Efficient Spatial and Temporal Learning with Sparse Spectral Gaussian Processes

Jeremy Sellier

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Statistical Science

University College London

March 19, 2024

Declaration of authorship

I, Jeremy Sellier, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to begin by expressing my deepest gratitude to my main supervisor, Professor Petros Dellaportas, whose guidance and mentorship have been invaluable throughout my Ph.D. journey.

I am appreciative of the funding and support I received through the EPSRC CASE studentship award, in collaboration with Shell Research Ltd. I extend my sincere thanks to Professor Philip Jonathan from Lancaster University and Shell Research, as well as Matthew Jones, from Shell Research, for their unwavering support. Their guidance and insights have been instrumental in the success of my research.

I am also immensely grateful to my partner, friends, and family for their unwavering support, encouragement, and understanding. Their presence and encouragement have been a constant source of strength throughout my academic pursuits.

Abstract

Gaussian Processes (GPs) have gained substantial attention within the fields of statistics and machine learning. Rooted in Bayesian principles, their appeal lies in their ability to provide a robust framework for conducting inference. Particularly, GPs excel in their ability to capture intricate data dependencies and offer a comprehensive representation of predictive uncertainty.

However, as datasets grow in size and complexity, Bayesian nonparametric models employing GPs face notable challenges. A key concern revolves around conducting inference for models with intractable likelihoods, including cases where the likelihoods cannot be feasibly evaluated. Furthermore, a practical challenge arises from the substantial computational demands associated with GPs, characterized by a cubic time complexity, making them less suitable for large datasets.

Two prominent examples of these challenges are particularly evident. First, in the case of Poisson process likelihoods used in spatial statistics, where likelihood computations involve the intractable integration of a random function across the input space. Secondly, in time series analysis, while GPs generalizes traditional linear models, their integration into Bayesian change point detection framework (BOCPD) exposes a notable limitation: a naive implementation incurs $O(n^5)$ complexity. In both these scenarios, the ability to conduct efficient inference, accurately discern underlying patterns, and seamlessly adapt to scaling demands becomes paramount.

This thesis focuses on advancing efficient and adaptable inference methods using GPs for spatial data analysis and time series change point detection. A central emphasis lies in exploring the underutilized potential of reduced-rank GPs, derived from the spectral properties of their kernel, within these domains. This sparse spectral representation of GPs provides significant computational benefits and introduces novel perspectives for addressing complex data inference challenges in these fields.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Statistical learning, at its core, is the process of extracting predictive insights from data through statistical inference. It forms the foundation for various modern machine learning techniques, offering a versatile toolkit for data modeling, automating decision-making, and uncovering valuable insights.

In this landscape, Gaussian processes (GPs) play a crucial role in contemporary probabilistic machine learning. Publicized widely by Rasmussen and Williams (2005), GPs represent a form of Bayesian nonparametrics that serve as a flexible alternative to traditional parametric models. The historical roots of GPs extend nearly a century back with the pioneering works of Kolmogorov (1941) and Wiener (1949). Over the years, GPs have found extensive applications across the entire spectrum of machine learning, encompassing supervised learning (Rasmussen and Williams, 2005), unsupervised learning (Lawrence, 2003), and optimization (Mockus et al., 1978; Jones, 2001). A deeper exploration of these machine learning classes can be found in reference texts such as MacKay (2002) and Murphy (2012).

One of the key factors contributing to the widespread adoption of GPs is their inherent flexibility in accommodating increasing model complexity, dictated by the available data. As non-parametric models, they refrain from imposing rigid assumptions on the underlying data's functional form, enabling them to model a wide range

of functions, thus rendering them adaptable to various problem domains. Moreover, being rooted in Bayesian principles, GPs offer the capability to seamlessly integrate prior beliefs into complex tasks. Additionally, GPs provide the essential ability to quantify uncertainty, which is indispensable for both prediction and decision-making processes.

GPs have emerged as powerful tools in tackling classical statistical problems, notably in regression tasks. In regression, every data point in the training set forms an independent input-output pair, where the input correlates with the output. The principal hurdle in regression is deducing the underlying function that links input to output, thereby facilitating predictions for future data points. In this regard, Bayesian inference serves as the cornerstone, typically achieved through computing a closed-form posterior distribution.

This thesis is primarily dedicated to addressing the inherent challenges encountered in conducting efficient Bayesian inference with GPs, especially when dealing with contemporary complex datasets. The focus lies particularly on spatial data analysis and time series change point detection, where the conventional assumptions of independent and identically distributed observations often do not hold.

Temporal data, or time series, comprises a sequence of measurements taken at specific time intervals, finding prevalence in fields such as economics, finance, meteorology, and environmental science. Time series data introduces temporal dependence, where values at one time point depend on previous time points, thereby deviating from the i.i.d. assumption. Furthermore, these datasets often exhibit non-stationarity, where generative parameters may fluctuate over time, introducing complexity into the modeling task. In our research, we particularly focus on addressing this issue of non-stationarity and aim to develop change point (CP) models that can effectively identify and incorporate shifts in stationarity as an integral part of the inference process.

Spatial data, on the other hand, revolves around observations associated with distinct geographic locations, proving pivotal in diverse fields such as geography, ecology,

epidemiology, neuroscience, and crime analysis. Spatial data often demonstrates spatial dependence and heterogeneity, showcasing intriguing phenomena like attraction and repulsion, resulting in unique modeling challenges.

GPs offer a unified and versatile Bayesian framework that proves invaluable for the analysis of both spatial and temporal data. While they have demonstrated success in various applications, specific challenges have impeded their widespread adoption, thus necessitating further research. These challenges include:

- **Intractability Issues:** In many complex scenarios, the task of making inferences with GPs becomes arduous due to the intricate challenge of deriving tractable posterior distributions. This issue is particularly pronounced in the context of spatial data when employed with GPs, where the Poisson likelihood used becomes itself intractable. These models are often referred to as "doubly-intractable" in the literature, which has led to the development of approximation methods. These methods encompass Markov Chain Monte Carlo algorithms (Brooks et al., 2011) or variational inference methods(Beal, 2003), and will be discussed further in the subsequent sections.

- **Scalability Problems:** Adapting GPs to handle extensive datasets remains a persistent challenge. Standard GPs exhibit cubic time complexity, denoted as $O(n^3)$, rendering them impractical for datasets with thousands of observations or more. This challenge is exacerbated when dealing with the complex datasets mentioned earlier, necessitating the development of specialized techniques to reduce computational complexity.

To address these challenges and harness the full potential of GPs while managing their computational demands, various strategies and techniques have been proposed in the literature. These strategies include sparse approximations (Smola and Bartlett, 2001; Csató and Opper, 2002; Quiñonero-Candela and Rasmussen, 2005), variational inference (Titsias, 2009a), and reduced-rank approximations to GP covariance functions (Williams and Seeger, 2001a; Rahimi and Recht, 2007). One particularly promising approach is the use of spectral sparse GPs, which involves the

spectral decomposition of covariance functions. This technique, initially introduced for kernel methods by Rahimi and Recht (2007), based on the spectral decomposition by Bochner (1932)r, has been further developed and adopted for GPs in the context of GP regression (Lázaro-Gredilla et al., 2010). While substantial progress has been made in the context of regression, there is room for further development, especially in the context of handling spatial-temporal data.

This thesis sets out to explore and develop advanced methods using the spectral sparse representation of GPs for both time series data and spatial data. The primary research objectives encompass the following:

- **Establishing a theoretical foundation:** At the core of this study, we seek to establish a comprehensive and coherent theoretical groundwork for the application of spectral sparse GPs to time series and spatial data. This includes the development of a deep understanding of the fundamental principles and mathematical frameworks governing GPs, spatial data, and time series models. The elucidation of these theoretical foundations is integral to demonstrating how spectral sparse GPs constitute an effective instrument for data analysis.

- **Addressing computational efficiency challenges:** One of the primary goals is to tackle the computational challenges associated with large datasets. This involves devising efficient algorithms and methods to ensure that spectral sparse GPs can be applied to substantial datasets without compromising computational performance.

- **Demonstrating practical utility:** Through practical applications, this research aims to showcase the real-world utility of spectral sparse GPs. The focus areas include spatiotemporal forecasting, spatial interpolation, and uncertainty quantification. These practical demonstrations will illustrate the versatility and effectiveness of spectral sparse GPs in various contexts.

- **Comparative analysis:** To provide a comprehensive perspective, a comparative analysis will be conducted to evaluate the advantages of spectral sparse GPs. This analysis will consider factors such as accuracy and computational

efficiency, highlighting the strengths and benefits of this approach in relation to alternative methods.

1.2 Thesis structure

This thesis is organized into two distinct research domains: Gaussian process models for spatial data analysis and for change point detection for time series data. We establish a clear division between two parts: Part I and Part II, each tailored to the specific nature of the data under examination. Part I encompasses Chapters 3 and 4, delving into point processes and spatial data exclusively. Part II comprises Chapters 5 and 6, dedicated to the analysis of time series data.

Chapter 2 The second chapter of this thesis acts as an informative preamble, offering motivation and a foundational understanding of GPs. Our main aim is to establish a robust theoretical framework for Bayesian nonparametrics, with a specific focus on GPs. We also delve into practical aspects associated with GPs, especially regarding Gaussian process regression. Lastly, we deliver a comprehensive exploration of kernels and their intrinsic connections to covariance functions, introducing pivotal concepts and theorems that will serve as fundamental building blocks for subsequent chapters.

Part I The initial segment of the thesis, encompassing Chapter 3 and Chapter 4, focuses on our primary contribution, which revolves around point process models modulated by GPs for spatial data.

Chapter 3 serves as a comprehensive introduction, providing essential background information on point processes in the context of spatial data analysis. These models face two significant challenges: the first is inherent intractability, complicating Bayesian inference, while the second challenge arises due to the computational costs involved when handling large datasets. In this chapter, we also conduct a review of the current state of research concerning spatial models incorporating GPs, shedding light on their responses to these challenges.

In Chapter 4, we present our first substantial contribution: a novel approach for Bayesian inference in permanental processes that model the Poisson intensity as the square of a GP. Permanental processes are known for their inherent tractability, making them a convenient choice. However, certain limitations persist, particularly in terms of computational challenges and the restriction of tractability advantages to Gaussian kernels. Our contribution was inspired by the shortcomings of existing models. Our approach combines *generalized kernels* with a Fourier feature-based representation of the GP, resulting in rapid and efficient inference without the need for numerical integration across the input space. Furthermore, it permits the design of versatile kernels and offers linear scalability as the number of events grows.

Part II In the second part of our study, we investigate Bayesian nonparametric in the context of change point detection for time series data.

Chapter 5 provides a foundation for time series modeling, with a specific focus on models based on GPs. Unlike conventional methods for time series analysis, GPs offer significant advantages due to their non-parametric nature. We also introduce and discuss the Bayesian online change point detection method (BOCPD), along with its extensions utilizing Gaussian processes, for identifying changes in statistical characteristics within time series data.

Chapter 6 introduces our contribution, a novel variant of BOCPD featuring a reduced-rank Student-t process as a nonparametric time series model. This innovative approach combines a Student-t process with dependent Student-t noise to model time series data, with a Hilbert space reduced-rank kernel approximation to address computational complexity. The Student-t process extends the flexibility beyond the traditional Gaussian Process, offering a more versatile alternative. Within the context of BOCPD, Student-t processes also demonstrate a lower likelihood of generating false alarms when detecting change points caused by outliers. To further enhance computational efficiency, our approach incorporates a convenient Hilbert space-based reduced-rank representation of the Student-t process kernel, derived from an eigenfunction expansion of the Laplace operator.

1.3 Contribution

In this section, we outlined the key contributions in each main chapter and their relevance to the central theme of this thesis. Notably, the work presented here, including data generation and data analysis, was carried out by the author. This thesis incorporates contributions from two published works, both of which I served as the primary author:

- Sellier, J. and Dellaportas, P. (2023). Sparse spectral Bayesian permanental process with generalized kernel. *In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, AISTATS 2023.

- Sellier, J. and Dellaportas, P. (2023). Bayesian online change point detection with Hilbert space approximate Student-t process. *In Proceedings of the 40th International Conference on Machine Learning*, ICML 2023.

Chapter 2

Bayesian Nonparametrics with Gaussian processes

In this inaugural chapter, our primary goal is to establish a strong foundational framework within the field of Bayesian nonparametrics. We will place particular emphasis on Gaussian processes, recognizing their enduring importance as essential analytical tools in the upcoming contributing chapters. Within this framework, we will clarify fundamental concepts and intrinsic properties that are fundamental to the subject matter and the remainder of this thesis.

In Section 2.1, we provide a formal definition and characterization of stochastic processes, with a specific focus on Gaussian processes. In Section 2.2, we cover the Bayesian nonparametrics formalism and explain the use of Gaussian distributions as priors. Section 2.3 shifts our attention to the practical application of Gaussian processes, particularly in the context of regression, while also addressing its limitations. This section is designed to provide readers with a clear understanding of how to apply Gaussian processes to real-world problems. Finally, in Section 2.4, we offer a comprehensive overview of kernel methods, exploring the interconnections between kernels and covariance functions within the context of Gaussian processes. Additionally, Section 2.4 will introduce an overview of various kernels commonly found in the literature, each accompanied by concise explanations.

While we strive to make this thesis as self-contained as possible, we assume that

readers have a foundational understanding of measure theory. For an in-depth discussion of standard formulation of probability in terms of measure theory, we recommend referring to established texts such as Billingsley (1995a) or Capinski and Kopp (2013). It is worth noting that while measure theory plays a pivotal role throughout the majority of this chapter, Sections 2.3 and 2.4 do not require measure theory and should be of standalone interest.

2.1 Stochastic processes

Stochastic processes serve as versatile mathematical framework for modeling and understanding random phenomena and systems. This section will provide a concise introduction to stochastic processes, with a focus on their application in defining Gaussian Processes. For a deeper understanding of stochastic processes, we recommand consulting references such as Doob (1991) or Brémaud (2020).

In our pursuit of comprehending stochastic processes, we delve into two fundamental methods of characterization, each offering unique insights into their intrinsic properties. In Section 2.1.1, we will discuss the characterization of stochastic processes through their finite-dimensional distributions and the Kolmogorov extension theorem. This method plays a pivotal role in defining Gaussian processes, laying the foundation for our discussion in Section 2.1.3 where we establish a formal definition of Gaussian processes.

In Section 2.1.2, we introduce the Karhunen-Loève theorem, which provides a rigorous framework for representing stochastic processes as infinite series comprised of orthogonal functions. This theorem not only offers deeper insights into the inherent characteristics of stochastic processes but also paves the way for valuable perspectives on Gaussian processes, as discussed in Section 2.4.3. Through these methodical explorations, we aim to provide a structured foundation for comprehending stochastic processes and, in particular, their application – Gaussian processes.

A stochastic process is defined as a collection of random variables, each indexed by a set $\mathcal{X}$. The formal definition is as follows:

Definition 2.1.1 (**Stochastic process**). *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *be a mesurable state space. Let* $\mathcal{X}$ *be an indexed set defined on the standard Borel space* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$*. A stochastic process is a collection* $f = \{f(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ *such that for each fixed* $\mathbf{x} \in \mathcal{X}$*, the function* $f(\mathbf{x}, \cdot) : \Omega \to \mathbb{R}$ *is a random variable from* $(\Omega, \mathcal{F}, \mathbb{P})$ *to* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*.*

In simpler terms, a stochastic process f defined on $\mathcal{X}$ is a collection of random variables $f(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}$, where we often use the notation $f(\mathbf{x})$ as a shorthand for $f(\mathbf{x}, \omega)$. The index set $\mathcal{X}$ can take on various forms, including finite, countably infinite, or uncountably infinite. In the literature, it is frequently represented as T because stochastic processes are often defined as random functions over time. For example, T could be a subset of the real line, either countable (for discrete-time stochastic processes) or uncountable (for continuous-time stochastic processes). However, here, we take a more general approach and consider $\mathcal{X}$ as a multi-dimensional index, specifically $\mathcal{X} = \mathbb{R}^n$ for $n \in \mathbb{N}$.

Furthermore, a stochastic process can be interpreted as a random function $f : \mathcal{X} \times \Omega \to \mathbb{R}$ that assigns the value $f(\mathbf{x}, \omega)$ for every pair $(\mathbf{x}, \omega) \in \mathcal{X} \times \Omega$. When ω is fixed, the function $f(\cdot, \omega) : \mathcal{X} \to \mathbb{R}$ represents a deterministic sample function, commonly referred to as a trajectory or realization.

The concept of finite-dimensional distribution is a fundamental aspect of stochastic processes, offering insights into the joint marginal distributions of a finite number of indices within the process. Consider a finite set of indices $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$, where $n \in \mathbb{N}$. The collective joint marginal distributions of these indices constitute the finite-dimensional distributions of the stochastic process.

Definition 2.1.2 (**Finite dimensional distribution**). *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $f : \mathcal{X} \times \Omega \to \mathcal{Y}$ *be a stochastic process. The finite-dimensional distributions of a stochastic process is the family distributions of the random variables* $(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n))$ *for all choices of* $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$ *and* $n \in \mathbb{N}$*.*

2.1.1 Kolmogorov extension theorem

Representing stochastic processes as a function-valued random variable offers the advantage of treating them as single random objects rather than collections of random variables. This perspective can be achieved by considering the stochastic process $f : \mathcal{X} \times \Omega \to \mathbb{R}$ as a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the space of functions $\mathcal{X} \to \mathbb{R}$ along with its associated σ-algebra. Alternatively, we can express this as f belonging to the product space $\mathbb{R}^{\mathcal{X}}$, allowing us to view functions as vectors with elements indexed by members of $\mathcal{X}$.

However, this approach presents a challenge when defining probability distributions over infinite-dimensional objects. To address this challenge, the Kolmogorov extension theorem asserts that, under specific conditions, the finite-dimensional distributions of a stochastic process can be used to uniquely determine its infinite-dimensional distributions. Furthermore, it enables us to extend a family of distributions defined for finite subsets of $\mathcal{X}$ to uniquely define a stochastic process across the entire space.

In this section, we closely follow the presentation outlined by Matthews (2016) and adopt some of their notation. To introduce the σ-algebra associated with the infinite-dimensional space $\mathbb{R}^{\mathcal{X}}$, we need to establish several definitions. Let $\mathcal{V}$ and $\mathcal{U}$ be two finite-dimensional subspaces of $\mathcal{X}$ such that $\mathcal{V} \subseteq \mathcal{U} \subset \mathcal{X}$. The Borel σ-algebra of $\mathbb{R}^{\mathcal{V}}$ is denoted by $\mathcal{B}(\mathbb{R}^{\mathcal{V}})$. The sub-collection of random variables indexed by $\mathcal{U}$, represented as $f_{\mathcal{U}}$, is defined as $f_{\mathcal{U}} = \{f(\mathbf{x}) : \mathbf{x} \in \mathcal{U}\}$.

Additionally, we consider a *projection map*, denoted as $\pi_{\mathcal{U}:\mathcal{V}}$, from $\mathcal{U}$ onto $\mathcal{V}$ defined as follows:

$$\begin{aligned} \pi_{\mathcal{U}:\mathcal{V}} : \mathbb{R}^{\mathcal{U}} &\to \mathbb{R}^{\mathcal{V}} \\ f_{\mathcal{U}} &\to \pi_{\mathcal{U}:\mathcal{V}}(f_{\mathcal{U}}) = f_{\mathcal{V}}. \end{aligned} \tag{2.1}$$

In particular, the projection map $\pi_{\mathcal{X}:\mathcal{V}}(f)$ results in $f_{\mathcal{V}}$. For simplicity, we use $\pi_{\mathcal{V}}(f)$ to denote $\pi_{\mathcal{X}:\mathcal{V}}(f)$.

We define a *cylinder set* for a finite $\mathcal{V}$ as the pre-image of $\pi_{\mathcal{X}:\mathcal{V}}$, i.e.

$$\pi^{-1}_{\mathcal{X}:\mathcal{V}}(E) = \{f \in \mathbb{R}^{\mathcal{X}} : \pi_{\mathcal{X}:\mathcal{V}}(f) \in E\}, \tag{2.2}$$

where E is a set belonging to $\mathcal{B}(\mathbb{R}^{\mathcal{V}})$.

The σ-algebra generated by all cylinder sets, denoted as $\mathcal{G}$, is referred to as the product σ-algebra. In other words, $\mathcal{G}$ is generated from all sets of the form $\pi^{-1}_{\mathcal{X}:\mathcal{V}}(E)$, where $\mathcal{V}$ is a finite subset of $\mathcal{X}$ and E is an element of $\mathcal{B}(\mathbb{R}^{\mathcal{V}})$. The product σ-algebra serves as the underlying σ-algebra for the measurable space of the probability measure over functions, as defined in the Kolmogorov theorem.

With the product σ-field $\mathcal{G}$ now defined, we can proceed to state the Kolmogorov extension theorem as follows:

Theorem 2.1.3 (**Kolmogorov extension Theorem**). *Let $\mathcal{X}$ be an indexed set. Let $\{\mathbb{P}_{\mathcal{V}} : \mathcal{V} \subset \mathcal{X}\}$ be a family of probability distributions each defined on their respective measurable space $(\mathbb{R}^{\mathcal{V}}, \mathcal{B}(\mathbb{R}^{\mathcal{V}}))$. If for every finite sets $\mathcal{V}$ and $\mathcal{U}$ such that $\mathcal{V} \subset \mathcal{U} \subset \mathcal{X}$, we have*

$$\mathbb{P}_{\mathcal{U}}(\pi^{-1}_{\mathcal{U}:\mathcal{V}}(E)) = \mathbb{P}_{\mathcal{V}}(E), \qquad \forall\, E \in \mathcal{B}(\mathbb{R}^{\mathcal{V}}) \tag{2.3}$$

then there exists a unique probability measure $\mathbb{P}_{\mathcal{X}}$ on the product σ-algebra with the property

$$\mathbb{P}_{\mathcal{X}}(\pi^{-1}_{\mathcal{X}:\mathcal{V}}(E)) = \mathbb{P}_{\mathcal{V}}(E), \qquad \forall\, E \in \mathcal{B}(\mathbb{R}^{\mathcal{V}}). \tag{2.4}$$

A detailed proof of this theorem can be found in Billingsley (1995b, Chapter 7).

If the set of distributions $\{\mathbb{P}_{\mathcal{V}} : \mathcal{V} \subset \mathcal{X}\}$ satisfies the marginalization condition outlined in Equation (2.3), it implies the existence of a unique probability measure on the product σ-field $\mathcal{G}$ that has these distributions as its marginal distributions.

Alternatively, we can express this as the existence of a unique stochastic process f with probability distribution $\mathbb{P}_{\mathcal{X}}$ and finite-dimensional distributions $\{\mathbb{P}_{\mathcal{V}} : \mathcal{V} \subset \mathcal{X}\}$. In this context, $\mathbb{P}_{\mathcal{V}}$ can be defined as:

$$\mathbb{P}_{\mathcal{V}}(E) := \mathbb{P}(f_{\mathcal{V}} \in E), \qquad \forall\, E \in \mathcal{B}(\mathbb{R}^{\mathcal{V}}) \tag{2.5}$$

for any $\mathcal{V} \subset \mathcal{X}$. In the subsequent sections, we will refer to $\mathbb{P}_{\mathcal{V}}$ as $\mathbb{P}_{f,\mathcal{V}}$ for any $\mathcal{V} \subset \mathcal{X}$to indicate its association with the process f. Moreover, we will simplify the notation further by writing $\mathbb{P}_{f,\mathcal{X}}$ as simply $\mathbb{P}_f$.

In essence, the Kolmogorov extension theorem provides a systematic way to extend these partial probability distributions to a unique probability measure defined over the entire sample space. This measure represents the comprehensive probability distribution for the stochastic process.

2.1.2 Karhunen–Loève theorem

Another pivotal aspect of stochastic processes lies in their representation through the *Karhunen–Loève theorem* (Loeve, 1978). This theorem offers a method to express a stochastic process as an infinite series of orthogonal functions, akin to the representation of deterministic functions via Fourier analysis.

To apply the Karhunen-Loève theorem, we need to determine two key statistical moments of the process. The first is the mean function $m : \mathcal{X} \to \mathbb{R}$, defined as $m(\mathbf{x}) := \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]$. The second is the covariance function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as $c(\mathbf{x}, \mathbf{x}') := \mathbb{E}_{\mathbb{P}}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right]$.

Let $L^2(\mathcal{X})$ represent the space of square-integrable functions defined on $\mathcal{X}$. The Karhunen-Loève theorem can be formulated as follows :

Theorem 2.1.4 (**Karhunen Loève theorem**). *Let f be a zero mean square integrable stochastic process defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with continuous covariance function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\{\Phi_i\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$ be the orthonormal eigenfunctions and non-negative eigenvalues of the Hilbert-Schmidt operator $T_k : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined as $T_k[f] = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')d\mathbf{x}'$. Then f can be represented as:*

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} w_i \Phi_i(\mathbf{x}) \tag{2.6}$$

whith convergence in $L^2(\mathcal{X}, \mathbb{P})$ and

$$w_i = \int_{\mathcal{X}} f(\mathbf{x})\Phi_i(\mathbf{x})d\mathbf{x}. \tag{2.7}$$

Furthermore, the random variables w_i have zero-mean, are uncorrelated and have variance λ_i i.e.

$$\mathbb{E}[w_i] = 0 \qquad \textit{and} \qquad \mathbb{E}[w_i w_j] = \delta_{i,j}\lambda_i \quad \forall i, j \in \mathbb{N}. \tag{2.8}$$

For detailed definitions, please refer to Berlinet and Thomas-Agnan (2004, section 3.2) and Steinwart (2017, Lemma 3.3 and 3.7). The requirement of f being square integrable is equivalent to $f \in \mathcal{L}(\mathcal{X}, \mathbb{P})$ or $E_{\mathbb{P}}[f(\mathbf{x})^2] < \infty$ for all $\mathbf{x} \in \mathcal{X}$. Such processes are also referred to as second-order stochastic processes.

The Karhunen-Loève theorem (KL) asserts that any square-integrable stochastic process can be expressed as a linear combination of uncorrelated random variables with a mean of 0. The coefficients of this linear combination, represented by $\{w_i\}_{i=1}^{\infty}$, are orthogonal in the probability space, while the deterministic functions, $\{\Phi_i\}_{i=1}^{\infty}$, are orthogonal in $L^2(\mathcal{X})$.

The KL eigenfunctions and eigenvalues can be employed to define a unique representation of the random process, which is often simpler and more amenable to analysis than the original process. The KL theorem is closely intertwined with the spectral theorem for compact self-adjoint operators. For a deeper understanding of this relationship and its connections to other relevant concepts, such as the Mercer theorem, please refer to section 2.4.2.

One of the primary applications of the KL representation is its ability to reduce the dimensionality of the random process. This is achieved by considering only the most significant eigenfunctions and eigenvalues, as the eigenvalues are ranked in decreasing order. This process is commonly known as Principal Component Analysis (PCA) and finds extensive utility in signal processing and image analysis.

2.1.3 Gaussian Processes

In this section, we delve into the formal definition of Gaussian processes (GPs) through their finite-dimensional distribution and the Kolmogorov theorem (2.1.3). The Kolmogorov theorem establishes a fundamental characterization of GPs. Essen-

tially, a GP is a real-valued stochastic process, denoted as $f : \mathcal{X} \times \Omega \to \mathbb{R}$, where all its finite-dimensional distributions follow Gaussian distributions.

Definition 2.1.5. *Let $(\mathcal{X}, \mathcal{G})$ be a measurable space and $\mathbf{f} = f(\mathbf{x})_{\mathbf{x}\in\mathcal{X}}$ be a stochastic process defined on this space. Then, $\mathbf{f}$ is a Gaussian process if and only if for any finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, the corresponding random variables $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ have a joint Gaussian distribution.*

In other words, for any finite $\mathcal{V} \subset \mathcal{X}$, there exists a Gaussian measure $\mathbb{P}\{\boldsymbol{\mu}_\mathcal{V}, \boldsymbol{\Sigma}_\mathcal{V}\}$ parametrized by a finite vector $\boldsymbol{\mu}_\mathcal{V}$ and a positive semi-definite matrix $\boldsymbol{\Sigma}_\mathcal{V}$ such that $f_\mathcal{V} = \{f(\mathbf{x}) : \mathbf{x} \in \mathcal{V}\}$ have the distribution $\mathbb{P}\{\boldsymbol{\mu}_\mathcal{V}, \boldsymbol{\Sigma}_\mathcal{V}\}$.

The mean vector $\boldsymbol{\mu}_\mathcal{V}$ is computed via a mean function $\mu : \mathcal{X} \to \mathbb{R}$. For any finite set of inputs $\mathcal{V} := \{\mathbf{x}_i\}_{i=1}^n$, $n \in \mathbb{N}$, the value of the mean function at each input $\mathbf{x}_i$ corresponds to the i-th component of $\boldsymbol{\mu}_\mathcal{V}$, i.e., $\boldsymbol{\mu}_{\mathcal{V},i} = \mu(\mathbf{x}_i)$ for $i = 1, \cdots, n$.

