considered Sion (the closest station overall) and Jungfraujoch (closest station located at a mountain peak). It appeared that the distribution of temperatures in Jungfraujoch is clearly unimodal and that the SLGP model managed to capture this uni-modality. It suggests that the absence of the second mode at Col du Grand St-Bernard might be a simple consequence of the relative proximity of the two stations, both in latitude/longitude and in altitude.

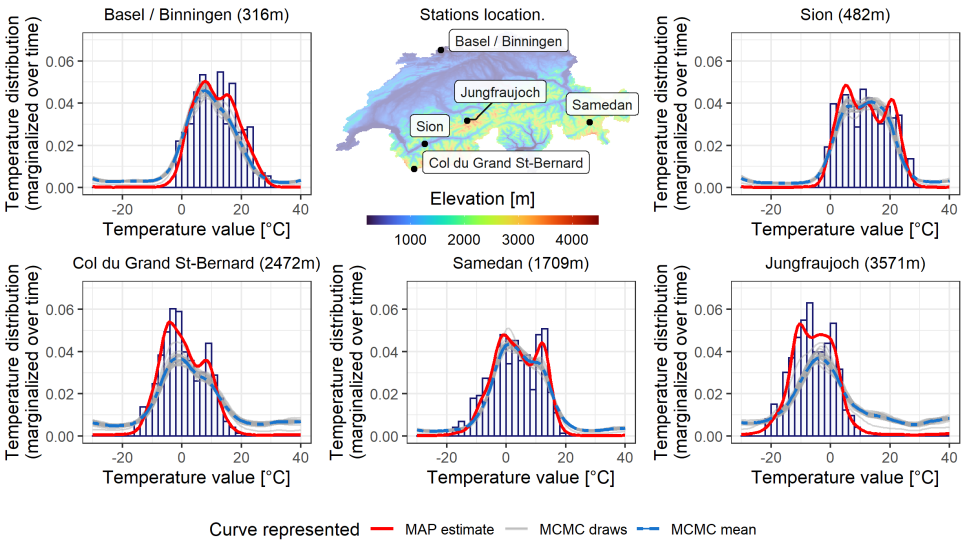


Figure 4.10: SLGP trained on 26 meteorological stations (365 observations each). We display for 5 stations in the training set: the histogram of the available data and curves obtained from SLGP estimation. For each station, we also specify its elevation above sea level and show its location.

Let us also note that the Jungfraujoch and the Col du Grand St-Bernard are on the north- respectively south-facing slope of their respective locations. Incorporating this information in the model might prove useful to yield better predictions, in particular at the considered stations.

We also note that for Jungfraujoch, the MAP estimate appears to follow the histogram more closely than the MCMC draws. It is possible that due to the flatness of the likelihood as represented in Figure 4.8, we misidentified the lengthscales which hinders the model's posterior expressivity.

We also make predictions at the three stations that we left out of the data set, to see whether the SLGP model manages to extrapolate at new locations. We observe, when comparing estimations performed at locations where data were available (Fig. 4.10) or not (Fig. 4.11) that the resulting random densities present a bit more variability at stations left out of the training set, a desirable feature. Other than that, the estimation seems reasonable for all stations.

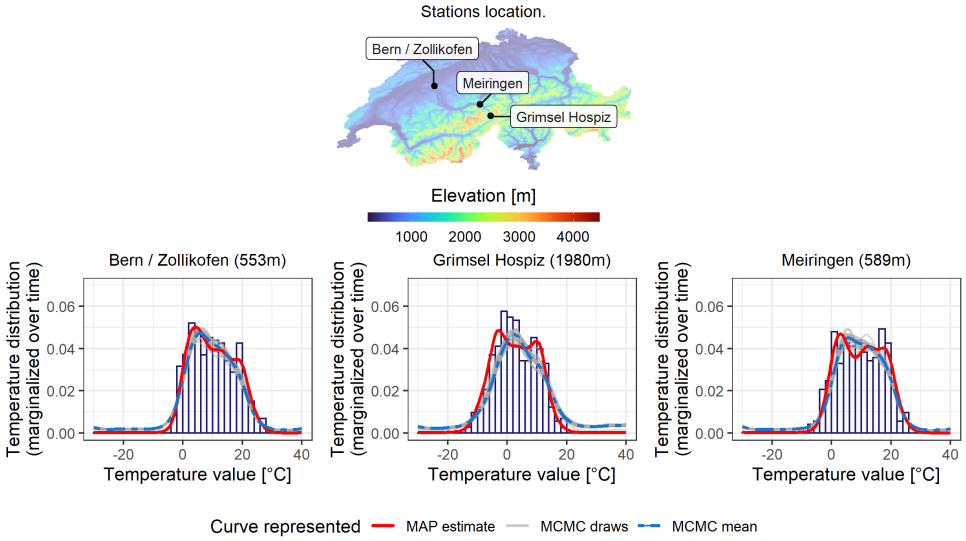


Figure 4.11: SLGP trained on 26 meteorological stations (365 observations each). We display for the 3 stations left out of the training set: the histogram of the available data and curves obtained from SLGP estimation. For each station, we also specify its elevation above sea level and show its location on a map of Switzerland.

So far, we displayed estimation results only at stations. However, SLGP modelling is a powerful tool that allows for predicting the density field over the whole domain (here, the whole of Switzerland). For plotting purposes, we can't display the full distributions, but we can easily represent moments or quantiles of

it, and their respective uncertainties. To illustrate further the full capabilities of our model, we display additional Figures. In Figure 4.12, we show the expected mean temperature across Switzerland, and its standard deviation under SLGP modelling. We make a similar plot in Figure 4.13 with the median temperature.

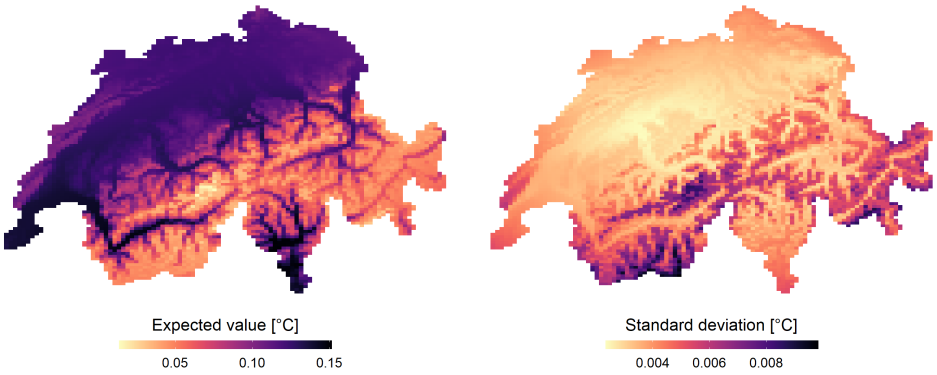


Figure 4.12: [Left] Expected mean temperature across Switzerland [Right] Standard deviation of mean temperature

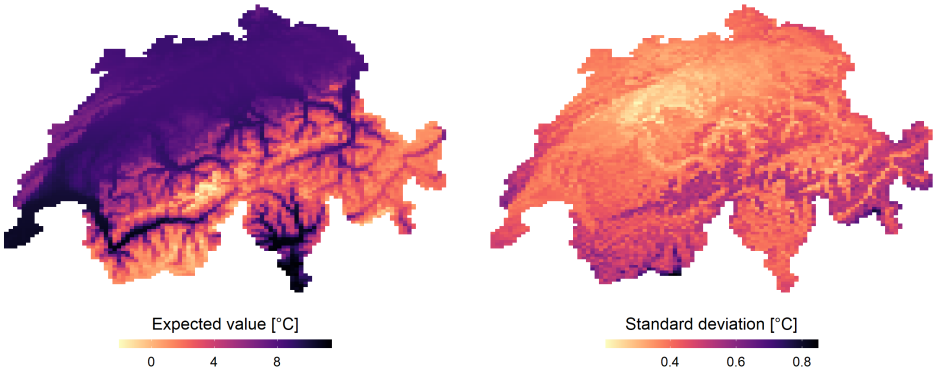


Figure 4.13: [Left] Expected median temperature across Switzerland [Right] Standard deviation of median temperature

We insist on the fact that the quantiles estimations are done simultaneously as a by-product of the SLGP models. To illustrate this, we show the behaviour of these quantiles of daily mean temperature across a slice of Switzerland in Figure 4.14

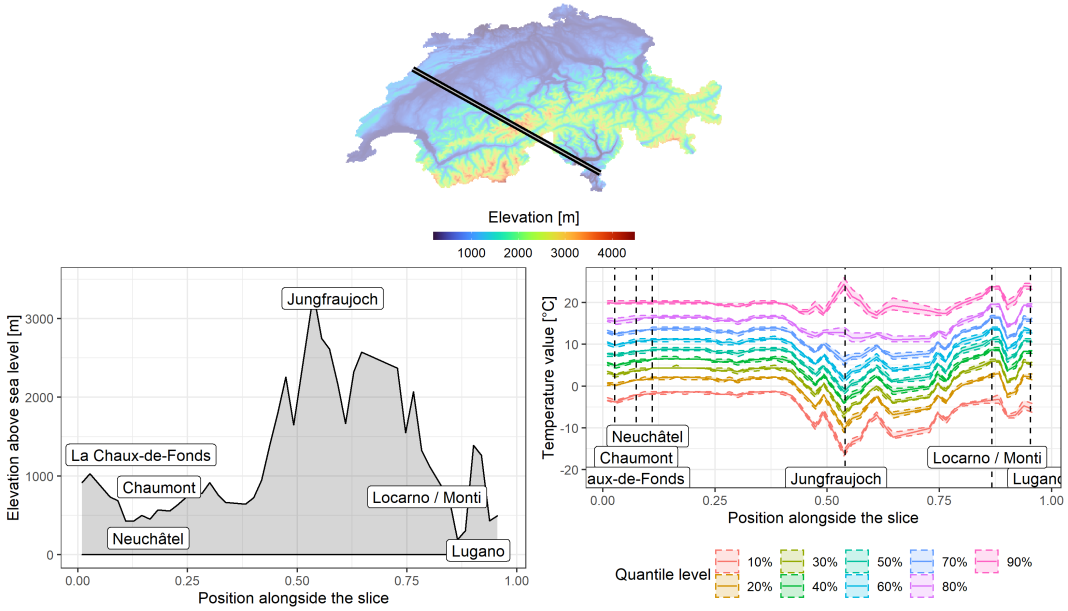


Figure 4.14: [Top] MAP of Switzerland and slice considered. [Bottom left] Elevation alongside the slice and Stations location. [Bottom right] Simultaneous quantile prediction (mean value and 10% quantile- 90% quantile bands) across a slice of Switzerland.

This application yields promising results, yet we noticed the presence of some artefacts in density field estimation. The specification of the topographical and other variables to be incorporated in the spatial index as well as the chosen families of covariance kernel appear to be of crucial importance regarding the resulting model and predictions. Also, the incorporation of trends appears as a meaningful avenue of research to be further explored to increase the realism of SLGP models.

Chapter 5

Accelerating inference with GP-based Modelling

One of the main challenges in statistical inference is to define suitable sequential design strategies. Choosing where to add observations is indeed crucial and a good design strategy must achieve a trade-off between exploration and exploitation of the input space so as to discover regions of interest while avoiding getting trapped in the vicinity of artefactual basins of attraction. Deriving such strategies requires anticipating (probabilistically) the effect of adding new observations.

Addressing this challenge generally boils down to studying the effect on a response of interest of varying some *decision* or *control* variables $\mathbf{x}$. Yet it is typically unrealistic to assume a deterministic relationship between $\mathbf{x}$ and the response, be it for instance because of uncertainty in other input variables or because the assumed response and/or observation generating processes themselves involve some randomness.

Relying on GP-based methods to derive probabilistic metamodels of the complex systems at hand is a common approach, as it allows for accounting for such uncertainty. We briefly review such approaches and propose extending them to SLGP-modelling, in Section 5.1 we address metamodel-powered Bayesian optimisation, while in Section 5.2 we do the same for stochastic inverse problems. In both contexts, the modelling capabilities of GP and SLGP modelling are examined and evaluated on a geophysical application.

Let us stress that benchmarking the performances of the approaches considered and obtaining representative results calls for numerous runs of numerical experiments with multiple seeds. To accomplish this, computations in the coming section were performed on UBELIX (`http://www.id.unibe.ch/hpc`), the HPC cluster at the University of Bern.

5.1 In Bayesian optimisation

This section is inspired and comprises elements presented in Gautier et al. (2021).

5.1.1 SLGP modelling in Stochastic Optimisation

We will denote by $p_0(\mathbf{x})$ the response's distribution at location $\mathbf{x}$. As in the previous Chapters, we will denote by t (and variations thereof) the response's values, and by $\mathbf{x}$ the index variables. We furthermore assume that an objective function $g(\mathbf{x}) = \rho(p_0(\mathbf{x}))$ depending on $\mathbf{x}$ through $p_0(\mathbf{x})$ is to be minimised, where ρ returns for any considered probability distribution a real-valued quantity such as a moment or a quantile with some given level.

The classical setting of Bayesian optimisation In Stochastic Optimisation (Ruszczynski and Shapiro, 2003; Birge and Louveaux, 2011; Prékopa, 2013), one considers that the distribution $p_0(\mathbf{x})$ is either known or needs to be estimated. Approximation procedures have been studied, including but not limited to the Robbins Monro procedures and further developments in Bayesian Approximation (Robbins and Monro, 1951; Mandt et al., 2017) or the multi-armed bandit paradigm (Thompson, 1933; Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012).

The most natural choice for the functional ρ is to consider the expectation. Yet, many other choices for ρ have been considered. These choices include quantiles (Rostek, 2010; Torossian et al., 2020), the conditional value at risk (Rockafellar and Uryasev, 2000) or the expectiles (Bellini and Di Bernardino, 2015). It is also possible to learn the unknown distribution $p_0(\mathbf{x})$, as in Hall et al. (2004); Efromovich (2010); Moutoussamy et al. (2015) but the approaches classically used require numerous replicates and struggle with sample heterogeneously scattered across space.

Bayesian Optimisation (BO) is often considered to be the most parsimonious approach allowing to tackle such challenge. Indeed, it leverages the potential of meta-modelling, especially Gaussian processes (GP) (Williams and Rasmussen, 2006), to keep a memory of explored points with the aim to explore decision space while keeping a parsimonious evaluation budget. First introduced by Mockus et al. (1978); Jones et al. (1998) in the noise free setting, it has latter been extended to stochastic black box optimisation (Frazier et al., 2009; Frazier, 2018; Srinivas et al., 2009; Picheny et al., 2013a; Hernández-Lobato et al., 2014; Jalali et al., 2017) and further sequential strategies have been studied (Risk and Ludkovski, 2018; Binois et al., 2019). However, existing approaches typically assume Gaussian response distributions $p_0(\mathbf{x})$.

We propose surrogating the response density $p_0(\mathbf{x})$ with a SLGP, allowing us to draw inspiration from the fertile literature of GP-based Bayesian Optimisation while setting ourselves free of the strong Gaussianity assumption.

SLGP-based Bayesian Optimisation: the model We recall that we are interested in minimising $g(\mathbf{x}) = \rho(p_0(\mathbf{x}))$, while we do not know the field $p_0(\mathbf{x})$. To achieve this goal, we leverage a dataset assumed to be obtained by independent sampling of the reference p_0 at some prescribed locations $\mathbf{x}_i$. By that, we mean that as in Section 4.1.1 our dataset consists in n couples of locations and observations $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$, where the $\mathbf{x}_i$ are in $[0,1]^{d_x}$. Moreover, we assume the t_i's are obtained by independent sampling of random variables T_i with respective densities $p_0(\mathbf{x}_i)$. The (random) vectors of observations are denoted by $\mathbf{T} = (T_i)_{1\leq i\leq n}$ and $\mathbf{t} = (t_i)_{1\leq i\leq n}$.

We call $(p_0(\mathbf{x}))_{\mathbf{x}\in D}$ a density field, i.e. a collection of pdf on $\mathcal{I}$ indexed by $\mathbf{x} \in D$, and surrogate it with a SLGP denoted by $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$. We then propose considering $(G_{\mathbf{x}})_{\mathbf{x}\in D} := (\rho(Y)_{\mathbf{x},\cdot})_{\mathbf{x}\in D}$, the random field obtained by applying ρ to the density valued field delivered by a SLGP model. $G_{\mathbf{x}}$ naturally induces a surrogate model for $g(\mathbf{x})$.

This model allows for more flexibility in the distributional assumptions on $p_0(\mathbf{x})$, but we can also leverage its stochastic nature to derive subsequent designs of experiments.

Indeed, conditioning $G_{\mathbf{x}}$ on observed data $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$ boils down to applying the functional ρ to the (random) density field obtained by conditioning the SLGP $(Y_{\mathbf{x},t})_{(\mathbf{x},t)\in D\times\mathcal{I}}$ on these same data $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$.

Therefore, we can directly apply the density field estimation presented in Chapter 4 to obtain draws from the conditional distribution of SLGP, and transform them by applying the functional of interest to obtain an draws of the conditional distribution of $G_{\mathbf{x}}$.

Remark. The process $G_{\mathbf{x}}$ remains uncertain knowing $\mathbf{T_n} = \mathbf{t_n}$ because of the conditional variability of Y.

SLGP-based Bayesian Optimisation: a criterion for data acquisition In the rest of this section, we will consider the particular case where ρ is the median, but the presented approach is not restricted to this choice and can be applied to arbitrary (measurable) mappings, potentially also mappings depending on $\mathbf{x}$.