The covariance matrix $\boldsymbol{\Sigma}_\mathcal{V}$ is constructed as a Gram matrix, with entries determined by evaluating a covariance function, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, on the set of inputs $\mathcal{V}$. The function k must be symmetric, i.e., $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and positive semi-definite, i.e., $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ for all $n \in \mathbb{N}, \mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{X}$ and non-zero $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$.

The GP, denoted as $\mathcal{GP}(\mu, k)$, is defined by its mean function μ and covariance function k. The notation $f(\mathbf{x}) \sim \mathcal{GP}(\mu, k)$ indicates that the function $f(\mathbf{x})$ is a sample from this GP.

To demonstrate the existance of the GP, we can easily verify that the familiy of multivariate Gaussian measures $\{\mathbb{P}\{\boldsymbol{\mu}_\mathcal{V}, \boldsymbol{\Sigma}_\mathcal{V}\} : \mathcal{V} \subset \mathcal{X}\}$ parametrized by μ anf k satisfies the marginalization property in Equation (2.3). This implies that for any $\mathcal{V} \subset \mathcal{U} \subset \mathcal{X}$, the measure we obtain by restricting $\mathbb{P}\{\boldsymbol{\mu}_\mathcal{U}, \Sigma_\mathcal{U}\}$ to the set $\mathcal{V}$ is $\mathbb{P}\{\boldsymbol{\mu}_\mathcal{V}, \Sigma_\mathcal{V}\}$ [1]. As a result, this family of Gaussian finite-dimensional distributions can be extended to a unique stochastic process, which is the GP.

[1] This holds due to the fact that the entries in the covariance matrix corresponding to $\mathbf{x}, \mathbf{x}' \in \mathcal{V}$ are equal in both $\boldsymbol{\Sigma}_\mathcal{U}$ and $\boldsymbol{\Sigma}_\mathcal{V}$, and the same holds true for the mean function.

2.2 Bayesian nonparametrics

In the preceding section, we introduced the concept of a stochastic process as an infinite-dimensional random valued function. In this section, we introduce Bayesian nonparametrics — a versatile framework designed to infer the distribution of such functions from a finite set of observations. Unlike parametric models that rely on a fixed number of parameters, Bayesian nonparametric models operate within an infinite-dimensional parameter space, affording them the flexibility and expressiveness needed to tackle a wide array of complex modeling tasks.

Bayesian nonparametric models have showcased their versatility across various domains, including regression, classification, clustering, and latent variable modeling. A substantial body of literature has explored the foundational concepts and motivations underlying these models, with notable contributions from Hjort et al. (2010), Orbanz and Teh (2011), and Ghahramani (2012).

This section reviews the fundamental theoretical principles of Bayesian nonparametric models. In Section 2.2.1, we present a generalized version of Bayes' theorem tailored to address the intricacies inherent in infinite-dimensional parameter spaces . Moving forward, Section 2.2.2 navigates the practical considerations essential for the evaluation of Bayesian nonparametric models, particularly when constrained to a finite subset of the parameter space. Lastly, we provide an insightful exploration of Gaussian processes as a prime exemplar of Bayesian nonparametric models in Section 2.2.3. Through these discussions, we aim to provide a comprehensive overview of Bayesian nonparametrics, offering both theoretical insights and practical guidance for leveraging this powerful framework in various modeling scenarios.

2.2.1 Generalized Bayes theorem

Applying Bayes' theorem in infinite dimensions presents a complex challenge. The conventional Bayesian framework relies on the Lebesgue measure, which poses difficulties when dealing with infinite-dimensional spaces. This leads to undefined prior and posterior densities for infinite objects, necessitating the development of a gener-

alized form of Bayes' theorem. In this section, we closely follow the insights presented by Schervish (1996) and Matthews (2016) to address this challenge.

In the previous section, we introduced a random function denoted as f, defined over an index set $\mathcal{X}$. We demonstrated that f can be defined on the product space $\mathbb{R}^{\mathcal{X}}$ along with its associated product σ-algebra $\mathcal{G}$, effectively forming a measurable space denoted as $(\mathbb{R}^{\mathcal{X}}, \mathcal{G})$. To perform Bayesian inference on f, we introduce a prior probability distribution $\mathbb{P}_f$ on the measurable space $(\mathbb{R}^{\mathcal{X}}, \mathcal{G})$. We also introduce a vector of observations $\mathbf{y} = \{y_i\}_{i=1}^n$ with elements in $\mathbb{R}$. Our objective is to infer the posterior distribution of f given $\mathbf{y}$, denoted by $\hat{\mathbb{P}}_f$.

To establish a probabilistic linkage between f and $\mathbf{y}$, we introduce a critical component, the likelihood function denoted as $p(\mathbf{y} \mid f)$. Under the assumption that $\mathbb{P}_f$ is absolutely continuous with respect to $\hat{\mathbb{P}}_f$ (Billingsley, 1995b, p. 422), we can use the Radon-Nikodym (RN) derivative of the posterior with respect to the prior to derive a more general form of Bayes' theorem, as proven in Stuart (2010, Theorem 6.31):

$$\frac{\partial \hat{\mathbb{P}}_f}{\partial \mathbb{P}_f}(u) = \frac{p(\mathbf{y}|f=u)}{p(\mathbf{y})}, \qquad \forall u \in \mathbb{R}^{\mathcal{X}}. \tag{2.9}$$

Here, $p(\mathbf{y})$ represents a crucial quantity known as the *marginal likelihood*, given by:

$$p(\mathbf{y}) := \int_{\mathbb{R}^{\mathcal{X}}} p(\mathbf{y}|f=u) d\mathbb{P}_f(u). \tag{2.10}$$

In the special case where f is finite and both $\mathbb{P}_f$ and $\hat{\mathbb{P}}_f$ are absolutely continuous with respect to the Lebesgue measure μ, we can recover the classical Bayes' theorem using the chain rule for RN derivatives as follows:

$$\begin{aligned}\frac{\partial \hat{\mathbb{P}}_f}{\partial \mu}(u) &= \frac{\partial \hat{\mathbb{P}}_f}{\partial \mathbb{P}_f}(u) \frac{\partial \mathbb{P}_f}{\partial \mu}(u) \\ &= \frac{p(\mathbf{y}|f=u)}{p(\mathbf{y})} \frac{\partial \mathbb{P}_f}{\partial \mu}(u).\end{aligned}$$

In the more general scenario where f is an infinite-dimensional function, we express the posterior measure $\hat{\mathbb{P}}_f$ as the measure induced by the RN derivative $\frac{\partial \hat{\mathbb{P}} f}{\partial \mathbb{P}_f}$, as

defined in Equation (2.9). Consequently, we have:

$$\hat{\mathbb{P}}_f(E) = \int_E \frac{p(\mathbf{y}|f=u)}{p(\mathbf{y})} d\mathbb{P}_f(u) \tag{2.11}$$

for all $E \in \mathcal{G}$.

2.2.2 Working with a finite subset of the parameter space

In practical applications of Bayesian nonparametric models, we often focus on a finite subset of the available parameter space. This subset is chosen to adequately explain the observed sample, and the excess dimensions are typically integrated out over the prior to compute the marginal likelihood.

In some cases, the likelihood function $p(\mathbf{y}|f)$ depends solely on the values of f at a limited set of data points denoted as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. Let $\mathbf{f}$ represent the evaluation of f at these data points, i.e., $\mathbf{f} := f_{\mathbf{X}} \triangleq \{f(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}}$. As a result, we can simplify the likelihood function to $p(\mathbf{y}|\mathbf{f})$.

The posterior measure, which expresses the updated belief about the parameter f after considering the observed data, can be calculated using the generalized Bayes formula in Equation (2.9) :

$$\frac{\partial \hat{\mathbb{P}} f}{\partial \mathbb{P} f}(u) = \frac{p(\mathbf{y}|\mathbf{f} = \pi_{\mathbf{X}}(u))}{p(\mathbf{y})}, \qquad \forall u \in \mathbb{R}^{\mathcal{X}} \tag{2.12}$$

where $\pi_{\mathbf{X}} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^n$ is a coordinate projection onto $\mathbf{X}$ such that $\pi_{\mathbf{X}}(u) := u_{\mathbf{X}} \triangleq \{u(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}}$ for all $u \in \mathbb{R}^{\mathcal{X}}$. The denominator, $p(\mathbf{y})$, represents the marginal likelihood of the observed data and can be calculated as a finite integral over the prior marginal probability measure of $\mathbf{f}$, denoted by $\mathbb{P}_{f_{\mathbf{X}}}$,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f}) d\mathbb{P}_{f_{\mathbf{X}}}(\mathbf{f}). \tag{2.13}$$

Equation (2.11) is transformed into a finite-dimensional integral term, as expressed

in Equation (2.14):

$$\hat{\mathbb{P}}_f(E) = \int_E \frac{p(\mathbf{y}|\mathbf{f} = \pi_{\mathbf{X}}(u))}{p(\mathbf{y})} d\mathbb{P}_f(u), \qquad \forall E \in \mathcal{B}(\mathbb{R}^X). \tag{2.14}$$

Inference For inference, we only need to evaluate the posterior in its finite-dimensional form. In prediction tasks, where we anticipate a set of predictive points denoted by $\mathbf{X}^*$, alongside the data points $\mathbf{X}$, we define the finite-dimensional vector $\mathbf{f}^*$ as the evaluation of the model's underlying function f at the predictive points $\mathbf{X}^*$. The projection of f onto $\mathbf{X} \cup \mathbf{X}^*$ is denoted as $f_{*\cup\mathbf{X}}$.

The equation for the posterior measure of the projected function $f_{*\cup\mathbf{X}}$, denoted as , $\hat{\mathbb{P}}_{f*\cup\mathbf{X}}$ can be expressed as:

$$\begin{aligned}\hat{\mathbb{P}}_{f_{*\cup\mathbf{X}}}(E) = \hat{\mathbb{P}}_f(\pi^{-1}_{*\cup\mathbf{X}}(E)) &= \int_{\pi^{-1}_{*\cup X}(E)} \frac{p(\mathbf{y}|\mathbf{f} = \pi_{\mathbf{X}}(u))}{p(\mathbf{y})}\, d\mathbb{P}_f(u) \\ &= \int_E \frac{p(\mathbf{y}|\mathbf{f} = \pi_{*\cup\mathbf{X}:\mathbf{X}}(u))}{p(\mathbf{y})}\, d\mathbb{P}_{f_{*\cup X}}(u)\end{aligned} \tag{2.15}$$

for all E in the σ-algebra associated with $\{\mathbf{X}^* \cup \mathbf{X}\}$. Importantly, this integral is of finite dimension, enabling efficient calculations in practical scenarios.

2.2.3 Gaussian process as a prior

The Gaussian process described in Section 2.1.3 provides a suitable prior for modeling functions. The finite-dimensional marginals of the underlying prior measure, denoted as $\mathbb{P}_f$, are Gaussian, characterized by a mean μ and covariance function k. Consequently, these marginals possess a well-defined density with respect to the Lebesgue measure.

Equation (2.15) reveals that the finite-dimensional form of the posterior measure $\hat{\mathbb{P}}_f$ at the input location $\mathbf{X}$ also possesses a density, denoted by $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$. This density can be expressed as follows:

$$p(\mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{p(\mathbf{y})}. \tag{2.16}$$

Here, $p(\mathbf{f}|\mathbf{X})$ represents the Gaussian density of the finite-dimensional measure $\mathbb{P}_{f_\mathbf{X}}$.

Similarly, the density of the posterior distribution of $f_{\cup \mathbf{X}}$ can be expressed as follows:

$$p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, \mathbf{X}, \mathbf{X}^*) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}, \mathbf{f}^*|\mathbf{X}, \mathbf{X}^*)}{p(\mathbf{y})}. \tag{2.17}$$

Here $p(\mathbf{f}, \mathbf{f}^*|\mathbf{X}, \mathbf{X}^*)$ denotes the Gaussian density of $\mathbb{P}_{f_{*\cup \mathbf{X}}}$ and $p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, \mathbf{X}, \mathbf{X}^*)$ represents the density of $\hat{\mathbb{P}}_{f_{*\cup X}}$.

Assuming a Gaussian process prior allows us to derive these standard forms of the Bayes formula, presented in Equations (2.16) and (2.17). Furthermore, these equations retain their tractability when we assume a Gaussian distribution for $p(\mathbf{y}|\mathbf{f})$, as discussed in the subsequent section, making them particularly convenient for practical applications.

2.3 Gaussian processes as inference mode

In the preceding section, we introduced GPs as a valuable Bayesian framework for modeling mapping functions. GPs have demonstrated their effectiveness in capturing intricate and non-linear relationships between input variables $\mathbf{X}$ and output variables $\mathbf{y}$ by incorporating a prior on the underlying mapping function that connects $\mathbf{X}$ and $\mathbf{y}$ and refining it through Bayesian inference as new data emerges. This versatility has led to their widespread adoption across a broad spectrum of machine learning applications, spanning supervised learning, unsupervised learning, and reinforcement learning.

This section provides a comprehensive overview of the practical aspects associated with Gaussian processes. In Section 2.3.1, we present the Gaussian process regression model, which serves as an illustrative example of the GP methodology. Additionally, in Section 2.3.2, we outline the various limitations and challenges inherent to GP modeling, offering insights into both its strengths and areas that require careful consideration.

2.3.1 Gaussian process regression

Gaussian processes have been particularily popular due to their conjugacy property. Specifically, when the likelihood function $p(\mathbf{y}|\mathbf{f})$ takes on a Gaussian form, the resulting posterior within the GP framework, in Equations (2.16) and (2.17), remains a GP. This property greatly simplifies the inference process. Such a scenario occurs when observations $\mathbf{y}$ correspond precisely to evaluations of a mapping function $f : \mathcal{X} \to \mathbb{R}$ at data points $\mathbf{X}$, including additive Gaussian noise.

In the context of regression, this relationship is typically represented by assuming that each observation in $\mathbf{y} = \{y_i\}_{i=1}^n$ arises from both f and a noise vector $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^n$, as follows:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \quad \text{for all } i = 1, \ldots, n. \tag{2.18}$$

Here, $\boldsymbol{\varepsilon}$ follows an independent and identically distributed Gaussian distribution with zero mean and variance σ_n^2, i.e. $\varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma_n^2)$. This leads to a likelihood function:

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_n^2 \mathbf{I}_n). \tag{2.19}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. In GP regression, we assume a GP prior over f in Equation (2.18), defined by a mean function μ and a covariance function k_θ with hyperparameters θ. These hyperparameters θ control the smoothness, scaling, and other general properties of the covariance and functions generated from the GP. It is common to set the prior mean function to zero, i.e., $\mu(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$. The inclusion of Gaussian noise simplifies the inference process, enhancing the efficiency and practicality of Gaussian process regression.

Posterior over the latent function In a Bayesian context, Equation (2.16) provides a means to compute the posterior distribution of the latent functions $\mathbf{f}$ and $\mathbf{f}^*$. When dealing with a Gaussian likelihood, both the posterior and predictive posterior can be determined analytically, as the prior and Gaussian likelihood form a conjugate pair. Consider the joint distribution of $\mathbf{f}$, $\mathbf{f}^*$ and $\mathbf{y}$ under this prior, which is described

by the following Gaussian distribution:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_n & \mathbf{K}_{n,n} & \mathbf{K}_{n,*} \\ \mathbf{K}_{n,n} & \mathbf{K}_{n,n} & \mathbf{K}_{n,*} \\ \mathbf{K}_{n,*}^\top & \mathbf{K}_{n,*}^\top & \mathbf{K}_{*,*} \end{bmatrix} \right) \tag{2.20}$$

where $\mathbf{K}_{n,n}$ is the $n \times n$ Gram matrix with i, j entries $k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{K}_{*,*}$ is the $n^* \times n^*$ Gram matrix with i, j entries $k(\mathbf{x}_i^*, \mathbf{x}_j^*)$ and $\mathbf{K}_{n,*}$ is the $n \times n^*$ Gram matrix with i, j entries $k(\mathbf{x}_i, \mathbf{x}_j^*)$.

Finally, the posterior distribution over the latent values $\mathbf{f}$ and $\mathbf{f}^*$ can be computed using the standard conditional rule of the Gaussian distribution, as detailed in Appendix B.1, resulting in:

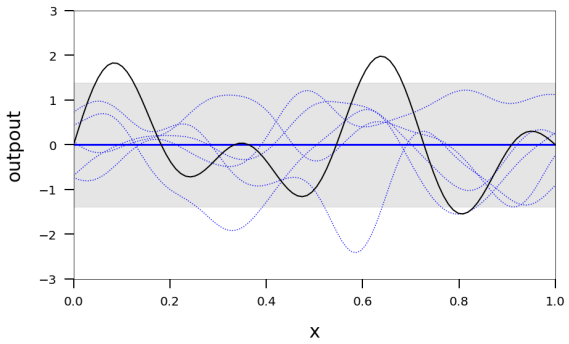
$$p(\mathbf{f}, \mathbf{f}^* | \mathbf{y}, \mathbf{X}, \mathbf{X}^*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \mu(\mathbf{X}) + \mathbf{K}_{n,n} \left(\mathbf{K}_{n,n} + \sigma_{n^2} \mathbf{I}_n \right)^{-1} (\mathbf{y} - \mu(\mathbf{X})) \\ \mu(\mathbf{X}^*) + \mathbf{K}_{*,n} \left(\mathbf{K}_{n,n} + \sigma_{n^2} \mathbf{I}_n \right)^{-1} (\mathbf{y} - \mu(\mathbf{X})) \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_{n,n})^{-1} \mathbf{K}_{n,*} & \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_{n,n})^{-1} \mathbf{K}_{n,*} \\ \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_{n,n})^{-1} \mathbf{K}_{n,*} & \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_{n,n})^{-1} \mathbf{K}_{n,*} \end{bmatrix} \right) \tag{2.21}$$

where $\mu(\mathbf{X})$ is the n-dimension vector with i-th entry $\mu_i = \mu(\mathbf{x}_i)$.

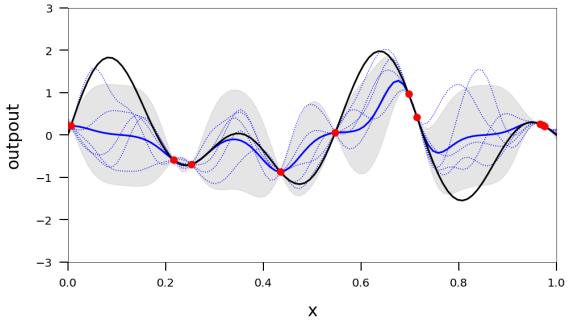
Given $p(\mathbf{f}, \mathbf{f}^* | \mathbf{y}, \mathbf{X}, \mathbf{X}^*)$, we can also calculate the predictive distribution of the latent function at test points $\mathbf{X}^*$ by marginalizing out $\mathbf{f}$ using:

$$p(\mathbf{f}^* | \mathbf{y}, \mathbf{X}, \mathbf{X}^*) = \int p(\mathbf{f}^* | \mathbf{f}, \mathbf{X}^*) p(\mathbf{f} | \mathbf{X}, \mathbf{y}) d\mathbf{f} \tag{2.22}$$

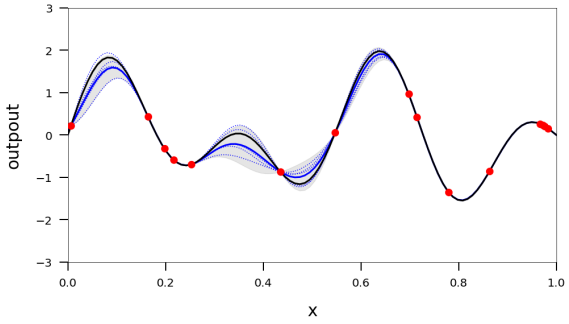
By once again applying the conditional rule of the Gaussian distribution, we can derive the posterior distribution $p(\mathbf{f}^* | \mathbf{y}, \mathbf{X}, \mathbf{X}^*)$, which is also Gaussian and charac-



(a) GP prior ($n = 0$)



(b) GP posterior ($n = 10$)



(c) GP posterior ($n = 15$)

Figure 2.1: Gaussian process prior and posterior of a one-dimensional function $f(x) = \sin(4\pi x) + \sin(7\pi x)$ (black) with Gaussian kernel. The blue line represents the GP posterior mean conditioned on n observations (red dot). The blue dashed lines represent 5 random function samples from the GP prior (a) or posterior (b,c). The shaded area represents the pointwise mean ± 1.96 standard deviation of each value (corresponding to a 95% confidence region). As the number of observations increases, the GP posterior increasingly better approximates the function. Specifically, (b) and (c) show the GP posterior for $n = 10$ and $n = 15$ observations, respectively.

terized by the following mean and variance expressions:

$$p(\mathbf{f}^*|\mathbf{y}, \mathbf{X}, \mathbf{X}^*) = \mathcal{N}(\mathbf{f}^*|\mu^*, \Sigma^*) \qquad \text{with} \tag{2.23}$$

$$\mu^* := \mu(\mathbf{X}) + \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2\mathbf{I}_n)^{-1}(\mathbf{y} - \mu(\mathbf{X})) \tag{2.24}$$

$$\Sigma^* := \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma_n^2\mathbf{I}_n)^{-1}\mathbf{K}_{n,*} \tag{2.25}$$

These results are discussed in detail in Rasmussen and Williams (2005, p.16). Notably, Equation (2.25) illustrates that the predictive variance comprises two components: the prior variance at the test points and a term representing the reduction in uncertainty resulting from the additional information provided by the observations.

In fact, we can represent the posterior as a single Gaussian process, referred to as the posterior GP. This GP has a mean function $\mu^* : \mathcal{X} \to \mathbb{R}$ and a covariance function $k^* : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as follows:

$$\mu^*(\mathbf{x}) = \mu(\mathbf{x}) + k(\mathbf{x}, \mathbf{X})(\mathbf{K}_{n,n} + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{y} - \mu(\mathbf{x})), \qquad \forall\, \mathbf{x} \in \mathcal{X} \tag{2.26}$$

$$k^*(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X})(\mathbf{K}_{n,n} + \sigma^2\mathbf{I}_n)^{-1}k(\mathbf{X}, \mathbf{x}), \qquad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{2.27}$$

where $k(\mathbf{x}, \mathbf{X}) = k(\mathbf{X}, \mathbf{x})^\top$ denotes the $1 \times n$ vector whose ith entry is $k(\mathbf{x}, \mathbf{x}_i)$ for $i = 1, \cdots, n$. These functions map inputs $\mathcal{X}$ to real values $\mathbb{R}$ and satisfy the properties of a Gaussian process.

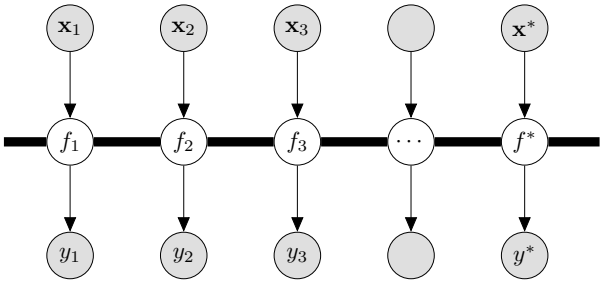
Marginal likelihood The marginal likelihood $p(\mathbf{y}|\mathbf{X})$ can be obtained by integrating out the latent function using the likelihood and the prior:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f} \tag{2.28}$$

In GP regression the marginal likelihood is available in closed form:

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}|0, \mathbf{K}_{n,n} + \sigma_n^2\mathbf{I}_n). \tag{2.29}$$

Figure 2.2: The arrows directions suggest directions of influence. Grey circles represent the observed variables. The bold horizontal bar represent a set of fully connected nodes. The observation y_i are equal to the corresponding latent variables f_i plus some observation noise and are conditionally independent of all other nodes given f_i.



The log marginal likelihood can then be calculated as:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top \left(\mathbf{K}_{n,n} + \sigma_n^2 \mathbf{I}_n\right)^{-1} \mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2 \mathbf{I}_n| - \frac{n}{2}\log 2\pi. \tag{2.30}$$

where $|\cdot|$ denotes the determinent of the matrix argument.

2.3.2 Practical challenges in GP modelling

In this section, we delve into the practical challenges frequently encountered during the implementation of GPs and explore the strategies devised to overcome them. Specifically, we discuss model selection, non-Gaussian likelihoods and scalability issue.

Model Selection Selecting the appropriate GP prior is a pivotal aspect of Gaussian Process modeling. This includes the choice of a covariance function family and the definition of its hyperparameters. This critical step is commonly referred to as "model selection" within the literature. In-depth discussions on different approaches can be found in Rasmussen and Williams (2005).

One approach is to employ full Bayesian inference (Williams and Rasmussen, 1995; Hensman et al., 2015), which entails placing priors on the hyperparameters and computing the posterior distribution over these parameters. However, implementing Bayesian inference can be challenging, as it necessitates the evaluation of multiple

integrals, which may not be analytically tractable depending on the model.

Another widely-used strategy is Empirical Bayes, also known as the maximum likelihood approach. This method optimizes the hyperparameters by maximizing the log-marginal likelihood of the data, denoted as $p(\mathbf{y}|\mathbf{X})$. This approach is numerically more stable, particularly as n (the number of data points) increases. Nevertheless, it diverges from traditional Bayesian methods because the prior is estimated from the data.

Cross-Validation (CV) presents an alternative option, employing resampling techniques to assess the model's predictive performance. The data is partitioned into M non-overlapping sets, with $M-1$ sets allocated for training the GP and the remaining set used for validation. This process is repeated with different validation sets iteratively. The model's performance on the validation sets acts as a proxy for generalization error, aiding in the selection of the optimal hyperparameters for the GP model.

Non-Gaussian Likelihoods Gaussian processes are renowned for their convenience when paired with Gaussian likelihoods, as they yield tractable Gaussian posteriors. However, the real world often presents scenarios where these Gaussian assumptions do not apply. For instance, in linear logistic regression models for binary classification, the likelihood takes the form of a sigmoid transformation applied to the latent function f. Similarly, when Gaussian processes are employed for count data, likelihoods may involve Poisson distributions.

To contend with such non-Gaussian likelihoods, approximate inference methods have been developed. These methods provide computationally feasible Gaussian approximations for non-Gaussian posteriors. Commonly used techniques include Monte Carlo sampling (Filippone et al., 2013; Havasi et al., 2018; Neal, 1997), Laplace Approximation (Williams and Barber, 1998; Flaxman et al., 2015a), Expectation Propagation (EP) (Minka, 2001; Hernandez-Lobato and Hernandez-Lobato, 2016), and Variational Bayes (Opper and Archambeau, 2009; Frigola et al., 2014a; Blei et al., 2017a; Tran et al., 2016; Sheth et al., 2015; Hensman et al., 2015). Comparative

studies of these approximate inference methods can be found in Kuss and Rasmussen (2006) and Nickisch and Rasmussen (2008).

Scalability Standard GPs have a cubic time complexity $O(n^3)$ which arises from the need to compute the inverse and determinant of the $n \times n$ kernel matrix $\mathbf{K}_{n,n}$ (as seen in Equations (2.24), (2.25), and (2.30)). This limitation makes them challenging to apply to datasets with thousands of observations or more. To address this, various approximations have been proposed in recent literature to enhance GP scalability while preserving prediction quality. These approaches are comprehensively reviewed by Liu et al. (2020).

These scalability approximations can be broadly categorized into two types: *global approximations* and *local approximations*. Global approximations aim to summarize the entire training set using a smaller set of support points, such as sparse methods, kernel approximation via a subset of the training data (Keerthi and Chu, 2005; Lawrence et al., 2002; Seeger, 2003), sparse kernel approximation (Buhmann, 2001; Gneiting, 2002; Melkumyan and Ramos, 2009; Wendland, 2004). In contrast, local approximations rely on multiple local experts that each cover a specific region (Datta et al., 2016; Gramacy, 2016; Liu et al., 2018; Rasmussen and Ghahramani, 2001; Samo and Roberts, 2015c; Park and Huang, 2016).

The most prevalent approach for mitigating the cubic complexity of GPs is through sparse approximation. In these methods, a set of $m \ll n$ inducing variables is employed to globally represent the GP posterior, substantially reducing the training complexity to $O(m^2 n)$. One approach involves a reduced-rank approximation of the kernel function, achieved by mapping the input space from $\mathbb{R}^d$ to a lower-dimensional space $\mathbb{R}^m$ ($m < n$) using a feature map ϕ. The approximated kernel function becomes $k(\mathbf{x}, \mathbf{x}') \approx \phi(\mathbf{x})^\top \phi(\mathbf{x}')$. Examples of algorithms in this category include the Nyström method (Williams and Seeger, 2001a) and random Fourier features (RFF) (Rahimi and Recht, 2007), which are discussed further in Sections 3.3.3 and 4.3.1.

Another popular class of sparse approximations involves modifying the GP prior to establish an augmented joint prior $p(\mathbf{f}, \mathbf{f}_m) = p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m)$, where $(\mathbf{f}_m, \mathbf{X}_m)$ repre-

sents a set of inducing pair. This can be accomplished through deterministic (Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger et al., 2003) or partially/fully independent (Snelson and Ghahramani, 2006; Quiñonero-Candela and Rasmussen, 2005) assumptions on the conditionals $p(\mathbf{f}|\mathbf{f}_m)$. However, these approaches transform the inducing points into additional kernel hyperparameters, which can lead to overfitting when jointly optimizing all unknown hyperparameters together. Further information on these methods can be found in Quiñonero-Candela and Rasmussen (2005).