In the spirit of robust optimisation, we will consider the problem of minimising an α-quantile of the random function $G_{\mathbf{x}}$. We note $Q_n(\mathbf{x})$ the α-quantile of the conditional distribution of $G_{\mathbf{x}}$ knowing $\{(\mathbf{x}_i, T_i)\}_{1\leq i\leq n}$. Note that $Q_n(\mathbf{x})$ is a random function as the observations $\mathbf{T}_n$ are left in random form.

Due to the computational cost of performing SLPG-based density field estimation, we focus on adding batches of observations at a candidate location. Note that whenever the batch size is set to 1, we boil down to the classical setting of step-wise adaptive design.

We denote $Q_{n+K}(\mathbf{x}; \mathbf{x}_{\text{new}})$, the α-quantile of $G_{\mathbf{x}}$ conditioned on past observations T_n and on a batch of K i.i.d. observations to be made at $\mathbf{x}_{\text{new}} \in D$. Quantifying the impact the choice $\mathbf{x}_{\text{new}}$ has on Q_{n+K} is the backbone of BO. One typically seek to find values that is likely to yield the most "improvement" when going from Q_n to Q_{n+K}. We propose a criterion quantifying the change that occurs when adding observations at a given location. Denoting $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathbf{T_n} = \mathbf{t_n}]$, we consider the following:

$$\text{EQI}_n(\mathbf{x}_{\text{new}}, K) = \mathbb{E}_n\left[\left(\min_{\mathbf{x}\in D} Q_n(\mathbf{x}) - \min_{\mathbf{x}\in D} Q_{n+K}(\mathbf{x}; \mathbf{x}_{\text{new}})\right)^+\right]. \tag{5.1}$$

Remark. This criterion was inspired by the Expected Quantile Improvement presented in Picheny et al. (2013b), which would write here as:

$$\mathbb{E}_n[(\min_{i\leq n} Q_n(\mathbf{x}_i) - Q_{n+1}(\mathbf{x}_{\text{new}}; \mathbf{x}_{\text{new}}))^+] \tag{5.2}$$

but is modified in the spirit of knowledge gradient approaches from Frazier et al. (2009) to account for improvements on the whole domain.

Of course, SLGP-based adaptative learning is not restricted to using this criterion, and one could apply any of the commonly encountered sampling scheme in GP-based BO to SLGP-based BO. The main remaining challenge lies in the evaluation of criteria such as the one in Equation 5.1, and we now propose a simulation-based approximation of it.

5.1.2 Simulation-based computation of criteria

Classically, in Sequential Uncertainty Reduction (SUR) (Bect et al. (2019)) approaches, it is assumed that the function of interest is a realisation of a GP. Under these assumptions, several criteria enjoy (semi-)analytical forms, favouring criterion optimisation and the implementation of design strategies.

However, in our situation, it does not appear feasible to obtain a closed-form formula for the considered EQI criterion and we therefore estimate it via stochastic simulation.

In order to quantify the effect of adding an observation T_{n+1} at a given location $\mathbf{x}_{\text{new}}$, one needs to study the probability density of a new observation T at $\mathbf{x}$ conditioned on $\mathbf{T_n} = \mathbf{t_n}$.

We consider Y, the SLGP induced by a finite-rank GP such as those in Definition 4.1.1. For each instance of $\boldsymbol{\epsilon}$ corresponds a SLGP, to emphasize this dependency we denote it, for some suitable family of basis functions F, $Y^{\boldsymbol{\varepsilon}}_{\mathbf{x},t} := \Psi\left(\boldsymbol{\varepsilon}^\top F(\mathbf{x}_i, t_i)\right)$. Note that if we keep the weights in random form, $Y^{\boldsymbol{\varepsilon}}_{\mathbf{x},t}$ is random through $\boldsymbol{\varepsilon}$, while if we plug-in a realisation $\boldsymbol{\varepsilon}$ of the weights, $Y^{\boldsymbol{\epsilon}}_{\mathbf{x},t}$ is itself a realisation of the SLGP.

In favour of the law of total probability, one finds that the posterior density of a new observation at $\mathbf{x}$ conditioned on the observed data is given by:

$$\pi(t|\mathbf{T_n} = \mathbf{t_n}) \propto \int \pi(t|\boldsymbol{\varepsilon} = \boldsymbol{\epsilon})\pi(\boldsymbol{\epsilon}|\mathbf{T_n} = \mathbf{t_n})\, d\boldsymbol{\epsilon} \tag{5.3}$$

$$\propto \int Y^{\boldsymbol{\epsilon}}_{\mathbf{x},t}\phi_p(\boldsymbol{\epsilon}) \prod_{i=1}^{n} Y^{\boldsymbol{\epsilon}}_{\mathbf{x}_i,t_i}\, d\boldsymbol{\epsilon} \tag{5.4}$$

where ϕ_p is the pdf of the p-variate standard normal distribution.

This motivates the following approach, which can be considered as a basic application of Sequential Monte Carlo (Doucet et al. (2001)), where we use a simple simulation-based particle filter to approximate an unknown future quantity:

1. The generative model given by the SLGP model is implemented as described in Section 4.3 and yields N realisations of an approximation of $\boldsymbol{\varepsilon}|\mathbf{T}_n = \mathbf{t}_n$, denoted thereafter by $(\boldsymbol{\epsilon}_{(j)})_{1\leq j\leq N}$.
   The density of a new observation at $\mathbf{x}$ (See Equation 5.4) is approximated by the mixture $\frac{1}{N}\sum_{j=1}^{N} Y^{\varepsilon_{(j)}}_{\mathbf{x},\cdot}$.

2. The impact of adding K observations at a given location $\mathbf{x}_{\text{new}}$ is estimated by doing M simulations:
   For each simulation, K realisations of the random variable $\widetilde{T}_{\text{new}}$ are independently drawn from the density $\frac{1}{N}\sum_{j=1}^{N} Y^{\varepsilon_{(j)}}_{\mathbf{x}_{\text{new}},\cdot}$ and the corresponding batch of response values is denoted $\widetilde{\mathbf{t}}^{(i)}_{\text{new}} = \left(\widetilde{t}^{(i),1}_{\text{new}}, ..., \widetilde{t}^{(i),K}_{\text{new}}\right)$.
   In light of Equation 5.4, the response density at $\mathbf{x}$ conditional on past data and on the simulated batch is given by:

$$\pi(t|\mathbf{T_n} = \mathbf{t_n}, \widetilde{\mathbf{T}}^{(i)}_{\text{new}} = \widetilde{\mathbf{t}}^{(i)}_{\text{new}}) \propto \int \pi(\boldsymbol{\epsilon}|\mathbf{T_n} = \mathbf{t_n}) Y^{\boldsymbol{\epsilon}}_{\mathbf{x},t} \prod_{\ell=1}^{K} Y^{\boldsymbol{\epsilon}}_{\mathbf{x}_{\text{new}},\widetilde{t}^{(i),\ell}_{\text{new}}}\, d\boldsymbol{\epsilon} \tag{5.5}$$

   Leveraging the fact that the $\boldsymbol{\epsilon}_j$'s are drawn from $\pi(\boldsymbol{\epsilon}|\mathbf{T_n} = \mathbf{t_n})$, we can approximate the integral in Equation 5.5 by the Monte Carlo sum $\sum_{j=1}^{N} Y^{\boldsymbol{\epsilon}_{(j)}}_{\mathbf{x},\cdot} w_{i,j}$,

with weights $w_{i,j}$ proportional to $\prod_{\ell=1}^{K} Y^{\boldsymbol{\epsilon}_{(j)}}_{\mathbf{x}_{\text{new}},\widetilde{t}^{(i),\ell}_{\text{new}}}$ and summing to one.

3. The density field distribution after adding K observations at $\mathbf{x}_{\text{new}}$ is approximated by:

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{K} Y^{\boldsymbol{\epsilon}_{(j)}}_{\mathbf{x},\cdot} w_{i,j} \tag{5.6}$$

In turn, this allows for computing a proxy to the future value of our quantity of interest by simply applying the functional ρ to each of the M plausible future fields $\sum_{j=1}^{K} Y^{\boldsymbol{\epsilon}_{(j)}}_{\mathbf{x},\cdot} w_{i,j}$.

We use this simulation-based method to compute any sampling criterion that we are interested in.

5.1.3 Benchmarking the SLGP for guiding stochastic optimisation

For all the coming applications, we leverage a zero mean GP $Z_{\mathbf{x},t} = \sum_{j=1}^{p} \sqrt{\lambda_j} e_j(\mathbf{x},t)\varepsilon_j,\ p \in \mathbb{N}$, with t and $\mathbf{x}$ being uni-variate. To ensure consistency with the rest of the document, we will keep the bold notation for $\mathbf{x}$. Our basis function are bi-variate Fourier functions of order $q > 0$: sine and cosine of $2\pi(\omega_1 t + \omega_2 \mathbf{x})$, where ω_1 and ω_2 are integers satisfying $-q \leq \omega_1, \omega_2 \leq q$. Then, we remove redundant terms as well as those irrelevant in the SLGP setting (i.e. functions independent of t, that would be simplified with the normalisation of the process).

Some other analytical applications: presenting the test-case.

In the analytical applications, we have $D = \mathcal{I} = [0,1]$ and consider four known density fields. The median functions of these fields appear in Figure 5.1 and are defined as $f_1(\mathbf{x}) = 0.25\sin(16\mathbf{x}+9) + 0.25\sin(4.8\mathbf{x}+2.7) + 0.625$ (minimum at $\mathbf{x}^* \approx 0.5095$) and $f_2(\mathbf{x}) = 0.15 + \frac{7}{72}\frac{1.1(10\mathbf{x}-5)^2 - 5(10\mathbf{x}-5)+6.1}{(10\mathbf{x}-5)^2+1}$ (minimum at $\mathbf{x}^* \approx 0.7414$).

Our first class of probability density fields, that we will refer to as "truncated Gaussian fields", writes as $h_{\mathbf{x}}\left(\frac{t-f_i(\mathbf{x})}{0.05}\right)$, $i \in \{1,2\}$, with $h_{\mathbf{x}}$ a symmetrically truncated standard Gaussian with thresholds $\pm\min\left(f_i(\mathbf{x}), 1-f_i(\mathbf{x})\right)$. The truncation ensures that the distribution remains symmetrical around its median and mean f_i. The second case - that we will refer to as "multi-modal field"- is of the same form yet with $h_{\mathbf{x}}$ median-0 but multi-modal such as represented in Figure 5.1.

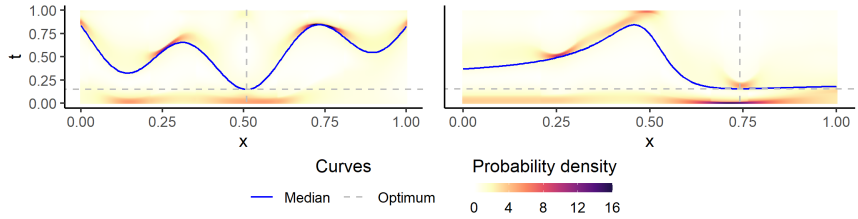


Figure 5.1: The two multi-modal reference density fields, their median and its global optimum.

We perform density field estimation as presented in Section 4.3 for such reference fields for different sample sizes and order of the GP. In this section, we are not yet in an adaptive setting: each time a new observation is added to the model, its location is determined randomly, with uniform distribution over the index set. Figure 5.2 displays the posterior mean field with two sample sizes for the truncated Gaussian field with median f_1.

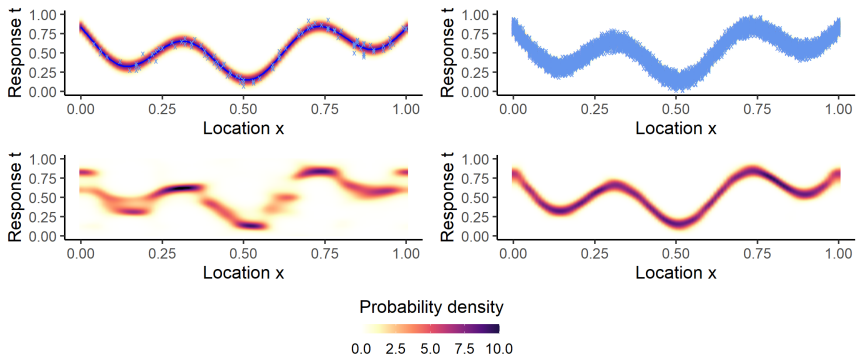


Figure 5.2: Results the truncated Gaussian field with median f_1, using 121 basis functions. True field and samples used (top), posterior mean field (bottom) for a respective sample size of 100 (left) and 10000 (right).

We observe in Figure 5.2 that a higher sample size seems to yield a better estimation. In order to quantify the prediction error, we ran the same experiment as in Section 4.4.2xcu for different sample sizes and GP's order. In Figure 5.3, we display the distribution of d^2_{ISH} between true and estimated fields for various sample sizes and SLGP orders. As both functions yielded close results, we show only the results for f_1. We see once again that the errors are comparable for small sample sizes. The order becomes limiting when more observations are available as those of the considered SLGPs relying on the smallest numbers of basis functions appear to struggle capturing small scale variations.

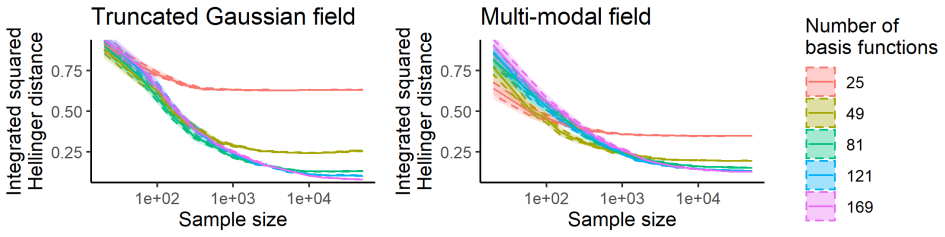


Figure 5.3: Integrated squared Hellinger distance distribution for different sample sizes and process orders, when the reference field has f_1 as its median.

Application in hydrogeology: presenting the dataset

We also consider a one dimensional contaminant problem.

In this application case, we want to localize the source of a contaminant propagating into a saturated aquifer when the geological structure is unknown. Indeed, characterisation of subsurface properties is very uncertain as soon the distance to the scarce samples increase. So, hydrogeologist must rely on the use of analogues and expert knowledge to generate an ensemble of plausible geological realisations that can be used to quantify prediction uncertainty. To keep the problem simple, the zone of interest of the aquifer is modelled as a 2D vertical section (10 meter deep and 5 meter wide) aligned with the main flow direction. At the domain inlet, the depth of the released contaminant (normalised location s in what follows) is the unknown of the problem. The reference observations consist of concentration breakthrough curves at different depths of the domain outlet.

The ensemble of plausible geological realisations and the geological references are multiple-point statistics realisations generated with the Deesse algorithm (Mariethoz et al. (2010)) that reproduce the complex features of braided-river aquifer models (Pirot et al. (2015)). The contaminant flow and transport is simulated under steady-state flow and fixed boundary conditions (constant hydraulic gradient) using the Maflot Matlab code (Künze and Lunati (2011)). The misfit between simulated and reference concentration breakthrough curves are normalised and denoted as the response t in what follows.

In Figure 5.5 we represent the data practitioners would obtain after running numerous simulations: to make this figure, we select one breakthrough curve as reference and run simulations for all the possible combinations between 200 plausible geological structures and 100 source depth, which gives us 10000 concentration breakthrough curves, and as many normalised misfits values. We clearly identify in these figures the non-gaussianity of data as well as the variations in modalities and skewness across index space, hence justifying the need to rely on flexible modelling such as the SLGP.

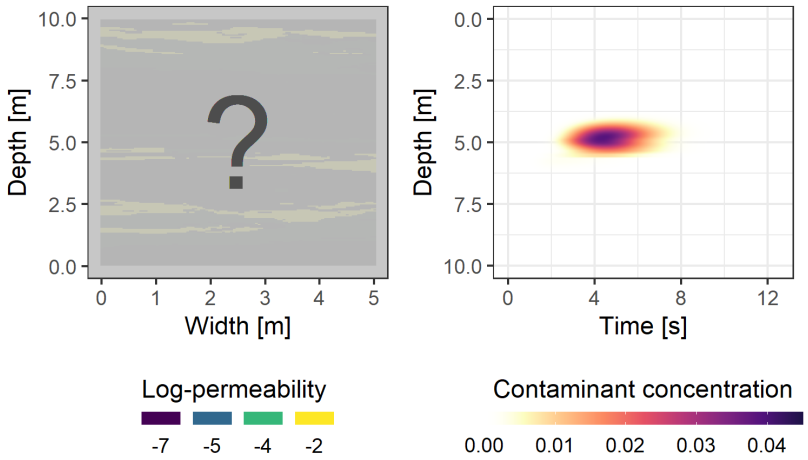


Figure 5.4: Typical situation: the geological structure and the source depth are unknown, and we only have access to the concentration breakthrough curves.

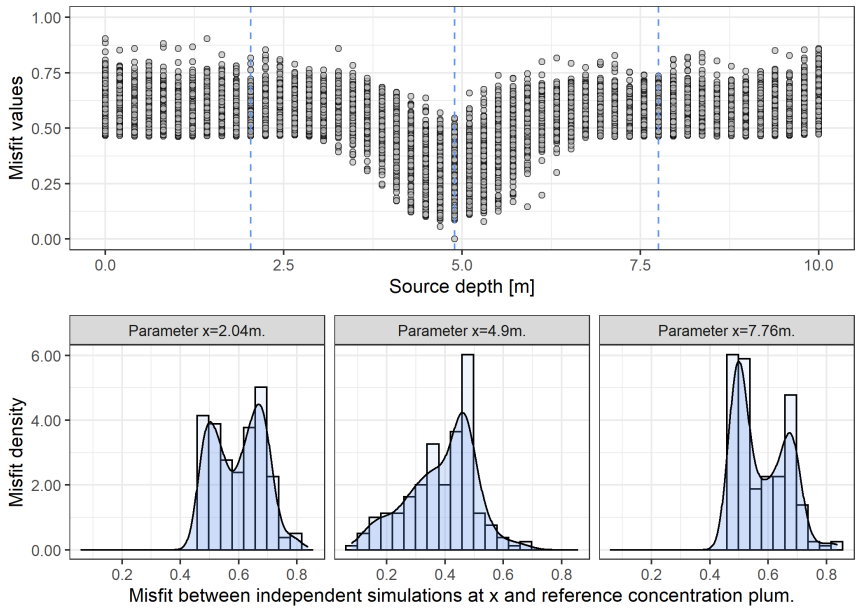
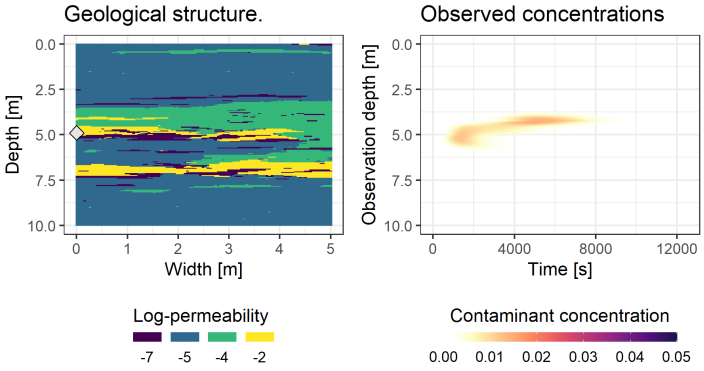


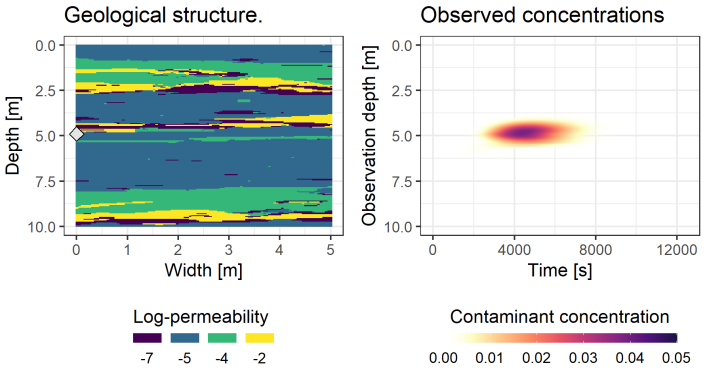
Figure 5.5: [Top] Misfits obtained when running simulations for 200 geological structures and 50 source depths. [Bottom] Misfit empirical distributions (histograms and kernel density estimators) at $\mathbf{x} \in \{2.04\text{m}, 4.9\text{m}, 7.76\text{m}\}$

A few selected simulations results are available in Figure 5.6 and show the simulation outputs and corresponding normalised misfits for various geological realisation and source depths. It clearly illustrates the dependency of the contaminant concentration curves (and hence misfits) in the underlying (considered

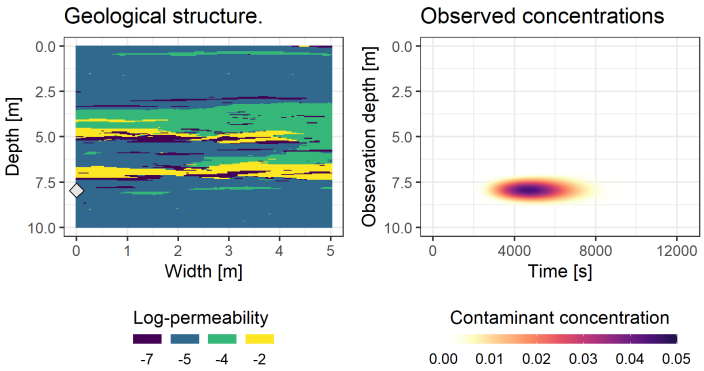
unknown) geological structure. This contributes to the changes in shape and modalities of the misfit distributions that we just observed.



(a) $x = 4.90$m, misfit ≈ 0.035



(b) $x = 4.90$m, misfit ≈ 0.35



(c) $x = 7.96$m, misfit ≈ 0.63

Figure 5.6: Simulations results: varying the choice of geological structures and source depth yields different simulated concentration breakthrough curves, and as such different misfit values.

Optimisation for hydrogeology

Throughout the section and as illustrated in Figure 5.7, we use two reference fields of misfit distributions that are obtained by fitting SLGP models to misfit data produced under two latent geologies.

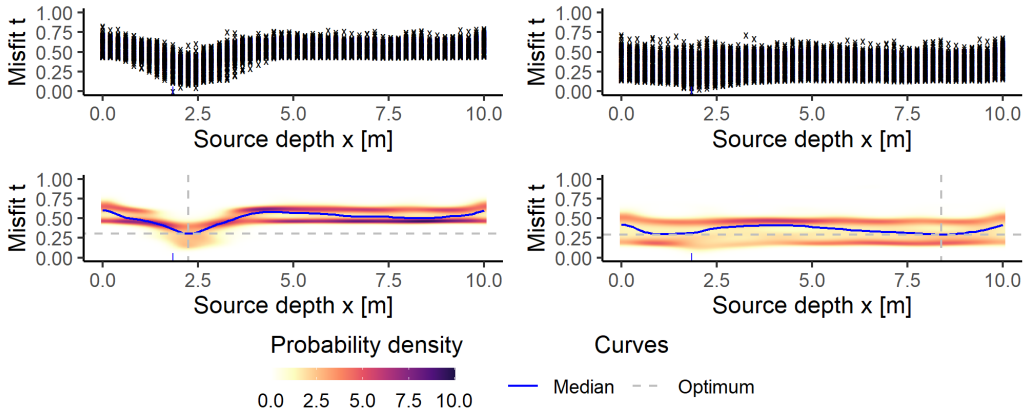


Figure 5.7: Misfit data (top) and posterior mean field (bottom) for two latent geological structures.

Using these two reference density fields to draw new samples, we follow the methodology introduced in part 4.3 and represent in Figure 5.8 the posterior mean field and its estimated median before and after running 25 steps of the algorithm on the first geological structure.

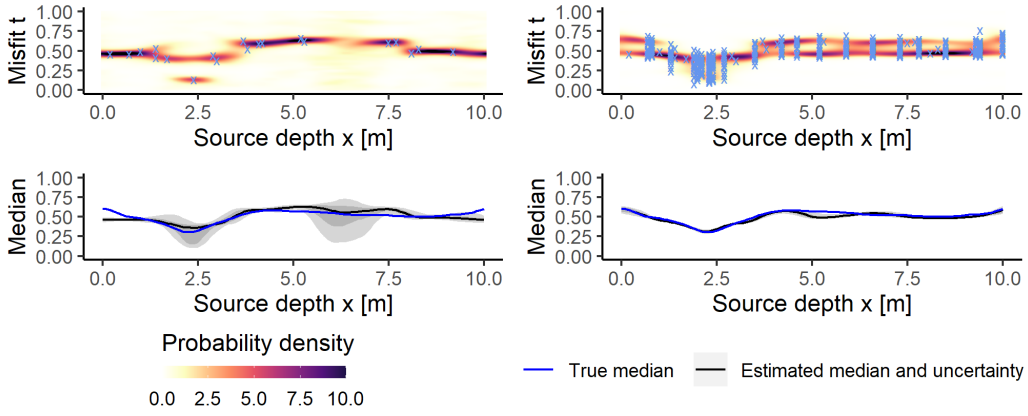


Figure 5.8: Results at the beginning of the algorithm [left] and after the 25th step [right].
Mean field estimated and samples available [top]; Estimated VS reference median [bottom].

In this first setting, the global minimum (at 2,24m) is easily found and the algorithm focuses on improving its estimation of the median. This corresponds to an exploitation-oriented approach.

With the second latent geological structure, we found out, as reflected in Figure 5.9 that our approach was also able to locate the minimum, this time by focusing on exploring by adding observations at new locations of interest.

For our small benchmark, the starting design consists in $n = 20$ data points $(\mathbf{x}_i, t_i)$ from the reference fields heterogeneously scattered across space. Their location are independent uniformly distributed across parameter space. At each step, observations are added in batches of 20 at the same location. We repeat 24 independent instances of the optimisation process for each strategy and each application and compare the performances in term of optimality gap (difference between real and estimated optimal medians). This approach is favoured due to the relatively high cost of one evaluation of the EQI criterion for SLGP, but we expect it to be detrimental to our GP-based competitors, as GPs would benefit more from having scattered observations rather than batches scattered over different points.

We compare different strategies for modelling the field and choosing the next sampling location. The value of the minimiser is inferred by modelling the function of interest with one of three models: the first one, that will be called homoskedastic GP consists in a GP regression where the observation noise level is assumed to be uniform throughout the domain. The second one, a GP regression with input dependent noise rates as in Kersting et al. (2007), Binois et al. (2018) will be called heteroscedastic GP. The last one is the SLGP model.
For each of these three models, we compare a non-adaptive approach (at each step, new observations are added at a location chosen uniformly at random) to an adaptive approach. The criteria used are: Approximated Knowledge Gradient (AKG) as implemented in the R package DiceOptim version 2.0.1 (Picheny et al., 2020) for the homoskedastic GP, the Expected improvement, as implemented in the R package hetGP version 1.2.1 (Binois and Gramacy, 2019) for the heteroscedastic GP, and the EQI criterion from Equation 5.1 for the SLGP. The results of the benchmark are shown in Figure 5.9.

Hyperparameters of the two GP's are estimated by maximum likelihood. For the adaptive SLGP, we decided to display the results obtained when minimising a 90% quantile of the random median, as our first experiments showed no strong sensitivity to the chosen level α. On a personal laptop with 16Go RAM and a simple R implementation, performing the SLGP density estimation took between 13 and 24 minutes depending on the sample size, and inferring the EQI with 150 simulations for 101 locations took 5 additional minutes.

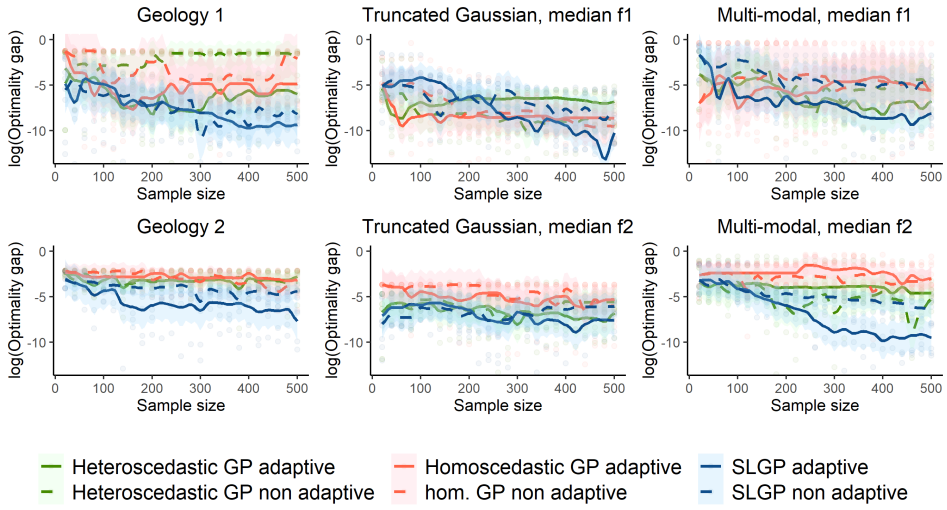


Figure 5.9: Median of the log optimality gap for the 6×6 considered strategies and test cases.

One notices that in the most complex situations, the sampling scheme based on GP modelling of the functional performs worse than the approaches based on SLGP modelling. We found out that GP based approaches tends to get trapped in local optima.

5.2 GP-based surrogating for stochastic inverse problems.

Inverse problems, as described in (Tarantola, 2005) involve finding an unknown parameter or function from indirect, noisy, or incomplete observations. These problems arise in a wide range of applications including machine learning, signal processing, image analysis (Bertero et al., 2021), but here our main focus will be inverse problems arising in natural sciences. The challenges of inverse problems include the presence of noise, non-uniqueness of solutions, and the need for regularization to prevent overfitting. Inverse problems can be broadly classified into two categories: deterministic and Bayesian (as addressed in (Stuart, 2010)). Deterministic inverse problems are solved by minimizing a criterion function that measures the misfit between the data and the model predictions, while Bayesian inverse problems are solved by updating a probability distribution over the parameters based on the data.

In the Bayesian framework, the solution to an inverse problem requires the computation of the posterior distribution of the parameters, given the observa-

tions and a model that connects the parameters to the observations. However, in many cases, the likelihood function is difficult or impossible to calculate, either because it is (considered to be) mathematically intractable or computationally infeasible to evaluate. The framework of Likelihood Free Inference (LFI) has been developed to address this issue. Approximate Bayesian Computation (ABC) methods Beaumont et al. (2009); Marin et al. (2012) have arguably become the most popular class of approaches to perform LFI. It draws it strength from the availability of complex simulation models that allow accurate modelling of complex phenomena (Herbel et al., 2017; Holden et al., 2018; McKinley et al., 2018; Weyant et al., 2013). ABC aims at identifying parameters leading to simulation results similar to observed data, by-passing in turn the need to evaluate the likelihood function.

This chapter deals with Bayesian inference and give a brief overview of LFI, with a focus on ABC methods. We provide a concise summary of commonly used acceleration techniques, with particular attention given to GP-based methods. We then propose two novel strategies within this framework. We then draw inspiration from the field of probabilistic forecasting to present a systematic approach for evaluating performance. We illustrate the applicability of the proposed framework on a reduced benchmark from hydrogeology.

5.2.1 Approximate Bayesian computation

We assume that the reader has some degree of familiarity with the basics of Bayesian inference, and can consult resources such as O'Hagan and Forster (2004); Robert (2007); Gelman et al. (1995) for further information. Similarly, for an overview of Likelihood Free Inference, the reader can refer to Marin et al. (2012); Hartig et al. (2011); Turner and Van Zandt (2012); Sisson et al. (2018). In this document, we will only provide the necessary information for it to be self-sufficient.

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

$$\pi[\mathbf{x}|t_{obs}] = \frac{\pi[t_{obs}|\mathbf{x}]\pi[\mathbf{x}]}{\int_D \pi[t_{obs}|\mathbf{x}']\pi[\mathbf{x}']\, d\mathbf{x}'} \propto \pi[t_{obs}|\mathbf{x}]\pi[\mathbf{x}] \tag{5.7}$$

It is common in practice to focus on the rightmost term called unnormalised posterior. Depending on the application, one can seek to approximate the whole posterior on $\mathbf{x}$ or instead focus on some summary statistics (most commonly the mean or variance). For a suitable function ρ, it can be expressed as:

$$\mathbb{E}_{\mathbf{x}|t_{obs}}[\rho(\mathbf{x})] = \int_D \rho(\mathbf{x})\pi[\mathbf{x}|t_{obs}]\, d\mathbf{x} \tag{5.8}$$

It is also frequent to use other summary statistics such as the Maximum A Posterior (MAP), obtained as:

$$\mathbf{x}^* \in \arg\max_{\mathbf{x}\in D} \pi[\mathbf{x}|t_{obs}] \tag{5.9}$$

No matter the end goal, be it computing the quantity in Equation 5.7, Equation 5.8 or Equation 5.9, the likelihood function is a crucial component in determining the probability of the data given the parameters and hence the posterior. However, in many cases, it happens to be intractable or too costly to evaluate, making it impossible to achieve exact Bayesian Inference. To overcome this problem, Approximate Bayesian Computation (ABC) is a widely used framework that provides an alternative approach for inferring the parameters of interest.

Approximate Bayesian Computation: key idea. 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 $\mathbf{x}$. These simulations, also called pseudo-data are assumed to be drawn exactly from the data-generating process, i.e. $t_{\mathbf{x}} \sim \pi[\cdot|\mathbf{x}]$. 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 $T_{\mathbf{x}}$ the random response with input $\mathbf{x}$ and viewing $\mathbf{x}$ as random with prior density π, the essence of ABC is to approximate the posterior as follows:

$$\pi[\mathbf{x}|t_{obs}] \approx \pi[\mathbf{x}|\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta] =: \pi_{ABC}[\mathbf{x}|t_{obs}; \delta], \tag{5.10}$$

where $\delta > 0$ is a prescribed "small enough" threshold. From now on, we shall refer to $\pi[\mathbf{x}|t_{obs}]$ as posterior or exact posterior and $\pi_{ABC}[\mathbf{x}|t_{obs}; \delta]$ as the ABC-posterior. Note that Bayes theorem can be applied to the ABC posterior and yields:

$$\pi_{ABC}[\mathbf{x}|t_{obs}; \delta] \propto \pi[\Delta(t_{obs}, t_{\mathbf{x}}) \leq \delta|\mathbf{x}]\pi[\mathbf{x}], \tag{5.11}$$

The term $\pi[\Delta(t_{obs}, t_{\mathbf{x}}) \leq \delta|\mathbf{x}]$ will prove to be of particular interest in the upcoming development, and we will refer to it as ABC-likelihood. Note that it depends

on the distribution of the dissimilarities between observations and simulations at parameter $\mathbf{x}$, and as such we will also call it misfit cumulative distribution function, or, when encountered in the form $\pi[\Delta(t_{obs}, t_{\mathbf{x}}) = \delta|\mathbf{x}]$ misfit density.