The Variational Sparse GP method, introduced by Titsias (2009a), combines the strengths of exact prior knowledge with variational approximation for the GP posterior distribution. This method approximates the exact GP posterior, $p(\mathbf{f}|\mathbf{y})$, with a variational distribution, $q(\mathbf{f}) = p(\mathbf{f}|\mathbf{f}_m)q(\mathbf{f}_m)$, by minimizing the KL divergence between the two distributions. The optimization process adjusts the pseudo-points and kernel hyperparameters jointly to maximize the evidence lower bound (ELBO). This approach uses inducing points as variational parameters, helping to address the overfitting problem associated with sparse approximations of the prior. Moreover, the variational distribution is typically assumed to be Gaussian, allowing for approximate inference even when the posterior is not available in closed form. Variational Sparse GP has garnered significant attention in recent literature, with various improvements and extensions proposed (Lázaro-Gredilla and Figueiras-Vidal, 2009; Hensman et al., 2017; Adam et al., 2020; van der Wilk et al., 2017; Hensman et al., 2013; Hoffman et al., 2013).

Finally, scalability can also be improved by leveraging the structure of the covariance function to achieve efficient matrix inversion with fast matrix-vector multiplication. For instance, the Kroenecker method Gilboa et al. (2013); Flaxman et al. (2015a) takes advantage of the tensor product structure of the kernel for multivariate cartesian product grid inputs. The Toeplitz method (Cunningham et al., 2008a) is another example that exploits the kernel matrix structure on a regularly spaced one-dimensional grid. While these methods offer enhanced efficiency, they are constrained by the requirement for grid-structured inputs, limiting their applicability to most datasets.

2.4 Kernels and covariance functions

As previously discussed, the selection and design of covariance functions play a pivotal role in the proper design of Gaussian process tasks. This section explores fundamental concepts related to covariance functions that will be utilized in subsequent chapters. Specifically, we introduce the concept of kernel functions and conduct a comprehensive investigation into their relationships with covariance functions within the framework of Gaussian processes.

In Section 2.4.1, we begin by introducing the foundational concepts of Reproducing Kernel Hilbert Spaces (RKHS) and kernel methods. Notably, we discuss the kernel trick, a cornerstone of machine learning, that empowers linear algorithms to discern non-linear patterns in data. This technique forms the basis for kernel methods as a distinct algorithmic class. Moving forward, Section 2.4.2 introduce the Mercer theorem, a critical tool for decomposing kernels into their eigenfunctions. The significance of this theorem becomes apparent in our exploration of the interplay between kernels and covariance functions within the context of Gaussian processes, a formal presentation of which is found in Section 2.4.3. Within Section 2.4.3, we introduce the KL-decomposition of a Gaussian process, providing valuable insights into its inherent properties and behavior. Finally, in Section 2.4.4, we provide a diverse array of various kernel types and offer comprehensive descriptions of their unique characteristics. Through these examples, we aim to illustrate the practical applications of the concepts elucidated throughout this chapter.

2.4.1 Reproducing kernel Hilbert space

In this section, we provide a concise introduction to Reproducing Kernel Hilbert Spaces (RKHS) and their associated reproducing kernels. RKHS, initially introduced by various authors including Aronszajn (1950) in seminal works, represent a class of Hilbert spaces of functions characterized by a unique reproducing kernel. For a more comprehensive understanding of the subject, we recommend referring to the works of Berlinet and Thomas-Agnan (2004) and Steinwart and Christmann (2007).

An RKHS is a Hilbert space $\mathcal{H}$ of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$ that possesses a distinctive property: the evaluations of two functions become pointwise close when these two functions are close enough in the norm of $\mathcal{H}$. This property is formalized through evaluation functionals $e_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}$, defined as $e_{\mathbf{x}}(f) = f(\mathbf{x})$, required to be continuous.

Definition 2.4.1 (**Reproducing kernel Hilbert space**). *A Hilbert space $\mathcal{H}$ of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing kernel Hilbert space (RKHS) if all evaluational functionals $e_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}$, defined by $e_{\mathbf{x}}(f) = f(\mathbf{x})$, are continuous.*

In Definition 2.4.1, the term "continuity"signifies that for any function f in $\mathcal{H}$, there exists a constant $c > 0$ such that $|e_{\mathbf{x}}(f)| \leq c||f||_{\mathcal{H}}$, where $||f||_{\mathcal{H}}$ denotes the norm induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, the inner product on $\mathcal{H}$, namely $||f||^2_{\mathcal{H}} := \langle f, f \rangle_{\mathcal{H}}$.

An RKHS is closely associated to a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ known as the *reproducing kernel*, which enables the evaluation of any function within the space.

Definition 2.4.2 (**Reproducing kernel**). *Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is considered a reproducing kernel of $\mathcal{H}$ if it satisfies*

- $\forall \mathbf{x} \in \mathcal{X}, \quad k(\cdot, \mathbf{x}) \in \mathcal{H}$
- $\forall \mathbf{x} \in \mathcal{X}, \forall h \in \mathcal{H}, \quad \langle h, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = h(\mathbf{x})$

In particular, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$k(\mathbf{x}, \mathbf{x}') = \langle k(\cdot, \mathbf{x}'), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}.$$

The second condition in Definition 2.4.2, known as the *reproducing property*, establishes that any evaluation functional in $\mathcal{H}$ is continuous. Additionally, the Reiz's theorem (Rudin, 1987, Theorem 4.12) guarantees the existence of a unique reproducing kernel in any RKHS. Therefore, $\mathcal{H}$ is an RKHS if and only if a reproducing kernel exists. As an alternative to Definition 2.4.1, RKHS may be defined as Hilbert spaces of functions with reproducing kernels. Moreover, the reproducing kernel associated

with $\mathcal{H}$ is unique, and as such, $\mathcal{H}$ is often denoted as $\mathcal{H}_k$ to indicate the specific reproducing kernel k associated with it.

The reproducing kernel k has two important properties derived from Definition 2.4.2 : it is symmetric, meaning that $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and positive definite, meaning that for any, $n \geq 1$, any $(w_1, \ldots, w_n) \in \mathbb{R}$ and $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0. \tag{2.31}$$

RKHS construction We have established the existence of a unique reproducing kernel function for every RKHS, which is both symmetric and positive definite. Conversely, the Moore-Aronszajn theorem (Aronszajn, 1950) provides a crucial insight by stating that any symmetric and positive definite function defines a unique RKHS. The theorem is presented as follows:

Definition 2.4.3 (**Moore-Aronsajn theorem**). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite function. Then there exists a unique RKHS $\mathcal{H}_k$ of functions on $\mathcal{X}$ for which k is a reproducing kernel. And the subspace $\mathcal{H}_0$ spanned by the functions $k(\cdot, \mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$ with the inner product*

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}'_j) \tag{2.32}$$

where $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_j k(\mathbf{x}, \mathbf{x}_i)$ and $g(\mathbf{x}) = \sum_{j=1}^{n} \beta_j k(\mathbf{x}, \mathbf{x}'_j)$, is a valid pre-RKHS [2].

Hence, a profound one-to-one correspondence exists between positive definite functions and RKHSs. Each positive definite function gives rise to an RKHS, and conversely, every RKHS is uniquely generated by a positive definite function.

Kernel trick The reproducing property of the RKHS establishes it as a valuable framework for various learning algorithms. However, in certain situations, working directly within the RKHS can be impractical. Fortunately, the "kernel trick" offers a solution by enabling the embedding of a given space into a larger RKHS using kernels.

[2] $\mathcal{H}_0$ being a pre-RKHS implies that $\mathcal{H}_k$ is the set of functions on $\mathcal{X}$ which are point-wise limits of Cauchy sequences in $\mathcal{H}_0$.

This embedding allows previously daunting operations to be transformed into more manageable tasks within the extended space.

The kernel trick is based on the idea that a symmetric, positive-definite function $k(\mathbf{x}, \mathbf{x}')$ can define an inner product in a real Hilbert space $\mathcal{H}'$ via a map $\phi : \mathcal{X} \to \mathcal{H}'$.

Theorem 2.4.4 (**Kernel trick**). *Let $\mathcal{X}$ be a non-empty space. Then for any symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a real Hilbert space $\mathcal{H}'$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}'}$ and a map $\phi : \mathcal{X} \to \mathcal{H}'$ such that*

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}'}. \tag{2.33}$$

The *feature map* ϕ transforms the original space $\mathcal{X}$ into the larger *feature space* $\mathcal{H}'$, aligning the inner product in $\mathcal{H}'$ with the kernel function k in the original space. This embedding empowers linear learning algorithms to effectively capture nonlinear functions or decision boundaries without explicitly mapping the data to a higher-dimensional space. Instead, these operations can be expressed in terms of evaluations of k, enhancing the efficiency and efficacy of the learning process.

Numerous algorithms have undergone "kernelization", which involves replacing their inner products with reproducing kernels, as extensively discussed in the works of Schölkopf and Smola (2018); Liu et al. (2010); Shawe-Taylor and Cristianini (2004). These kernelized algorithms encompass a wide range of applications, including the kernel perceptron (Aizerman et al., 1964), support vector machines, principal component analysis (Schölkopf et al., 1998), ridge regression (Caponnetto, 2007), and many others.

2.4.2 The Mercer decomposition

In this section, we introduce the Mercer's theorem (Mercer, 1909), a fundamental result with wide applications in machine learning and various data analysis domains. The theorem offers both a necessary and sufficient condition for a positive semi-definite kernel function to be expressible as a series decomposition. Mercer's theorem holds a pivotal role in bridging stochastic processes with kernel methods and serves

as a crucial tool for our subsequent discussions.

Before we delve into the main discussion, we establish some preliminary considerations. We assume that the domain $\mathcal{X}$ is a compact metric space, and the kernel k is a continuous function on $\mathcal{X}$. We define the space of square-integrable functions with respect to a finite Borel measure μ on $\mathcal{X}$ as $L^2(\mathcal{X}, \mu)$[3]. Additionally, we assume that $||k||_{L^2(\mathcal{X},\mu)} < \infty$. A kernel with these properties is known as a *Mercer kernel*. We consider a Hilbert-Schmidt integral operator $T_k : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ defined as :

$$T_k[h](\cdot) = \int_{\mathcal{X}} k(\mathbf{x}, \cdot)h(\mathbf{x})d\mu(\mathbf{x}), \quad \forall h \in L^2(\mathcal{X}, \mu). \tag{2.34}$$

The operator T_k inherits various properties from k. In particular, T_k is a positive, self-adjoint, compact operator. Applying the spectral theorem for self-adjoint, compact operators (Steinwart and Christmann, 2007, Theorem A.5.13) to T_k, we conclude that there exists at most a countable orthonormal set $\{\Phi_i\}_{i\in I}$ of $L^2(\mathcal{X}, \mu)$ and a family $\{\lambda_i\}_{i\in I} \subset \mathbb{R}$ of strictly positive values converging to 0 such that:

$$T_k[h](\cdot) = \sum_{i\in I} \lambda_i \langle \Phi_i, h\rangle_{L^2(\mathcal{X},\mu)} \Phi_i(\cdot), \quad \forall h \in L^2(\mathcal{X}, \mu). \tag{2.35}$$

In fact, $\{\lambda_i\}_{i\in I}$ and $\{e_i\}_{i\in I}$ are respectively the eigenvalues and eigenfunctions of T_k satisfying the eigenvalue problem $T_k[\Phi_i](\cdot) = \lambda_i\Phi_i(\cdot)$, for all $i \in \mathbb{N}$. The orthogonality of the eigenfunctions reduces to

$$\int_{\mathcal{X}} \Phi_i(\mathbf{x})\Phi_j(\mathbf{x})d\mu(\mathbf{x}) = \delta_{i,j}, \quad \forall\, i, j \in \mathbb{N}. \tag{2.36}$$

where $\delta_{i,j}$ defined as the Kronecker delta.

Finally, the Mercer's theorem (Mercer, 1909), states that the kernel k has a representation in terms of $\{\lambda_i\}_{i\in I}$ and $\{\Phi_i\}_{i\in I}$.

Theorem 2.4.5 (**Mercer's Theorem**). *Let k be a continuous positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a compact metric space $\mathcal{X}$, and let μ be a finite Borel measure on*

[3]Strictly here each $f \in L^2(\mathcal{X}, \mu)$ represents a class of functions that are equivalent everywhere with respect to μ.

$\mathcal{X}$ with supp$[v] = \mathcal{X}$. Then k has a representation

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i \in I} \lambda_i \Phi_i(\mathbf{x}) \Phi_i(\mathbf{x}'), \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.37}$$

where the convergence is absolute and uniform.

This version of the Theorem can be found in Steinwart and Christmann (2007, Theorem 4.49) or Cucker and Zhou (2007, Theorem 4.10), with a proof provided by Cucker and Zhou (2007) and Riesz and Nagy (1990). Equations 2.35 and 2.37 are heavily utlized in machine learning. It is worth noting that the theorem requires $\mathcal{X}$ to be compact, which is a restrictive assumption. However, Steinwart and Scovel (2012) have established several Mercer type series representations under weaker assumptions on $\mathcal{X}$.

Furthermore, while the expansion in Equation (2.37) depends on the choice of measure μ, the kernel k in the left-hand side is unique and independent of μ. Therefore, while different choices of μ may lead to different eigenfunctions and eigenvalues, they all yield different representations of the same underlying kernel k.

Mercer representation of a RKHS In the previous section, we established that a feature space and feature map for the kernel k can be given by $\mathcal{H}_0 = \mathcal{H}_k$ and $\phi(\mathbf{x}) = k(\cdot, \mathbf{x})$, respectively, for all $\mathbf{x} \in \mathcal{X}$. However, the Mercer decomposition offers an alternative representation of the feature map for k. Equation (2.37) can be re-expressed as

$$k(\mathbf{x}, \mathbf{x}') = \langle \{\sqrt{\lambda_i}\Phi_i(\mathbf{x})\}_{i \in I}, \{\sqrt{\lambda_i}\Phi_i(\mathbf{x}')\}_{i \in I} \rangle_{\ell^2(\mathbb{R})}, \tag{2.38}$$

where $\ell^2(\mathbb{R})$ is the space of square-summable sequences. This allows us to express k using Equation (2.33) with a feature space $\mathcal{H}' = \ell^2(\mathbb{R})$ and a corresponding explicit feature map $\phi : \mathcal{X} \to \ell^2(\mathbb{R})$ defined as $\phi(\mathbf{x}) = \{\sqrt{\lambda_i}\Phi_i(\mathbf{x})\}_{i \in I}$. This feature map is well-defined since $\sum_{i \in I} |\sqrt{\lambda_i}\Phi_i(\mathbf{x})|^2 = k(\mathbf{x}, \mathbf{x}) < \infty$.

Furthermore, the Mercer representation provides an explicit characterization of

the RKHS $\mathcal{H}_k$ associated with a continuous kernel k on a compact metric space $\mathcal{X}$ (Steinwart and Christmann, 2007, Theorem 4.51). Specifically, we have

$$\mathcal{H}_k := \left\{ h := \sum_{i \in I} a_i \lambda_i^{1/2} \Phi_i : \quad \{a_i\}_{i \in I} \in \ell^2(\mathbb{R}) \right\}. \tag{2.39}$$

The inner product of $\mathcal{H}_k$ is given by $\langle g, h \rangle_{\mathcal{H}_k} = \sum_i \alpha_i \beta_i$ for $g := \sum_i b_i \lambda_i^{1/2} \Phi_i$ and $h := \sum_i b_i \lambda_i^{1/2} \Phi_i$.

2.4.3 Connection between kernels and Gaussian processes

In this section, we establish a fundamental connection between reproducing kernels and covariance functions of stochastic processes, specifically GPs. This connection provides valuable insights into the relationship between kernel methods and probabilistic modeling.

Loève's Theorem, originally presented by Loeve (1978), establishes a precise link between positive definite functions and the covariance functions of second-order stochastic processes [4]:

Theorem 2.4.6 (**Loève 's Theorem**). *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance function of a second-order stochastic process if and only if it is positive definite.*

In other words, any positive definite function can be interpreted as the covariance function of a suitable stochastic process. This theorem underpins the bridge between kernels and stochastic processes. For more details and the formal proof, one can refer to Berlinet and Thomas-Agnan (2004, Theorem 27).

On a compact space $\mathcal{X}$, the connection between the KL expansion of a process f (discussed in Theorem 2.1.4) and the Mercer decomposition of its continuous covariance function k is immediately apparent. The KL-expansion utilizes the eigensystem $(\Phi_i, \lambda_i)_{i \in I}$ of T_k (as defined in Equation (2.34)), which is also the one used in the Mercer representation of k.

When $I = \mathbb{N}$ in Equation (2.34), the convergence in $\mathcal{L}^2(\mathcal{X}, \mathbb{P})$ of the KL expansion

[4] A stochastic process f is said to be second-order if its expected squared value, $E[f(\mathbf{x})^2]$, is finite for all $\mathbf{x}$ in the space $\mathcal{X}$

of f can be deduced from the Mercer decomposition of its covariance function k. Indeed for all $n \in \mathbb{N}$, we can establish that

$$\mathbb{E}\left[\left|f(\mathbf{x}) - \sum_{i}^{n} w_i \Phi_i(\mathbf{x})\right|^2\right] = k(\mathbf{x}, \mathbf{x}) - \sum_{i=1}^{n} \lambda_k \Phi_i(\mathbf{x})^2.$$

This expression tends to 0 as $n \to \infty$ by Mercer's theorem.

KL expansion of Gaussian processes Theorems like 2.1.4 give us an elegant means to express a GP, $f \sim \mathcal{GP}(0, k)$, through an infinite decomposition:

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} w_i \Phi_i(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{2.40}$$

Here, the convergence occurs in the mean square, and the random coefficients w_i follow a Gaussian distribution: $w_i \sim \mathcal{N}(w_i|0, \lambda_i)$. Indeed, the Mercer theorem implies that $\mathrm{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$. The expression in Equation (2.40) is commonly known as the KL expansion of GPs. It provides an alternative representation of GPs and complements the finite distribution-based characterization discussed in Section 2.1.3.

A practical application of this expansion is the construction of finite-dimensional approximations to GPs. By employing i.i.d standard normal random variables $(w_1, \cdots, w_n)$, where $w_i \sim \mathcal{N}(w_i|0, \lambda_i)$, we can create a truncated KL expansion,

$$\sum_{i=1}^{n} w_i \Phi_i(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{2.41}$$

This finite-dimensional Gaussian process approximates the original GP $f \sim \mathcal{GP}(0, k)$.

Multiple view of GP Equation (2.40) provides an alternative perspective on GP regression. We can interpret GP regression as a Bayesian regression problem, featuring an infinite number of basis functions and a Gaussian prior on the weights, as detailed in Rasmussen and Williams (2005, Section 2.1). This approach is sometimes referred to as the "weight" view of GP regression, in contrast to the "function" view described in Section 2.3.1.

The model can be written as:

$$\mathbf{y} = \sum_{i=1}^{\infty} z_i \lambda_i^{1/2} \Phi_i(\mathbf{x}) + \boldsymbol{\varepsilon}, \qquad \forall \mathbf{x} \in \mathcal{X}. \tag{2.42}$$

where the weights are drawn from a Gaussian prior distribution with $z_i \sim \mathcal{N}(z_i|0,1)$ for $i \in \mathbb{N}$ and the basis functions are defined as $\{\lambda_i^{1/2}\Phi_i\}_{i=1}^{\infty}$. Remarkably, this formulation exhibits mathematical equivalence with Gaussian process regression expressed in Section (2.3) via the decomposition in Equation 2.4.3. Fundamentally, this correspondence can be seen as an instance of the *kernel trick*, where the kernel function takes the form $k(\mathbf{x},\mathbf{x}') = \langle\{\sqrt{\lambda_i}\Phi_i(\mathbf{x})\}_{i\in I}, \{\sqrt{\lambda_i}\Phi_i(\mathbf{x}')\}_{i\in I}\rangle \ell^2(\mathbb{R})$. Consequently, GP regression emerges as a kernelized version of Bayesian linear regression.

2.4.4 Examples of kernels

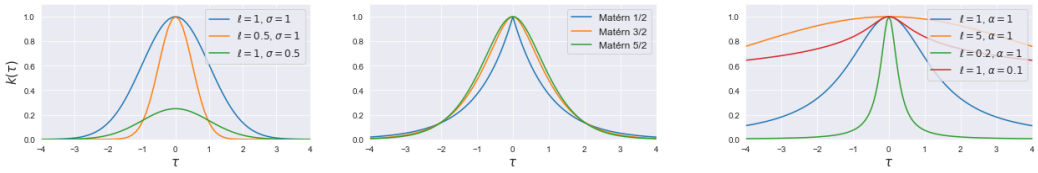
In this section, we present popular choices of valid kernels, that are widely used in the literature. Further examples of kernels can be found in the works of Schölkopf and Smola (2002a); Rasmussen and Williams (2005).

One commonly studied class is Stationary kernels, which maintain their properties when translated. This property is expressed as $k(\mathbf{x},\mathbf{x}') = k(\mathbf{x}-\mathbf{x}')$, ensuring that for any constant shift $\alpha \in \mathbb{R}$, $k(\mathbf{x}+\alpha,\mathbf{x}'+\alpha) = k(\mathbf{x}-\mathbf{x}')$ holds.

A more specific type of kernel, known as isotropic kernels, is defined as $k(\mathbf{x},\mathbf{x}') = k(||\mathbf{x}-\mathbf{x}||_2)$. This implies that the covariance function depends solely on the distance between points, without regard to their relative orientations. Some common examples of isotropic kernels include:

Squared exponential kernel The *squared exponential* kernel (SE), also referred to as the Gaussian kernel or Gaussian radial basis function (RBF) kernel in literature, is a widely used covariance function in Gaussian processes. It is expressed as:

$$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{||(\mathbf{x}-\mathbf{x}') \oslash \boldsymbol{\ell}||_2^2}{2}\right), \quad \forall \mathbf{x},\mathbf{x}' \in \mathcal{X} \tag{2.43}$$



(a) SE kernel (b) Matérn kernel (c) Rational quadratic kernel

Figure 2.3: Distance plot of the main kernel functions. Here, $\tau = ||\mathbf{x} - \mathbf{x}'||$. The Matérn covariance functions utilize $\sigma = 1$ and $\ell = 1$. The Rational Quadratic functions employ $\sigma = 1$ as well. The plot highlights the variations in the distance functions for these different kernel types.

In this equation, σ^2 represents a strictly positive scaling parameter, and ℓ is an arbitrary real-length scale parameter.

The squared exponential kernel generates covariance functions that exhibit infinite differentiability. Consequently, GPs with this kernel produce functions with mean square derivatives of all orders, yielding highly smooth outputs. The length-scale parameter ℓ controls the smoothness of the functions: larger ℓ values result in smoother functions with less rapid changes. Meanwhile, σ dictates the average deviation from the mean.

However, it's essential to recognize that the squared exponential kernel's extreme smoothness is based on strong assumptions, making it potentially unsuitable for all applications, as discussed by Stein in Stein (1999). Nevertheless, despite these limitations, the squared exponential kernel remains a popular choice due to its user-friendly nature and versatility.

Matérn kernel The Matérn kernel, initially proposed in spatial statistics by Matérn Matérn (1960) and further developed by Stein (Stein, 1999, section 2.7), offers a versatile family of kernels. It is defined as follows:

$$k_\nu(\mathbf{x}, \mathbf{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||_2}{\ell} \right)^\nu K_\nu(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||_2}{\ell}), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.44}$$

Here, ν, ℓ, and σ are strictly positive parameters, $\Gamma(\cdot)$ denotes the Gamma function, and $K_\nu(\cdot)$ is a modified Bessel function (Abramowitz and Stegun, 1965, section 9.6).

The Matérn family of kernels finds its origins in the study of spatial forest organization by Matérn and has since become a cornerstone in spatial statistics. Particularly, when ν takes on half-integer values, i.e., $\nu = p + 1/2$ where $p \in \mathbb{N}$, the Matérn kernel simplifies into a product of an exponential function and a polynomial function of order p. The most commonly employed cases in the literature are for $p = 0, 1, 2$:

$$
\begin{aligned}
k_{\frac{1}{2}}(\mathbf{x}, \mathbf{x}') &= \sigma^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||_2}{\ell}\right), \\
k_{\frac{3}{2}}(\mathbf{x}, \mathbf{x}') &= \sigma^2 \exp\left(1 + \frac{\sqrt{3}||\mathbf{x} - \mathbf{x}'||_2}{\ell}\right)\left(-\frac{\sqrt{3}||\mathbf{x} - \mathbf{x}'||_2}{\ell}\right), \\
k_{\frac{5}{2}}(\mathbf{x}, \mathbf{x}') &= \sigma^2 \exp\left(1 + \frac{\sqrt{5}||\mathbf{x} - \mathbf{x}'||_2}{\ell} + \frac{5||\mathbf{x} - \mathbf{x}'||_2}{3\ell^2}\right)\left(-\frac{\sqrt{5}||\mathbf{x} - \mathbf{x}'||_2}{\ell}\right).
\end{aligned}
$$

Specifically, $k_{\frac{1}{2}}$ is known as the Laplace or exponential kernel.

The degree of smoothness induced by the Matérn kernel is determined by the parameter ν. In general, it results in functions that are $[\nu] - 1$ times differentiable. Similarly, a GP with a Matérn kernel with parameter ν is also $[\nu] - 1$ times differentiable. As ν increases, both the function and the Gaussian process become smoother. For instance, when $\nu = 1/2$, the kernel functions and the process exhibit rough behavior, whereas as $\nu \to \infty$, the kernel converges to the infinitely smooth radial basis function (RBF) kernel.

It's important to note that distinguishing between values of ν less than or equal to $7/2$ can be challenging without prior knowledge about the function's differentiability. In practice, the values $\nu = 1/2, 3/2, 5/2$ are often selected to represent different levels of differentiability (Rasmussen and Williams, 2005, section 4).

Rational quadratic kernel The *rational quadratic* kernel, given by :

$$
k(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \frac{||\mathbf{x} - \mathbf{x}'||_2^2}{2\alpha\ell^2}\right)^{-\alpha}, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.45}
$$

where α and ℓ are positive parameters, produces relatively smooth function priors when used in GPs. It can be conceptualized as an infinite sum of squared exponential kernels, each with a different lengthscale. The weighting between these lengthscales

is governed by the parameter α. Notably, as α tends toward infinity, the rational quadratic kernel converges to the exponentiated quadratic kernel (Rasmussen and Williams, 2005).

Non stationnary kernel An example of a non-stationary kernel is the polynomial kernel family, defined as $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + b)^n$, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ where $b > 0$ and $n \leq 1$.

Operation with kernels Complex kernels can be constructed by combining or modifying existing kernels while preserving the positive semi-definiteness property. For instance, if k_1 and k_2 are properly defined real kernels, then the sum $k(\mathbf{x}, \mathbf{x}') = \alpha k_1(\mathbf{x}, \mathbf{x}') + \beta k_2(\mathbf{x}, \mathbf{x}')$ for $\alpha, \beta \geq 0$ is a valid kernel. Similarly, the product $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$ is also a kernel. Additionally, the convolution operation $k(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{X},\mathcal{X}} k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{z}, \mathbf{z}')k_1(\mathbf{z}', \mathbf{x}')d\mathbf{z}d\mathbf{z}'$ is another valid kernel construction method. For further details on kernel construction, refer to Rasmussen and Williams (2005).

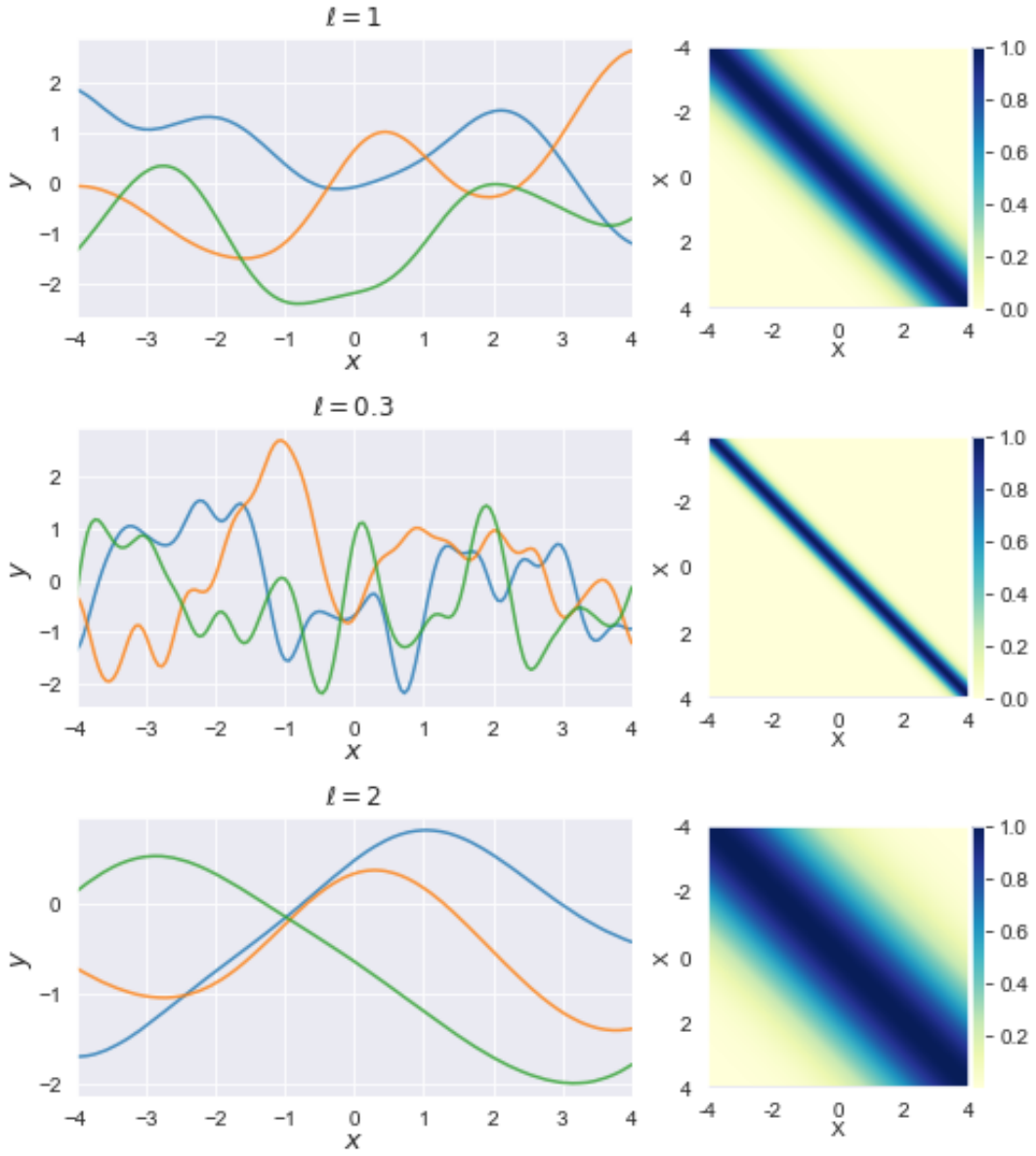


Figure 2.4: Visualization of Gaussian process with Squared Exponential kernel (SE) : On the left-hand side panels, we present three realizations of random functions drawn from Gaussian process priors with SE covariance functions. These functions have a fixed $\sigma = 1$, and different ℓ values. On the right-hand side panels, we showcase the corresponding covariance matrices. This analysis sheds light on the behavior of Squared Exponential kernels under varying ℓ values.

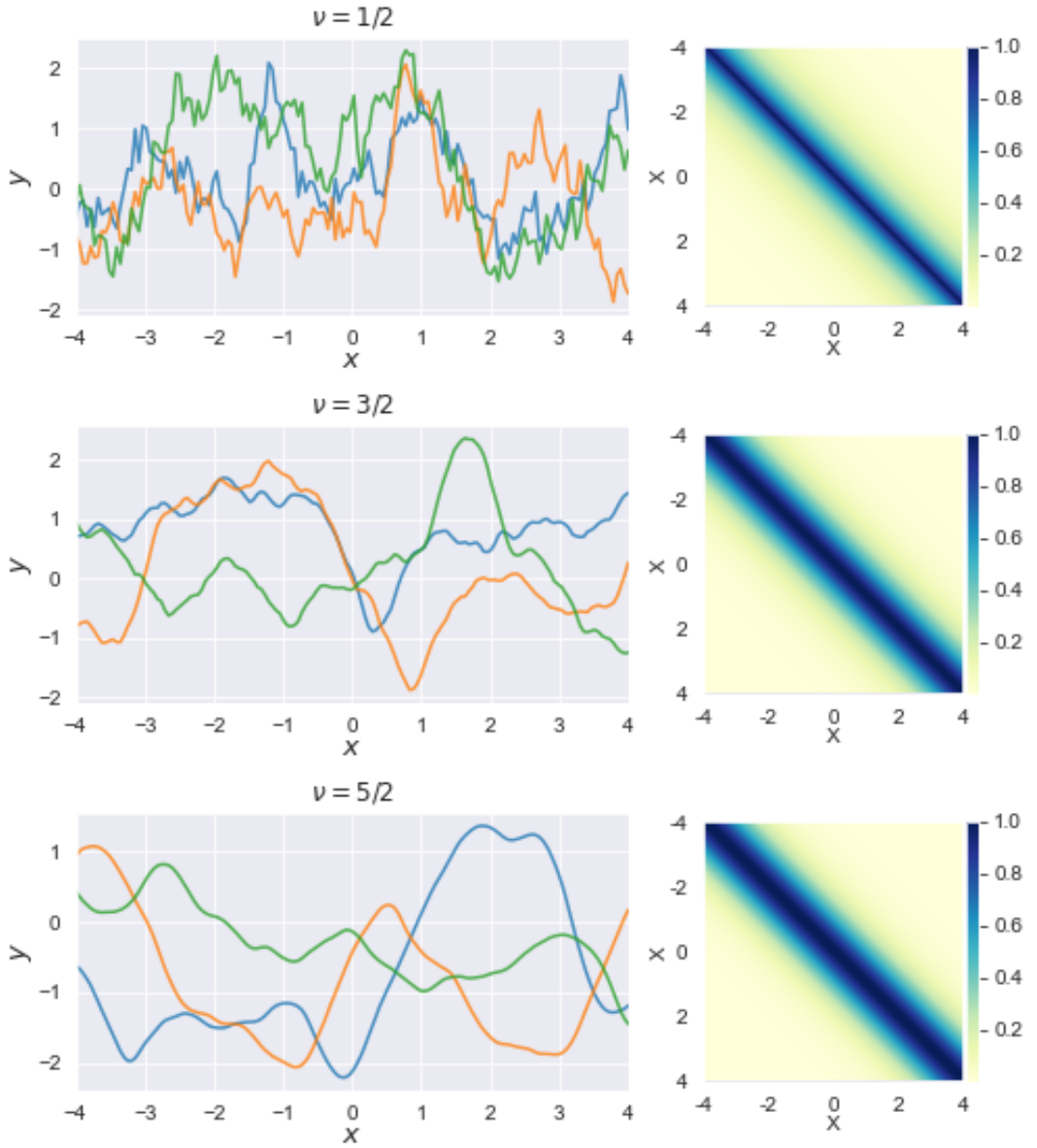


Figure 2.5: Visualization of Gaussian process with Matérn kernel : On the left-hand side panels, we present three realizations of random functions drawn from Gaussian process priors with Matérn covariance functions. These functions have a fixed $\sigma = 1$, fixed $\ell = 1$, and different ν values. On the right-hand side panels, we showcase the corresponding covariance matrices. This analysis sheds light on the behavior of Matérn kernels under varying ν values.

2.5 Summary

In this introductory chapter, we have laid the essential theoretical groundwork for the subsequent chapters. We introduced the fundamental principles of Bayesian nonparametrics, outlined the formalism of Gaussian processes, and emphasized the pivotal role that kernels play in this context. Our primary aim was to establish a comprehensive understanding of the interplay between these key concepts. In the forthcoming chapters, these foundational elements will be extensively employed and relied upon.

Part I

Chapter 3

Background review : Gaussian process modulated spatial Cox processes

Spatial point pattern data constitutes a set of discrete points denoting the precise spatial locations of observed events within a two-dimensional plane. These patterns are pervasive across diverse academic disciplines, including but not limited to ecology, epidemiology, geology, and urban planning. Spatial point processes represent a specialized class of stochastic models designed to analyze and model such point pattern data. These processes inherently generate a finite set of spatial points on a finite space, thereby facilitating the systematic examination of event distribution.

This chapter provides a comprehensive overview of the theoretical foundations that underpin the analysis of point pattern data through the lens of spatial point processes. While our objective does not encompass exhaustive coverage, its principal mission is to furnish the reader with the requisite insights necessary for a comprehensive understanding of the forthcoming contributions.

Section 3.1 introduces the fundamental concepts, definitions, and core characteristics of point processes. In our pursuit of utmost clarity and self-sufficiency, we allocate substantial attention to the comprehensive exposition of these foundational

properties. This includes exploring key statistical metrics such as the intensity function and pair correlation. The intensity function plays a pivotal role in defining the distribution and pseudo-likelihood functions of point processes. In many model implementations, there is a focus on modeling the intensity function itself. Additionally, the pair correlation function enables us to elucidate second-order relationships between point patterns, such as attraction and repulsion.

Section 3.2 marks a shift in our focus towards two pivotal model classes: the Poisson and Cox point processes. The Poisson process assumes a foundational role, while the Cox process represents a generalized and more adaptable approach. Notably, the Poisson process presupposes a constant event rate, whereas the Cox process, a more versatile model, accommodates dynamic event intensities. This versatility makes the Cox process an invaluable tool for addressing complex real-world scenarios, where event rates may vary considerably.

In the last Section 3.3, we will introduce a notable advancement: Cox processes modulated with GPs. In this approach, a GP function acts as a prior for its intensity function. To ensure the non-negativity of the intensity function, a positive transformation is necessary for the GP prior. This paradigm shift lays the groundwork for our forthcoming research. To provide a comprehensive context for our work, we delve into various implementations of Gaussian Cox processes, distinguished by different transformation functions. These include the log Gaussian Cox process Møller et al. (1998), which relies on the exponential transformation, and the sigmoidal Cox process Adams et al. (2009), which leverages a sigmoid transformation. Finally, we introduce the permanental process Shirai and Takahashi (2003); McCullagh and Møller (2006) defining the Poisson process intensity in terms of the square of a GP. We will focus on inference and computational challenges associated with these models. These implementations hold particular significance as they will serve as benchmarks against which we can assess the contributions of our own research in the subsequent chapters.

In this chapter, we exclusively examine cases where the intensity function is solely dependent on the point locations. However, it's important to note that there are proposals beyond the scope of our study that extend this paradigm, which we briefly

mention for informational purposes. First, there are efforts to extend the analysis to spatio-temporal data by incorporating a time dependency into the intensity function. Common approaches include kernel smoothing methods or the log Gaussian Cox process Brix and Diggle (2001); Diggle et al. (2005, 2013). Secondly, researchers may wish to explore the relationship between intensity and covariates to understand factors influencing events. Classical models found in literature include the Cox proportional hazard model Cox (1972); Cygu et al. (2021), which models the log of the intensity function as a linear combination of covariates, as well as various kernel-based intensity estimators Baddeley et al. (2012); Guan (2008). Additionally, the log Gaussian Cox process Yue and Loh (2010) has been examined as a Bayesian alternative.

3.1 Introduction to Point processes

In this section, we delve into the fundamental concepts of point processes and provide formal definitions. Additionally, we explore two critical summary statistics for spatial point processes: the intensity function, which quantifies the expected number of points per unit volume, and the pair correlation, which measures the extent of attraction or repulsion among points. These statistics play a pivotal role in characterizing and distinguishing various classes of point processes throughout the remainder of this chapter. For readers interested in learning more about statistics for spatial point processes, we recommend consulting recent textbooks such as Diggle (2003a), Gelfand et al. (2010), and Moller and Waagepetersen (2003).

3.1.1 Point process definition

Point processes are stochastic models used to describe the random distribution of points in a multi-dimensional space. In this context, both the number of points and their specific locations are treated as random variables. We typically consider two cases: the event case (with dimension $d = 1$), where points correspond to isolated events, and the spatial case (with dimension $d = 2$), where points represent the positions of objects or events within a two-dimensional space.

Consider a bounded metric state space $\mathcal{X} \subset \mathbb{R}^d$, where $d \geq 1$, equiped with an

adequate σ-field $\mathcal{B}$. A realization of a point process on $\mathcal{X}$ is represented as a set of points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, with $N \in \mathbb{N}^+$ and $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \cdots, N$. This set $\mathbf{X}$ is termed "locally finite" if, for any bounded Borel set $B \in \mathcal{B}$, the number of points in $\mathbf{X}$ that fall within B is almost surely finite.

Formally, this condition can be expressed as $|\mathbf{X}_B| < \infty$ for all $B \in \mathcal{B}_0$, where $\mathbf{X}_B := \mathbf{X} \cap B$ represents the restriction of $\mathbf{X}$ to the set B, $\mathcal{B}_0$ is the class of bounded Borel sets in $\mathcal{X}$, and $|\cdot|$ denotes the cardinality of a set. The family of all locally finite point configurations is denoted by N_{lf}. The notation used in this discussion follows Moller and Waagepetersen (2003) and Gelfand et al. (2010).

Now that we have introduced the key concepts above, we can present two equivalent formal definitions of point processes:

Random locally finite point configurations A point process $\mathbf{X}$ can be formally defined as a random locally finite point configurations. In essence, it is represented as a random variable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(N_{\text{lf}}, \mathcal{N}_{\text{lf}})$. Here, $\mathcal{N}_{lf}$ is the smallest σ-algebra generated by all locally finite configurations. A more detailed treatment can be found in Daley and Vere-Jones (2003) or Moller and Waagepetersen (2003, Section B.2).

Random count function Alternatively, a point process can be defined using the count function $N : \mathcal{B}_0 \to \mathbb{N}^+$. In this context, $N(B)$ denotes the count of points from the point process $\mathbf{X}$ located within the set B. Consequently, we can express $N(B)$ as $N(B) := |\mathbf{X}_B|$ for all $B \in \mathcal{B}_0$. The measurability of $\mathbf{X}$ is equivalent to the count $N(B)$ being a random variable for any $B \in \mathcal{B}_0$. In fact, a point process can be regarded as a stochastic process of count variables $N := \{N(B)\}_{B \in \mathcal{B}_0}$.

It's important to note that the two definitions of point processes presented here are equivalent, and the choice of which one to use depends on the specific problem or application at hand. In this document, we will use the notation N or $\mathbf{X}$ interchangeably to refer to the same point process. To simplify the notation, we will also use $\mathbf{X} \subset \mathcal{X}$ instead of $\mathbf{X} \in N_{\text{lf}}$ to denote a point configuration, and $B \subseteq \mathcal{X}$ instead of

$B \in \mathcal{B}$ for a Borel set. This slight abuse of notation does not affect the underlying meaning.

Throughout this chapter, we also adopt a simplifying assumption regarding point processes, considering them to be *simple*. Under this assumption, the occurrence of multiple events at the exact same location is not allowed. Specifically, for all $\mathbf{x} \in \mathcal{X}$, we have $N(\mathbf{x}) \in \{0, 1\}$ with almost sure certainty. This simplicity assumption is applicable to many significant classes of point processes, including Poisson processes and Cox processes, which we will delve into in later sections.

3.1.2 Moment measures and intensity function

Moment measures are vital quantities used to characterize the behavior of point processes. They provide valuable insights into the spatial distribution and clustering patterns of points, aiding our understanding of underlying spatial arrangements. This discussion provides a concise overview of moment measures, with a specific focus on two critical measures: the intensity (first moment) and the pair correlation function (second moment).

Moment measures Let $B_1, \cdots, B_n \subseteq \mathcal{X}$ be not necessarily disjoint Borel sets, where $n \in \mathbb{N}^+$. Consider a point process $\mathbf{X}$ on $\mathcal{X}$ with a count function N. The nth order *moment measure* $M^{(n)}$ of $\mathbf{X}$ is defined as:

$$M^{(n)}(B_1 \times \cdots \times B_n) := \mathbb{E}\left[N(B_1) \cdots N(B_n)\right] \tag{3.1}$$

where $B_1 \times \cdots \times B_n$ represents a product set.

Additionally, the nth order *factorial moment measure* $M_{[n]}$ is defined as:

$$M_{[n]}(B_1 \times \cdots \times B_n) := \mathbb{E}\left[\sum_{\{\mathbf{x}_1, \cdots \mathbf{x}_n\} \in \mathbf{X}}^{\neq} \prod_{i=1}^{n} \mathbf{1}\{\mathbf{x}_i \in B_i\}\right] \tag{3.2}$$

where the sum is taken over n-tuples of pairwise distinct points in $\mathbf{X}$.

Factorial moment measures are of significant importance in point process theory.

They allow us to calculate the expectation of a function summed over a point process, a relationship commonly known as the Campbell theorem (Campbell, 1909). More precisely, for a non-negative measurable function h on the product space $\mathcal{X}^n$, the Campbell theorem states:

$$\mathbb{E}\left[\sum_{\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}\in\mathbf{X}}^{\neq} h(\mathbf{x}_1,\cdots,\mathbf{x}_n)\right] = \int_{\mathcal{X}}\cdots\int_{\mathcal{X}} h(\mathbf{u}_1,\cdots,\mathbf{u}_n)M_{[n]}(d\mathbf{u}_1,\cdots,d\mathbf{u}_n). \quad (3.3)$$

For further details, refer to Moran (1968, pp. 417–423).

If the nth order moment measure $M_{[n]}$ of a point process $\mathbf{X}$ is absolutely continuous with respect to the Lebesgue measure, it admits an n-order product density $\lambda^{(n)} : \mathcal{X}^n \to [0,\infty]$ such that Equation (3.3) can be written as

$$\mathbb{E}\left[\sum_{\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}\in\mathbf{X}}^{\neq} h(\mathbf{x}_1,\cdots,\mathbf{x}_n)\right] = \int_{\mathcal{X}}\cdots\int_{\mathcal{X}} h(\mathbf{u}_1,\cdots,\mathbf{u}_n)\lambda^{(n)}(\mathbf{u}_1,\cdots,\mathbf{u}_n)d\mathbf{u}_1,\cdots,d\mathbf{u}_n. \quad (3.4)$$

Here, $\lambda^{(n)}$ is the n-order *joint intensity function* of $\mathbf{X}$. When $\mathbf{x}_1,\cdots,\mathbf{x}_n$ are distinct, $\lambda^{(n)}(\mathbf{x}_1,\cdots,\mathbf{x}_n)$ can be heuristically interpreted as the probability of observing one point in each of the infinitesimal small volumes $d\mathbf{x}_1,\cdots,d\mathbf{x}_n$. The n-order intensity function implicitly defines a complete probability model for the point process, serving as the fundamental building block for constructing the likelihoods and probability distributions required for point process data analysis.

Intensity function In our analysis, we will primarily focus on the first two orders of the factorial moment measure. For the first order ($n = 1$), the first factorial moment measure can defined as:

$$M_{[1]}(B) = E[N(B)] = \int_B \lambda(\mathbf{u})d\mathbf{u}, \qquad \forall\, B \subseteq \mathcal{X}. \quad (3.5)$$

Here, $\lambda(\cdot) := \lambda^{(1)}(\cdot)$ is known as the intensity function, representing the average number of points per unit area, volume, or higher-dimensional space at a given location in the space. Alternatively, it can be interpreted as the instantaneous probability of point occurrence around a location $\mathbf{x} \in \mathcal{X}$. These concepts are formalized using the

limit:

$$\lambda(\mathbf{x}) = \lim_{|d\mathbf{x}|\to+0} \frac{\mathbb{P}\{N(d\mathbf{x}) \neq 0\}}{|d\mathbf{x}|} = \lim_{|d\mathbf{x}|\to+0} \frac{\mathbb{E}\left[N(d\mathbf{x})\right]}{|d\mathbf{x}|}. \tag{3.6}$$

Please note that this interpretation implicitly assumes that the point process $\mathbf{X}$ is simple.

Pair correlation For the second order ($n = 2$), the second-order moment measure $M_{[2]}$ is expressed using the the Campbell theorem as:

$$M_{[2]}(B_1 \times B_2) = \int_{B_1}\int_{B_2} \lambda^{(2)}(\mathbf{u}_1, \mathbf{u}_2) d\mathbf{u}_1\mathbf{u}_2, \qquad \forall\, B_1, B_2 \subseteq \mathcal{X} \tag{3.7}$$

where $\lambda^{(2)}$ is the second-order product density. Additionally, we can define the second moment measure $M^{(2)}(B_1 \times B_2) := \mathbb{E}[N(B_1)N(B_2)]$ and relate it to $M_{[2]}$ as follows:

$$M^{(2)}(B_1 \times B_2) = M_{[2]}(B_1 \times B_2) + M^{(1)}(B_1 \cap B_2), \quad \forall B_1, B_2 \subseteq \mathcal{X}. \tag{3.8}$$

This enables us to define the covariance between two sets, B_1 and B_2:

$$\begin{aligned}
\operatorname{cov}(N(B_1), N(B_2)) &= \mathbb{E}\left[N(B_1)N(B_2)\right] - \mathbb{E}[N(B_1)]\mathbb{E}[N(B_2)] \\
&= M^{(2)}(B_1 \times B_2) - \int_{B_1}\int_{B_2} \lambda(\mathbf{u}_1)\lambda(\mathbf{u}_2) d\mathbf{u}_1 d\mathbf{u}_2 \\
&= \int_{B_1\cap B_2} \lambda(\mathbf{u}) d\mathbf{u} + \int_{B_1}\int_{B_2} \lambda(\mathbf{u}_1)\lambda(\mathbf{u}_2)(\rho(\mathbf{u}_1, \mathbf{u}_2) - 1) d\mathbf{u}_1 d\mathbf{u}_2.
\end{aligned} \tag{3.9}$$

Here, λ is the intensity function, and $\rho(\mathbf{x}_1, \mathbf{x}_2)$ is the pair correlation function, which satisfies:

$$\rho(\mathbf{x}_1, \mathbf{x}_2) := \frac{\lambda^{(2)}(\mathbf{x}_1, \mathbf{x}_2)}{\lambda(\mathbf{x}_1)\lambda(\mathbf{x}_2)}, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}. \tag{3.10}$$

Equation (3.9) comprises two terms: the first term represents the variance of the number of points in both sets, while the second term accounts for the additional variance resulting from the interaction between the points.

In summury, we presented in this section the first two moments of a point process, from which we derive two key functions: the intensity function and the pair

correlation. These functions play a crucial role in the analysis of spatial point processes. Indeed, spatial point patterns often exhibit *inhomogeneity* and *clustering* (or *aggregation*), which refer to the spatial variation of point density and the stochastic dependence among points, respectively. These phenomena are fundamentally distinct, although they are difficult to disentangle.

Inhomogeneity is typically quantified using the intensity function λ. By studying how λ varies across the study region, we gain valuable insights into the spatial distribution of points and can pinpoint areas of interest with high or low point densities.

On the other hand, clustering is quantified using the correlation function, ρ (Møller and Waagepetersen, 2016; Dvořák et al., 2019). When $\rho = 1$, points are independent, implying no clustering. If $\rho > 1$, it suggests clustering or aggregation i.e. points are attracted to one another. Conversely, $\rho < 1$ indicates inhibition or repulsion, where points actively avoid each other, leading to dispersion. ρ unveils the intricate spatial dynamics within point patterns, crucial for understanding spatial dependencies among points.

3.2 Poisson and Cox processes

In the field of point processes, the Poisson process and Cox process are considered two of the most significant models. Their relevance extends across a multitude of real-world point pattern applications, encompassing seismic activity (Gardner and Knopoff, 1974), epidemiology (Diggle, 2003b; Banerjee et al., 2003), neuroscience (Cunningham et al., 2008b) and crime incident locations (Grubesic and Mack, 2008; Flaxman et al., 2019). In this section, we will delve deeper into the Poisson process and the Cox process, examining their distinct characteristics and applications.

3.2.1 Poisson processes

Poisson processes are a class of simple point processes that play a foundational role in the study of spatial point patterns. They are characterized by "complete spatial randomness", which means that the points are distributed independently and uniformly

throughout the spatial domain.

Definition 3.2.1 (Poisson processes)**.** *Let $\mathcal{X}$ denote a spatial domain, and let Λ be a locally finite measure expressed as $\Lambda(B) = \int_B v(\mathbf{x})d\mathbf{x}$, $B \subseteq \mathcal{X}$, where $v : \mathcal{X} \to \mathbb{R}^+$. A point process $\mathbf{X}$ on $\mathcal{X}$ is a Poisson process if it satisfies the following conditions:*

1. *For any measurable subset $B \subseteq \mathcal{X}$, $N(B)$, follows a Poisson distribution with mean parameter $\Lambda(B)$.*

2. *Given $N(B) = n$, the n points in B are independent and identically distributed with a density function proportional to v.*

The measure Λ introduced in this definition coincides with the *first moment measure* $M^{(1)}$ of the point process $\mathbf{X}$. Specifically, for any bounded subset B of $\mathcal{X}$, we have $E[N(B)] = \Lambda(B)$. Consequently, the function v in this context is the first-order intensity function of $\mathbf{X}$, denoted by λ. Henceforth, we will use λ in place of v. Moreover, the second condition of the definition implies that, given $\{N(B) = n\}$, the n points of $\mathbf{X}_B$ are independently and identically distributed with a density equal to $\lambda(\mathbf{x})/\Lambda(B)$ for all $\mathbf{x} \in B$.

If the intensity function λ remains constant throughout space, the Poisson process $\mathbf{X}$ is referred to as *homogeneous*. In this scenario, points are uniformly distributed within any region of space, and the expected point count is proportional to the region's size. Conversely, if the intensity function exhibits arbitrary variations across space, $\mathbf{X}$ is termed an *inhomogeneous* Poisson process.

The Poisson process exhibits independent scattering, which means that if B_1 and B_2 are disjoint subsets of $\mathcal{X}$, then $\mathbf{X}_{B_1}$ and $\mathbf{X}_{B_2}$ are independent entities (Moller and Waagepetersen, 2003, Proposition 3.2). This property suggest a lack of interaction or complete spatial randomness among points. It is also reflected in the n-order intensity function, which satisfies

$$\lambda^{(n)}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \lambda(\mathbf{x}_1) \ldots \lambda(\mathbf{x}_n) \tag{3.11}$$

for all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and the pair correlation function ρ being equal to one.

Algorithm 1 *Thinning algorithm (W. and Shedler, 1979) for Homogeneous Poisson process simulation* : Simulate data from a Poisson process on region B with random $\lambda(\cdot)$.

input: Region B, Upper-bound λ^*, $\mathbf{X} = \emptyset$
$N \sim \text{Poisson}(\lambda^*|\mathbf{B}|)$
$\{\mathbf{x}\}_{i=1}^{N} \sim \text{Uniform}(\mathbf{B})$
for $i = 1$ **to** N **do**
  $u \sim \text{Uniform}[0, 1]$
  Retain or thin with probability $\lambda(\cdot)/\lambda^*$*:*
  if $u < \lambda(\mathbf{x})/\lambda^*$ **then**
    $\mathbf{X} = \mathbf{X} \cup \{\mathbf{x}\}$
  end if
end for
return $\mathbf{X}$

Thinning and simulation

Definition 3.2.1 offers a straightforward approach to simulating a homogeneous Poisson process: start by simulating $N(B) \sim \text{Poisson}(\Lambda(B))$ for a specified region of interest, denoted as B, and then generate $N(B)$ independent points distributed uniformly within B.

To simulate inhomogeneous processes, a distinct strategy is required. Here, we employ *independent thinning* to derive a new point process, denoted as $\mathbf{X}_{\text{thin}}$. In this process, each point $\mathbf{x}$ in the original $\mathbf{X}$ is included in $\mathbf{X}_{\text{thin}}$ with a probability given by $p_{\text{thin}}(\cdot)$, where $p_{\text{thin}} : \mathcal{X} \to [0, 1]$. Crucially, the decision to include or exclude each point is made independently. If the initial process $\mathbf{X}$ is a Poisson process with intensity λ, then $\mathbf{X}_{\text{thin}}$ also follows a Poisson process, albeit with an intensity of $\lambda \cdot p_{\text{thin}}$ (Moller and Waagepetersen, 2003, Proposition 3.7).

Notably, when dealing with an inhomogeneous Poisson process $\mathbf{X}$ characterized by an intensity function λ bounded by a constant λ^*, it can be represented as an independent thinning of a homogeneous Poisson process with a constant intensity of λ^*. The retention probabilities, denoted as $p_{\text{thin}}(\cdot)$, can be defined as $p_{\text{thin}}(\cdot) = \lambda(\cdot)/\lambda^*$. This insightful perspective leads to the development of the thinning algorithm (W. and Shedler, 1979), which is an efficient method for simulating inhomogeneous Poisson processes. We present the algorithm in detail in Algorithm 1.

Poisson process distribution

The n-th order intensity function, denoted as $\lambda^{(n)}$ and defined in Equation , holds significant utility as it allows us to express the distribution of a Poisson process.