ABC in practice The most basic ABC algorithm, the ABC rejection sampler, described in Pritchard et al. (1999); Tavaré et al. (1997), can be summarised by the following pseudo code:

Algorithm 6: ABC rejection sampler

input : Prior distribution $\pi[\mathbf{x}]$, simulation model $\pi[t_{\mathbf{x}}|\mathbf{x}]$, threshold δ, number of steps T

for $i \leftarrow 1$ **to** T **do**

  Draw $\mathbf{x}_i$ from $\pi[\mathbf{x}]$

  Simulate y_i from $\mathcal{F}_{\mathbf{x}_i}$

  Accept $\mathbf{x}_i$ if $\Delta(t_{obs}, y_i) \leq \delta$

output: Parameters $\mathbf{x}_i$ that have been accepted

There are two main challenges in this formulation:

- The dissimilarity measure Δ is crucial for the estimation, and it typically requires expert knowledge to design it in a way that accurately reflects the important features of the system. As pointed out in Marin et al. (2012), comparing observations and simulations element by element can be ineffective, and it is often recommended to use distances between low-dimensional summary statistics instead. However, finding sufficient summary statistics is often not possible and leads to the frequent use of insufficient statistics, which results in a loss of information and added error in ABC methods. For a more in-depth review on summary statistics selection methods for ABC, the reader can refer to Sisson et al. (2018). Note that in this thesis, we decided to assume that suitable summary statistics (and hence dissimilarity measure) were provided, and we do not concern ourselves with these choices.

- The threshold parameter δ plays a central role in balancing the trade-off between numerical efficiency and precision in the posterior estimation. A larger value of δ will result in accepting most if not all simulated data, producing samples from the prior. On the other hand, decreasing δ towards 0 increases the accuracy but at the cost of making the algorithm less efficient, as more simulations are required to obtain enough accepted parameters.

In practice, practitioners often use variations of Algorithm 6. One such variation is to run the algorithm until a fixed number of accepted simulations is

reached. However, this approach can be impractical due to time or cost constraints. Another common approach is to run a fixed number of simulations, store all discrepancies, and set δ to a small quantile of them after the fact (as done in Beaumont et al. (2002)).

No matter the algorithm considered, the core of standard ABC methods lies in the approximation of the ABC posterior in Equation 5.10 by a Monte-Carlo sum of retained parameter values. This approximation results in three sources of information loss: the use of (often) insufficient summary statistics, the error introduced by non-zero δ and the Monte Carlo error.

In this thesis, we shall not concern ourselves with the first two sources of uncertainty as one remains a problem with no satisfactory general solution and the other is unavoidable, as it constitutes the "approximate" in ABC. Instead, we will focus on other methods for approximating the ABC posterior as well as improving sampling efficiency.

5.2.2 Accelerating inference in ABC

Common approaches in accelerating inference

As we briefly mentioned when discussing the choice of δ, computational efficiency is a challenge of ABC. Poor choices of algorithm, small values of δ and a prior that is substantially broader than the posterior are three causes of computational inefficiency and cause most simulations to be rejected.

To address this problem, several approaches aiming at more efficient ABC posterior sampling have been developed. We give here a brief review of the principles behind the methods we found to be the most noteworthy.

Markov Chain Monte Carlo ABC. First proposed by Marjoram et al. (2003), it uses a Markov chain to explore the parameter space, and employs the acceptance/rejection step of ABC to control the chain's stationary distribution (whereas a Metropolis-Hastings implementation would require to evaluate the intractable exact posterior). The convergence properties of MCMC-ABC have been studied in Andrieu and Roberts (2009). However, in practice, MCMC-ABC can suffer from poor mixing, where the algorithm can get stuck for a long time after a point in a far-off region of the parameter space is accepted.

Sequential and population Monte Carlo ABC. These classes, that we later denote SMC-ABC and PMC-ABC, group several methods, such as those introduced in Beaumont et al. (2009); Bonassi and West (2015); Del Moral et al. (2012); Drovandi and Pettitt (2011); Lenormand et al. (2013); Sisson et al. (2007); Toni et al. (2009).These methods rely on replacing draws of $\mathbf{x}$ from the

prior with draws from an adapted proposal density. These algorithms are widely used and have been shown to be more efficient than standard ABC, but they still require a large number of simulations to obtain accurate results.

Synthetic Likelihood In the synthetic likelihood methods, Price et al. (2018); Wood (2010), the summary statistics distribution is assumed to stem from a parametric family (usually Gaussian). It is a parametric approach that yields a probabilistic prediction of the ABC-posterior and can be leveraged for sequential design of experiments.

The approach we present in this thesis belongs to the synthetic likelihood class. It leverages continuity assumption in the summary statistics/likelihood to reduce the simulation cost of the inference and is called GP-ABC. We present the main ideas of it in Section 5.2.2 before presenting our adaption dubbed SLGP-ABC in Section 5.2.2.

GP-ABC, and guiding data acquisition

GP-ABC is an approach that was first introduced in Wilkinson (2014) and models the summary statistics likelihood using a Gaussian Process. It utilizes a sequential history matching process where the GP is used to eliminate regions of the parameter space that are unlikely to yield interesting values. New simulations are then selected in the remaining plausible regions. The final step involves using a Metropolis Hastings algorithm to sample from the posterior and perform the inference of the ABC-posterior. The approach has been shown to be efficient in high-dimensional problems, where the number of parameters is large and traditional ABC methods may struggle. However, it requires having access to (estimates of) the log-likelihood, which are typically obtained through Monte-Carlo sum based on numerous replicates.

To overcome this limitation, further developments in the field have been proposed such as surrogating the summary statistics directly with a GP as presented in works by Meeds and Welling (2014) or Jabot et al. (2014). Another related framework is BOLFI (Bayesian optimisation for likelihood-free inference, Gutmann et al. (2016)) , which surrogates the misfit (or log-misfit) with a Gaussian Process. We assume that a dataset $\mathcal{D}_{1:n} := \{\mathbf{x}_i, \Delta_i\}_{1\leq i\leq n}$, where $\Delta_i := \Delta(t_{obs}, t_i)$ is available, and we use the notation $\mathbf{\Delta}$ for the concatenated vector of Δ_i values. The Δ_i's are modelled as a (noisy) function of $\mathbf{x}$ with a (noisy) Gaussian Process Regression (GPR). Assuming that the observation noise is denoted τ^2, and the prior used was $\Delta_i = Z_{\mathbf{x}_i} + \varepsilon_i$ $(1 \leq i \leq n)$ with $Z \sim \mathcal{GP}(0, k)$ and the ε_is i.i.d. $\mathcal{N}(0, \tau^2)$, we have $Z_{\mathbf{x}}^{(n)} := Z_{\mathbf{x}}|\mathcal{D}_{1:n} \sim \mathcal{GP}(m_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}'))$, with for any

$\mathbf{x}, \mathbf{x}' \in D$:

$$m_n(\mathbf{x}) := k(\mathbf{x}, \mathbf{X})k(\mathbf{X}, \mathbf{X})^{-1}\mathbf{\Delta} \tag{5.12}$$

$$k_n(\mathbf{x}, \mathbf{x}') := k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X})\left[k(\mathbf{X}, \mathbf{X}) + \mathrm{Diag}(\tau^2)\right]^{-1} k(\mathbf{x}', \mathbf{X})^\top \tag{5.13}$$

where $k(\mathbf{x}, \mathbf{X})$ denotes the vector $(k(\mathbf{x}, \mathbf{x}_i))_{1\leq i\leq n}$, $k(\mathbf{X}, \mathbf{X})$ is the matrix $(k(\mathbf{x}_i, \mathbf{x}_j))_{1\leq i,j\leq n}$ and $\mathrm{Diag}(\tau^2)$ is the diagonal matrix with elements τ^2 on the diagonal.

This model yields a predictive distribution for new misfit values obtained from the model. It is then used to obtain an estimation of the ABC likelihood:

$$\hat{\pi}_{\text{GP-BOLFI}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta | \mathbf{x}, \mathcal{D}_{1:n}] = \frac{\int_{-\infty}^{\delta} e^{-\frac{(u-m_n(\mathbf{x}))^2}{2(k_n(\mathbf{x},\mathbf{x})+\tau^2)}} du}{\sqrt{2\pi(k_n(\mathbf{x}, \mathbf{x}) + \tau^2)}} \tag{5.14}$$

Similarly, when modelling the log-misfit one uses a dataset $\mathcal{D}_{1:n} := \{\mathbf{x}_i, \log \Delta_i\}_{1\leq i\leq n}$, and the previous equation can easily be adapted. Since it does not yield any significant difference (other than δ being replaced with $\log \delta$ in the right-hand term of the previous equation), we shall focus here on equations where the misfit is modelled directly.

Guiding data acquisition: an insight from Bayesian Optimisation Leveraging this uncertainty if a key feature of the BO LFI framework as compared to other GP-ABC methods. It fully considers the probabilistic nature of the GP-ABC posterior and uses Bayesian Optimisation techniques to select new data points. In particular, in Gutmann et al. (2016) the authors guide the selection of new data points with the Lower Confidence Bound (LCB) acquisition function defined as:

$$\mathrm{LSB}(\mathbf{x}) := m_n(\mathbf{x}) - \beta\sqrt{k_n(\mathbf{x}, \mathbf{x})} \tag{5.15}$$

where β is a trade-off parameter, destined to balance between evaluating the GP where the mean is small or where the uncertainty is large. Such exploration-exploitation tradeoff is common in BO settings.

After gathering more data, the final step within the BOLFI setting consists in using MCMC to sample from the resulting model-based ABC posterior.

These methods resulted in improvement of the computational efficiency by several orders of magnitude and produced reasonable yet conservative ABC posterior approximations. The use of BO acquisition functions in this setting was motivated by the fact that regions with small discrepancy usually correspond to those with non-negligible likelihood. However, the criterion used are not explicitly designed for estimating the ABC posterior and do not always work

as intended, as demonstrated in Järvenpää et al. (2019). In fact, these methods tended to stay stuck in unsuitable regions of space, and the need for more appropriate approaches was noted by the authors.

Guiding data acquisition: ABC-specific framework The authors of Järvenpää et al. (2019) instead proposed ABC-oriented framework that fully leverage the probabilistic nature of GP regression. Indeed, after modelling the misfit with a GP as in the previous subsection, rather than considering the estimation in Equation 5.14, they instead focused on a predictive distribution of the ABC-likelihood:

$$\hat{\pi}_{\mathrm{GP}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta | \mathbf{x}, \mathcal{D}_{1:n}] = \frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\delta} e^{-\frac{\left(u - Z_{\mathbf{x}}^{(n)}\right)^2}{2\tau^2}} \, du \tag{5.16}$$

First, note that one can easily consider the GP-ABC posterior, which retains a probabilistic nature due to the residual uncertainty on $Z_{\mathbf{x}}|\mathcal{D}_{1:n}$:

$$\hat{\pi}_{\text{GP-ABC}}[\mathbf{x}|\mathcal{D}_{1:n}] \propto \pi[\mathbf{x}]\hat{\pi}_{\mathrm{GP}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta | \mathbf{x}, \mathcal{D}_{1:n}] \tag{5.17}$$

Note that since we consider that the threshold δ is fixed, we decided to make it implicit in the previous equation and all those who follow, so as to shorten notations. It is possible to derive some summary statistics of this GP-ABC posterior, as pointed out in Järvenpää et al. (2019):

$$\mathbb{E}_{Z^{(n)}}\left[\hat{\pi}_{\mathrm{GP}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta | \mathbf{x}, \mathcal{D}_{1:n}]\right] = \int_{-\infty}^{a_n(\mathbf{x})} e^{-u^2/2} \, du \tag{5.18}$$

$$\text{where } a_n(\mathbf{x}) := \frac{\delta - m_n(\mathbf{x})}{\sqrt{\tau^2 + k_n(\mathbf{x}, \mathbf{x})}} \tag{5.19}$$

$$\mathrm{Med}_{Z^{(n)}}\left[\hat{\pi}_{\mathrm{GP}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta | \mathbf{x}, \mathcal{D}_{1:n}]\right] = \frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\delta} e^{-\frac{(u - m_n(\mathbf{x}))^2}{2\tau^2}} \, du \tag{5.20}$$

where Med denotes the median. These formulae are useful, but they do not capture uncertainty on the surrogated posterior.

Denoting by $\mathcal{D}^*$ future data to be collected at a vector of locations $\mathbf{X}^*$, and shortening the notations by using $\hat{\pi}_{GP-ABC}[\mathbf{x}]$ instead of $\hat{\pi}_{\text{GP-ABC}}[\mathbf{x}|\mathcal{D}_{1:n}]$, we have:

Expected Integrated Variance (EIV)

$$L_{\mathrm{EIV}}(\mathbf{x}^*) := \mathbb{E}_{t^*|\mathbf{x}^*}\left[\int_D \mathrm{Var}_{\mathrm{GP}|\mathcal{D}_{1:n} \cup \mathcal{D}^*}\left(\hat{\pi}_{GP-ABC}[\mathbf{u}]\right) d\mathbf{u}\right] \tag{5.21}$$

Expected Integrated Mean Absolute Deviation (EIMAD)

$$L_{\mathrm{EIMAD}}(\mathbf{X}^*) := \mathbb{E}_{t^*|\mathbf{X}^*}\left[\int_D \mathbb{E}_{\mathrm{GP}|\mathcal{D}_{1:n}\cup\mathcal{D}^*}\left[\mathrm{MAD}(\hat{\pi}_{GP-ABC}[\mathbf{u}])\right] d\mathbf{u}\right] \tag{5.22}$$

where $\mathrm{MAD}(\hat{\pi}_{GP-ABC}[\mathbf{x}]) := |\hat{\pi}_{GP-ABC}[\mathbf{x}] - \mathrm{Med}(\hat{\pi}_{GP-ABC}[\mathbf{x}])|$.

Both these criteria focus on reducing the surrogated ABC posterior variability. At first glance, one could assume that the efforts are focused towards reducing uncertainty without truly seeking to identify promising regions of parameter space. However, since it is targeted towards the ABC posterior rather than the whole misfit distribution, it will naturally gather data that will be beneficial for inversion.

Due to the specific form of the GP-ABC posterior and the fact that they involve simple functionals thereof, one can show that they can be reformulated to obtain analytical expressions. Indeed, by letting T be the Owen's T function (Owen, 1956), and integrating with respect to Lebesgue measure, we have:

$$\begin{aligned} L_{\mathrm{EIV}}(\mathbf{X}^*) = 2\int_D \pi[\mathbf{u}]^2 \Bigg[& T\left(a_n(\mathbf{u}), \frac{\sqrt{\tau^2 + k_n(\mathbf{u},\mathbf{u}) - c_n(\mathbf{u},\mathbf{X}^*)}}{\sqrt{\tau^2 + k_n(\mathbf{u},\mathbf{u}) + c_n(\mathbf{u},\mathbf{X}^*)}}\right) \\ & - T\left(a_n(\mathbf{u}), \frac{\sqrt{\tau^2}}{\sqrt{\tau^2 + 2k_n(\mathbf{u},\mathbf{u})}}\right)\Bigg] d\mathbf{u} \end{aligned} \tag{5.23}$$

$$L_{\mathrm{EIMAD}}(\mathbf{X}^*) = 2\int_D \pi[\mathbf{u}]^2 \left[T\left(a_n(\mathbf{u}), \frac{\sqrt{k_n(\mathbf{u},\mathbf{u}) - c_n(\mathbf{u},\mathbf{X}^*)}}{\sqrt{k_n(\mathbf{u},\mathbf{u}) + c_n(\mathbf{u},\mathbf{X}^*)}}\right)\right] d\mathbf{u} \tag{5.24}$$

where $c_n(\mathbf{x},\mathbf{X}^*) := k_n(\mathbf{x},\mathbf{X}^*)[k_n(\mathbf{X}^*,\mathbf{X}^*) + \mathrm{Diag}(\tau^2)]^{-1}k_n(\mathbf{X}^*,\mathbf{x})$.

We will draw inspiration from this setting to derive two novel approaches.

HetGP-ABC First, we propose applying the GP-ABC framework with Gaussian Processes that do not necessarily have an homoskedastic variance. We use a plug-in of the heteroskedastic GP models from Binois et al. (2018) as implemented in the package Binois and Gramacy (2019), and use the same criteria EIV and EIMAD for data acquisition.

SLGP-ABC as a synthetic likelihood method

The GP-ABC frameworks allows for a straightforward extension, by using SLGP-modelling for the misfit distribution. Indeed, given a dataset $\mathcal{D}_{1:n} := \{\mathbf{x}_i, \Delta_i\}_{1\leq i\leq n}$, where Δ_i are realisations of $\Delta(t_{obs}, T_i)$, a SLGP can be leveraged to learn the

probability density of Δ_i indexed by $\mathbf{x}$. Let us consider a SLGP Y and denote by $Y^{(n)} := Y|\mathcal{D}_{1:n}$ the SLGP conditioned on the available data. Then, we have:

$$\hat{\pi}_{\text{SLGP}}[\Delta(t_{obs}, T_{\mathbf{x}}) \leq \delta|\mathbf{x}, \mathcal{D}_{1:n}] \propto \int_0^{\delta} Y^{(n)}_{\mathbf{x},u}\, du \tag{5.25}$$

When modelling the log-misfit, one has a similar expression. Under this model, we would have a SLGP $\tilde{Y}$ that once conditioned on data would be written as $\tilde{Y}^{(n)} := Y|\left\{\mathbf{x}_i, \log \Delta_i\right\}_{1\leq i\leq n}$ and the SLGP-ABC posterior is easily obtained as:

$$\hat{\pi}_{SLGP-ABC}[\mathbf{x}|\mathcal{D}_{1:n}] \propto \pi[\mathbf{x}] \int_0^{\log \delta} \tilde{Y}^{(n)}_{\mathbf{x},u}\, du \tag{5.26}$$

The criteria presented in the framework of GP-ABC are easily transposed to this new framework.

Expected Integrated Variance (EIV)

$$L_{\text{EIV}}(\mathbf{x}^*) := \mathbb{E}_{t^*|\mathbf{x}^*}\left[\int_D \text{Var}_{\text{SLGP}|\mathcal{D}_{1:n}\cup\mathcal{D}^*}\left(\hat{\pi}_{SLGP-ABC}[\mathbf{u}]\right) d\mathbf{u}\right] \tag{5.27}$$

Expected Integrated Mean Absolute Deviation (EIMAD)

$$L_{\text{EIMAD}}(\mathbf{X}^*) := \mathbb{E}_{t^*|\mathbf{X}^*}\left[\int_D \mathbb{E}_{\text{SLGP}|\mathcal{D}_{1:n}\cup\mathcal{D}^*}\left[\text{MAD}(\hat{\pi}_{SLGP-ABC}[\mathbf{u}])\right] d\mathbf{u}\right] \tag{5.28}$$

Within the SLGP modelling framework, we do not enjoy analytical expressions for L_{EIV} and L_{EIMAD}, as opposed to the ones available for the GP-ABC framework. We will rely on simulation-based estimations of the criteria, as already presented for Bayesian Optimisation in Section 5.1.2.

However, we hope that the improved flexibility of the SLGP models compared to that of GP models will result in better modelling of some misfit distributions on application cases. Quantifying this change calls for inquiries in the field of probabilistic forecasting.

5.2.3 Evaluating the performances: scoring of distributions

Depending on the inference approach employed, one either obtains a sample from the posterior (in standard ABC, MCMC-ABC, SMC-ABC or PMC-ABC) or a probabilistic prediction of it (as in GP-ABC or SLGP-ABC). Evaluating the predictive performance of an approach requires being able to compare samples drawn from a distribution (e.g. from the ABC-posterior), a deterministic

distribution (either the ABC posterior or the true posterior) and a probabilistic prediction of the ABC-posterior. As such, it naturally calls for investigations in the field of scoring probabilistic forecasts of densities.

In probabilistic forecasting, the process of evaluating the accuracy of a forecast is known as scoring. This is done by assigning a numerical value to each probabilistic forecast distribution. It is common to use a proper scoring rule, i.e. rules that give a lower score to forecasts that are more accurate. There are different types of scoring rules, each one tailored to different types of forecast distributions and applications.

General setting of scoring We start by giving formal definitions of the framework that we informally described. We relay here definitions and properties stated in Steinwart and Ziegel (2021). Let us consider $(X, \mathcal{F})$ a measurable space and $\mathcal{M}_\infty(X)$ the class of all probability measures on X.

Definition 5.2.1 (Scoring rule)**.** For $\mathcal{P} \subset \mathcal{M}_\infty(X)$, a scoring rule is a function $\mathcal{S} : \mathcal{P} \times X \to [-\infty, \infty]$ such that the integral $\int \mathcal{S}(P, x)\, dQ(x)$ exists for all $P, Q \in \mathcal{P}$

Definition 5.2.2 (Propriety of a scoring rule)**.** A scoring rule $\mathcal{S}$ is called proper if:

$$\int \mathcal{S}(P, x)\, dP(x) \leq \int \mathcal{S}(Q, x)\, dP(x) \text{ for all } P, Q \in \mathcal{P} \tag{5.29}$$

It is called strictly proper is equality in the previous equation implies $P = Q$.

We refer the reader to Gneiting and Raftery (2007) for more details on scoring rules and their propriety, and shall now focus on a broad class of scoring rules called kernel scores.

Definition 5.2.3 (Kernel scores)**.** The kernel score $\mathcal{S}_k$ associated with a conditionally positive definite measurable "kernel" k is the scoring rule defined by:

$$\mathcal{S}_k(P, x) := -\int k(\mathbf{u}, x)\, dP(\mathbf{u}) + \frac{1}{2}\int\int k(\mathbf{u}, \mathbf{v})\, dP(\mathbf{u})\, dP(\mathbf{v}) + \frac{1}{2}k(x, x) \tag{5.30}$$

We present here the definition from Gneiting and Raftery (2007) where k is only required to be conditionally p.d., but adapted in the flavour of Ziegel et al. (2022) where the addition of the term $\frac{1}{2}k(x, x)$ ensures the non-negativity of the scoring rule without affecting any of its other properties.

The framework considered leverages the versatility of kernel methods. As proved in Theorem 4 of Gneiting and Raftery (2007), the kernel score is a proper rule relative to Borel probability measures on a Hausdorff space. Additionally, a careful choice of the underlying kernel can directly induce the strict propriety of the kernel score, as pointed out in Steinwart and Ziegel (2021).

Scoring (probabilistic) probability distributions In the current work, we want to design a scoring rule on the space of probability distributions on $\mathcal{I}$. Thankfully, we noted in Chapter 2, Proposition 2.2.4 that the negative squared MMD induced by a kernel k on $\mathcal{I} \times \mathcal{I}$, is itself a conditionally positive kernel.

This setting calls for additional notations, so as to avoid confusion between the various quantities at hand. First, note that in definition 5.2.1, we considered a space X, here chosen to be the space of probability distributions on $\mathcal{I}$. As noted when introducing Random (Probability) Measures in Section 2.3, the space of probability distributions on $\mathcal{I}$ is a measurable space when equipped with the σ-algebra from Definition 2.3.1. Elements of X were denoted with the general notation x in the previous definitions, but we will from now-on prefer the notations P or Q to emphasize that they are probability distributions. We will also prefer the notation $\mathbb{P}$ to refer to a distribution on X (previously denoted P). The corresponding scoring rule is:

$$\begin{aligned}\mathcal{S}_{\mathrm{MMD}}(\mathbb{P}, Q) :=& \mathbb{E}_{P\sim\mathbb{P}}\left[\mathrm{MMD}^2(P, Q)\right] - \frac{1}{2}\mathrm{MMD}^2(Q, Q) \\ &- \frac{1}{2}\mathbb{E}_{P,P'\sim\mathbb{P}}\left[\mathrm{MMD}^2(P, P')\right]\end{aligned} \tag{5.31}$$

As noted when introducing kernel embedding earlier in this document, it is possible to express the MMD induced by k as an explicit function of k. After replacing MMD^2 with in Equation 5.31 with its expression given in Equation 2.27 and accounting for terms that cancel out, we obtain the reformulation:

$$\begin{aligned}\mathcal{S}_{\mathrm{MMD}}(\mathbb{P}, Q) =& \mathbb{E}_{Y,Y'\sim Q}\left[k(Y, Y')\right] + \mathbb{E}_{P,P'\sim\mathbb{P}}\left[\mathbb{E}_{X\sim P,X'\sim P'}\left[k(X, X')\right]\right] \\ &- 2\mathbb{E}_{P\sim\mathbb{P}}\left[\mathbb{E}_{X\sim P,Y\sim Q}\left[k(X, Y)\right]\right]\end{aligned} \tag{5.32}$$

Using this reformulation, we can use the scoring rule on application cases.

Remark. Further theoretical inquiries are still required in order to establish such scoring rules are proper. Furthermore, in upcoming developments, it might be worth considering kernels and scoring rules inspired by those in Ziegel et al. (2022).

Using the kernel scoring rule for scoring ABC posteriors In practice, one does not have access to $\mathbb{P}$, and practitioners want to compare $K \geq 1$ (approximate) draws from it to a reference distribution π_{ref}.

We will focus on our current setting, and denote by $\left(\pi_{\mathrm{ABC}}^{(i)}\right)_{1\leq i\leq K}$ the realisations of the ABC posterior under a probabilistic modelling.

Depending on the considered setting, we have different ways of approximating the scoring rule.

- When the reference distribution π_{ref} is available, we have:

$$\begin{aligned}\mathcal{S}_{\text{MMD}}(\mathbb{P}, \pi_{\text{ref}}) =& \frac{2}{K(K-1)} \sum_{1\leq i<j\leq K} \int\int k(\mathbf{u},\mathbf{v})\, d\pi^{(i)}_{\text{ABC}}(\mathbf{u})\, d\pi^{(j)}_{\text{ABC}}(\mathbf{v}) \\ &- \frac{2}{K} \sum_{1\leq i\leq n} \int\int k(\mathbf{u},\mathbf{v})\, d\pi_{\text{ref}}(\mathbf{u})\, d\pi^{(i)}_{\text{ABC}}(\mathbf{v}) \\ &+ \int\int k(\mathbf{u},\mathbf{v})\, d\pi_{\text{ref}}(\mathbf{u})\, d\pi_{\text{ref}}(\mathbf{v})\end{aligned} \tag{5.33}$$

  Such a setting is possible when considering either the true posterior $\pi[\mathbf{x}|t_{obs}]$ or the ABC-posterior $\pi_{ABC}[\mathbf{x}|t_{obs};\delta]$ of an analytical test case or of a simple problem. Note that we have $K(K+1)/2+1$ double integrals to compute.

- When the reference distribution π_{ref} is not available, but we have access to a sample of i.i.d. samples from it that we denote by $(\mathbf{x}^{(i)})_{1\leq i\leq m}$, we can use the empirical distribution $\pi_{\text{emp}} := \frac{1}{m}\sum_{1\leq i\leq m} \delta_{\mathbf{x}^{(i)}}$ as a reference:

$$\begin{aligned}\mathcal{S}_{\text{MMD}}(\mathbb{P}, \pi_{\text{emp}}) =& \frac{2}{K(K-1)} \sum_{1\leq i<j\leq K} \int\int k(\mathbf{u},\mathbf{v})\, d\pi^{(i)}_{\text{ABC}}(\mathbf{u})\, d\pi^{(j)}_{\text{ABC}}(\mathbf{v}) \\ &- \frac{2}{Km} \sum_{1\leq i\leq K} \sum_{1\leq j\leq m} \int k(\mathbf{u},\mathbf{x}^{(j)})\, d\pi^{(i)}_{\text{ABC}}(\mathbf{u}) \\ &+ \frac{2}{m(m-1)} \sum_{1\leq i<j\leq m} k(\mathbf{x}^{(i)},\mathbf{x}^{(j)})\end{aligned} \tag{5.34}$$

  Such a setting is possible when comparing SLGP-ABC to the standard rejection-sampling algorithm. Note that we have $K(K-1)/2$ double integrals, and mK simple ones to compute.

Although one could define as many MMD-based scoring rules are there are kernels, here we will solely focus on the ones relying on Matérn kernels. As it was mentioned in Section 2.2.4, Matérn kernels are characteristic kernels on $\mathbb{R}^d$ ($d \geq 1$) and this ensures that the resulting MMDs are metrics.

5.2.4 Benchmarking the GP-based approaches for accelerating inverse problems

Benchmark setup We go back to the unidimensional contaminant localization problem presented in Section 5.1.3. In order to perform our experiments in a controlled setting where the ABC posterior is perfectly known, we decided to model the misfit distributions such as those presented in Figure 5.7 with a SLGP. All further samples will be drawn from this field, and the ABC posterior derived from it. We used MAP estimators of the fields, and performed the optimisation similarly to what we did in application 4.4.3. In particular, we followed the methodology from Section 4.3 with hyperparameters being determined with a grid-search. Also note that for simplicity, we re-normalised the index space (for it to span over $[0, 1]$ instead of $[0m, 10m]$).

We arbitrarily set the ABC-threshold to 0.15. We represent in Figure 5.10 the scatter plots of simulated misfits used to train the fields used for ABC, the fields, and subsequent ABC-posteriors.

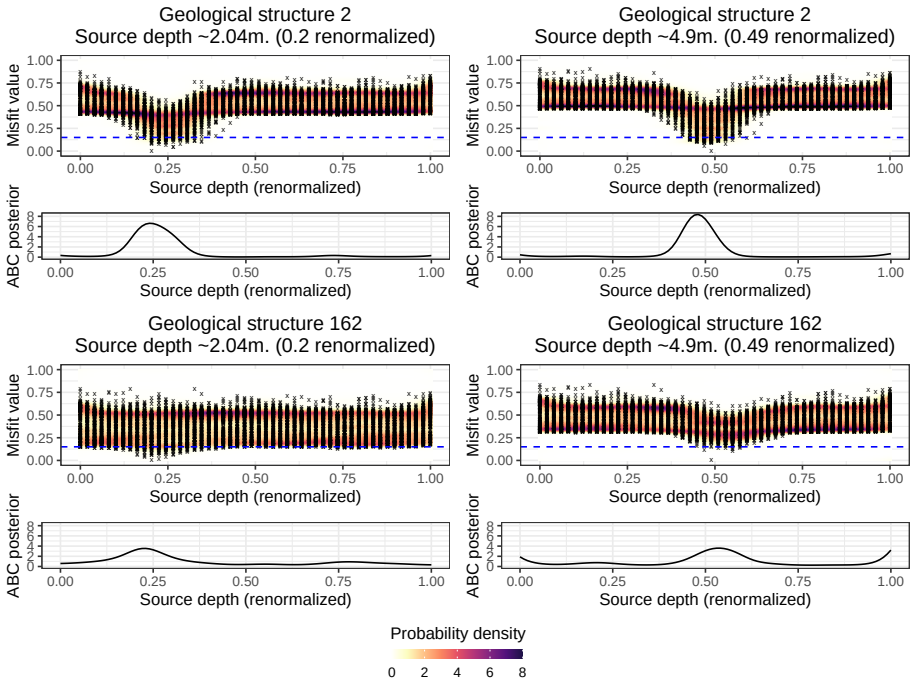


Figure 5.10: Scatter plots, fitted fields and ABC posterior for various reference geological structures and release depth.

We ran the experiments on UBELIX (`http://www.id.unibe.ch/hpc`), the HPC cluster at the University of Bern. For each ABC strategy considered

(homoskedastic GP-ABC, heteroskedastic GP-ABC, SLGP-ABC), and various sampling strategies (random according to the prior, EIMAD, EIV), we repeated our the experiment 50 times while varying the random seed each time.

The scoring rule considered is a MMD-based scoring rule, with the MMD kernel being an exponential kernel (i.e. Matérn 1/2) with lengthscale set to 0.1. This hyper-parameter was selected based on our knowledge of the real ABC-posteriors and is expected to accurately capture variations in the posteriors and samples thereof.

We compared several modelling, all implemented in R:

- Homoskedastic GP-ABC relies on the GP regression as implemented in the kergp package (Deville et al., 2021). We used a Matérn 5/2 kernel, and hyperparameters are estimated by MLE.
- Heteroskedastic GP-ABC relies on the GP regression as implemented in the hetGP package (Binois and Gramacy, 2019). We also used a Matérn 5/2 kernel, and hyperparameters are estimated by MLE.
- SLGP-ABC relies on our implementation of SLGP density field estimation (Gautier, Athénaïs, 2023). We used a Random Fourier Features approach based on the Matérn 5/2 kernel with 75 random frequencies sampled (i.e. 150 basis functions). Hyperparameters were given to the model with a lengthscale selected to be equal to 0.15.

Benchmark results The reduced benchmark's results are displayed in Figure 5.11. We recall that this score is negatively oriented and as such needs to be minimized to reach better performances. We notice that for our hydrogeological applications, the heteroskedastic GP-ABC consistently performs similarly or better than the homoskedastic GP-ABC.

Interestingly enough, performances of the SLGP can directly be related to the modality of the misfit distribution field in the area of interest. Indeed, for geological structure 2 and a normalised value of the source of 0.2 (top left panel in Figures 5.10 and 5.11), the misfit distribution was unimodal around $x = 0.2$. This might explain the better performances of homGP-ABC and hetGP-ABC. On the other hand, for geological structure 162 and a normalised value of the source of 0.49 (bottom right panel), the distribution field presented multimodality in this region, and SLGP-ABC clearly outperforms homGP-ABC and performs similarly to hetGP-ABC. Moreover, whereas the hetGP-ABC approach benefits from adaptive design of experiment in all the other settings, it is not the case in this later, suggesting that the adaptive strategies considered might be ill-suited for misspecified hetGPs (especially in the event of multimodality), and calling for further enquiries.

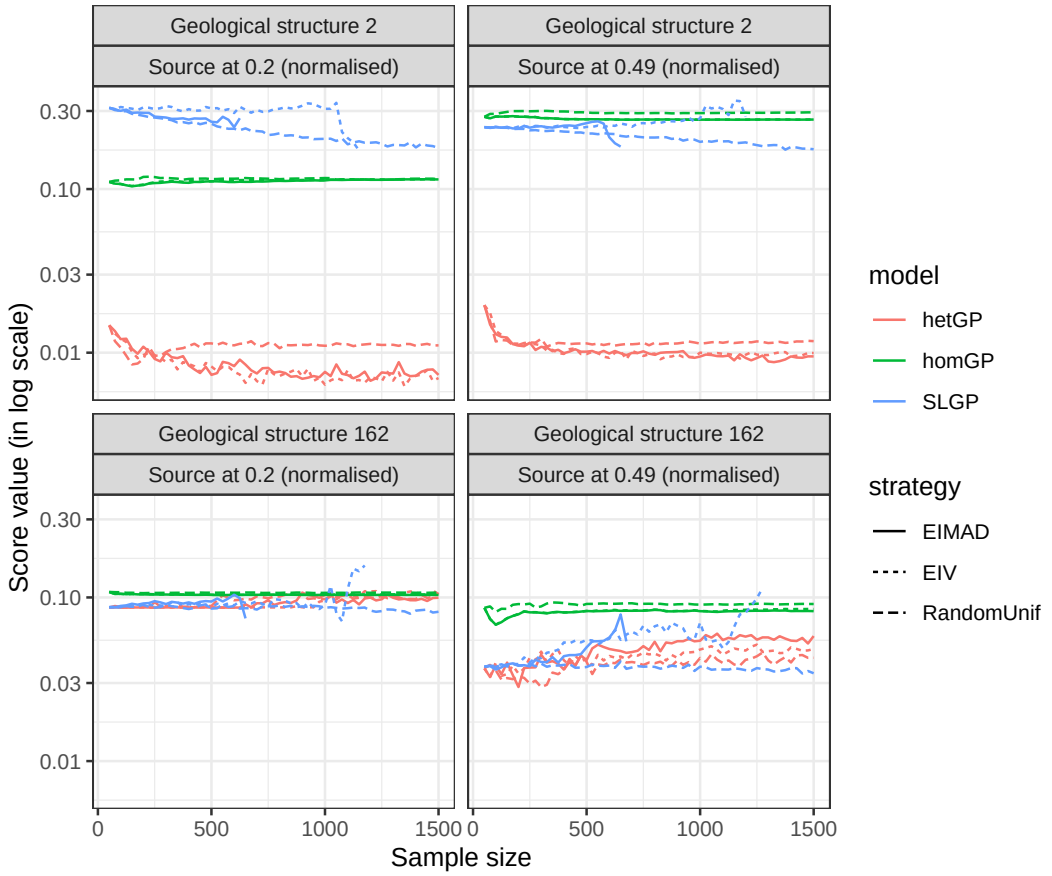


Figure 5.11: Evolution of the median score value in the benchmark.