Proposition 3.2.2. *Let $B \subseteq \mathcal{X}$. If $\mathbf{X}$ is a Poisson process on $\mathcal{X}$ with an intensity function $\lambda(\cdot)$ and $\Lambda(B) < \infty$ then,*

$$\mathbb{E}\left[h(\mathbf{X}_B)\right] = \sum_{n=0}^{\infty} \frac{\exp(-\Lambda(B))}{n!} \int_B \cdots \int_B h(\{\mathbf{x}_i\}_{i=1}^n) \prod_{i=1}^n \lambda(\mathbf{x}_i)\, d\mathbf{x}_1 \cdots d\mathbf{x}_n. \tag{3.12}$$

Particularly, for all $F \in \mathcal{N}_{fl}$:

$$\mathbb{P}(\mathbf{X}_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\Lambda(B))}{n!} \int_F \cdots \int_F \prod_{i=1}^n \lambda(\mathbf{x}_i)\, d\mathbf{x}_1 \cdots d\mathbf{x}_n. \tag{3.13}$$

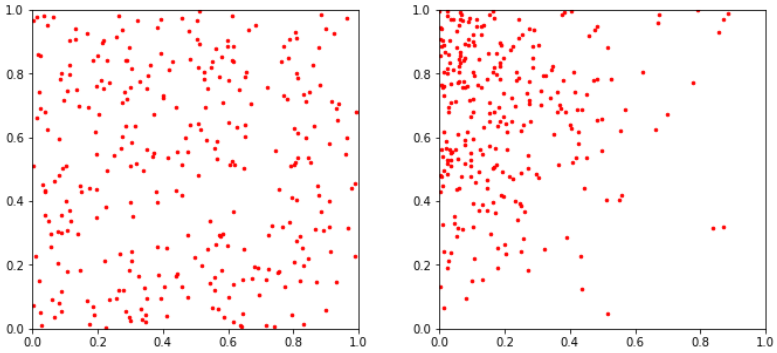
The derivation of Equation (3.12) follows directly from the tower property of conditional expectation and the definition of a Poisson process. Specifically:

$$\begin{aligned}
\mathbb{E}[h(\mathbf{X}_B)] &= \mathbb{E}\left[\mathbb{E}\left[h(\mathbf{X}_B)|N(B) = n\right]\right] \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N(B) = n)\mathbb{E}[h(\mathbf{X}_B)|N(B) = n] \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N(B) = n) \int_B \cdots \int_B h(\{\mathbf{x}_i\}_{i=1}^n) \prod_{i=1}^n \frac{\lambda(\mathbf{x}_i)}{\Lambda(B)} d\mathbf{x}_1 \cdots d\mathbf{x}_n.
\end{aligned}$$

The rest follows from the fact that $N(B) \sim \text{Poisson}(\Lambda(B))$. Equation (3.13) is obtain from Equation (3.12) when $h(\{\mathbf{x}_i\}_{i=1}^n) := \mathbf{1}\{\{\mathbf{x}_i\}_{i=1}^n \in F\}$.

Poisson process density and likelihood

In some instances, it proves valuable to establish the density of a Poison process with respect to another Poisson process. Consider two Poisson processes, $\mathbf{X}_1$ and $\mathbf{X}_2$, each characterized by its corresponding intensity functions, λ_1 and λ_2. If the distribution of $\mathbf{X}_1$ is absolutely continuous with respect to the distribution of $\mathbf{X}_2$ for all $F \in \mathcal{N}_{lf}$, then by the Radon-Nikodym theorem (Billingsley, 1995b, p.422) there



(a) Homogeneous Poisson process (b) Inhomogeneous Poisson process

Figure 3.1: Realization of homogeneous (left) and inhomogeneous (right) spatial Poisson process on $\mathcal{X} = [0, 1]^2$. In both case the expected number of points is 300. For the inhomogeneous Poisson process, $\lambda(x_1, x_2) \propto x_2 \exp(-5x_1)$.

exists a function $p : N_{fl} \to [0, \infty)$ such that

$$\mathbb{P}(\mathbf{X}_1 \in F) = \mathbb{E}\left[\mathbf{1}_{\{\mathbf{X}_2 \in F\}} p(\mathbf{X}_2)\right], \quad \forall F \in \mathcal{N}_{fl} \tag{3.14}$$

This function, p, assumes the role of the density of $\mathbf{X}_1$ with respect to $\mathbf{X}_2$. To maintain consistency with Proposition 3.2.2, it must satisfy:

$$p(\mathbf{X}) = \exp(\Lambda_2(B) - \Lambda_1(B)) \prod_{\mathbf{x} \in \mathbf{X}} \frac{\lambda_1(\mathbf{x})}{\lambda_2(\mathbf{x})}, \quad \text{for all } \mathbf{X} \subseteq B. \tag{3.15}$$

Here, Λ_1 and Λ_2 represent the first moment functions of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Regrettably, Poisson processe distributions do not always exhibit absolute continuity with respect to each other. However, when the space $\mathcal{X}$ is bounded, Poisson processes consistently demonstrate absolute continuity with respect to the unit rate Poisson process, where $\lambda = 1$ (Møller and Waagepetersen, 2004, Proposition 3.8). Consequently, in such cases, a Poisson process with intensity λ possesses a density with respect to the unit rate Poisson process, expressed as:

$$p(\mathbf{X}_B) = \exp(|B| - \Lambda(B)) \prod_{\mathbf{x} \in \mathbf{X}_B} \lambda(\mathbf{x}) \tag{3.16}$$

Then for all $F \in \mathcal{N}_{fl}$,

$$\mathbb{P}(\mathbf{X}_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-|B|)}{n!} \int_F \cdots \int_F p(\{\mathbf{x}_i\}_{i=1}^n) d\mathbf{x}_1 \cdots d\mathbf{x}_n. \tag{3.17}$$

As a consequence, for a finite point configurations $\mathbf{X}_B = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, where $N \in \mathbb{N}^+$, we can define a pseudo-likelihood, represented as the logarithm of its density:

$$\log p(\mathbf{X}_B) = |B| - \int_B \lambda(\mathbf{x}) d\mathbf{x} + \sum_{i=1}^{N} \log \lambda(\mathbf{x}_i). \tag{3.18}$$

The pseudo-likelihood formulation in Equation (3.18) is a widely adopted approach in the literature, especially when the explicit density expression is unknown. In this formulation, we conveniently represent the density in terms of the intensity function. Subsequently, the intensity function can be estimated through established statistical techniques, including maximum likelihood estimation based on Equation (3.18) or Bayesian methods.

Finite point processes constructed from Poisson processes

While Poisson processes may not always be the ideal choice for modeling real-world phenomena, they serve as a valuable foundation for constructing more structured point processes. This concept is particularly applicable to *finite point processes*, which almost certainly possess finite realizations. This characteristic implies that the spatial domain $\mathcal{X}$ is bounded, and the point count $N(\mathcal{X})$ is finite almost surely.

Finite point process distributions are in general characterized by a discrete probability distribution that specifies the probability of having a given number of points within a certain region B i.e. $\{\mathbb{P}(N(B) = n)\}_{n \in \mathbb{N}^+}$ for all $B \subseteq \mathcal{X}$; and a family of joint probability densities $\rho^{(n)}$ for the spatial locations of points, given that there are exactly n points present (Daley and Vere-Jones, 2003, Chapter 5). However, deriving closed-form expressions for these distributions is often a challenging endeavor.

An alternative approach to constructing finite point processes involves explicitly defining their density function, denoted as p, with respect to the unit Poisson process.

This approach parallels the method used to define the Poisson process itself. By utilizing this technique, the distribution of a finite point process can be expressed in a manner similar to what is demonstrated in Equation (3.17), employing the predefined density function p instead. For a more comprehensive understanding, we recommend Daley and Vere-Jones (2003, Chapter 5).

3.2.2 Cox processes

The Poisson process can be extended by considering the intensity as a realization of a stochastic process. This extended framework, initially explored by Cox (1955) and known as the Cox process, provides a richer probabilistic model.

In essence, a Cox process is constructed based on a non-negative stochastic process represented as $z = \{z(\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$. If the conditional distribution of $\mathbf{X}$, given z, conforms to a Poisson process with intensity function z, then $\mathbf{X}$ is referred to as a Cox process driven by z.

Definition 3.2.3 (Cox Process)**.** *Suppose that $z = \{z(\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ is a non-negative random field such that with probability one, $\mathbf{x} \to z(\mathbf{x})$ is a locally integrable function. If $[\mathbf{X}|z]$ is a Poisson process over $\mathcal{X}$ with intensity function λ, then $\mathbf{X}$ is said to be a Cox process driven by z.*

To characterize the Cox process, we condition on z and leverage the properties of the Poisson process. The n-order intensity function of a Cox process $\mathbf{X}$ driven by z is expressed as:

$$\lambda^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \mathbb{E}\left[z(\mathbf{x}_1) \cdots z(\mathbf{x}_n)\right], \qquad \forall\, \mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{X}. \tag{3.19}$$

Here, the expectations are calculated with respect to z. Specifically, the first order density function is $\lambda(\mathbf{x}) = \mathbb{E}\left[z(\mathbf{x})\right]$ for all $\mathbf{x} \in \mathcal{X}$, and the second order intensity function is expressed as $\lambda^{(2)}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E}\left[z(\mathbf{x}_1)z(\mathbf{x}_2)\right]$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$.

Advantages Cox processes offer several advantages over deterministic approaches, including:

1. *Bayesian treatment of Poisson processes:* Cox processes enable naturally a Bayesian treatment of Poisson processes, allowing the incorporation of prior knowledge about the intensity function. By encoding prior beliefs regarding the functional form of the intensity function through a prior distribution, we can estimate a posterior distribution and quantify uncertainty.

2. *Higher-order spatial dependence properties*: While Poisson processes model first-order inhomogeneity, they fail to model second or higher-order spatial dependence properties. Cox processes provide a more flexible class of models capable of accommodating the aggregation or inhibition of events (i.e., when $\rho \neq 1$). This makes them well-suited for capturing complex spatial patterns. This flexibility is evident in the Cox process pair correlation function:

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbb{E}\left[z(\mathbf{x}_1)z(\mathbf{x}_2)\right]}{\mathbb{E}\left[z(\mathbf{x}_1)\right]\mathbb{E}\left[z(\mathbf{x}_2)\right]}, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}. \tag{3.20}$$

   In general, ρ does not equal 1, indicating the presence of higher-order spatial dependence properties.

3. *Flexibility:* Compared to other structured approaches, Cox processes remain flexible. One can exploit the properties of the Poisson process by conditioning, allowing for the modeling of complex spatial patterns that may not be captured by other approaches.

Pseudo-likelihood When the space is restricted to a bounded set $B \subseteq \mathcal{X}$, the Poisson process $[\mathbf{X}|z]$ can be described by a density with respect to the unit rate Poisson process, as discussed in Section 3.2.1. For a finite point configuration $\mathbf{X}_B$, we can derive a density similar to Equation (3.18):

$$p(\mathbf{X}_B) = \mathbb{E}\left[\exp\left(|B| - \int_B z(\mathbf{u})d\mathbf{u}\right)\prod_{\mathbf{x}\in\mathbf{X}_B} z(\mathbf{x})\right]. \tag{3.21}$$

Here, the expectation is taken with respect to z. This expression can be used to estimate the unknown intensity function i.e. the distribution of z, from data using

maximum likelihood or Bayesian methods. However, obtaining an explicit expression for Equation (3.21) is often infeasible, and the integral $\int_{\mathcal{X}} z(\mathbf{u})d\mathbf{u}$ can be challenging to compute.

Throughout the remainder of this document, we will use the notation $\lambda(\mathbf{x})$ interchangeably to refer to either the random field driving the Cox process i.e., $z(\mathbf{x})$, or the deterministic first-order intensity function $\lambda^{(1)}(\mathbf{x}) = \mathbb{E}[z(\mathbf{x})]$. Depending on the context of the discussion, it should be clear whether λ is stochastic or known.

In summary, this section provided a brief introduction to Poisson processes and Cox processes, emphasizing the derivation of their distributions, densities, and pseudo-likelihoods using an intensity function. In the upcoming section, we will delve into a specific category of Cox processes that regulate the stochasticity of the intensity function by incorporating a GP prior.

3.3 Gaussian Cox processes

A flexible extension of the Poisson process involves the incorporation of a stochastic Poisson intensity through the utilization of a nonparametric Gaussian process prior. This results in the popular *Gaussian Cox process model*, which has demonstrated remarkable efficacy across a wide range of applications.

Gaussian Cox process models leverage the capabilities of GPs, offering a natural and comprehensive framework for Bayesian inference on intensity functions. It distinguishes itself from alternative methodologies by combining the advantages of nonparametric modeling and the intrinsic capacity to quantify the uncertainty in intensity estimation. In this section, we will delve deeper into the concept of Cox processes modulated by GPs, providing more detailed discussions and insights.

Double intractability Inference with the Gaussian Cox process model is challenging due to the *doubly-intractable* likelihood, requiring integration of an infinite-dimensional random function over the input domain.

To illustrate this, consider a Gaussian Cox process, denoted as $\mathbf{X}$, driven by a

positive random field, z, defined as $z(\mathbf{x}) = g \circ f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Here, g represents a non-negative transformation function $g : \mathcal{X} \mapsto \mathbb{R}^+$, and f is a latent function. We then introduce a GP prior on f, denoted as $f \sim \mathcal{GP}(0, k)$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines the kernel function.

Furthermore, we assume the existence of a finite point configuration $\mathbf{X}_B = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, with $N \in \mathbb{N}^+$, observed within a finite region $B \subset \mathcal{X}$. Conditioned on the latent function f, the likelihood function is expressed as:

$$p(\mathbf{X}_B|f) = \exp\left(-\int_B g(f(\mathbf{x}))\, d\mathbf{x}\right) \prod_{i=1}^{N} g(f(\mathbf{x}_i)). \tag{3.22}$$

where f represents the infinite-dimensional object. Additionally, the latent posterior $p(f|\mathbf{X}_B)$ can be represented as:

$$p(f|\mathbf{X}_B) = \frac{\exp\big(-\int_B g(f(\mathbf{x}))\, d\mathbf{x}\big)\big[\prod_{i=1}^{N} g(f(\mathbf{x}_i))\big]p(f)}{\int \exp\big(-\int_B g(f(\mathbf{x}))\, d\mathbf{x}\big)\big[\prod_{i=1}^{N} g(f(\mathbf{x}_i))\big]p(f)\, df}. \tag{3.23}$$

Notably, the likelihood expression in Equation (3.22) requires evaluating an integral of f over the full domain B, making it "doubly-intractable".

Positive transformation choices In addition, a positive transformation must be applied to the GP prior to ensure the intensity function remains non-negative. Different choices of transformation function are found in the literature. A first classical approach relies on the exponential transformation, resulting in the *log Gaussian Cox process* proposed by Møller et al. (1998). This model usually requires numerical approximation of the integral by discretization over the input space (Møller et al., 1998; Diggle et al., 2013), a computationally-intensive procedure which scales poorly with the dimensionality of the input domain. A second approach uses a sigmoid transformation and an input space augmentation via *thinning*, to construct an Markov Chain Monte Carlo sampler (Adams et al., 2009; Gunter et al., 2014), eliminating the need for likelihood integration. In practice however, this approach is computationally intractable for large problems. Other works include the use of the *Relu* (Ko and Seeger, 2016) and softplus (Seeger and Bouchard, 2012; Park et al., 2014) as transformation

functions. Finally, Lopez-lopera et al. (2019) introduce a finite approximation where positiveness conditions is imposed directly on the GP.

Another approach exploits the so-called *permanental process*, defining the Poisson process intensity in terms of the square of a GP (McCullagh and Møller, 2006; Lloyd et al., 2015). It enables analytical computation of the intensity integral when coupled with a variational inference scheme with inducing points (similar to Titsias 2009b) and has received considerable recent attention (Lian et al., 2015; Flaxman et al., 2017; John and Hensman, 2018).

In the following sections, we provide deeper descriptions of these diverse approaches and the corresponding inference methods that have been proposed in the literature.

3.3.1 Log-Gaussian Cox processes

The *log-Gaussian Cox process* (LGCP), introduced by Møller et al. (1998), is an instance of a Gaussian Cox process where the intensity is modeled through an exponential transformation. Specifically, it defines the intensity as

$$\lambda(\mathbf{x}) := \exp(f(\mathbf{x})) \quad \text{with } f \sim \mathcal{GP}(0, k), \tag{3.24}$$

for all locations $\mathbf{x} \in \mathcal{X}$.

Computational grid To address the intractability of the integral, a common approach is to discretize the region of consideration $B \subset \mathcal{X}$ into a regular lattice of m non-overlapping regions $\{A_i\}_{i=1}^m$, each having centroids $\{\mathbf{c}_i\}_{i=1}^m$. The LGCP model is then approximated with a piecewise constant intensity function on this lattice, i.e., $\lambda(\mathbf{x}) \approx \exp(f_i)$ for all $\mathbf{x} \in A_i$, where $f_i := f(\mathbf{c}_i)$ for $i = 1, \cdots, m$. Given $\{f_i\}_{i=1}^m$, the number of points in each region A_i follows a Poisson distribution with a mean of $\int_{A_i} \lambda(\mathbf{x}) d\lambda(\mathbf{x}) = |A_i| \exp(f_i)$. In particular, assuming a point pattern $\mathbf{X}_B = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, the conditional log-likelihood of the model can be computed as

follows:

$$\log p(\mathbf{X}_B|f) \approx -\sum_{i=1}^{m} |A_i| \exp(f_i) + n_i f_i + C \tag{3.25}$$

where n_i denotes the number of points observed in A_i and C is a constant.

Inference Inference for LGCP models is challenging since the posterior, induced by Equation (3.25), is not tractable. Two commonly used methods for addressing this challenge are sampling with Markov Chain Monte Carlo (MCMC) techniques and employing Integrated Nested Laplace Approximation (INLA).

While this chapter doesn't delve into detailed explanations of MCMC methods, those interested can find comprehensive overviews in Gilks et al. (1995) and Gamerman and Lopes (2006). Several MCMC sampling techniques have been introduced for estimating the latent posterior distribution within the LGCP framework. These include the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996; Møller et al., 1998; Brix and Diggle, 2001; Diggle et al., 2005), its Riemann manifold variant (mMALA) (Girolami and Calderhead, 2011; Diggle et al., 2013), and elliptic sampling (Murray et al., 2010; Leininger and Gelfand, 2017).

On the other hand, Integrated Nested Laplace Approximation (INLA), introduced by Rue et al. (2009) and H. and L. (2005), offers a different approach. INLA is a statistical method that provides fast and accurate approximate inference for Bayesian models with latent Gaussian variables. It combines numerical integration and Laplace approximation to estimate posterior distributions for model parameters and latent variables. INLA has been proposed as an alternative approach for LGCP models (Illian et al., 2012) where the latent GP is approximated by a Gaussian Markov random field (GMRF) with a sparse precision matrix on a fine lattice. A study conducted by Taylor and Diggle (2014) compared the performance of MCMC schemes and INLA, particularly for spatial LGCP models, taking into account scenarios where hyperparameters are treated as known values.

The choice of the grid is critical to strike a balance between computational complexity and approximation accuracy. As the cell sizes approach zero, the method

converges to the true solution (Waageptersen, 2004, corollary A.1). However, handling the potentially dense covariance matrix over $\{\mathbf{c}_i\}_{i=1}^m \cup \mathbf{X}$ becomes challenging, often limiting the lattice to only a few points. This limitation can impact the quality of the likelihood approximation and, consequently, the accuracy of the inference.

3.3.2 Sigmoidal Cox processes

The *sigmoidal Cox Poisson process* (SGCP), a novel class of Gaussian Cox processes introduced by Adams et al. (2009), defines the intensity function as a sigmoidal transformation of a latent function f governed by a Gaussian Process (GP) prior. In explicit terms, the intensity function is mathematically represented as:

$$\lambda(\mathbf{x}) := \lambda^* \cdot \sigma(f(\mathbf{x})) \quad \text{with } f \sim \mathcal{GP}(0, k), \tag{3.26}$$

for all locations $\mathbf{x} \in \mathcal{X}$. In this equation, $\sigma(\cdot)$ represents the *sigmoid function*, defined as:

$$\sigma(\mathbf{x}) := \frac{1}{1 + \exp^{-\mathbf{x}}}.$$

The inclusion of the positive constant $\lambda^* > 0$ in Equation (3.26) plays a critical role, acting as an upper bound for the intensity function λ. This is essential because the sigmoid function inherently constrains values to the range between 0 and 1.

The key advantage of this sigmoidal specification lies in its capacity to generate asymptotically exact samples from the intensity function posterior without requiring any approximations. Instead of conducting MCMC inference directly on the posterior, the approach involves working with an augmented posterior that eliminates the need for integration.

Modified thinning To obtain a sample from an SGCP model, we employ a modified version of the *thinning algorithm*, as outlined in Algorithm 1. This adaptation takes advantage of the fundamental fact that $\lambda(\cdot)$ is bounded by λ^*. The algorithm's workflow begins with the sampling of $m + n$ points, denoted as $\hat{\mathbf{X}}_{m+n} = \{\hat{\mathbf{x}}_i\}_{i=1}^{m+n}$, drawn from a Poisson distribution characterized by a constant intensity of λ^*. Sub-

sequently, we simulate the function f using the GP prior, with input points derived from $\hat{\mathbf{X}}_{m+n}$. This step provides an evaluation of the intensity at these sampled points. Finally, we can apply a thinning procedure, partitioning $\hat{\mathbf{X}}_{m+n}$ into two distinct subsets: one containing the retained (observed) points, with probability $\sigma(f(\cdot))$, denoted as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, and rejected latent points $\mathbf{X}_m = \{\mathbf{x}_i^{(m)}\}_{i=1}^m$. As outlined by the *independent thinning property* described in Section 3.2.1, this process precisely yields an asymptotically exact sample from $\lambda(\cdot)$.

Exact MCMC via augmentation Inspired by the simulation algorithm, Adams et al. (2009) proposed to augment the variable set to include the latent points $\mathbf{X}_m$. This augmentation results in a new Bayesian hierarchical model formulation. For a finite point configuration $\mathbf{X}_B := \{\mathbf{x}_i\}_{i=1}^N$, the joint posterior of this augmented model can be expressed as follows:

$$\begin{aligned}\log\ p(\mathbf{X}_B, \mathbf{X}_m, m, \mathbf{f}_{N+m}|\lambda^*) &= -\lambda^*|B| + (N+m)\log(\lambda^*) \\ &+ \sum_{i=1}^{N} \log(\sigma(f(\mathbf{x}_i)) + \sum_{i=1}^{m} \log(-\sigma(f(\mathbf{x}_i^{(m)}))) + \log(p(\mathbf{f}_{N+m}))\end{aligned} \tag{3.27}$$

Here, $p(\mathbf{f}_{N+m})$ denotes the joint prior distribution of the latent function f at the input points $\hat{\mathbf{X}}_{N+m} := \mathbf{X}_B \cup \mathbf{X}_m$. We also used that, conditioned on the knowledge of $\hat{\mathbf{X}}_{N+m}$, for all $\mathbf{x} \in \hat{\mathbf{X}}_{N+m}$, the probability of a point being observed is determined by $\sigma(f(\mathbf{x}))$, whereas the probability of the point being latent is described by $1-\sigma(f(\mathbf{x}))$.

To sample from the posterior distribution of interest, $p(\mathbf{f}_N|\mathbf{X}_B, \lambda^*)$, we can employ a MCMC approach on the joint posterior specified in Equation (3.27). This MCMC procedure involves three transitions, addressing the number of thinned events m, their respective locations $\mathbf{X}_m$, and the latent function $\mathbf{f}_{N+m}$. After these transitions, we discard the latent points, enabling us to conduct MCMC inference on the model without relying on numerical approximations.

In their investigation, Adams et al. (2009) demonstrated that the SGCP, with its garantee of producing asymptotically exact sampling, outperformed the LGCP with discretization (using 10, 25, and 100 bins) for relatively small datasets containing

approximately 200 points. However, they also highlighted a limitation of the SGCP method: its impracticality for datasets comprising several thousand points or more. This limitation arises from poor scalability concerning both the dimension of the domain and the size of the data. The inclusion of thinned events within the GP leads to a computational cost of $O((N + m)^3)$ in each MCMC step, and the number of thinned events, denoted as m, tends to grow exponentially with the dimension of the space.

Nevertheless, researchers have extended the SGC model to enhance its performance and versatility. For example, Gunter et al. (2014) proposed an adaptive thinning method that replaces the global upper bound with a more efficient piece-wise function, reducing the need for thinned points. Another notable extension involves incorporating a latent marked Poisson process and a Polya–Gamma random variable, as suggested by Donner and Opper (2018a), to establish a likelihood representation. To approximate the posterior distribution in this extended framework, the authors utilize variational inference and sparse Laplace methods, significantly reducing the computational complexity of the model.

3.3.3 Permanental processes

The *permanental process* is a more flexible class of Gaussian Cox processes, which provides an exception to the general rule that analytical expressions for likelihoods are not available. This process is obtained by defining the intensity as a square of Gaussian processes, i.e.,

$$\lambda(\mathbf{x}) := |f(\mathbf{x})|^2 \quad \text{with } f \sim \mathcal{GP}(0, k). \tag{3.28}$$

Permanental processes were first introduced by Shirai and Takahashi (2003), and further studied by McCullagh and Møller (2006). They serve as a natural counterpart to the more extensively investigated determinantal point processes, which have been the focus of substantial research in both mathematical and physics literature. While there are certain resemblances, permanental processes possess distinct and unique

properties. McCullagh and Møller (2006) present a comprehensive examination of permanental processes, offering detailed insights into their density and moment properties.

When considering the intensity prior on a GP having a mean 0, squaring has the effect of redistributing more probability mass around the central location of the GP prior and towards extreme events. These effects arise from the inherent amplification process in squaring, which accentuates values closer to zero and magnifies extreme values. Furthermore, the transformation of negative values into positive ones reallocates more mass to their corresponding positive counterparts.

The purpose of this section is to establish a foundational understanding of permanental processes and their tractability properties, which are essential for comprehending our proposed Bayesian inference scheme on permanental processes in the following chapter. Firstly, we introduce the concept of permanental processes and discuss their key features. Following that, we provide an overview of existing inference methodologies from the literature, which will serve as benchmarks for evaluating the effectiveness of our contributions.

A permanent point process, denoted as $\mathbf{X}$, possesses a closed-form, non-negative n-th order intensity function defined as follows:

$$\begin{aligned}\lambda^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n) :&= \mathbb{E}\left[f(\mathbf{x}_1)^2 \cdots f(\mathbf{x}_n)^2\right] \\ &= \operatorname{per}\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]_{1 \le i,j \le n}\end{aligned} \tag{3.29}$$

where per$[\cdot]$ denotes here the $n/2$-weighted permanent of its matrix argument, defined as:

$$\operatorname{per}\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]_{1 \le i,j \le n} := \sum_{\pi} \left(\frac{n}{2}\right)^{\#\pi} \prod_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_{\pi(i)}). \tag{3.30}$$

In the above equations, $\pi(\cdot)$ represents a permutation of $1, \ldots, n$, and the summation is performed over all permutations, with $\#\pi$ denoting the number of cycles in π. The validity of this result is established in McCullagh and Møller (2006, Theorem 1). The term "permanental" is derived from the use of the permanent function in this

representation. The construction of the permanental process bears a resemblance to that of the determinantal process, with a notable distinction being that Equation (3.29) employs the permanental function instead of the determinantal function.

Integral expression The permanental process offers distinct tractability properties that make it particularly appealing. It leverages the Mercer decomposition of the kernel function k (as detailed in Theorem 2.4.5) and the Karhunen-Loève decomposition of f (as described in Theorem 2.1.4), which together allow for a closed-form expression of the intensity integral.

Indeed, from the Mercer's theorem (2.4.5), when the space $\mathcal{X}$ is compact and k is continuous and positive, it can be represented as:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \Phi_i(\mathbf{x}) \Phi_i(\mathbf{x}'), \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{3.31}$$

In this context, $\{\Phi_i\}_{i=1}^{\infty}$ denotes the orthogonal eigenfunctions in $L^2(\mathcal{X}, \mu)$ and $\{\lambda_i\}_{i=1}^{\infty}$ their corresponding eigenvalues. Then, similar to the formulation in (2.40), the GP $f(\mathbf{x}) \sim \mathcal{GP}(0, k)$ has a KL infinite decomposition of the form

$$f(\mathbf{x}) = \sum_{i=0}^{\infty} w_i \Phi_i(\mathbf{x}). \tag{3.32}$$

where $\mathbf{w} = (w_1, w_2, \ldots) \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Lambda})$ and $\mathbf{\Lambda}$ is a diagonal covariance matrix with i-entries λ_i for $i = 1, 2, \ldots$. Consequently, the integral of the intensity can be expressed as:

$$\begin{aligned} \int_{\mathcal{X}} f(\mathbf{x})^2 \, d\mu(\mathbf{x}) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \int_{\mathcal{X}} w_i w_j \Phi_i(\mathbf{x}) \Phi_j(\mathbf{x}) d\mu(\mathbf{x}) \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} w_i w_j \langle \Phi_i, \Phi_j \rangle_{L^2(\mathcal{X}, \mu)} \\ &= \sum_{i=0}^{\infty} w_i^2 < \infty \quad \text{a.s.} \end{aligned} \tag{3.33}$$

It's important to note that this integral expression holds for a generic compact space $\mathcal{X}$ and a measure μ, capitalizing on the orthogonality of $\{\Phi_i\}_{i=1}^{\infty}$ within $L^2(\mathcal{X}, \mu)$.