Overall, we conclude that the gain in flexibility obtained by plugging a heteroskedastic GP in the standard homoskedastic GP-ABC framework lead to improved performances compared to the homGP-framework. It appears that the SLGP-ABC only yields significant improvements when the underlying misfit distribution is far from Gaussian in the regions of interest for ABC, but so far it does not appear to out-perform a finely tuned heteroskedastic GP.

Possible areas of improvement for SLGP-ABC and upcoming work The heavy computational machinery underlying the SLGP methods suggest that a finer tuning is harder to reach than for GPR methods, and as such, improvements on the implementation could yield performance gains.

Another interesting direction stems from an observation made while performing smaller scales experiments for preliminary results. Using the MAP estimate of the misfit distribution field provided better estimation of the ABC posterior

than using draws from the posterior. This naturally makes us consider the use of Laplace approximation in our misfit distribution field estimation.

Moreover, it is worth noting that so far we only used one (proper) scoring rule. Upcoming development will include extending the benchmark by including more test functions and several scoring rules, to see how robust these current results are.

5.2.5 Conclusion

In this section, we have provided an overview of the ABC framework, with a specific focus on the GP-ABC methods. These methods consist in surrogating the misfits with a GP, enabling practitioners to derive sequential design of experiments and accelerate the process of scientific discovery. We have proposed two adaptations of the GP-ABC framework, one involving the use of a heteroskedastic GP to surrogate the misfit distribution, and the other involving the direct use of the SLGP to surrogate the misfit distribution.

Furthermore, we have introduced a principled approach to evaluating the performance of the ABC framework, which involves the use of kernel scoring rules to quantify the quality of ABC posterior estimations.

Finally, we have presented a hydrogeological benchmark to demonstrate the effectiveness of the proposed methods. Our experiments have shown that the hetGP-ABC approach consistently outperformed the homGP-ABC baseline, highlighting the potential of this model. We have also identified potential avenues for future research, such as exploring adaptations of the SLGP-ABC approach that could further improve performance.

Chapter 6

Discussions and perspectives

We conclude this manuscript with a brief summary of the main contributions of this thesis, a discussion on the advantages of the proposed framework compared to other work in the field, and a few perspectives for upcoming developments.

This work provides a comprehensive overview of the SLGP model, showcasing its potential for use in applications in statistical inference. Beginning with a synthesis of the important background properties in Chapter 2, we then delve into the intricacies of the model itself in Chapter 3. Here, we review the kinds of stochastic processes that make up SLGP, we characterise them, quantify their spatial regularity properties, and discussing the posterior consistency of the induced prior.

This contribution to the theoretical aspects of the SLGP model was further strengthened by the practical considerations and implementation guidelines outlined in Chapter 4. This chapter demonstrates that sample-based estimation of SLGPs is both possible and has favourable properties. Using unconditional realisations, we show the sharpness of the bounds derived from the study of the spatial regularity of SLGPs, and through analytical test cases, we demonstrate the posterior consistency of our models in a controlled setting. Finally, we showcase the applicability of our models to higher dimensions through an analysis of a 3D meteorological dataset.

The final chapter, Chapter 5, highlights the versatility of SLGP models by illustrating their ease of adaptation to any setting where GPs are used as surrogate models. Through a simulation-based approach to computing data-acquisition criteria and the application of our approaches to reduced benchmarks inspired by natural sciences, we demonstrate the broad applicability of SLGP models to a range of real-world problems.

In sum, this work presents a valuable resource for those interested in exploring the potential of SLGP models in statistical inference.

The main strength of our contribution to the flourishing field of statistical

modelling for guiding discovery in the sciences lies in the combination of its high flexibility, applicability in moderate data regimes, and probabilistic nature. This makes it an attractive choice for many scientific applications.

Furthermore, we believe that due to their similarities and links to Gaussian Process (GP) models, SLGP models can benefit from the theoretical results and advancements made in the GP community.

However, the relatively high numerical complexity of SLGP models is a challenge that must be addressed in order to make the model more scalable and accessible to a wider range of users. Recent efforts in using variational methods for GP models hold promise for adapting these techniques to SLGP models, which would be a possible direction in increasing the efficiency of SLGP models.

Although not presented in the current document, we identified a promising direction for SLGP-powered metamodelling. Indeed, one can consider a Laplace approximation of the SLGP posterior. Such an approximation only requires performing the MAP estimation, but does not require MCMC runs; and can be used to perform further inference. Not only it heavily reduces the computational cost of SLGP fitting, but it also provides the advantage of yielding approximations of the posterior that are also SLGP-distributed and can rely on previous results in the field. We intend to continue exploring this venue and discuss its applicability for sequential designs in upcoming work.

In addition to these suggestions, further work towards evaluating the predictive performance of our model is needed. So far, we mostly relied on a squared integrated Hellinger distance for analytical settings where the reference field was known, and on qualitative and visual validation for the meteorological application. Comparing samples to predicted density fields calls for investigations in the field of scoring probabilistic forecasts of (fields of) densities. Therefore, evaluating the performances of the SLGP under several kernel settings, and comparing them to each other and to baseline methods is another main research direction.

Finally, this work is just the beginning of our exploration of the rich framework of SLGPs. Further extensive benchmarking with various toy and real-world simulation models will deepen our understanding of the practical limitations of SLGP models and help to identify areas for improvement. So far, implementation choices and suggestions were done mostly thanks to expert knowledge or trial and errors. We would greatly benefit from development in the methodology, as this might reduce the occurrence of artefacts, thus improving the quality of predictions. A more efficient implementation would also allow us to use the SLGP model at higher scales (i.e. with more data points and higher dimensions).

Bibliography

Olav Kallenberg. Random measures, theory and applications, volume 1. Springer, 2017.

Patrick Billingsley. Probability and measure. John Wiley & Sons, 2008.

Michael Scheuerer. A comparison of models and methods for spatial interpolation in statistics and numerical analysis. PhD thesis, Niedersächsische Staats-und Universitätsbibliothek Göttingen, 2009.

Iosif Il'ich Gihman and Anatolij Vladimirovič Skorohod. The theory of stochastic processes I. Springer, Berlin, Heidelberg, New York, 1974.

Christopher K. I. Williams and Carl Edward Rasmussen. Gaussian processes for machine learning, volume 2. MIT press Cambridge, MA, 2006.

Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. Journal of Functional Analysis, 1(3):290–330, 1967.

Martin Hairer. An introduction to stochastic PDEs. Lecture notes and arXiv:0907.4178, 2009.

Vladimir Igorevich Bogachev. Gaussian measures, volume 62 of Mathematical Surveys and Monographs. American Mathematical Soc., 1998.

Michel Ledoux. Isoperimetry and Gaussian analysis. In Lectures on probability theory and statistics, pages 165–294. Springer, 1996.

Balram S. Rajput and Stamatis Cambanis. Gaussian processes and Gaussian measures. The Annals of Mathematical Statistics, pages 1944–1952, 1972.

Cédric Travelletti and David Ginsbourger. Disintegration of Gaussian Measures for Sequential Assimilation of Linear Operator Data. arXiv preprint arXiv:2207.13581, 2022.

Michael L. Stein. Interpolation of spatial data: some theory for kriging. Springer Science & Business Media, 1999.

Nachman Aronszajn. Theory of reproducing kernels. Transactions of the American mathematical society, 68(3):337–404, 1950.

George S. Kimeldorf and Grace Wahba. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. The Annals of Mathematical Statistics, 41(2):495–502, September 1970. doi: 10.1214/aoms/1177697089. URL https://doi.org/10.1214/aoms/1177697089.

Bernhard Schölkopf and Alexander J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, June 2018. ISBN 9780262256933. doi: 10.7551/mitpress/4175.001.0001. URL https://doi.org/10.7551/mitpress/4175.001.0001.

Saburou Saitoh and Yoshihiro Sawano. Theory of Reproducing Kernels and Applications, volume 44. Springer, January 2016. doi: 10.1007/978-981-10-0530-5.

Ingo Steinwart and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

Holger Wendland. Scattered data approximation, volume 17. Cambridge university press, 2004.

Salomon Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. Mathematische Annalen, 108(1):378–410, 1933.

Alexander Khinchin. Korrelationstheorie der stationären stochastischen Prozesse. Mathematische Annalen, 109(1):604–615, 1934.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. Constructive Approximation, 35:363–417, 2012.

Alain Berlinet and Christine Thomas-Agnan. Reproducing Kernel Hilbert Space in Probability and Statistics. Springer US, January 2004. ISBN 978-1-4613-4792-7. doi: 10.1007/978-1-4419-9096-9.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007. Proceedings 18, pages 13–31. Springer, 2007.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research, 11: 1517–1561, 2010.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning, 10(1-2):1–141, 2017.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773, 2012.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, et al. Kernel methods for measuring independence. Journal of Machine Learning Research, 6(70):2075–2129, 2005. URL `http://jmlr.org/papers/v6/gretton05a.html`.

Danica J. Sutherland. Scalable, flexible and active learning on distributions. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA PITTSBURGH United States, 2016.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. The Journal of Machine Learning Research, 17(1):5272–5311, 2016.

Aad W. van der Vaart and J. Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. The Annals of Statistics, 36 (3):1435–1463, 2008.

Kari Karhunen. Zur spektraltheorie stochastischer prozesse. Ann. Acad. Sci. Fennicae, AI, 34, 1946.

Kari Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung: akademische Abhandlu Sana, 1947.

Michel Loeve. Functions aleatoires du second ordre. Processus stochastique et mouvement Brownien, pages 366–420, 1948.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In Advances in neural information processing systems, pages 1177–1184, 2008.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In Advances in neural information processing systems, pages 1313–1320, 2009.

Bertil Matérn. Spatial variation. Lecture Notes in Statistics. Springer New York, 1960.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. arXiv preprint arXiv:1807.02582, 2018.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In 21st Annual Conference on Learning Theory (COLT 2008), pages 111–122. Omnipress, 2008.

Athénaïs Gautier and David Ginsbourger. Continuous logistic Gaussian random measure fields for spatial distributional modelling, 2021.

Alex L. Rojas, Christopher R. Genovese, Christopher J. Miller, Robert Nichol, and Larry Wasserman. Conditional density estimation using finite mixture models with an application to astrophysics, 2005.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika, 83(1): 189–206, 1996.

Peter Hall, Rodney C. L. Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. Journal of the American Statistical association, 94(445):154–163, 1999.

Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. Biometrika, 91(4):819–834, 2004.

Xujia Zhu and Bruno Sudret. Emulation of stochastic simulators using generalized lambda models. arXiv preprint arXiv:2007.00996, 2020.

John Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160, 1982.

Juan José Egozcue, José Luis Díaz-Barrero, and Vera Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, 22(4):1175–1182, 2006.

R. Talská, Alessandra Menafoglio, Jitka Machalová, Karel Hron, and E. Fišerová. Compositional regression with functional response. Computational Statistics & Data Analysis, 123:66–85, 2018.

Adityanand Guntuboyina and Bodhisattva Sen. Nonparametric shape-restricted regression. Statistical Science, 33(4):568–594, 2018.

Sonia Jain and Radford M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of computational and Graphical Statistics, 13(1):158–182, 2004.

Stephen G. Walker. Sampling the Dirichlet mixture model with slices. Communications in Statistics—Simulation and Computation®, 36(1):45–54, 2007.

Omiros Papaspiliopoulos and Gareth O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika, 95(1):169–186, 2008.

Thomas Peter Minka. A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology, 2001.

David B. Dunson and Ju-Hyun Park. Kernel stick-breaking processes. Biometrika, 95(2):307–323, 2008.

David B. Dunson, Natesh Pillai, and Ju-Hyun Park. Bayesian density regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):163–183, 2007.

Yeonseung Chung and David B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association, 104(488):1646–1660, 2009.

Jim E. Griffin and Mark F. J. Steel. Order-based dependent Dirichlet processes. Journal of the American statistical Association, 101(473):179–194, 2006.

Lorenzo Trippa, Peter Müller, and Wesley Johnson. The multivariate beta process and an extension of the Polya tree model. Biometrika, 98(1):17–34, 2011.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. Advances in neural information processing systems, 30, 2017.

Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. Conditional density estimation with neural networks: Best practices and benchmarks. arXiv preprint arXiv:1903.00954, 2019.

George Papamakarios. Neural density estimation and likelihood-free inference. arXiv preprint arXiv:1910.13233, 2019.

Alejandro Jara and Timothy E. Hanson. A class of mixtures of dependent tail-free processes. Biometrika, 98(3):553–566, 2011.

Christian Donner and Manfred Opper. Efficient Bayesian inference for a Gaussian process density model. arXiv preprint arXiv:1805.11494, 2018.

Surya T. Tokdar, Yu M. Zhu, and Jayanta K. Ghosh. Bayesian density regression with logistic Gaussian process and subspace projection. Bayesian analysis, 5 (2):319–344, 2010.

Athénaïs Gautier, David Ginsbourger, and Guillaume Pirot. Goal-oriented adaptive sampling under random field modelling of response probability distributions. ESAIM: Proceedings and Surveys, 71:89–100, 2021. doi: 10.1051/proc/202171108. URL https://doi.org/10.1051/proc/202171108.

Peter J. Lenk. The logistic normal distribution for Bayesian, nonparametric, predictive densities. Journal of the American Statistical Association, 83(402): 509–516, 1988.

Peter J. Lenk. Towards a practicable Bayesian nonparametric density estimator. Biometrika, 78(3):531–543, 1991.

Tom Leonard. Density estimation, stochastic processes and prior information. Journal of the Royal Statistical Society, Series B: Methodological, 40(2):113–132, 1978.

Surya T. Tokdar and Jayanta K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. Journal of statistical planning and inference, 137(1):34–42, 2007.

Surya T. Tokdar. Towards a faster implementation of density estimation with logistic Gaussian process priors. Journal of Computational and Graphical Statistics, 16(3):633–655, 2007.

Debdeep Pati, David B. Dunson, and Surya T. Tokdar. Posterior consistency in conditional distribution estimation. Journal of multivariate analysis, 116: 456–472, 2013.

Jean-Marc Azaïs and Mario Wschebor. Level sets and extrema of random processes and fields. John Wiley & Sons, 2009.

Lorraine Schwartz. On bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 4(1):10–26, 1965.

Gautier, Athénaïs. Github repository for the SLGP. `https://github.com/AthenaisGautier/SLGP/`, 2023. Accessed on 31.01.2023.

James O. Berger, Victor De Oliveira, and Bruno Sansó. Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association, 96(456):1361–1374, 2001.

Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of Gaussian random fields. Journal of the American Statistical Association, 114(525):445–452, 2019.

Murray Rosenblatt. Remarks on a multivariate transformation. The annals of mathematical statistics, 23(3):470–472, 1952.

Delphine Dupuy, Céline Helbert, and Jessica Franco. DiceDesign and DiceEval: Two R Packages for Design and Analysis of Computer Experiments. Journal of Statistical Software, 65(11):1–38, 2015. URL `https://www.jstatsoft.org/v65/i11/`.

Aad W. van der Vaart and J. Harry van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. The Annals of Statistics, 37(5B):2655–2675, 2009.

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with R-INLA. Journal of statistical software, 63:1–25, 2015.

Stan Development Team. RStan: the R interface to Stan, 2022. URL `https://mc-stan.org/`. R package version 2.21.5.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6):1087–1092, 1953.

W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97–109, 1970.

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. Bernoulli, pages 223–242, 2001.

Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: efficient adaptive MCMC. Statistics and computing, 16:339–354, 2006.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. Statistics and computing, 18:343–373, 2008.

Radford M. Neal. Regression and classification using Gaussian process priors. Bayesian statistics, 6:475, 1998.

Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. MCMC methods for diffusion bridges. Stochastics and Dynamics, 8(03):319–350, 2008.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. Statistical Science, 28(3):424–446, August 2013. ISSN 0883-4237. doi: 10.1214/13-sts421. URL http://dx.doi.org/10.1214/13-STS421.

Daniel Rudolf and Björn Sprungk. On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm. Foundations of Computational Mathematics, 18:309–343, 2018.

Yuxin Chen, David Keyes, Kody J. H. Law, and Hatem Ltaief. Accelerated dimension-independent adaptive Metropolis. SIAM Journal on Scientific Computing, 38(5):S539–S565, 2016.

Gareth O. Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. Methodology and computing in applied probability, 4: 337–357, 2002.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. Physics letters B, 195(2):216–222, 1987.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In Conference on learning theory, pages 793–797. PMLR, 2018.