Likelihood expression We consider an observed finite point configuration within a finite subset $B \subset \mathcal{X}$, represented as $\mathbf{X}_B = \{\mathbf{x}_i\}_{i=1}^N$. Assuming a kernel for f with an explicit Mercer representation with respect to B and μ, defined as the Lebesgue measure, we can derive a closed-form density for the conditional likelihood $p(\mathbf{X}_B)$ of a permanental process. Starting from the general density expression in Equation (3.21), and incorporating the permanental specification into it, we obtain:

$$\begin{aligned} p(\mathbf{X}_B) &= \mathbb{E}\left[\exp\left(-\sum_{i=0}^{\infty} w_i^2\right)\left(\prod_{i=1}^{N}\left(\sum_{j=0}^{\infty} w_j \Phi_j(\mathbf{x}_i)\right)^2\right)\right] \\ &= \exp^{|B|-\frac{1}{2}D} \operatorname{per}\left[\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)\right]_{1\le i,j\le N} \end{aligned} \tag{3.34}$$

In this equation, $\{\Phi_i\}_{i=1}^{\infty}$ are assumed orthogonal in $L^2(B)$. Additionally, $\tilde{k}$ denotes the modified kernel:

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \frac{\lambda_i}{(1+\lambda_i)} \Phi_i(\mathbf{x})\Phi_i(\mathbf{x}'), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{3.35}$$

Here the series D represents a convergent sum, defined as $D = \sum_{i=1}^{\infty} \log(1+\lambda_i)$ (McCullagh and Møller, 2006, section 2.3). A detailed proof is available in (McCullagh and Møller, 2006, Theorem 2).

However, two significant limitations become apparent:

1. The likelihood expression in Equation (3.34) necessitates a Mercer representation for k in $L^2(B)$, which, for most kernel choices, is not explicitly available. This issue will be explored in more detail in the upcoming sections.

2. Computing the permanent terms per$[\tilde{k}]$ is typically infeasible, even for relatively small matrices. In fact, it is a well-established result that there is no available deterministic polynomial-time method for the exact computation of permanents for general matrices (Valiant, 2006).

Despite these challenges, numerous approximate inference strategies have emerged in the literature to tackle the computational complexities associated with the permanental process while maintaining its tractability. In this context, we highlight two

recent and promising inference methods from the literature: a variational approach known as VBPP and a Laplace-based method that relies in the Mercer representation of the kernel, referred to as LBPP.

Inference with variational approximation (VBPP)

Variational inference (VI) methods, as introduced by Jordan et al. (1999), have emerged as a pivotal tool in Bayesian inference, offering a powerful means of approximating otherwise intractable probability distributions. VI achieves this by approximating an untractable distribution through a carefully selected "variational distribution" chosen from a family of tractable distributions. This selection process hinges on optimizing the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the target distribution and the variational distribution, resulting in significantly faster computations compared to traditional methods like MCMC sampling. Consequently, VI has found widespread utility across diverse applications, including Bayesian regression, matrix factorization, and deep learning. For a more comprehensive introduction to VI, readers can delve into the recent survey by Blei et al. (2017b).

Lloyd et al. (2015) introduced variational methods for permanental process which they term *Variational Bayes for Point Processes*(VBPP). This approach has garnered substantial attention in the machine learning literature Lian et al. (2015); Flaxman et al. (2017); John and Hensman (2018). VBPP extends a more general VI framework initially developed for sparse GPs with inducing points by Titsias (2009a). In the following section, we provide a short overview of VI within the context of the Permanental process. For a more detailed exposition of variational sparse GP with inducing points, readers are directed to the work of Titsias (2009a).

For the permanental process, we assume an observed point configuration in B denoted as $\mathbf{X}_B = \{\mathbf{x}_i\}_{i=1}^N$, $N \in \mathbb{N}^+$. To approximate the Gaussian process with inducing points, we adopt the sparse approximation proposed by Titsias (2009a). Specifically, we introduce a set of inducing points $\mathbf{X}_m = \{\mathbf{x}_i^{(m)}\}_{i=1}^m$ in B and the corresponding inducing variables $\mathbf{f}_m := f(\mathbf{X}_m)$. We assume a variational distribution

at the inducing variables, denoted as $q(\mathbf{f}_m)$, that follows a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{S}$.

Following the developments in Section 2.3.1, we proceed to define a variational GP as $q(f) = \int_B p(f|\mathbf{f}_m)q(\mathbf{f}_m)d\mathbf{f}_\mathbf{m}$. This definition implies $q(f) \sim \mathcal{GP}(\bar{m}, \bar{k})$ with mean function $\bar{m}$ and kernel $\bar{k}$ derived using the Gaussian distribution's conditioning rule and expressed as below:

$$\begin{aligned} \bar{m}(\mathbf{x}) &= k(\mathbf{x}, \mathbf{X}_m)\mathbf{K}_{m,m}^{-1}\mathbf{m}, \quad \forall \mathbf{x} \in B \\ \bar{k}(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X}_m)\mathbf{K}_{m,m}^{-1}k(\mathbf{X}_m, \mathbf{x}) \\ &\qquad + k(\mathbf{x}, \mathbf{X}_m)\mathbf{K}_{m,m}^{-1}\mathbf{S}\mathbf{K}_{m,m}^{-1}k(\mathbf{X}_m, \mathbf{x}), \quad \forall\, \mathbf{x}, \mathbf{x}' \in B. \end{aligned} \tag{3.36}$$

Variational bound The objective of the VI approach is to minimize the KL divergence between the posterior GP $p(f|\mathbf{X}_B)$ and the variational process $q(f)$, from the following:

$$\begin{aligned} \mathcal{KL}(q(f)||p(f|\mathbf{X}_B)) &:= -\mathbb{E}_{q(f)}\left[\frac{\log p(f|\mathbf{X}_B)}{q(f)}\right] \\ &= \log(p(\mathbf{X}_B)) - \left(\mathbb{E}_{q(f)}\left[\log p(\mathbf{X}_B|f)\right] - \mathcal{KL}(q(f)||p(f))\right). \end{aligned}$$

Thus, minimizing $\mathcal{KL}(q(f)||p(f|\mathbf{X}_B))$ with respect to $q(\cdot)$ is equivalent to maximizing the evidence lower bound $\mathcal{L}_{\text{ELBO}}$, given by

$$\mathcal{L}_{\text{ELBO}} := \mathbb{E}_{q(f)}\left[\log(p(\mathbf{X}_B|f))\right] - \mathcal{KL}(q(f)||p(f)).$$

In a Permanental process context, from the expression of $p(\mathbf{X}_B|f)$ given in Equation (3.22), the ELBO can be rewritten as:

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q(f)}\left[\int_B f(\mathbf{x})^2 d\mathbf{x}\right] + \sum_{i=1}^{N} \mathbb{E}_{q(f)}\left[\log(f(\mathbf{x}_i)\right] - \mathcal{KL}(q(f)||p(f)) \tag{3.37}$$

In this equation, the first terms represents the likelihood of the observed data under the variational distribution $q(f)$, while the last $\mathcal{KL}$ divergence term serves as a complexity penalty that encourages $q(f)$ to be close to the prior Gaussian process

$p(f)$.

As suggested by Matthews (2016), the $\mathcal{KL}$ divergence term in Equation (3.37) simplifies into a tractable KL divergence between two Gaussian distributions at the inducing points $\mathbf{X}_m$. Additionally, the sum-of-expectation term $\sum_{i=1}^{N} \mathbb{E}_{q(f)} [\log(f(\mathbf{x}_i))]$ can be evaluated analytically when $q(f) \sim \mathcal{GP}(\bar{m}, \bar{k})$ (Lloyd et al., 2015). However, the first term involving an expectation over an integral in Equation (3.37) is generally intractable.

Gaussian kernel case Specifically, the first term in Equation (3.37) can be computed as:

$$\begin{aligned}\mathbb{E}_{q(f)}\left[\int_B f(\mathbf{x})^2 d\mathbf{x}\right] &= \int_B \mathbb{E}_{q(f)}\left[f(\mathbf{x})^2\right] d\mathbf{x} \\ &= \int_B \mathbb{E}_{q(f)}\left[f(\mathbf{x})\right]^2 d\mathbf{x} + \int_B \mathrm{Var}_{q(f)}\left[f(\mathbf{x})\right] d\mathbf{x} \\ &= \mathbf{m}^\top \mathbf{K}_{m,m}^{-1} \mathbf{\Psi} \mathbf{K}_{m,m}^{-1} \mathbf{m} - \mathrm{tr}(\mathbf{K}_{m,m}^{-1} \mathbf{\Psi}) + \mathrm{tr}(\mathbf{K}_{m,m}^{-1} \mathbf{S}\, \mathbf{K}_{m,m}^{-1} \mathbf{\Psi}).\end{aligned} \tag{3.38}$$

In this equation, $\mathbf{\Psi}$ is a $m \times m$ matrix equal to $\int k(\mathbf{X}_m, \mathbf{x}) k(\mathbf{x}, \mathbf{X}_m) d\mathbf{x}$. Generally, $\mathbf{\Psi}$ is not tractable, but Lloyd et al. (2015) provide a closed-form solution for $\mathbf{\Psi}$ for the separable Gaussian kernel, as detailed in Equation (C.3).

Inference To perform inference, the variational parameters $(\mathbf{m}, \mathbf{S})$, together with the GP hyperparameters Θ, are jointly optimized to minimize $\mathcal{L}_{\mathrm{ELBO}}$. This optimization process can be executed using standard optimization methods. Upon completion of the optimization, the posterior process $p(f|\mathbf{X}_B)$ is approximated by $q(f)$ as determined by Equations (3.36) and the knowledge of $(\mathbf{m}, \mathbf{S})$.

Laplace Bayesian Permanental Processes (LBPP)

In their work, Walder and Bishop (2017) proposed a novel Bayesian approach for the permanental process that leverages connections with reproducing kernel Hilbert spaces (RKHS) and a Laplace approximation to estimate the intensity posterior. This approach, known as the *Laplace-Based Bayesian Permanental Process* (LBPP),

presents a significantly more efficient alternative to existing Bayesian inference methods. Moreover, LBPP can also be seen as a Bayesian extension of a frequentist approach introduced by Flaxman et al. (2017). To provide a better understanding of LBPP, we will first briefly explain the approach of Flaxman et al. (2017), and then introduce LBPP.

Intensity estimation with reproducing kernel In their work, Flaxman et al. (2017) developed a non-probabilistic regularization algorithm for estimating the intensity function of a of an inhomogenous Poisson process, by exploiting the properties of RKHS (as introduced in Section 2.4.1). They estimate the intensity function $\lambda(\cdot) = f(\cdot)^2$ using a function $f \in \mathcal{H}_k$, where $\mathcal{H}_k$ denotes the RKHS associated with a kernel function k that has a Mercer representation defined in Equation (3.31). The estimation of f is performed by minimizing the negative likelihood with a regularization term, given by:

$$\hat{f} = \arg\min_{\mathcal{H}_k} \Big\{ -\sum_{i=1}^{N} \log(f(\mathbf{x}_i)) + \int_B f(\mathbf{x})^2 d\mathbf{x} + ||f||^2_{\mathcal{H}_k} \Big\}. \tag{3.39}$$

Equation (3.39) defines the objective of a regularized *Empirical Risk Minimization* problem (ERM). This problem combines a Poisson likelihood with a regularization term $||f||^2_{\mathcal{H}_k}$. Notably, Flaxman et al. (2017)'s work demonstrates that this objective can be reformulated as an equivalent, tractable ERM problem over a RKHS. In particular, we can reframe the objective function, denoted as $J[f]$, as follows:

$$\begin{aligned} J[f] &= -\sum_{i=1}^{N} \log(f(\mathbf{x}_i)) + ||f||_{L^2(B)} + ||f||^2_{\mathcal{H}_k} \\ &= -\sum_{i=1}^{N} \log(f(\mathbf{x}_i)) + ||f||^2_{\mathcal{H}_{\tilde{k}}} \end{aligned} \tag{3.40}$$

where $\mathcal{H}_{\tilde{k}}$ is a new RKHS associated with the exact modified kernel $\tilde{k}$ given in Equation (3.34). Assuming that the modified kernel $\tilde{k}$ can be computed, the problem initially formulated in Equation (3.39), which operates within the RKHS $\mathcal{H}_k$, is equivalent to optimizing the problem $J[f]$ presented in Equation (3.40) within the RKHS

$\mathcal{H}_{\tilde{k}}$.

By invoking the representer theorem (Kimeldorf and Wahba, 1971) in conjunction with Equation (3.40), the solution $\hat{f}$ can be succinctly expressed as $\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i \tilde{k}(\mathbf{x}, \mathbf{x}_i)$ for all $\mathbf{x} \in B$, where $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1, \cdots, \hat{\alpha}_N]^\top \in \mathbb{R}^N$. For a deeper understanding of ERM within RKHS and the representer theorem, we refer you to Schölkopf and Smola (2002b, section 4.2). As a result, we are presented with a finite-dimensional problem, wherein we seek to determine $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^N$ such that:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\arg\min} \Big\{ -\sum_{i=1}^{N} \log(\boldsymbol{\alpha}^\top \tilde{k}(\mathbf{X}, x_i) + \boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_{n,n} \boldsymbol{\alpha}) \Big\} \tag{3.41}$$

In this equation, $\tilde{\mathbf{K}}_{n,n}$ denotes an $N \times N$ matrix with i, j entries $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$, and $\tilde{k}(\mathbf{X}, \mathbf{x}_i)$ represents an N-dimensional vector with j entries $k(\mathbf{x}_j, \mathbf{x}_i)$. Solving the optimization problem in Equation 3.41 can be efficiently accomplished using standard numerical optimization techniques.

Bayesian approach with Laplace approximation In contrast, Walder and Bishop (2017) propose a Bayesian approach that models the intensity function of a Permanental process by establishing a connection with RKHS, akin to the approach presented by Flaxman et al. (2017).

By employing the Karhunen-Loève (KL) representation of the GP introduced in Equation in Equation (3.32), the Permanental process can be redefined as a linear combination of basis functions, as depicted below:

$$\lambda(\mathbf{x}) := \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} w_i w_j \Phi_i(\mathbf{x}) \Phi_j(\mathbf{x}) \quad \text{with } \mathbf{w} \sim \mathcal{N}(\mathbf{w}|0, \Lambda). \tag{3.42}$$

This representation implies that inferring the latent function f is equivalent to inferring its random coefficients $\mathbf{w}$, with a Gaussian prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|0, \Lambda)$.

Substituting the Gaussian process expression with fixed hyperparameters Θ into the expression for the Permanental process likelihood in Equation (3.34), we arrive

at the following form:

$$\log p(\mathbf{X}_B|\mathbf{w},\Theta) = -\sum_{i=1}^{\infty} w_i^2 + \sum_{i=1}^{N} \log\big(|\sum_{j=0}^{\infty} w_j \Phi_j(\mathbf{x}_i)|^2\big). \tag{3.43}$$

The joint distribution over $\mathbf{w}$ and $\mathbf{X}_B$ can, in turn, be expressed as:

$$\begin{aligned}\log p(\mathbf{w},\mathbf{X}_B|\Theta) &= \log p(\mathbf{X}_B|\mathbf{w},\Theta) + \log p(\mathbf{w}) \\ &= \log p(\mathbf{X}_B|\mathbf{w},\Theta) - \frac{1}{2}\mathbf{w}^\top(\mathbf{I}+\Lambda^{-1})\mathbf{w} + C\end{aligned} \tag{3.44}$$

for some constant C.

In their research, Walder and Bishop (2017) introduced a Laplace approximation method to handle the intractable posterior distribution $p(\mathbf{w}|\mathbf{X},\Theta)$. This method involves a second-order approximation of the logarithm of the posterior distribution, where the Taylor expansion is truncated after the second-order term. The resulting approximation yields a Gaussian distribution $\mathcal{N}(\mathbf{w}|\hat{\mathbf{w}},\mathbf{Q})$, with a mean $\hat{\mathbf{w}}$ equal to the mode of the posterior distribution, i.e.

$$\hat{\mathbf{w}} := \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}_B,\Theta)$$

and a covariance matrix equal to the inverse of the Hessian of the negative log-posterior evaluated at the mode.

To establish a connection with Flaxman et al. (2017), we start by considering the stationary equation that the mode $\hat{\mathbf{w}}$ must satisfy, i.e., $\nabla_{\mathbf{w}}\ \log p(\mathbf{w}|\mathbf{X}_B,\Theta)|_{\mathbf{w}=\hat{\mathbf{w}}} = 0$. This implies that

$$\hat{\mathbf{w}} = \big(\mathbf{I}+\Lambda^{-1}\big)^{-1}\left(\sum_{i=1}^{N}\frac{\mathbf{\Phi}(\mathbf{x}_i)}{\mathbf{w}^\top\mathbf{\Phi}(\mathbf{x}_i)}\right). \tag{3.45}$$

With this stationary solution in place, we can compute the predictive mean:

$$\mathbb{E}\left[f(\mathbf{x})|\mathbf{X}_B\right] = \boldsymbol{\Phi}(\mathbf{x})^\top \hat{\mathbf{w}} \tag{3.46}$$

$$= \sum_{i=1}^{N} \frac{2}{\hat{\mathbf{w}}^\top \boldsymbol{\Phi}(\mathbf{x}_i)} \cdot \boldsymbol{\Phi}(\mathbf{x}_i)^\top \left(\mathbf{I} + \Lambda^{-1}\right)^{-1} \boldsymbol{\Phi}(\mathbf{x})$$

$$:= \sum_{i=1}^{N} \hat{\alpha}_i \tilde{k}(\mathbf{x}, \mathbf{x}_i) \tag{3.47}$$

where $\hat{\alpha}_i = 2/\hat{\mathbf{w}}^\top \boldsymbol{\Phi}(\mathbf{x}_i)$ for $i = 1, \cdots, N$.

Notably, this computation leads to the same $\tilde{k}$ as observed in Equations (3.34) and (3.41). Remarkably, Walder and Bishop (2017) demonstrated that optimizing $\hat{\mathbf{w}}$ is equivalent to optimizing $\boldsymbol{\alpha}$ similarly to the problem posed in Equation (3.41). This intriguing connection reveals that LBPP's predictive mean aligns with the intensity estimation approach with RKHS by Flaxman et al. (2017), despite their conceptual differences. In fact, the Bayesian Laplace method can be perceived as a probabilistic extension of the intensity estimation with RKHS approach.

Arbitrary domain In both Walder and Bishop (2017) and Flaxman et al. (2017), the feasibility of computing the integral term $\int_B f(\mathbf{x})d\mathbf{x}$ and the expression for $\tilde{k}$ depend on the availability of an explicit Mercer decomposition for k with respect to $L^2(B)$. Regrettably, such a decomposition is not generally available, which limits the applicability of these methods to arbitrary domains and kernels. For instance, while Flaxman et al. (2017) derived a specific representation for the Gaussian kernel in $L^2(\mathbb{R}^2, \mu)$ when μ is a Gaussian measure, this approach does not extend well to other kernel types or common scenarios where B is a finite subset of $\mathbb{R}^2$ and μ is the Lebesgue measure.

To address the challenge of more general domains, Flaxman et al. (2017) and Walder and Bishop (2017) introduce the Nyström method. This approach offers a low-rank approximation of the kernel k by utilizing a subset of $m \leq N$ uniformly sampled data points denoted as $\mathbf{X}_m = \{\mathbf{x}_i^{(m)}\}_{i=1}^m$. Specifically, the Nyström method approximates the kernel integral operator T_k (as defined in Equation (2.34)) through discretization over $\mathbf{X}_m$. This leads to an approximation of the eigenvalues and eigen-

functions of the Mercer decomposition (Theorem 2.4.5), in terms of the eigenvectors $\{\Phi_i^{(m)}\}_{i=1}^m$ and eigenvalues $\{\lambda_i^{(m)}\}_{i=1}^m$ of the $m \times m$ Gram matrix $\mathbf{K}_{m,m}$ with entries $k(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$. The detailed formulation for these approximated eigenfunctions and related eigenvalues will be presented in the forthcoming chapter. This delay is purposeful, as it has a direct correspondence to our forthcoming contribution. For further details about the Nyström method, please refer to Rasmussen and Williams (2005, Chap. 4.3, 8).

However, employing the Nyström method with these approximated finite-dimensional eigenvectors leads to a partial loss of the tractability property associated with the permanental process. Specifically, the analytical expression for the integral term $\int_B f(\mathbf{x})d\mathbf{x}$ can only be obtained under the Nyström method for certain standard kernel types, most notably the Gaussian kernel. For a precise and comprehensive computation of this integral, please refer to the upcoming chapter.

Reflection invariance and nodal lines

Generally speaking, the use of a square-rootl in the permanental process can lead to posterior distributions featuring artifacts known as "nodal lines" (John and Hensman, 2018). These nodal lines emerge due to the non-injective nature of the square link function, wherein different values of ± f can produce the same intensity. This phenomenon can create regions where the latent function modes alternates between positive and negative values, causing zero-crossings and artificially driving the intensity to zero. Figure 3.2 visually illustrates this issue in a 1D context, depicting the known intensity in black alongside model fits with only positive or alternating signs for the latent functions.

These effects tends to manifest when regions with a high number of events alternate with those of low activity. As a result, the artificial suppression of intensity may initially appear inconsequential. However, depending on the problem's configuration, nodal lines can also appear in moderate intensity region, though less frequently. This can potentially introduce issues in terms of model interpretation and sampling. In Figure 3.2, this is illustrated by the first crossing line represented by the "∘" symbol.

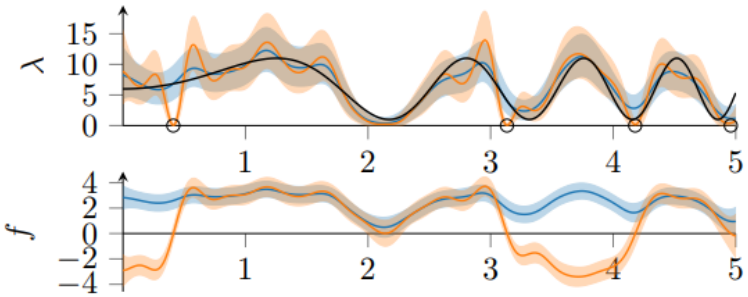


Figure 3.2: Examples of nodal lines for a 1D Toy example in John and Hensman (2018). The figure shows the effect of nodal lines for a 1D intensity function (black), a model fit (orange) over a realization where $N = 10$ and a better model fit without 0 crossing. The graph is taken from John and Hensman (2018, section 4.5).

In the subsequent chapter, we will discuss and introduce an empirical solution proposed by (John and Hensman, 2018) within our model to address this issue. (John and Hensman, 2018) demonstrated how these effects can be mitigated by incorporating an offset term into the prior. However, it's important to acknowledge that these observations are primarily grounded in empirical evidence. The precise delineation between performance enhancement and the mitigation of nodal lines through this prior reparametrization may lack a firmly established theoretical foundation, prompting the need for further exploration.

3.4 Summary

In this chapter, we provided a comprehensive overview of GP-modulated Cox processes. We began by introducing the fundamental concepts of point processes, emphasizing key elements such as intensity functions, densities, and pseudo-likelihoods. Our goal was to elucidate their formulation with justification. We also introduced Poisson processes, Cox processes, and Gaussian Cox processes in a general context.

We further explored various classes of Gaussian Cox processes, including the logarithmic and sigmoid versions. Finally, we introduced the Permanental process and its two different implementations: VBPP and LBPP, which play crucial roles in the subsequent chapter.

Despite the Permanental process offering compelling advantages, particularly in

terms of tractability, its applicability does face constraints that merit consideration:

1. One significant limitation, as extensively discussed in this chapter, pertains to the tractability advantage of the integral term, which is limited to specific standard kernel types. Notably, both VBPP and LBPP with Nyström rely exclusively on the Gaussian kernel. While these methods yield efficient inference for the Gaussian kernel, other kernel types may not enjoy the same benefits. These constraints are particularly limiting because the choice of kernel for a GP model profoundly impacts its performance on a given task. This aspect has been highlighted as a primary drawback of using the Permanental process in the existing literature (Aglietti et al., 2019).

2. Another limitation associated with the reflection invariance of the Permanental process involves the emergence of artifacts known as "nodal lines" (John and Hensman, 2018), as described earlier, within posterior distributions.

In the upcoming chapter, we present our contributions, specifically designed to address or mitigate these challenges.

Chapter 4

Sparse spectral Bayesian Permanental process with generalized kernel

In this chapter, we introduce a novel scheme for Bayesian inference on permanental processes which models the Poisson intensity as the square of a Gaussian process. Combining *generalized kernels* and a Fourier features-based representation of the Gaussian process with a Laplace approximation to the posterior, we achieve a fast and efficient inference that does not require numerical integration over the input space, allows kernel design and scales linearly with the number of events. Our method builds and improves upon the state-of-the-art Laplace Bayesian point process benchmark of Walder and Bishop (2017), demonstrated on both synthetic, real-world temporal and large spatial data sets.

4.1 Introduction

In the previous chapter, we introduced a promising instance of the Gaussian Cox process known as the permanental process (McCullagh and Møller, 2006), obtained by defining the intensity as the square of a GP. Walder and Bishop (2017) propose a Laplace Bayesian point process (LBPP) method, a fast inference schemes for the permanental process that relies on the Mercer decomposition of the Gaussian process kernel and a Laplace approximation to the intensity posterior. They show significant speed improvement compared to variational Bayesian inference. Inference based on the Laplace approximation has already been proposed in the context of a Gaussian Cox process by Cunningham et al. (2008a), Illian et al. (2012), and Flaxman et al. (2015b).

However, the tractability properties of the permanental process used by Lloyd et al. (2015) and Walder and Bishop (2017) only holds for certain standard types of kernels such as the squared exponential kernel, which encodes restrictive assumptions about the form of the function we are modelling. In general, the choice of kernel determines almost all the generalization properties of a Gaussian process model and profoundly affects its performance on a given task (Rasmussen and Williams, 2005). Approaches have been proposed in recent years to achieve more expressible kernels either by a composition of simple analytical forms (Duvenaud et al., 2011, 2013) or more flexibly through a spectral representation (Lázaro-Gredilla et al., 2010; Wilson and Adams, 2013; Samo and Roberts, 2015a).

In this chapter, we build on the LBPP approach of Walder and Bishop (2017), introducing an alternative fast Laplace-based inference exploiting spectral representation of kernels and random Fourier features (RFFs). Our approach, the Sparse Spectral Permanental Process (SSPP), retains the tractability properties of the permanental process, whilst being able to adapt to a broader range of stationary kernels. Furthermore, our method works with *generalized stationary spectral kernels* (Samo and Roberts, 2015a), to our knowledge, the most general class of expressible spectral kernels, that can approximate any stationary kernels to arbitrary precision.

Following John and Hensman (2018), we also include a mean constant to mitigate the effect of *nodal lines* observed for the permanental process, resulting from the non-injective nature of the squared transformation. Our approach shows systematic improvement in accuracy in synthetic and real-world data sets.

4.2 Preliminaries : Permanental processes

In the context of a Poisson process occurring in a continuous domain $\mathcal{X}$, the inference process involves estimating an intensity function $\lambda(\mathbf{x}) : \mathcal{X} \mapsto \mathbb{R}^+$. This chapter focuses on permanental process models, where the intensity λ is defined as $\lambda(\cdot) := f(\cdot)^2$, for a function $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$, with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ being the positive-definite covariance function for f.

We assume $\mathbf{X}_B = \{\mathbf{x}_i\}_{i=1}^N$ to be a realization of N observations in a finite region $B \subset \mathcal{X}$. In this chapter, we employ a change of notation, referring to $\mathbf{X}_B$ as $\mathbf{X}$ for clarity and convenience, as there is no ambiguity in this context.

4.2.1 Integral expression via Mercer Theorem

Here, we briefly recall the basic components of permanental processes discussed in Section 3.3.3. The GP covariance function k has a Mercer decomposition (Theorem (2.4.5)) on $(\mathcal{X}, \mu)$, if it can be written as

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \Phi_i(\mathbf{x}) \Phi_i(\mathbf{x}') \quad \text{for } \mathbf{x}, \mathbf{x}' \in \mathcal{X}$$

where $\{\lambda_i\}_{i=1}^{\infty}$ is a sequence of summable, non-negative, non-increasing *eigenvalues*, and $\{\Phi_i(\cdot)\}_{i=1}^{\infty}$ is a set of mutually-orthogonal, unit-norm *eigenfunctions* with respect to the inner product $\langle u, v \rangle = \int_{\mathcal{X}} u(\mathbf{x}) v(\mathbf{x}) d\mu(\mathbf{x})$. Subsequently, following Equation 3.32, $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ can be reformulated as an equivalent linear form

$$f(\mathbf{x}) = \sum_{i=0}^{\infty} w_i \Phi_i(\mathbf{x})$$

where $\mathbf{w} = (w_1, w_2, \ldots)^\top \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Lambda})$ and $\mathbf{\Lambda}$ is a diagonal covariance matrix with entries λ_i, $i = 1, 2, \ldots$. Further, it can be shown that $\text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = \mathbf{\Phi}(\mathbf{x})^\top \mathbf{\Lambda} \mathbf{\Phi}(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$, where $\mathbf{\Phi}(\cdot)$ is a vector with entries $\Phi_i(\cdot)$, $i = 1, 2, \ldots$. Referring to Equation (3.33), the integral of the intensity can be expressed as:

$$\int_{\mathcal{X}} f(\mathbf{x})^2 \, d\mu(\mathbf{x}) = \sum_{i=0}^{\infty} w_i^2.$$

4.2.2 Approximate Bayesian inference

In our case, to make the reformulation of the integral $\int_B f(\mathbf{x})^2 d\mathbf{x}$ possible as in Equation (3.33), the kernel for f requires an explicit Mercer representation with respect to B and μ defined as the Lebesgue measure ; this is not available for most choices of kernel. In such cases, the Nyström method can be used to approximate the eigenfunctions and eigenvalues of the Mercer decomposition. Both Flaxman et al. (2017) and Walder and Bishop (2017) adopt the Nyström approach in the context of the Permanental Cox process with Gaussian kernel. Walder and Bishop (2017) further propose a Bayesian inference scheme based on a Laplace approximation for a non-GP likelihood. We provide a quick review of the LBPP model using Nyström method proposed by Walder and Bishop (2017). We also provide in Proposition 4.2.1, an expression for the integral term $\int_B f(\mathbf{x})^2 d\mathbf{x}$ under Nyström approximation that is not directly available in Walder and Bishop (2017).

In Section 3.3.3, we discussed the Nyström method (Rasmussen and Williams, 2005, Chap. 4.3, 8), which offers a way to approximate the function f based on a reduced-rank approximation for k. The method achieves this by considering a subset of $m \leq N$ data points $\mathbf{X}_m = \{\mathbf{x}_i^{(m)}\}_{i=1}^n$ sampled uniformly from the original data. The eigenvalues and eigenfunctions of the Mercer decomposition are approximated using the eigenvectors $\mathbf{u}_i^{(m)}$ and eigenvalues $\lambda_i^{(m)}$ of $\mathbf{K}_{m,m}$, the Gram matrix with i, j entry $k(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$. Thus, f can be approximated by

$$f(\mathbf{x}) \approx \sum_{i=1}^{m} \hat{\lambda}_i^{\frac{1}{2}} w_i^{(m)} \hat{\Phi}_i(\mathbf{x}) \tag{4.1}$$

where $\mathbf{w}^{(m)} = (w_1^{(m)}, \ldots, w_m^{(m)})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ and

$$\hat{\lambda}_i := \frac{1}{m}\lambda_i^{(m)} \tag{4.2}$$

$$\hat{\Phi}_i(\cdot) := \frac{\sqrt{m}}{\lambda_i^{(m)}} k(\cdot, \mathbf{X}^{(m)})^\top \mathbf{u}_i^{(m)}. \tag{4.3}$$

We can reformulate Equation (4.1) by substituting the expressions for for $\{\hat{\lambda}_i\}_{i=1}^m$ and $\{\hat{\Phi}(\cdot)_i\}_{i=1}^m$ from Equations (4.2) and (4.3):

$$\begin{aligned} f(\mathbf{x}) &\approx k(\mathbf{x}, \mathbf{X}_m)^\top \sum_{i=1}^m \frac{w_i^{(m)}}{\sqrt{\lambda_i^{(m)}}} \mathbf{u}_i^{(m)} \\ &= k(\mathbf{x}, \mathbf{X}_m)^\top \mathbf{U}^{(m)} {\mathbf{\Lambda}^{(m)}}^{-\frac{1}{2}} \mathbf{w}^{(m)} \\ &:= {\mathbf{w}^{(m)}}^\top \boldsymbol{\varphi}^{(m)}(\mathbf{x}) \end{aligned} \tag{4.4}$$

where $\mathbf{U}^{(m)}$ is the $m \times m$ matrix of eigenvectors $\{\mathbf{u}_i^{(m)}\}_{i=1}^m$ as columns, and $\boldsymbol{\varphi}^{(m)}(\cdot) := \left[k(\cdot, \mathbf{X}_m)^\top \mathbf{U}^{(m)} {\mathbf{\Lambda}^{(m)}}^{-\frac{1}{2}}\right]^\top$ denotes the new features vector.

Integral calculation In Proposition 4.2.1 we express the integral term $\int_B f(\mathbf{x})^2 d\mathbf{x}$ under the Nyström approximation with the Gaussian kernel k, both because it is not available in Walder and Bishop (2017) and to demonstrate the similarities with the corresponding derivation of our proposed method in Proposition 4.3.3.

Proposition 4.2.1. *Under the GP approximation (4.4) with the Gaussian kernel k, the integral expression $\int_B f(\mathbf{x})^2 \, d\mathbf{x}$ can be written as*

$$\int_B f(\mathbf{x})^2 \, d\mathbf{x} = {\mathbf{w}^{(m)}}^\top \mathbf{M}^{(m)} \mathbf{w}^{(m)}$$

where $\mathbf{M}^{(m)}$ is a $n \times n$ matrix defined as

$$\mathbf{M}^{(m)} := {\mathbf{\Lambda}^{(m)}}^{-\frac{1}{2}} \left[{\mathbf{U}^{(m)}}^\top \mathbf{\Psi}^{(m)} \, \mathbf{U}^{(m)}\right] {\mathbf{\Lambda}^{(m)}}^{-\frac{1}{2}}$$

and $\mathbf{\Psi}^{(m)}$ is a $n \times n$ matrix given in Section C.1 of Appendix C.

We provide a proof of this result in Section C.1 of Appendix C.

4.3 Model

Motivated in part by the shortcomings of the Nyström approach proposed by Walder and Bishop (2017), we now present an alternative LBBP approach to inference for the permanental process. In contrast to the Mercer approach, it is based on a sparse spectral representation of a GP, exploiting random Fourier features (RFFs, Rahimi and Recht, 2007) for reduced-rank kernel expression. As a result, it provides a tractable expression for the integral of the intensity over the input domain.

Our spectral approach works for any bounded, continuous and shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') := k(\mathbf{x} - \mathbf{x}')$ that satisfies the condition of Bochner's theorem (see Theorem 4.3.1) and admits a finite dimensional feature space representation or approximation. In contrast, the variational inference approach of Lloyd et al. (2015) and the LBPP with Nyström, yield an analytical integral expression for a limited choice of kernels, like the Gaussian kernel. Furthermore, we are able to adapt our method to *generalized stationary spectral kernels* (Samo and Roberts, 2015a) which generalize two other classes of expressible spectral kernels, the *sparse spectrum kernels* (Lázaro-Gredilla et al., 2010) and the *mixture spectral kernels* (Wilson and Adams, 2013). These two kernels have been proven to be able to approximate any bounded continuous stationary kernels to arbitrary precision.

We also address the issue of *nodal lines* discussed in John and Hensman (2018). This problem arises since the inverse link function $\lambda(\cdot) := f(\cdot)^2$ is not injective, with $\pm f(\cdot)$ producing the same intensity. Therefore, regions of negative and positive f must exhibit zero-crossings, where the intensity is artificially forced to zero, despite the underlying intensity being positive. Following John and Hensman (2018), we add an offset parameter β to the intensity function $\lambda(\cdot) := (f(\cdot) + \beta)^2$ corresponding to an initial value for the prior mean of the GP, to alleviate the problem. Notably, this β parameter is considered a model hyperparameter and is fitted to the data accordingly.

4.3.1 Sparse spectral kernels

In this section, we briefly present two families of spectral kernels, *sparse spectrum kernels* (Lázaro-Gredilla et al., 2010) and *mixture spectral kernels* (Wilson and Adams, 2013) that have been proposed in recent years for kernel design. They are both known to be dense in the family of stationary kernels, implying that they can approximate any stationary kernel to an arbitrary precision given sufficient spectral components.

Spectral kernels are constructed via the Bochner's theorem (Bochner, 1932), which states that any bounded, continuous and shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') := k(\boldsymbol{\tau})$ with $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, is the inverse Fourier transform of a bounded positive measure.

Theorem 4.3.1. *(Bochner) An integrable function $k : \mathbb{R}^d \to \mathbb{C}$ is the covariance function of a weakly stationary mean square continuous random process on $\mathbb{R}^d$ if and only if it can be represented as*

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^d} \exp(i\mathbf{z}^\top \boldsymbol{\tau}) d\mu(\mathbf{z}) \tag{4.5}$$

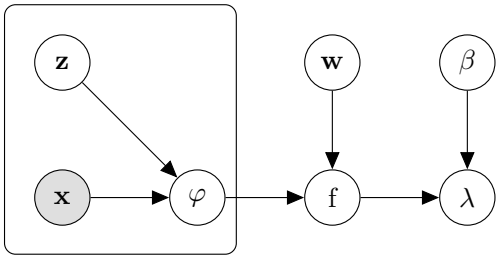
where $\mu(\mathbf{z})$ is a positive definite measure.

Sparse spectrum kernels *Sparse spectrum kernels* can be obtained by setting μ in Equation (6.13) to be a positive discrete symmetric measure $\mu_{ss} = \sum_{k=1}^{K} \frac{a_k}{2} (\delta_{\omega_k} + \delta_{-\omega_k})$[1] where $a_k > 0$ and δ_{ω_k} denotes the Dirac measure centred at the point spectral frequencies $w_k \in \mathbb{R}^d$ for $k = 1, \cdots, K$. Note that as such, μ_{ss} is singular with respect to the Lebesgue measure and does not admit a density. Through Equation (6.13), we obtained the *sparse spectrum kernel*, spanned by the trigonometric functions $\{\cos(\omega_k^\top \mathbf{x})\}_{k=1}^K$. A major challenge is that a direct optimization of the linear coefficients $\{a_k\}_{k=1}^K$ and the frequencies $\{\omega_k\}_{k=1}^K$ often leads to over-fitting as illustrated by Lázaro-Gredilla et al. (2010) in the context of GP regression.

Mixture spectral kernels Wilson and Adams (2013) consider the case when μ is absolutely continuous with respect to the Lebesgue measure and admits a spec-

[1] Note that positive finite discrete measures are weakly dense in the space of all positive finite measure (Hu and Papageorgiou, 2013).

Figure 4.1: Illustration of the SSPP model. The arrows directions suggest directions of influence.



tral density $S(\cdot)$. In that case, $S(\cdot)$ and the kernel function k are Fourier duals of each other. *Mixture spectral kernels* model the spectral density S as a mixture of independent Gaussian densities with non-zero mean. Since mixtures of Gaussians are dense in the set of all distribution functions (Plataniotis and Hatzinakos, 2001), the resulting dual of this set is dense in the family of continuous stationary kernels.

Random Fourier features Random Fourier features, as initially proposed by Rahimi and Recht (2007), leverage a randomized lower-dimensional feature mapping to achieve scalability. This approach is grounded in Bochner's theorem, allowing the kernel function k to be reformulated as:

$$k(\mathbf{x} - \mathbf{x}') \approx \frac{\sigma^2}{r} \sum_{k=1}^{r} \exp(i\mathbf{z}_k^\top (\mathbf{x} - \mathbf{x}')) \tag{4.6}$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_r$ in $\mathbb{R}^d$ are independent samples from the distribution with density $S(\cdot)$, for some integer $r > 0$ and $\sigma > 0$. Here, we have assumed that μ in Equation (6.13) is absolutely continuous with respect to the Lebesgue measure and has a spectral density $S(\cdot)$. Equation (6.13) is then approximated using Monte Carlo integration. We also treat the scale parameter σ of the kernel function separately for convenience.

We thus obtain a kernel approximation

$$k(\mathbf{x} - \mathbf{x}') \approx \boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}') \tag{4.7}$$

where $\boldsymbol{\varphi}^{(r)}$ is an explicit feature mapping $\boldsymbol{\varphi}^{(r)} : \mathcal{X} \to \mathbb{R}^r$ such that

$$\boldsymbol{\varphi}^{(r)}(\mathbf{x}) = \frac{\sigma}{\sqrt{r}} \left[\exp(i\mathbf{z}_1^\top \mathbf{x}), \ldots, \exp(i\mathbf{z}_r^\top \mathbf{x})\right]^\top. \tag{4.8}$$

We may obtain a $2r$-sized real-valued mapping that satisfies Equation (4.6) using

$$\begin{aligned}\boldsymbol{\varphi}^{(r)}(\mathbf{x}) = \frac{\sigma}{\sqrt{r}}[&\cos(\mathbf{z}_1^\top \mathbf{x}), \ldots, \cos(\mathbf{z}_r^\top \mathbf{x}), \\ &\sin(\mathbf{z}_1^\top \mathbf{x}), \ldots, \sin(\mathbf{z}_r^\top \mathbf{x})]^\top\end{aligned} \tag{4.9}$$

where $\mathbf{z} \sim S(\mathbf{z})$. The derivation of Equation (4.9) is provided in Section C.2 of Appendix C.

RFF methods share a close relationship with *sparse spectrum kernel*. Examining Equation (4.9), we observe RFF as a particular instance of a *sparse spectrum kernel*, wherein the frequencies ω_k are randomly sampled from a distribution rather than being optimized. However, it's important to note that while RFF methods provide scalability, they do not inherently adress the challenge of adaptively learning the spectral measure μ from the available data.

RFF are widely recognized for their efficacy in approximating various isotropic kernels, such as the Gaussian, Laplace, or Cauchy kernels. However, empirical evidence suggests that kernels with spectral densities exhibiting multimodal behavior or sharp edges, such as oscillatory kernels, pose a greater challenge for accurate approximation. Addressing this challenge often requires employing a larger number of random features to ensure precise capture.

From a theoretical perspective, several studies have advanced our understanding of the optimal number of features necessary to effectively approximate the kernel matrix (Rahimi and Recht, 2007; Sriperumbudur and Szabó, 2015; Sutherland and Schneider, 2015) and to maintain the statistical properties of the original method across diverse learning tasks (Rahimi and Recht, 2007; Li et al., 2021). Specifically, for the former, research has demonstrated that $O(\epsilon^{-2} \log |\mathcal{X}|)$ features are enought to achieve an approximation accuracy of ϵ with respect to the L^∞ error (Sriperumbudur

and Szabó, 2015). Regarding the latter, learning with RFF and Lipschitz continuous loss functions necessitates $\Omega(\sqrt{n}\log(n))$ features to prevent any loss of learning accuracy (Li et al., 2021). Notably, fewer features are typically sufficient when dealing with the Gaussian kernel in both cases.

Additionally, several works have extended the initial formulation proposed by Rahimi and Recht (2007) through alternative Monte Carlo sampling techniques. These methods aim to either enhance the approximation quality (Yu et al., 2016) or reduce the time complexity (Le et al., 2013). Other approaches include Quasi-Monte Carlo sampling (Yang et al., 2014; Lyu, 2017) and Quadrature methods (Dao et al., 2017).

Note that the above RFF formulation is usually only suitable for stationary kernels, which satisfy the conditions set forth by Bochner's theorem. This restricts its applicability to many nonstationary kernels. However, several explicit random feature map approximations have been proposed for some specific nonstationary kernels. Examples include the generalized Gaussian kernel(Vempati et al., 2010), additive kernels Vedaldi and Zisserman (2012), Polynomial kernels on the unit sphere (Pennington et al., 2015). Additionally, some authors Genton (2001); Samo and Roberts (2015a); Ton et al. (2018) have introduced Fourier features for nonstationary kernels based on a generalization of the Bochner's theorem proposed by Yaglom (Yaglom, 1987).

4.3.2 Generalized stationary kernels

One advantage of our method is that it can work with the *generalized stationary kernels* (Samo and Roberts, 2015a), that are dense in the family of stationary kernel and admits *sparse spectrum kernels* and *mixture spectral kernels* as special cases. *Generalized kernels* can also account for different degree of differentiability of the latent function. *Sparse spectrum kernels* and *mixture spectral kernels* are more limited in a sense that, when used as covariance functions, they yield infinite differentiability of the corresponding stochastic process, which might be unrealistic for certain learning tasks (Stein, 1999).

Definition 4.3.2. *(Generalized stationary kernel) Let g be a stationary kernel g :*

$\mathbb{R}^d \to \mathbb{R}$ *such that* $g(0) = 1$. *A generalized kernel* k_{GS} *with* $K \in \mathbb{N}^+$ *components takes the form*

$$k_{GS}(\boldsymbol{\tau}) = \sum_{k=1}^{K} \sigma_k^2 \; g(\boldsymbol{\tau} \odot \boldsymbol{\gamma}_k) \cos(\boldsymbol{\omega}_k^\top \boldsymbol{\tau}) \tag{4.10}$$

where $\boldsymbol{\omega}_k \in \mathbb{R}^d$, $\boldsymbol{\gamma}_k \in \mathbb{R}^{+d}$, $\sigma_k > 0$ *for* $k = 1, \cdots, K$ *and* $\odot$ *denotes the element-wise Hadamard product.*

The parameters $\{\boldsymbol{\gamma}_k\}_{k=1}^K$ are used as inverse input scales. When $\{\boldsymbol{\gamma}_k\}_{k=1}^K$ are set to zero, we retrieve the *sparse spectrum kernels*. The *spectral mixture kernels* corresponds to a special case where $g(\boldsymbol{\tau}) = \exp(-||\boldsymbol{\tau}||_2^2/2)/\sqrt{2\pi}$.

The degree of smoothness of a zero-mean GP with kernel k_{GS} is determined by the kernel g. Samo (2017) proposes learning the differentiability of the underlying latent function, by setting g to be a Matérn kernel with different parameter values ν from $\frac{1}{2} + i, i = 0, \cdots 2$. The case of $i = 0$ corresponds to continuity and the case of $i > 0$ to i times differentiability.

Finite-dimensional Feature Space Approximation For our methodology, we are interested in having a reduced rank representation for k_{GS} similar to Equation (4.9). Any consistent RFF approximation of g in Equation (4.10) such that $g(\mathbf{x}-\mathbf{x}') \approx \boldsymbol{\varphi}_g^{(r)}(\mathbf{x})^\top \boldsymbol{\varphi}_g^{(r)}(\mathbf{x}')$ where $\boldsymbol{\varphi}_g^{(r)}$ is an explicit feature mapping $\boldsymbol{\varphi}_g : \mathcal{X} \to \mathbb{R}^r$, would results in a finite-dimensional feature space approximation for k_{GS} (Samo, 2017). In that case,

$$k_{GS}(\mathbf{x}, \mathbf{x}') \approx \sum_{k=1}^{K} h_k(\mathbf{x})^\top h_k(\mathbf{x}')$$

with

$$h_k(\mathbf{x}) = \sigma_k \; \boldsymbol{\varphi}_g^{(r)}(\mathbf{x} \odot \boldsymbol{\gamma}_k) \otimes \begin{bmatrix} \cos(\boldsymbol{\omega}_k^\top \mathbf{x}) \\ \sin(\boldsymbol{\omega}_k^\top \mathbf{x}) \end{bmatrix} \tag{4.11}$$

for $k = 1, \ldots, K$, where $\otimes$ denotes the Kronecker product.

To be consistent with previous notations, we define $k_{GS}(\mathbf{x}, \mathbf{x}') \approx \boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}')$

Table 4.1: Common Stationary distance-dependent kernels and their duals, where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is a modified Bessel function.

NAME	KERNEL FUNCTION $k(\boldsymbol{\tau})$	SPECTRAL DENSITY $S(\mathbf{z})$
Gaussian	$\exp\left(-\frac{\|\|\boldsymbol{\tau}\oslash\boldsymbol{\ell}\|\|^2}{2}\right)$	$\frac{\prod_{i=1}^d \ell_i}{(2\pi)^{d/2}} \exp\left(-\frac{\|\|\mathbf{z}\odot\boldsymbol{\ell}\|\|^2}{2}\right)$
Matérn(ν)	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\boldsymbol{\tau}}{\ell}\right)^\nu K_\nu(\frac{\sqrt{2\nu}\boldsymbol{\tau}}{\ell})$	$\frac{\Gamma(\nu+d/2)\ell^d}{\Gamma(\nu)(2\nu\pi)^{d/2}}\left(1+\frac{\ell}{2\nu}\|\|\mathbf{z}\|\|^2\right)^{-(\nu+d/2)}$

where $\boldsymbol{\varphi}^{(r)}(\mathbf{x})$ is a feature mapping $\boldsymbol{\varphi}^{(r)}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{4rK}$ satisfying

$$\boldsymbol{\varphi}^{(r)}(\mathbf{x}) = [h_1(\mathbf{x})^\top, \cdots, h_K(\mathbf{x})^\top]^\top. \tag{4.12}$$

Comparing the RFF in Equation (4.8) with those in Equation (4.12), we observe that the latter appears as a vector comprising both *'cos-sin'* and *'cos-cos'* product terms. These terms are parametrized by d-dimensional vectors $\{\boldsymbol{\omega}_k\}_{k=1}^K$ and $\{\boldsymbol{\gamma}_k\}_{k=1}^K$, which serve as hyperparameters and are optimized during fitting. Of particular note, when all $\{\boldsymbol{\gamma}_k\}_{k=1}^K$ are set to vectors of ones entries, $\{\boldsymbol{\omega}_k\}_{k=1}^K$ to zero vectors, and the scale parameter σ_k to $1/\sqrt{K}$, we achieve a finite-dimensional feature space that delivers an equivalent kernel approximation to that of RFF in Equation (4.7).

4.3.3 Sparse spectral Permanental processes (SSPP)

Using RFFs for the GP approximation leads to a so-called sparse spectrum GP, first proposed by Lázaro-Gredilla et al. (2010) in the context of GP regression. Sparse spectrum GPs are GPs defined with the kernel induced by the feature map in Equations (4.9) or (4.12), $k^{(r)}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}')$.

The resulting approximate Gaussian process can be written in terms of a new

r-size latent vector

$$f^{(r)}(\mathbf{x}) \approx {\mathbf{w}^{(r)}}^{\top} \boldsymbol{\varphi}^{(r)}(\mathbf{x}) \text{ where } \mathbf{w}^{(r)} \sim \mathcal{N}(\mathbf{w}^{(r)}|\mathbf{0}, \mathbf{I}_r). \tag{4.13}$$

We define a permanental process with the spectral approximation of Equation (4.13), in which the intensity vector $\mathbf{f}$ follows a similar *linear form* with feature vector given by Equations (4.9) or (4.12). We refer to it as the *Sparse Spectral Permanental Process* (SSPP) when using the feature map (4.9) or *Generalized Sparse Spectral Permanental Process* (GSSPP) when using the feature map (4.12).

In both cases, a tractable expression can be obtained for the integral term in the likelihood, now defined as $\int_B \lambda(\mathbf{x})\ d\mathbf{x} := \int_B (f(\mathbf{x}) + \beta)^2\ d\mathbf{x}$, as follows:

Proposition 4.3.3. *(with proof in sections C.2 and C.3 of Appendix C) Under the GP approximation of Equation (4.13), the integral expression* $\int_B \lambda(\mathbf{x})\ d\mathbf{x}$ *can be expressed as*

$$\int_B \lambda(\mathbf{x})\ d\mathbf{x} = \mathbf{w}^{(r)\top}\mathbf{M}^{(r)}\mathbf{w}^{(r)} + 2\beta\ \mathbf{w}^{(r)\top}\mathbf{m}^{(r)} + \beta^2|B| \tag{4.14}$$

where $\mathbf{M}^{(r)}$ *is an* $r \times r$ *matrix with* i, j *entries defined as*

$$\mathbf{M}^{(r)}_{i,j} := \int_{B^2} \boldsymbol{\varphi}^{(r)}_i(\mathbf{x})\boldsymbol{\varphi}^{(r)}_j(\mathbf{x})\, d\mathbf{x} \tag{4.15}$$

for $i, j = 1, \cdots, r$ *and* $\mathbf{m}^{(r)}$ *is a* r*-vector with entries*

$$\mathbf{m}^{(r)}_i := \int_B \boldsymbol{\varphi}^{(r)}_i(\mathbf{x})\, d\mathbf{x}. \tag{4.16}$$

for $i = 1, \cdots, r$*. Final expressions for* $\mathbf{M}^{(r)}$ *and* $\mathbf{m}^{(r)}$ *are provided in Appendix C.*

The solution of Proposition 4.3.3 shares similarities with Warren et al. (2022) approach, who utilized Random Fourier Features (RFF) in Bayesian Quadrature (BQ). However, our result also covers the calculation of $\int f^2(\mathbf{x})d\mathbf{x}$ and applies to the feature map of a generalized kernel, with the feature map of standard RFF being a specific instance.

4.4 Inference

Adopting the sparse spectral GP, $f^{(r)}$ assumes the linear form of Equation (4.13) for weight vector $\mathbf{w}^{(r)}$ with independently-distributed standard Gaussian elements. Moreover, the integral in Equation (4.14) reduces to a quadratic form. We also define the model hyperparameters Θ to be the parameters of the kernel function together with the offset term β. More precisely, for SSPP, the hyperparameters consist of $\Theta := (\sigma, \boldsymbol{\ell}, \beta)$, while for GSSPP, $\Theta := (\{\sigma_k\}_{k=1}^K, \{\boldsymbol{\omega}_k\}_{k=1}^K, \{\boldsymbol{\gamma}_k\}_{k=1}^K, \beta)$.

The (non-log) likelihood function in Equation (3.22) therefore becomes

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{w}^{(r)}, \Theta) = \\ &- \mathbf{w}^{(r)\top}\mathbf{M}^{(r)}\mathbf{w}^{(r)} - 2\beta\ \mathbf{w}^{(r)\top}\mathbf{m}^{(r)} - \beta^2|B| \\ &+ \sum_{i=1}^{N} \log(|\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i) + \beta|^2). \end{aligned} \tag{4.17}$$

where $\mathbf{M}^{(r)}$ and $\mathbf{m}^{(r)}$ are the matrix and vector terms from proposition 4.3.3. We can compute the log-likelihood function in $\mathcal{O}(r^2N)$, i.e. linearly in N. The log of the joint-distribution over $\mathbf{w}^{(r)}$ and $\mathbf{X}$ is then

$$\begin{aligned}\log p(\mathbf{w}^{(r)}, \mathbf{X}|\Theta) &= \log p(\mathbf{X}|\mathbf{w}^{(r)}, \Theta) + \log p(\mathbf{w}^{(r)}) \\ &= \log p(\mathbf{X}|\mathbf{w}^{(r)}, \Theta) - \frac{1}{2}\mathbf{w}^{(r)\top}\mathbf{w}^{(r)} + C \end{aligned} \tag{4.18}$$

for some constant C, where $p(\mathbf{w}^{(r)}) = \mathcal{N}(\mathbf{w}^{(r)}|\mathbf{0}, \mathbf{I}_r)$ denotes the prior distribution over $\mathbf{w}^{(r)}$.

4.4.1 Laplace approximation

The latent posterior $p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$ induced from Equation (4.18) is approximated using Laplace's method. In this context, Laplace's method refers to a Gaussian approximation obtained from a second-order Taylor expansion of $\log p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$ around the mode of the posterior. This yields

$$\begin{aligned} p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta) &\approx \mathcal{N}(\mathbf{w}^{(r)}|\hat{\mathbf{w}}^{(r)}, \mathbf{Q}) \\ &:= q(\mathbf{w}^{(r)}|\mathbf{X}, \Theta) \end{aligned} \tag{4.19}$$

where $\hat{\mathbf{w}}^{(r)} := \arg\max_{\mathbf{w}^{(r)}} p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$ is the mode of the latent posterior and $\mathbf{Q}$ is chosen to be the negative inverse Hessian of the true posterior at that point.

The gradient and the Hessian of the true posterior with respect to $\mathbf{w}^{(r)}$ are

$$\begin{aligned}\nabla_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}|\mathbf{X},\Theta) = &- (2\mathbf{M}^{(r)} + \mathbf{I}_r)\mathbf{w}^{(r)} \\ &- 2\beta\ \mathbf{m}^{(r)} + 2\sum_{i=1}^{N} \frac{\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)}{{\mathbf{w}^{(r)}}^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i) + \beta} \\ \nabla^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}|\mathbf{X},\Theta) = &- (2\mathbf{M}^{(r)} + \mathbf{I}_r) \\ &- 2\sum_{i=1}^{N} \frac{\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)^\top}{({\mathbf{w}^{(r)}}^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i) + \beta)^2}.\end{aligned}$$

The mode $\hat{\mathbf{w}}^{(r)}$ must satisfy the stationary constraint

$$\nabla_{\mathbf{w}^{(r)}}\ \log p(\mathbf{w}^{(r)}|\mathbf{X},\Theta)|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}} = 0,$$

that implies

$$\hat{\mathbf{w}}^{(r)} = \left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1} \left(\sum_{i=1}^{N} \frac{\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)}{{\mathbf{w}^{(r)}}^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i) + \beta} - \beta\ \mathbf{m}^{(r)}\right). \tag{4.20}$$

Equation (4.20) cannot be solved analytically. Instead, we estimate $\hat{\mathbf{w}}^{(r)}$ iteratively using Newton-Raphson method, with step

$$\mathbf{w}^{(r)^{\text{new}}} = \mathbf{w}^{(r)} - (\nabla^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}))^{-1} \nabla_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}). \tag{4.21}$$

The precision matrix $\mathbf{Q}^{-1}$ is then given by $-\nabla^2_{\mathbf{w}^{(r)}}\ \log p(\mathbf{w}^{(r)}|\mathbf{X},\Theta)|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}}$.

4.4.2 Model selection

We first derive a marginal likelihood approximation similar to Walder and Bishop (2017, Section 4.1.6).

$$\begin{aligned}
&\log p(\mathbf{X}|\Theta) \\
&\quad = \log p(\hat{\mathbf{w}}^{(r)}, \mathbf{X}|\Theta) - \log p(\hat{\mathbf{w}}^{(r)}|\mathbf{X}, \Theta) \\
&\quad \approx \log p(\hat{\mathbf{w}}^{(r)}, \mathbf{X}|\Theta) - \log q(\hat{\mathbf{w}}^{(r)}|\mathbf{X}, \Theta) \\
&\quad = -\hat{\mathbf{w}}^{(r)^\top}\mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)} - 2\beta\ \hat{\mathbf{w}}^{(r)^\top}\mathbf{m}^{(r)} - \beta^2|B| \\
&\qquad + \sum_{i=1}^{N} \log(|\hat{\mathbf{w}}^{(r)^\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i) + \beta|^2) \\
&\qquad - \frac{1}{2}\hat{\mathbf{w}}^{(r)^\top}\hat{\mathbf{w}}^{(r)} + \frac{1}{2}\log|\mathbf{Q}| + \frac{N}{2}\log(2\pi)
\end{aligned} \tag{4.22}$$

since the quadratic term of $\log q(\hat{\mathbf{w}}^{(r)}|\mathbf{X}, \Theta)$ cancels out.