Swiss Federal Office of Meteorology and Climatology MeteoSwiss. Climatological Network - Daily Values , 2019. URL https://opendata.swiss/en/dataset/klimamessnetz-tageswerte. Accessed on 01.10.2021.

Swiss Federal Office of Topography swisstopo. The digital height model of Switzerland with a 200m grid, 2019. URL https://opendata.swiss/en/dataset/das-digitale-hohenmodell-der-schweiz-mit-einer-maschenweite-von-200-m. Accessed on 01.10.2021.

A. Ruszczynski and A. Shapiro. Stochastic programming (handbooks in operations research and management science), 2003.

John R. Birge and Francois Louveaux. Introduction to stochastic programming. Springer Science & Business Media, 2011.

András Prékopa. Stochastic programming, volume 324. Springer Science & Business Media, 2013.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. Ann. Math. Statist., 22(3):400–407, September 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. The Journal of Machine Learning Research, 18(1):4873–4907, 2017.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.

Nicolo Cesa-Bianchi and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning, 5(1):1–122, 2012.

Marzena Rostek. Quantile maximization in decision theory. The Review of Economic Studies, 77(1):339–371, 2010.

Léonard Torossian, Victor Picheny, Robert Faivre, and Aurélien Garivier. A review on quantile regression for stochastic computer experiments. Reliability Engineering & System Safety, page 106858, 2020.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of Conditional Value-at-Risk. Journal of Risk, 2:21–41, 2000.

Fabio Bellini and Elena Di Bernardino. Risk Management with Expectiles. European Journal of Finance, May 2015. doi: 10.1080/1351847X.2015.1052150.

Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association, 99(468):1015–1026, 2004.

Sam Efromovich. Dimension reduction and adaptation in conditional density estimation. Journal of the American Statistical Association, 105(490):761–774, 2010.

Vincent Moutoussamy, Simon Nanty, and Benoît Pauwels. Emulators for stochastic simulation codes. ESAIM: Proceedings and Surveys, 48:116–155, 2015.

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of Bayesian methods for seeking the extremum. Towards global optimization, 2 (117-129):2, 1978.

Donald Jones, Matthias Schonlau, and William Welch. Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization, 13:455–492, December 1998. doi: 10.1023/A:1008306431147.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. INFORMS journal on Computing, 21(4): 599–613, 2009.

Peter I. Frazier. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995, 2009.

Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. Structural and Multidisciplinary Optimization, 48(3):607–626, September 2013a. doi: 10.1007/s00158-013-0919-4.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In Advances in neural information processing systems, pages 918–926, 2014.

Hamed Jalali, Inneke Van Nieuwenhuyse, and Victor Picheny. Comparison of Kriging-based algorithms for simulation optimization with heterogeneous noise. European Journal of Operational Research, 261(1):279 – 301, 2017. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2017.01.035. URL http://www.sciencedirect.com/science/article/pii/S037722171730070X.

Jimmy Risk and Michael Ludkovski. Sequential design and spatial modeling for portfolio tail risk measurement. SIAM Journal on Financial Mathematics, 9 (4):1137–1174, 2018.

Mickaël Binois, Jiangeng Huang, Robert B. Gramacy, and Mike Ludkovski. Replication or exploration? Sequential design for stochastic simulation experiments. Technometrics, 61(1):7–23, 2019.

Victor Picheny, David Ginsbourger, Yann Richet, and Gregory Caplin. Quantile-based optimization of noisy computer experiments with tunable precision. Technometrics, 55(1):2–13, 2013b.

Julien Bect, François Bachoc, and David Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. Bernoulli, 25(4A):2883–2919, 2019.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. Sequential Monte Carlo Methods in Practice. Information Science and Statistics. Springer New York, 2001. ISBN 9780387951461.

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

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

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

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most Likely Heteroscedastic Gaussian Process Regression. In Proceedings of the 24th International Conference on Machine Learning, ICML'07, pages 393–400, 2007.

Mickael Binois, Robert B. Gramacy, and Mike Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. Journal of Computational and Graphical Statistics, 27(4):808–821, 2018.

Victor Picheny, David Ginsbourger, and Olivier Roustant. DiceOptim: Kriging-Based Optimization for Computer Experiments, 2020. URL `https://CRAN.R-project.org/package=DiceOptim`. R package version 2.0.1.

Mickael Binois and Robert B. Gramacy. hetGP: Heteroskedastic Gaussian Process Modeling and Design under Replication, 2019. URL `https://CRAN.R-project.org/package=hetGP`. R package version 1.1.2.

Albert Tarantola. Inverse Problem Theory and Methods for Model Parameter Estimation. EngineeringPro collection. Society for Industrial and Applied Mathematics, 2005. ISBN 9780898715729.

Mario Bertero, Patrizia Boccacci, and Christine De Mol. Introduction to inverse problems in imaging. CRC press, 2021.

Andrew M. Stuart. Inverse problems: a Bayesian perspective. Acta numerica, 19:451–559, 2010.

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

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

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.

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.

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.

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.

Anthony O'Hagan and Jonathan J. Forster. Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference, second edition, volume 2B. Arnold, 2004. URL `https://eprints.soton.ac.uk/46376/`.

Christian P. Robert. The Bayesian choice: from decision-theoretic foundations to computational implementation, volume 2. Springer, 2007.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian data analysis. Chapman and Hall/CRC, 1995.

Florian Hartig, Justin M. Calabrese, Björn Reineking, Thorsten Wiegand, and Andreas Huth. Statistical inference for stochastic simulation models–theory and application. Ecology letters, 14(8):816–827, 2011.

Brandon M. Turner and Trisha Van Zandt. A tutorial on approximate Bayesian computation. Journal of Mathematical Psychology, 56(2):69–85, 2012.

Scott A. Sisson, Yanan Fan, and Mark Beaumont. Handbook of approximate Bayesian computation. CRC Press, 2018.

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.

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

Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate Bayesian computation in population genetics. Genetics, 162(4):2025–2035, 2002.

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.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics, 37(2):697–725, 2009.

Fernando V. Bonassi and Mike West. Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation. Bayesian Anal., 10(1):171–187, March 2015. doi: 10.1214/14-BA891. URL https://doi.org/10.1214/14-BA891.

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.

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

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

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.

Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. 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.

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.

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

Richard Wilkinson. Accelerating ABC methods using Gaussian processes. In Artificial Intelligence and Statistics, pages 1015–1023. PMLR, 2014.

Edward Meeds and Max Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, pages 393–400. Corvallis, Oregon: AUAI Press, 2014.

Franck Jabot, Guillaume Lagarrigues, Benoît Courbaud, and Nicolas Dumoulin. A comparison of emulation methods for approximate bayesian computation. arXiv preprint arXiv:1412.7560, 2014.

Michael U. Gutmann, Jukka Cor, and er. Bayesian optimization for likelihood-free inference of simulator-based statistical models. Journal of Machine Learning Research, 17(125):1–47, 2016. URL `http://jmlr.org/papers/v17/15-017.html`.

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.

Donald B. Owen. Tables for computing bivariate normal probabilities. The Annals of Mathematical Statistics, 27(4):1075–1090, 1956.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. Applied and Computational Harmonic Analysis, 51:510–542, 2021.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association, 102 (477):359–378, 2007.

Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures. arXiv preprint arXiv:2206.07588, 2022.

Yves Deville, David Ginsbourger, and Olivier Roustant. Contributors: Nicolas Durrande. kergp: Gaussian Process Laboratory, 2021. URL `https://CRAN.R-project.org/package=kergp`. R package version 0.5.5.

Appendix A

Appendices

A.1 Basics of posterior consistency in the Bayesian literature

This Section aims at introducing the basic ideas and tools necessary to establish posterior consistency of a prior distribution. Posterior consistency is a desirable property as it ensures that provided enough data, a well-specified Bayesian model will recover the true data generating process.

We consider a parameter space Θ, which does not need to be euclidean. Assuming that we have available observations, noted $\mathbf{Y}^{(n)} = \{\mathbf{Y}_i\}_{i=1}^n$. The probability distribution of $\mathbf{Y}^{(n)}$ is assumed to be controlled by a parameter θ and is noted $P_\theta^{(n)}$. Let Π be a prior over Θ.

Definition A.1.1 (Weak posterior consistency)**.** It is said that the prior Π achieves weak posterior consistency at $\theta_0 \in \Theta$ with respect to a given topology if for any weak neighbourhood U of θ_0:

$$\Pi\left[U|\mathbf{Y}^{(n)}\right] \xrightarrow[n\to\infty]{} 1 \tag{A.1}$$

almost surely under $P_{\theta_0}^{(n)}$

Remark. It is common, but not necessary, to consider the observations $\mathbf{Y}^{(n)}$ to be independent, as it simplifies the expression of the distribution $P_{\theta_0}^{(n)}$.

This definition is quite general, and it is often difficult to prove consistency in a general setting. One really indispensable result for non-parametric and semi-parametric problems is Schwartz's theorem. It leverages the Kullback-Leibler divergence. Schwartz's method consists in a general method for establishing consistency and writes as follows:

Theorem A.1.1 (Schwartz, 1965)**.** *Let* $\{f_\theta : \theta \in \Theta\}$ *be a class of densities and let* $\mathbf{Y}_1, \mathbf{Y}_2, ...$ *be i.i.d. with density* f_{θ_0}*, where* $\theta_0 \in \Theta$*. Suppose for every neighbourhood* U *of* θ_0*, there is a test for* $\theta = \theta_0$ *against* $\theta \notin U$ *with power strictly greater than the size. Let* Π *be a prior on* Θ *such that for every* $\epsilon > 0$*:*

$$\Pi\left[\theta : KL(f_{\theta_0}, f_\theta) < \epsilon\right] > 0 \tag{A.2}$$

Then the posterior is consistent at θ_0

Remark. Schwartz's theorem gives a sufficient condition for weak posterior consistency, however it is not a necessary condition.

Remark. If Θ is itself a class of densities with $f_\theta = \theta$, then the condition on existence of tests in Schwartz's theorem is satisfied if Θ is endowed with the topology of weak convergence.
More generally, existence of a uniformly consistent estimator implies the existence of such a test.

Finally, note that consistency is defined with respect to a given topology. Here, we define some topologies that will be relevant in this thesis.

Definition A.1.2 (Weak neighbourhood of a conditional density)**.** We define the weak convergence neighbourhood of a density field p_0 of pdfs on $\mathcal{I}$ indexed by D through a sub-base. It is given for any bounded continuous function $g : D \times \mathcal{I} \to \mathbb{R}$ and any $\epsilon > 0$ by:

$$V_{\epsilon,g} = \left\{ f \in \mathcal{F}_d(D;\mathcal{I}), \left| \int_{D\times\mathcal{I}} gf - \int_{D\times\mathcal{I}} gf_0 \right| < \epsilon \right\} \tag{A.3}$$

A weak neighbourhood base is formed by finite intersections of neighbourhoods of the above type.

Definition A.1.3 (Weak neighbourhood of a joint density)**.** We define the weak convergence neighbourhood of a joint density h_0 on $D \times \mathcal{I}$ through a sub-base. It is given for any bounded continuous function $g : D \times \mathcal{I} \to \mathbb{R}$ and any $\epsilon > 0$ by:

$$W_{\epsilon,g} = \left\{ h \in \mathcal{F}(D \times \mathcal{I}), \left| \int_{D\times\mathcal{I}} gh - \int_{D\times\mathcal{I}} gh_0 \right| < \epsilon \right\} \tag{A.4}$$

A weak neighbourhood base is formed by finite intersections of neighbourhoods of the above type.

A.2 Complete proofs

For the proof of proposition 2.1.3, we will need to bound the entropy number coming into play in Dudley's theorem.

Lemma 7. *For $d \geq 1$, and I a convex, compact subset of $\mathbb{R}^d$, we recall that if $\epsilon \geq diam(I)$, then $N(\epsilon, I, \|\cdot\|_\infty) = 1$. Let Vol be the volume, and B_1^d the d-dimensional unit ball, we have:*

$$\left(\frac{1}{\epsilon}\right)^d \frac{Vol(I)}{Vol(B_1^d)} \leq N(\epsilon, I, \|\cdot\|_\infty) \tag{A.5}$$

Additionally, if $\epsilon < diam(I^d)$:

$$N(\epsilon, I, \|\cdot\|_\infty) \leq \left(\frac{4}{\epsilon}\right)^d \frac{Vol(I)}{Vol(B_1^d)} \tag{A.6}$$

Proof of proposition 2.1.3. We consider the canonical pseudo metric associated to k, as defined in 2.1.9. Dudley's integral theorem gives:

$$\mathbb{E}\left[\|Z\|_\infty\right] \leq 24 \int_0^\infty \sqrt{\log\left(N(\epsilon, I, d_Z)\right)}\, d\epsilon \tag{A.7}$$

We note that for $\epsilon \geq \sup_{\mathbf{y},\mathbf{y}' \in I} d_Z(\mathbf{y}, \mathbf{y}')$, then $N(\epsilon, I, d_Z) = 1$.

It follows from the Hölder condition that $N(\epsilon, I, d_Z) \leq N\left((\epsilon/K)^{2/\beta}, I, \|\cdot\|_\infty\right)$.

$$\mathbb{E}\left[\|Z\|_\infty\right] \leq 24 \int_0^{\sup d_Z(\mathbf{y},\mathbf{y}')} \sqrt{\log N\left((\epsilon/K)^{2/\beta}, I, \|\cdot\|_\infty\right)}\, d\epsilon \tag{A.8}$$

Applying Lemma 7 and using the fact that for all $a > 0, x > 0, \log(a/x) \leq a/x$, we have:

$$\mathbb{E}\left[\|Z\|_\infty\right] \leq 24 \int_0^{\sup d_Z(\mathbf{y},\mathbf{y}')} \sqrt{\log \frac{C}{\epsilon}^{2d/\beta}}\, d\epsilon \tag{A.9}$$

$$\leq 24 \int_0^{\sup d_Z(\mathbf{y},\mathbf{y}')} \sqrt{\frac{2d}{\beta} \log \frac{C}{\epsilon}}\, d\epsilon \tag{A.10}$$

$$\leq 24 \sqrt{\frac{2d}{\beta}} \int_0^{\sup d_Z(\mathbf{y},\mathbf{y}')} \sqrt{\frac{C}{\epsilon}}\, d\epsilon < \infty \tag{A.11}$$

where $C := K 4^{\beta/2} \dfrac{\mathrm{Vol}(I)}{\mathrm{Vol}(B_1^d)} > 0$. The convergence of the integral induces that Z admits a version with sample path almost surely bounded and uniformly continuous on (O, d_k).

Then, as d_Z is continuous with respect to $\|\cdot\|_\infty$, it also induces that Z admits a version with sample path almost surely bounded and uniformly continuous on $(I, \|\cdot\|_\infty)$. □

Proof of Proposition 3.3.2. For fixed $(\mathbf{x}, \mathbf{x}') \in D^2$ and $Z \sim \mathcal{GP}(0, k)$, we note $d_{\mathbf{x},\mathbf{x}'}$ the canonical semi-metric associated to $Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}$, defined by:

$$d^2_{\mathbf{x},\mathbf{x}'}(t,t') = \mathbb{E}\left[([Z_{\mathbf{x},t} - Z_{\mathbf{x}',t}] - [Z_{\mathbf{x},t'} - Z_{\mathbf{x}',t'}])^2\right] \forall (t,t') \in \mathcal{I}^2 \tag{A.12}$$

Using the Hölder condition on k we note that we have simultaneously:

$$d^2_{\mathbf{x},\mathbf{x}'}(t,t') \leq 3C\|\mathbf{x} - \mathbf{x}'\|^{\alpha_1}_\infty \forall (t,t') \in \mathcal{I}^2 \tag{A.13}$$

$$d^2_{\mathbf{x},\mathbf{x}'}(t,t') \leq 4C\|t - t'\|^{\alpha_2}_\infty \forall (t,t') \in \mathcal{I}^2 \tag{A.14}$$

By Dudley's theorem, we can write :

$$\begin{aligned} M(\mathbf{x},\mathbf{x}') &\leq 24 \int_0^\infty \sqrt{\log(N(\epsilon, \mathcal{I}, d_{\mathbf{x},\mathbf{x}'}))}\, d\epsilon \\ &\leq 24 \int_0^{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})} \sqrt{\log(N(\epsilon, \mathcal{I}, d_{\mathbf{x},\mathbf{x}'}))}\, d\epsilon \end{aligned} \tag{A.15}$$

where $D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})$ stands for the diameter of $\mathcal{I}$ with respect to the canonical semi-metric associated to $d_{\mathbf{x},\mathbf{x}'}$. As $d^2_{\mathbf{x},\mathbf{x}'}(t,t') \leq 4C\|t - t'\|^{\alpha_2}_\infty$, we can combine the bounds stated in Lemma (7) with the inequality

$$N(\epsilon, \mathcal{I}, d_{\mathbf{x},\mathbf{x}'}) \leq N\left(\left(\frac{\epsilon}{4C}\right)^{2/\alpha_2}, \mathcal{I}, \|\cdot\|_\infty\right)$$

It follows that

$$M(\mathbf{x},\mathbf{x}') \leq 24\sqrt{\frac{2d_t}{\alpha_2}} \int_0^{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})} \sqrt{\log\left(\frac{K}{\epsilon}\right)}\, d\epsilon \tag{A.16}$$

where $K := C4^{1+\alpha_2/2}\left(\frac{\mathrm{Vol}(\mathcal{I})}{\mathrm{Vol}(B_1^{d_t})}\right)^{\alpha_2/(2d_t)}$ and $B_1^{d_t}$ stands for the d_t-dimensional unit ball for $\|\cdot\|_\infty$. To further compute the right-hand term, we introduce the error function, defined as $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$.

$$\int_0^{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})} \sqrt{\log\left(\frac{K}{\epsilon}\right)}\, d\epsilon = \left[\epsilon\sqrt{\log\frac{K}{\epsilon}} - \frac{\sqrt{\pi}}{2} K \mathrm{erf}\left(\sqrt{\log\frac{K}{\epsilon}}\right)\right]_0^{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})} \tag{A.17}$$

$$= D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})\sqrt{\log\frac{K}{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})}} + K\int_{\sqrt{\log(K/D_{\mathbf{x},\mathbf{x}'}(\mathcal{I}))}}^\infty e^{-t^2}\, dt \tag{A.18}$$

Since for $y > 0$, $\frac{2}{\sqrt{\pi}} \int_y^\infty e^{-t^2}\, dt \leq e^{-y^2}$, we also have:

$$M(\mathbf{x}, \mathbf{x}') \leq 24\sqrt{\frac{2d_t}{\alpha_2}} \left(D_{\mathbf{x},\mathbf{x}'}(\mathcal{I}) \sqrt{\log \frac{K}{D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})}} + \frac{\sqrt{\pi} D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})}{2K} \right) \tag{A.19}$$

Then, for any $0 < \delta < \dfrac{\alpha_1}{2}$, D being compact and considering that

$$y \left(\log(K/y)\right)^{1/2} \underset{y \to 0}{=} o\left(y^{1-2\delta/\alpha_1}\right)$$

we can conclude that there exists K_δ such that:

$$M(\mathbf{x}, \mathbf{x}') \leq \quad K_\delta \frac{1}{(3C)^{1/2-\delta/\alpha_1}} \left(D_{\mathbf{x},\mathbf{x}'}(\mathcal{I})^{1-2\delta/\alpha_1} \right) \tag{A.20}$$

Finally, by Equation 3.37, we have $D_{\mathbf{x},\mathbf{x}'}(\mathcal{I}) \leq \sqrt{3C} \|\mathbf{x} - \mathbf{x}'\|^{\alpha_1/2}$, and we can conclude that:

$$M(\mathbf{x}, \mathbf{x}') \leq K_\delta \|\mathbf{x} - \mathbf{x}'\|^{\alpha_1/2-\delta} \tag{A.21}$$

□

Proof of Theorem 3.3.5. Let us consider $(\mathbf{x}, \mathbf{x}') \in D^2$ and $\gamma > 0$. By Lemma 3, there exists two constants $C_{KL}, C_{TV} > 0$ such that:

$$\begin{aligned} \mathbb{E}\left[d_H((Y_{\mathbf{x},\cdot}), Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty^\gamma e^{\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty \gamma/2}\right] \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{KL} \mathbb{E}[f_1(\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty)] \\ \mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{TV} \mathbb{E}[f_2(\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty)] \end{aligned} \tag{A.22}$$

where $f_1(x) = x^{2\gamma} (1 + x)^\gamma e^{\gamma x}$ and $f_2(x) = x^{2\gamma} (1 + x)^{2\gamma} e^{\gamma x}$.

We consider the three functions, defined for $\gamma, M, y > 0$:

$$\begin{aligned} f_{H,\gamma,M}(y) &= (My)^\gamma e^{\frac{M\gamma}{2} y} \\ f_{KL,\gamma,M}(y) &= (My)^{2\gamma} (1 + My)^\gamma e^{M\gamma y} \\ f_{TV,\gamma,M}(y) &= (My)^{2\gamma} (1 + My)^{2\gamma} e^{M\gamma y} \end{aligned} \tag{A.23}$$

Then, if we consider $M(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty\right]$, the previous inequalities can be rewritten as:

$$\begin{aligned} \mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \mathbb{E}\left[f_{H,\gamma,M(\mathbf{x},\mathbf{x}')} \left(\frac{\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty}{M(\mathbf{x}, \mathbf{x}')} \right) \right] \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{KL} \cdot \mathbb{E}\left[f_{KL,\gamma,M(\mathbf{x},\mathbf{x}')} \left(\frac{\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty}{M(\mathbf{x}, \mathbf{x}')} \right) \right] \\ \mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{TV} \cdot \mathbb{E}\left[f_{TV,\gamma,M(\mathbf{x},\mathbf{x}')} \left(\frac{\|Z_{\mathbf{x},\cdot} - Z_{\mathbf{x}',\cdot}\|_\infty}{M(\mathbf{x}, \mathbf{x}')} \right) \right] \end{aligned} \tag{A.24}$$

By Fernique theorem (cf proposition 2.1.6), there exists universal constant α, $K > 0$, as well as $C_{H,\gamma,M}, C_{KL,\gamma,M}$ and $C_{TV,\gamma,M} > 0$ such that:

$$\begin{aligned} \mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{H,\gamma,M(\mathbf{x},\mathbf{x}')} \cdot K \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq C_{KL} \cdot C_{KL,\gamma,M(\mathbf{x},\mathbf{x}')} \cdot K \\ \mathbb{E}\left[d_{TV}((Y_{\mathbf{x},\cdot}),(Y_{\mathbf{x}',\cdot})_)^\gamma\right] &\leq C_{TV} \cdot C_{TV,\gamma,M(\mathbf{x},\mathbf{x}')} \cdot K \end{aligned} \tag{A.25}$$

Detailed expressions of $C_{H,\gamma,M}, C_{KL,\gamma,M}$ and $C_{TV,\gamma,M}$ are given below this proof, and were derived with the tightness of our bounds in mind. We note that these coefficients seen as functions of M are continuous, strictly positive for any $M > 0$ and that:

$$\begin{aligned} C_{H,\gamma,M} &\underset{M\to 0}{\sim} M^\gamma \left(\frac{\gamma}{2\alpha}\right)^{\gamma/2} \exp\left\{-\frac{\gamma}{2}\right\} \\ C_{KL,\gamma,M} &\underset{M\to 0}{\sim} M^{2\gamma} \left(\frac{\gamma}{\alpha}\right)^{\gamma} \exp\left\{-\gamma\right\} \\ C_{TV,\gamma,M} &\underset{M\to 0}{\sim} M^{2\gamma} \left(\frac{\gamma}{\alpha}\right)^{\gamma} \exp\left\{-\gamma\right\} \end{aligned} \tag{A.26}$$

This equivalence allows us to state that for a given $\gamma > 0$, there exists a rank $M_0 > 0$ and a constant $\kappa_{\gamma,1} > 1$ such that for any $M < M_0$:

$$C_{H,\gamma,M} \leq \kappa_{\gamma,1} M^\gamma, \quad C_{KL,\gamma,M} \leq \kappa_{\gamma,1} M^{2\gamma}, \quad C_{TV,\gamma,M} \leq \kappa_{\gamma,1} M^{2\gamma} \tag{A.27}$$

We also observe that if M is bounded, as $C_{H,\gamma,M}, C_{KL,\gamma,M}$ and $C_{TV,\gamma,M}$ seen as function of M are continuous and strictly positive, there exists a constant $\kappa_{\gamma,2} > 0$ such that for values of $M \geq M_0$:

$$C_{H,\gamma,M} \leq \kappa_{\gamma,2} M^\gamma, \quad C_{KL,\gamma,M} \leq \kappa_{\gamma,2} M^{2\gamma}, \quad C_{TV,\gamma,M} \leq \kappa_{\gamma,2} M^{2\gamma} \tag{A.28}$$

Combining these two observations, and $M(\mathbf{x}, \mathbf{x}')$ being bounded, we conclude that for any $\gamma > 0$ there exist κ_γ such that:

$$\begin{aligned} \mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma M(\mathbf{x},\mathbf{x}')^\gamma \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma M(\mathbf{x},\mathbf{x}')^{2\gamma} \\ \mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] &\leq \kappa_\gamma M(\mathbf{x},\mathbf{x}')^{2\gamma} \end{aligned} \tag{A.29}$$

This argument relies on an equivalence at zero. Therefore, it ensures that the convergence rates are not degraded when bounding $C_{H,\gamma,M}, C_{KL,\gamma,M}$ and $C_{TV,\gamma,M}$, and that our bounds are still tight.

Finally, using proposition 3.3.2, stating that for all $\delta > 0$, there exists K_δ such that:

$$M(\mathbf{x}, \mathbf{x}') \leq K_\delta \|\mathbf{x} - \mathbf{x}'\|_\infty^{\alpha_1/2-\delta} \tag{A.30}$$

We conclude that for all $\gamma > 0, \delta > 0$, there exists a constant $K_{\gamma,\delta}$ such that:

$$\begin{array}{l}\mathbb{E}\left[d_H(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x}-\mathbf{x}'\|_\infty^{\gamma\alpha_1/2-\delta} \\ \mathbb{E}\left[KL(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x}-\mathbf{x}'\|_\infty^{\gamma\alpha_1-\delta} \\ \mathbb{E}\left[d_{TV}(Y_{\mathbf{x},\cdot}, Y_{\mathbf{x}',\cdot})^\gamma\right] \leq K_{\gamma,\delta}\|\mathbf{x}-\mathbf{x}'\|_\infty^{\gamma\alpha_1-\delta}\end{array} \tag{A.31}$$

□

Finding the constants in Proof of Theorem 3.3.5. For fixed M, $\gamma > 0$, we consider the following three functions, for $x \geq 0$:

$$\begin{array}{rl} f_{H,\gamma,M}(x) & = (Mx)^\gamma e^{\frac{M\gamma}{2}x} \\ f_{KL,\gamma,M}(x) & = (Mx)^{2\gamma}(1+Mx)^\gamma e^{M\gamma x} \\ f_{TV,\gamma,M}(x) & = (Mx)^{2\gamma}(1+Mx)^{2\gamma} e^{M\gamma x}\end{array} \tag{A.32}$$

For $\alpha > 0$, we look for constants $C_{H,\gamma,M}, C_{KL,\gamma,M}, C_{TV,\gamma,M}$, satisfying:

$$\begin{array}{rl} f_{H,\gamma,M}(x) & \leq C_{H,\gamma,M} e^{\alpha x^2} \\ f_{KL,\gamma,M}(x) & \leq C_{KL,\gamma,M} e^{\alpha x^2} \\ f_{TV,\gamma,M}(x) & \leq C_{TV,\gamma,M} e^{\alpha x^2}\end{array} \tag{A.33}$$

such constants satisfy:

$$\begin{array}{l} \sup\limits_{x\geq 0}\ g_{H,\gamma,M}(x) := f_{H,\gamma,M}(x)e^{-\alpha x^2} \leq C_{H,\gamma,M} \\ \sup\limits_{x\geq 0}\ g_{KL,\gamma,M}(x) := f_{KL,\gamma,M}(x)e^{-\alpha x^2} \leq C_{KL,\gamma,M} \\ \sup\limits_{x\geq 0}\ g_{TV,\gamma,M}(x) := f_{TV,\gamma,M}(x)e^{-\alpha x^2} \leq C_{TV,\gamma,M}\end{array} \tag{A.34}$$

Studying the variations of $g_{H,\gamma,M}(x)$ simply involves finding the roots of a degree 2 polynomial and yields that this function attains its supremum at:

$$x_{H,\gamma,M} = \frac{M\gamma + 2\sqrt{M^2\gamma^2 + 8\alpha\gamma}}{8\alpha} \tag{A.35}$$

Therefore a valid upper bound for :

$$C_{H,\gamma,M} = (Mx_{H,\gamma,M})^\gamma \exp\left\{M\gamma\frac{x_{H,\gamma,M}}{2} - \alpha x_{H,\gamma,M}^2\right\} \tag{A.36}$$

This constant is optimal in the sense that it is the smallest constant satisfying inequality A.34.

However, studying the variations of $g_{KL,\gamma,M}(x)$ and $g_{TV,\gamma,M}(x)$ is longer as it involves finding the roots of third degree polynomials. In order to simplify the constant, we use the simple property $1 + x \leq e^x$ and introduce the bounds :

$$\begin{array}{l} g_{KL,\gamma,M}(x) \leq (Mx)^{2\gamma} e^{2M\gamma x - \alpha x^2} =: h_{KL,\gamma,M}(x) \\ g_{TV,\gamma,M}(x) \leq (Mx)^{2\gamma} e^{3M\gamma x - \alpha x^2} =: h_{TV,\gamma,M}(x)\end{array} \tag{A.37}$$

These inequalities are tight at $Mx = 0$.

Studying the variations of $h_{KL,\gamma,M}(x)$ and $h_{TV,\gamma,M}(x)$ simply involves finding the roots of a degree 2 polynomial and yields that this function attains their supremum at:

$$\begin{aligned} x_{KL,\gamma,M} &= \frac{M\gamma + \sqrt{M^2\gamma^2 + 4\alpha\gamma}}{2\alpha} \\ x_{TV,\gamma,M} &= \frac{3M\gamma + \sqrt{9M^2\gamma^2 + 16\alpha\gamma}}{4\alpha} \end{aligned} \tag{A.38}$$

Therefore, we can take the bounds:

$$C_{KL,\gamma,M} = \left(Mx_{KL,\gamma,M}\right)^{2\gamma} \exp\left\{2M\gamma x_{KL,\gamma,M} - \alpha x_{KL,\gamma,M}^2\right\} \tag{A.39}$$

$$C_{TV,\gamma,M} = \left(Mx_{TV,\gamma,M}\right)^{2\gamma} \exp\left\{x_{TV,\gamma,M} - \alpha x_{TV,\gamma,M}^2\right\} \tag{A.40}$$

These bounds are tight around zero. □

A.3 Additional figures

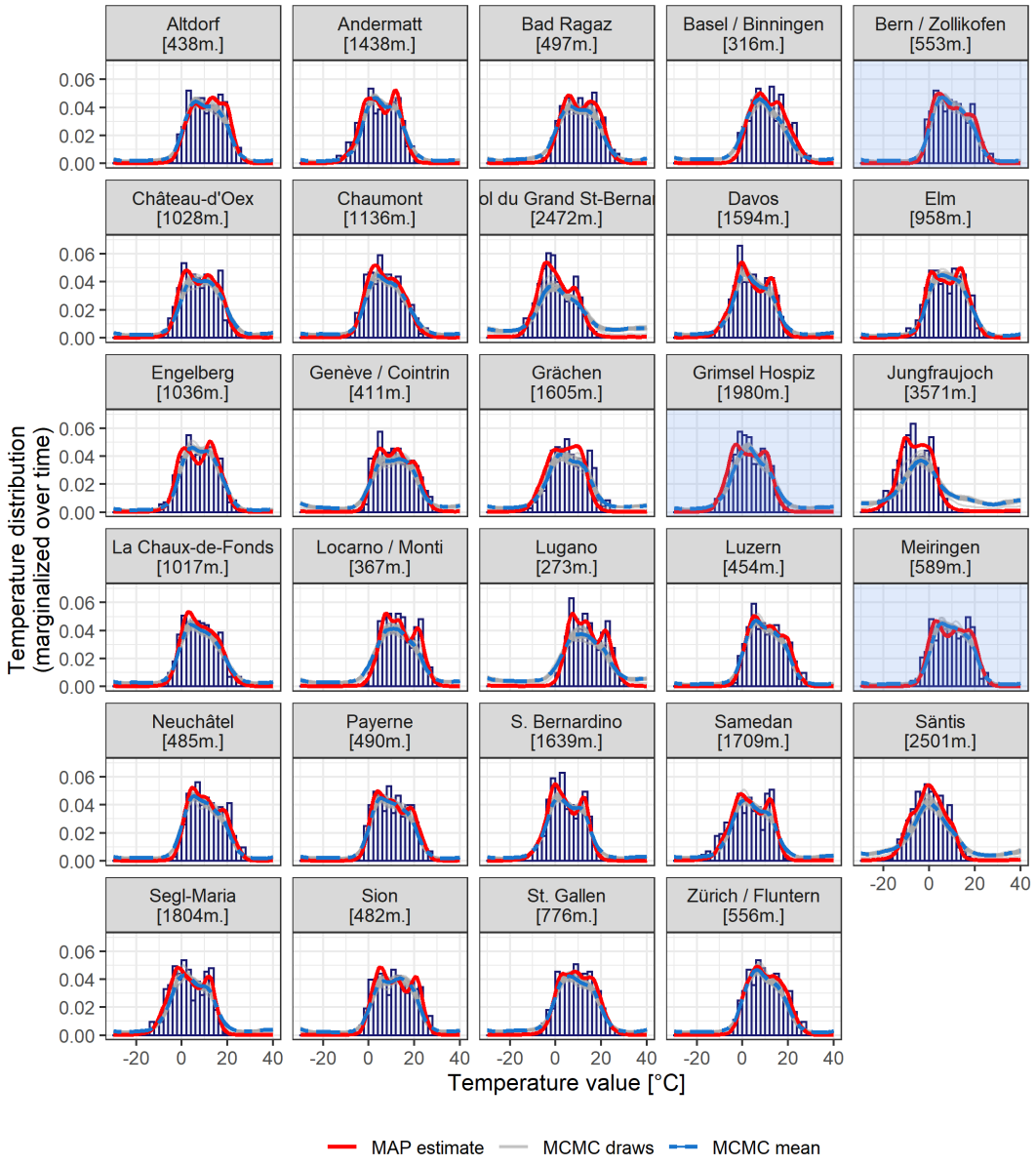


Figure A.1: Histograms and SLGP-based estimation at each of the 29 Stations present in the data-set, the stations located in the canton of Bern are in blue.

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Gautier / Athénaïs

Registration Number: 118-136-820

Study program: Statistics

Bachelor ☐ Master ☐ Dissertation ☑

Title of the thesis: Modelling and predicting distribution-valued fields with applications to inversion under uncertainty

Supervisor: Prof. Dr. David Ginsbourger

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of September 5th, 1996 and Article 69 of the University Statute of June 7th, 2011 is authorized to revoke the doctoral degree awarded on the basis of this thesis.
For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with theses submitted by others.

At Bern, the 29/03/2023

Place/Date

Signature