We tune the hyperparameters Θ by maximizing Equation (4.22). The model selection is facilitated by the fact that the gradient of the marginal likelihood in Equation (4.22) with respect to Θ can be easily expressed. The terms $\mathbf{M}^{(r)}$, $\mathbf{m}^{(r)}$ and $\boldsymbol{\varphi}^{(r)}(\cdot)$ are functions of the hyperparameters Θ. The mode $\hat{\mathbf{w}}^{(r)}$ is also a function of Θ.

The partial derivatives of the marginal likelihood with respect to Θ is obtained using the chain rule,

$$\begin{aligned}
\nabla_{\Theta_i} \log p(\mathbf{x}|\Theta) = &\left.\frac{\partial \log p(\mathbf{x}|\Theta)}{\partial \Theta_i}\right|_{explicit} \\
&+ \sum_{j=1}^{r} \frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \hat{w}_j} \frac{\partial \hat{w}_j}{\partial \Theta_i}.
\end{aligned} \tag{4.23}$$

Expressions for the terms $\frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \Theta_j}$, $\frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \hat{w}_j}$ and $\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_i}$ above are given in the next section, requiring a full mode search within each iterative hyperparameters update. In the current work, we note that assuming $\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_i} = 0$ and alternating independent updates for the mode in Equation (4.21) and the hyperparameters in Equation (4.23) provides faster and yet acceptable results.

Algorithm 2 Compute the mode $\hat{\mathbf{w}}^{(r)}$ and the hyperparameters Θ

input: data $\mathbf{X}$
initialize Θ_0 and $\hat{\mathbf{w}}_0^{(r)}$
for $t = 1$ **to** T **do**
  compute $\mathbf{M}^{(r)}$ and $\boldsymbol{\varphi}^{(r)}(\mathbf{X})$
  $\hat{\mathbf{w}}_t^{(r)} := \text{mode}(p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta), \hat{\mathbf{w}}_{t-1}^{(r)})$ {locate posterior mode using Equation (4.21) with initial value $\hat{\mathbf{w}}_{t-1}^{(r)}$}
  $\hat{\mathbf{f}} := \boldsymbol{\varphi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}_t^{(r)} + \beta$
  $\mathbf{V} := \text{diag}(\hat{\mathbf{f}}^{-1})\boldsymbol{\varphi}^{(r)}(\mathbf{X})$
  $\mathbf{Q}^{-1} := 2\mathbf{M}^{(r)} + I_r + 2\mathbf{V}^\top\mathbf{V}$ {precision matrix $\mathbf{Q}^{-1} = -\nabla^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}}$}
  compute the gradient $\mathbf{g}$ from Algorithm 8
  $\Theta_t \leftarrow \text{update}(\Theta_{t-1}, \mathbf{g})$
end for
return $\hat{\mathbf{w}}_T^{(r)}$ (mode) and Θ_T (hyperparameters)

Marginal likelihodd derivatives

We now express the gradient of the marginal likelihood $\log p(\mathbf{X}|\Theta)$ with respect to the hyperparameters Θ. Using the chain rule,

$$\nabla_{\Theta_i} \log p(\mathbf{X}|\Theta) = \frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \Theta_i}\Big|_{explicit} + \sum_{j=1}^{r} \frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \hat{w}_j} \frac{\partial \hat{w}_j}{\partial \Theta_i}. \tag{4.24}$$

The first term of Equation (4.24) can be solved as

$$\frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \Theta_j}\Big|_{explicit} = -\hat{\mathbf{w}}^{(r)\top} \frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_j} \hat{\mathbf{w}}^{(r)} + 2\sum_{i=1}^{N} \frac{\hat{\mathbf{w}}^{(r)\top} \frac{\partial \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)}{\partial \Theta_j}}{\hat{\mathbf{w}}^{(r)\top} \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)} + \frac{1}{2}\text{tr}(\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \Theta_j}) \tag{4.25}$$

where $\mathbf{Q}^{-1} = -\nabla^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}}$ is the precision matrix expressed in

Algorithm 3 Compute the mode $\hat{\mathbf{w}}^{(r)}$ and the hyperparameters Θ with independent updates

1: **input:** data $\mathbf{X}$
2: initialize Θ_0 and $\hat{\mathbf{w}}_0^{(r)}$
3: **for** $t = 1$ **to** T **do**
4: compute $\mathbf{M}^{(r)}$ and $\boldsymbol{\varphi}^{(r)}(\mathbf{X})$
5: $\hat{\mathbf{w}}_t^{(r)} \leftarrow \hat{\mathbf{w}}_t^{(r)}$ {update the posterior mode using one iteration of Equation (4.21)}
6: $\hat{\mathbf{f}} := \boldsymbol{\varphi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}_t^{(r)} + \beta$
7: $\mathbf{V} := \mathrm{diag}(\hat{\mathbf{f}}^{-1})\boldsymbol{\varphi}^{(r)}(\mathbf{X})$
8: $\mathbf{Q}^{-1} := 2\mathbf{M}^{(r)} + I_r + 2\mathbf{V}^\top\mathbf{V}$ {precision matrix }
9: compute the gradient $\frac{d\mathbf{p}}{d\Theta}$ from Algorithm 8 (line 18)
10: $\Theta_t \leftarrow \mathrm{update}(\Theta_{t-1}, \frac{d\mathbf{p}}{d\Theta})$
11: **end for**
12: **return** $\hat{\mathbf{w}}_T^{(r)}$ (mode) and Θ_T (hyperparameters)

Section 4.4.1. The last term of Equation (4.25) can be expressed as

$$\begin{aligned}
\frac{1}{2}\mathrm{tr}(\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \Theta_j}) &= -\frac{1}{2}\mathrm{tr}(\frac{\partial \mathbf{Q}^{-1}}{\partial \Theta_j}\mathbf{Q}) \\
&= -\mathrm{tr}\left(\frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_j}\mathbf{Q} + \frac{\partial}{\partial \Theta_j}\left[\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_i)^\top}{(\mathbf{w}^\top\boldsymbol{\varphi}(\mathbf{x}_i))^2}\right]\mathbf{Q}\right) \\
&= -\mathrm{tr}\left(\frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_j}\mathbf{Q}\right) - \sum_{i=1}^{N}\frac{2}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i))^2}\left[\boldsymbol{\varphi}(\mathbf{x}_i)^\top\mathbf{Q}\frac{\partial\, \boldsymbol{\varphi}(\mathbf{x}_i)}{\partial \Theta_j}\right] \\
&\quad + 2\sum_{i=1}^{N}\frac{\hat{\mathbf{w}}^{(r)\top}\frac{\partial \boldsymbol{\varphi}(\mathbf{x}_i)}{\partial \theta_j}}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i))^3}\left[\boldsymbol{\varphi}(\mathbf{x}_i)^\top\mathbf{Q}\,\boldsymbol{\varphi}(\mathbf{x}_i)\right] \qquad (4.26)
\end{aligned}$$

The $\frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \hat{w}_j}$ terms of Equation (4.24) is

$$\begin{aligned}
\frac{\partial \log p(\mathbf{X}|\Theta)}{\partial \hat{w}_j} &= -\underbrace{\frac{\partial \log p(\hat{\mathbf{w}}^{(r)}, \mathbf{X}|\Theta)}{\partial \hat{w}_j}}_{=0} + \frac{1}{2}\mathrm{tr}(\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \hat{w}_j}) \\
&= -\frac{1}{2}\mathrm{tr}(\frac{\partial\, \mathbf{Q}^{-1}}{\partial \hat{w}_j}\mathbf{Q}) \\
&= -\mathrm{tr}\left(\frac{\partial}{\partial \hat{w}_j}\left[\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_i)^\top}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i))^2}\right]\mathbf{Q}\right) \\
&= 2\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}_j(\mathbf{x}_i)}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i))^3}\left[\boldsymbol{\varphi}(\mathbf{x}_i)^\top\mathbf{Q}\,\boldsymbol{\varphi}(\mathbf{x}_i)\right] \qquad (4.27)
\end{aligned}$$

Algorithm 4 Compute the log marginal likelihood derivatives

input: $\mathbf{X}$ (inputs), spectral locations and hyper-parameters Θ, posterior mode $\hat{\mathbf{w}}^{(r)}$, precision matrix $\mathbf{Q}^{-1}$, $r \times r$ "integral" matrix $\mathbf{M}^{(r)}$, $N \times r$ features matrix $\boldsymbol{\varphi}^{(r)}(\mathbf{X})$

compute $\nabla \mathbf{M}^{(r)}$, the $\dim(\tilde{\Theta}) \times r \times r$ tensor of partial derivatives of $\mathbf{M}^{(r)}$ with respect to $\tilde{\Theta}$

compute $\nabla \boldsymbol{\varphi}^{(r)}(\mathbf{X})$, the $\dim(\tilde{\Theta}) \times N \times r$ tensor of partial derivatives of $\boldsymbol{\varphi}^{(r)}(\mathbf{X})$ with respect to $\tilde{\Theta}$

$\hat{\mathbf{f}} := \boldsymbol{\varphi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}^{(r)}$

$\mathbf{L} = \text{cholesky}(\mathbf{Q}^{-1})$ {Solve $\mathbf{L}\mathbf{L}^\top = \mathbf{Q}^{-1}$}

$\mathbf{C}_\varphi := \mathbf{L} \setminus \boldsymbol{\varphi}^{(r)}(\mathbf{X})$

$\mathbf{C}_{\nabla\varphi} := \mathbf{L} \setminus \nabla\boldsymbol{\varphi}^{(r)}(\mathbf{X})$

$\mathbf{r}_\varphi := (\mathbf{C}_\varphi * \mathbf{C}_\varphi)\mathbf{1}_r$

$\mathbf{R}_{\nabla\varphi} = (\mathbf{C}_\varphi * \mathbf{C}_{\nabla\varphi})\mathbf{1}_r$

$\mathbf{S} := \nabla\boldsymbol{\varphi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}^{(r)}$

$\mathbf{E} := (\mathbf{S}\ \text{diag}(\hat{\mathbf{f}}^{-2})) * \boldsymbol{\varphi}^{(r)}(\mathbf{X})$.

$\mathbf{s}_{integral} = \hat{\mathbf{w}}^{(r)\top} \nabla\mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)}$

$\mathbf{s}_{data} = 2\ (\mathbf{S}\ \text{diag}(\hat{\mathbf{f}}^{-1}))\mathbf{1}_N$

$\mathbf{s}_1 = \text{tr}(\mathbf{L}^\top \setminus (\mathbf{L} \setminus \nabla\mathbf{M}^{(r)}))$

$\mathbf{s}_2 = 2\ (\mathbf{R}_{\nabla\varphi}\ \text{diag}(\hat{\mathbf{f}}^{-2}))\mathbf{1}_N$

$\mathbf{s}_3 = 2(\mathbf{S}\ \text{diag}(\hat{\mathbf{f}}^{-3}))\mathbf{r}_\varphi$

$\mathbf{v} := \text{diag}(\hat{\mathbf{f}}^{-1})\nabla\boldsymbol{\varphi}^{(r)}(\mathbf{X}) - \mathbf{1}_N^\top \mathbf{E}$ {Equation (4.28)}

$\frac{d\mathbf{p}}{d\tilde{\Theta}} := \mathbf{s}_{data} - \mathbf{s}_{integral} - \mathbf{s}_1 - \mathbf{s}_2 + \mathbf{s}_3$ { Equation (4.25)}

$\frac{d\mathbf{p}}{d\mathbf{w}} := 2\boldsymbol{\varphi}^{(r)}(\mathbf{X})^\top \text{diag}(\hat{\mathbf{f}}^{-3})\mathbf{r}_\varphi$ {Equation (4.27)}

$\frac{d\mathbf{W}}{d\tilde{\Theta}} := 2(\ \mathbf{L} \setminus (\mathbf{L}^\top \setminus (\nabla\mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)} - \mathbf{v})))$ {Equation (4.29)}

$\mathbf{g} = \frac{d\mathbf{p}}{d\tilde{\boldsymbol{\Theta}}} + \frac{d\mathbf{W}}{d\tilde{\Theta}} \frac{d\mathbf{p}}{d\mathbf{w}}$ {Equation (4.24)}

return $\mathbf{g}$ ($\dim(\tilde{\Theta})$-vector of partial derivatives)

where, in the first line, we have imposed stationarity using $\nabla_{\mathbf{w}^{(r)}}\ \log p(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}} = 0$.

We finally express the last terms $\frac{\partial\hat{\mathbf{w}}^{(r)}}{\partial\Theta_i}$ of Equation (4.24). From the expression

of $\hat{\mathbf{w}}^{(r)}$ in Equation (4.20). We obtain

$$\begin{aligned}
\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_j} &= \frac{\partial}{\partial \Theta_j}\left[\left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1}\left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i)}\right)\right] \\
&= \frac{\partial}{\partial \Theta_j}\left[\left(\mathbf{M}^{(r)} + \frac{1}{2}I_r\right)^{-1}\right]\left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i)}\right) \\
&\quad + \left(\mathbf{M}^{(r)} + \frac{1}{2}I_r\right)^{-1}\left[\frac{\partial}{\partial \Theta_j}\left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i)}\right) + \frac{\partial}{\partial \hat{\mathbf{w}}^{(r)}}\left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i)}\right)\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_j}\right] \\
&= -\left(\mathbf{M}^{(r)} + \frac{1}{2}I_r\right)^{-1}\left[\frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_i}\,\hat{\mathbf{w}}^{(r)} - \,\hat{\mathbf{v}} - \left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_i)^\top}{(\mathbf{w}^\top\boldsymbol{\varphi}(\mathbf{x}_i))^2}\right)\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_i}\right]
\end{aligned}$$

where

$$\hat{\mathbf{v}} := \frac{\partial}{\partial \Theta_j}\left(\sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i)}\right) = \sum_{i=1}^{N}\frac{\frac{\partial \boldsymbol{\varphi}(\mathbf{x}_i)}{\partial \Theta_j}}{\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i)} - \sum_{i=1}^{N}\frac{\hat{\mathbf{w}}^{(r)\top}\frac{\partial \boldsymbol{\varphi}(\mathbf{x}_i)}{\partial \Theta_j}}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}(\mathbf{x}_i))^2}\,\boldsymbol{\varphi}(\mathbf{x}_i) \quad (4.28)$$

Thus

$$\begin{aligned}
\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \Theta_i} &= -\left[\mathbf{M}^{(r)} + \frac{1}{2}I_r + \sum_{i=1}^{N}\frac{\boldsymbol{\varphi}(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_i)^\top}{(\mathbf{w}^\top\boldsymbol{\varphi}(\mathbf{x}_i))^2}\right]^{-1}\left[\frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_i}\,\hat{\mathbf{w}}^{(r)} - \,\hat{\mathbf{v}}\right] \\
&= -2\mathbf{Q}\left[\frac{\partial \mathbf{M}^{(r)}}{\partial \Theta_i}\,\hat{\mathbf{w}}^{(r)} - \,\hat{\mathbf{v}}\right]
\end{aligned} \quad (4.29)$$

Implementation Details The implementation of the marginal log likelihood partial derivatives with respect to the hyperparameters $\tilde{\Theta}$ is shown in Algorithm 4. This implementation has been preferred over an autodiff method for performance reasons. However, the output has been thoroughly validated against an autodiff approach.

4.5 Predictive distribution

To form predictive distributions, we assume that the latent posterior is approximated by $q(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$ as in Equation (4.19).

4.5.1 Predictive intensity distribution

For some $\mathbf{x}^* \in B$, the predictive distribution of $f(\mathbf{x}^*)$ can be deduced from Equations (4.13) and (4.19) to be

$$f(\mathbf{x}^*)|\mathbf{X}, \Theta \sim \mathcal{N}(f(\mathbf{x}^*)|\mu^*(\mathbf{x}^*), \sigma^*(\mathbf{x}^*)) \tag{4.30}$$

where

$$\mu^*(\mathbf{x}^*) := \hat{\mathbf{w}}^{(r)^\top} \boldsymbol{\varphi}^{(r)}(\mathbf{x}^*) \tag{4.31}$$

and

$$\sigma^*(\mathbf{x}) := \boldsymbol{\varphi}^{(r)}(\mathbf{x}^*)^\top \mathbf{Q} \boldsymbol{\varphi}^{(r)}(\mathbf{x}^*). \tag{4.32}$$

Given $\lambda(\cdot) = (f(\cdot) + \beta)^2$ and Equation (4.30) above, we can also derive the predictive distribution of the intensity function

$$\lambda(\mathbf{x}^*)|\mathbf{X}, \Theta \sim \mathit{Gamma}(\lambda(\mathbf{x}^*)|a^*(\mathbf{x}^*), b^*(\mathbf{x}^*)) \tag{4.33}$$

with parameters

$$a^*(\mathbf{x}^*) = \frac{\left(\mu^*(\mathbf{x}^*)^2 + \sigma^*(\mathbf{x}^*)^2\right)^2}{2\sigma^*(\mathbf{x}^*)^2 \left(2\mu^*(\mathbf{x}^*)^2 + \sigma^*(\mathbf{x}^*)^2\right)} \mathit{and} \tag{4.34}$$

$$b^*(\mathbf{x}^*) = \frac{\mu^*(\mathbf{x}^*)^2 + \sigma^*(\mathbf{x}^*)^2}{2\sigma^*(\mathbf{x}^*)^2 \left(2\mu^*(\mathbf{x}^*)^2 + \sigma^*(\mathbf{x}^*)^2\right)}. \tag{4.35}$$

4.5.2 Predictive expected log-likelihood

For a training set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and an held-out test set $\mathbf{X}^* = \{\mathbf{x}_i^*\}_{i=1}^{N^*}$, we can derive from Equation (3.22), an approximation for the expected predictive log-likelihood

$$\begin{aligned} \mathbb{E}\left[\log p(\mathbf{X}^*|\mathbf{X})\right] \approx\ & \mathbb{E}_{\mathbf{w}^{(r)}}\left[-\int_B (\mathbf{w}^{(r)^\top} \boldsymbol{\varphi}^{(r)}(\mathbf{x}) + \beta)^2 d\mathbf{x}\right] \\ & + \sum_{i=1}^{N^*} \mathbb{E}_{\mathbf{w}^{(r)}}\left[\log(\mathbf{w}^{(r)^\top} \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i^*) + \beta)^2)\right] \end{aligned}$$

where $\mathbf{w}^{(r)} \sim q(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$. The expectation over the integral term can be solved analytically. The sum-of-expectation can be expressed using Pochhammer series, that we approximate in practise by interpolation of a look-up table of precomputed values. This is very similar to Lloyd et al. (2015, section 4.3). We provide more details in Section C.4 of Appendix C.

4.6 Experiments

We benchmark the SSPP scheme introduced in Sections 4.4 and 4.5 against a Nyström-based implementation of the LBPP scheme of Walder and Bishop (2017), a frequentist kernel smoothing approach with edge correction (KS) (Diggle, 1985) and the variational inference scheme for point processes proposed by Lloyd et al. (2015) (VBPP). We test the algorithms on three 1D synthetic data sets and three real data sets (one in 1D and two in 2D). Typically, experiments conducted under the SSPP scheme do not directly assess the stability of the RFF due to the absence of controls fixing a dataset while rerunning the RFF. However, observing consistent gains across multiple datasets can be considered indicative of its stability.

4.6.1 Benchmarks settings

To benchmark the algorithms, we use the following settings. Our KS implementation uses standard kernel density estimation with truncated normal kernels to account for domain knowledge. The kernel bandwidth parameter is estimated via grid search using the leave-one-out log average likelihood objective of Lloyd et al. (2015). We used a publicly-available implementation of VBPP (https://github.com/st–/vbpp). We adopt fixed inducing points on a grid over $|B|$. For consistency, we also used a constant offset β for both LBPP and VBPP implementations.

4.6.2 Performance metrics

The average *test expected log-likelihood* $\mathcal{L}_{\text{test}} := \mathbb{E}[\log p(\mathbf{X}^*|\mathbf{X})]$ is used as an evaluation metric. This is generally difficult to compute for point process models, but is available for SSPP and LBPP (see Section 4.5). For the synthetic experiment we

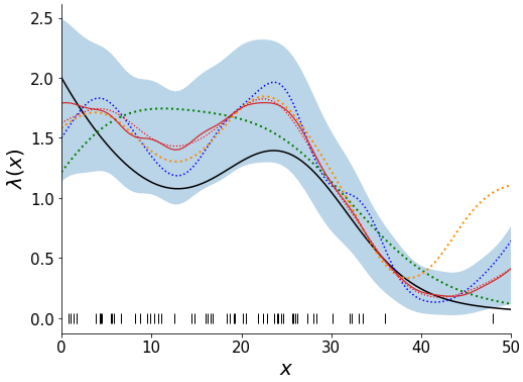
also consider the normalized $\mathcal{L}_2$ norm to the known ground truth intensity function i.e. $\mathcal{L}_2 := \frac{1}{|B|}\left(\int_B \mathbb{E}[(\lambda^*(\mathbf{x}) - \lambda_{\text{truth}}(\mathbf{x}))^2]\, d\mathbf{x}\right)^{\frac{1}{2}}$ where $\lambda^*(\mathbf{x})$ denotes the predictive intensity.

4.6.3 Synthetic dataset

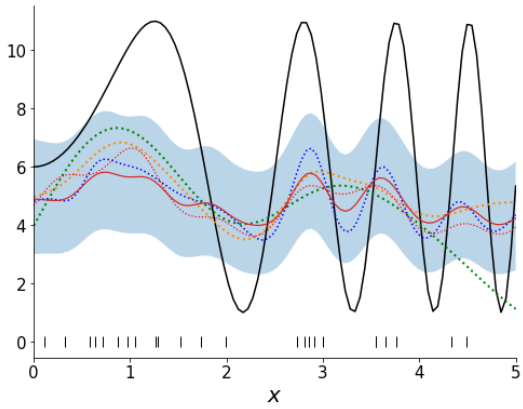
Three 1D simulated examples from Adams et al. (2009) are considered. The corresponding intensities are defined as $\lambda_1(x) = 2\exp(-x/15) + \exp(-((x-25)/10)^2)$ on the interval $[0, 50]$ for approximately 47 events per sample, $\lambda_2(x) = 5\sin(x^2) + 6$ on the interval $[0, 5]$ for approximately 36 events per sample and $\lambda_3(x)$ is a piecewise linear function shown in Figure 4.2 on the interval $[0, 100]$ for approximately 225 events per sample.

These intensity functions have been considered previously in the context of Gaussian Cox process by Samo and Roberts (2015b), Donner and Opper (2018b), John and Hensman (2018) and Aglietti et al. (2019). We train the models on 10 independent samples generated from the ground truth, and evaluate the performance of each using 50 test sets sampled independently from the ground truth. We use the acronyms GSSPP-SE and GSSPP-m(ν) to refer to the generalized kernel variants of SSPP as in Equation (4.10) with g set to be the Gaussian kernel and Matérn kernel with parameter ν respectively.

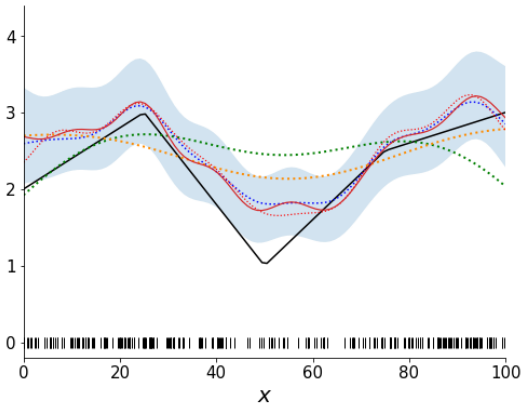
We report optimal performance across models for sets of spectral points or inducing points of size (denoted by p) ranging from 15 to 100. Results are given in Table 4.2 and the mean predictive intensities displayed in Figure 4.2. GSSPP and SSPP outperform the other methods in terms of both $\mathcal{L}_{\text{test}}$ and $\mathcal{L}_2$ for λ_1 and λ_2. Compared to LBPP, GSSPP and SSPP perform better consistently, but with slightly increase execution times. In a similar manner to findings for VBPP, GSSPP and SSPP fitting remains up to three orders of magnitude faster than alternative MCMC-based methods (Adams et al., 2009), see Lloyd et al. (2015) for comparison.



(a) $\lambda_1(x)$



(b) $\lambda_2(x)$



(c) $\lambda_3(x)$

Figure 4.2: Mean predictive intensity of the three toy intensity functions λ_1, λ_2 and λ_3 defined as in Adams et al. (2009). Solid colored lines represent the predictive mean. The solid black lines shows the ground truth. The shaded areas are the 80% credible region of the GSSPP model.

Table 4.2: Performance of GSSPP, SSPP, KS, VBPP and LBPP schemes on three samples of synthetic data. Values in bold-face refer to best performance, which corresponds to lower values of $\mathcal{L}_2$, but higher values of $\mathcal{L}_{test}$.

	$\lambda_1(x)$			$\lambda_2(x)$			$\lambda_3(x)$		
	$\mathcal{L}_2$	$\mathcal{L}_{\text{test}}$	time(s)	$\mathcal{L}_2$	$\mathcal{L}_{\text{test}}$	time(s)	$\mathcal{L}_2$	$\mathcal{L}_{\text{test}}$	time(s)
GSSPP-SE	0.74	105.92	1.32	0.83	52.62	1.59	1.62	**842.05**	2.94
GSSPP-m12	0.71	104.23	1.88	0.86	48.58	1.42	1.98	838.17	2.54
GSSPP-m32	**0.68**	106.05	1.79	0.78	51.65	1.56	1.75	840.80	2.53
GSSPP-m52	0.69	**106.12**	1.71	0.84	52.77	1.59	1.68	841.77	2.32
SSPP	0.78	105.19	0.63	**0.74**	56.01	0.69	**1.60**	835.35	1.05
KS	1.10	102.49	0.09	0.89	**58.07**	0.08	3.22	834.68	0.19
VBPP	0.72	104.65	1.98	0.85	51.11	1.15	1.63	838.65	1.52
LBPP	0.81	103.29	0.06	0.96	51.76	0.07	1.87	833.31	0.13

4.6.4 Real datasets

In the following sections, we present the results of applying our proposed method to the real-world datasets.

Coal data set The classic *coal mine accidents* data set is a well-known collection of 191 fatal coal-mining accidents that occurred in Britain between March 15, 1875, and March 22, 1962. Originally compiled by MaguireE et al. (1952), the data set was later updated and refined by Jarrett (1979) in a revised format. This data set has been widely used to study various problems, such as changepoint analysis and nonhomogeneous Poisson processes (Fearnhead, 2006a; Carlin et al., 1992; Adams et al., 2009; Lloyd et al., 2015). Researchers have noted that the frequency of coal mine accidents varied over time due to changes in safety regulations and historical developments.

Our analysis revealed a clear correlation between the historical introduction of safety regulations and the inferred intensity of coal mine disasters. Specifically, we observed a significant decline in the rate of disasters from 1870 to 1890, coinciding

with the passage of several acts by the UK parliament to improve mine worker safety. Notably, these acts included the Coal Mines Regulation Acts of 1872 and 1887. Similarly, we found that the inferred intensity of disasters decreased after 1950, which could be attributed to the implementation of further safety regulations under the Mines and Quarries Act of 1954. Previous studies on this data set have also reported similar findings (Fearnhead, 2006a; Carlin et al., 1992; Lloyd et al., 2015).

For this data set, we evaluate predictive performance for the competing inference schemes using 100 random partitions of the sample into train and test subsets ($\mathbf{X}$ and $\mathbf{X}^*$) of approximately equal size. Figure 4.3 shows the predicted mean intensity with credible intervals. Results are presented in Table 4.3.

Table 4.3: Results on Coal data experiment with standard errors in brackets.

	Coal data (1D)	
	$\mathcal{L}_{\text{test}}$	time(s)
GSSPP-SE	**224.44** (±0.57)	1.56
GSSPP-m12	220.80 (±0.85)	1.84
GSSPP-m32	224.25 (±0.55)	1.58
GSSPP-m52	223.84 (±0.54)	1.25
SSPP	221.23 (±0.86)	0.64
KS	219.50 (±0.33)	0.11
VBPP	221.19 (±1.34)	1.75
LBPP	218.68 (±0.87)	0.16

Bei data set The *Bei* data set contains information about the location of 3605 trees of the species *Beilschmiedia pendula* (Lauraceae) in a 1000 by 500 meter rectangular sampling region in the tropical rainforest of Barro Colorado Island. This data set is part of a much larger data set that includes hundreds of thousands of trees belonging to thousands of species (Hubbell and Foster, 1983). The Bei data set has already been

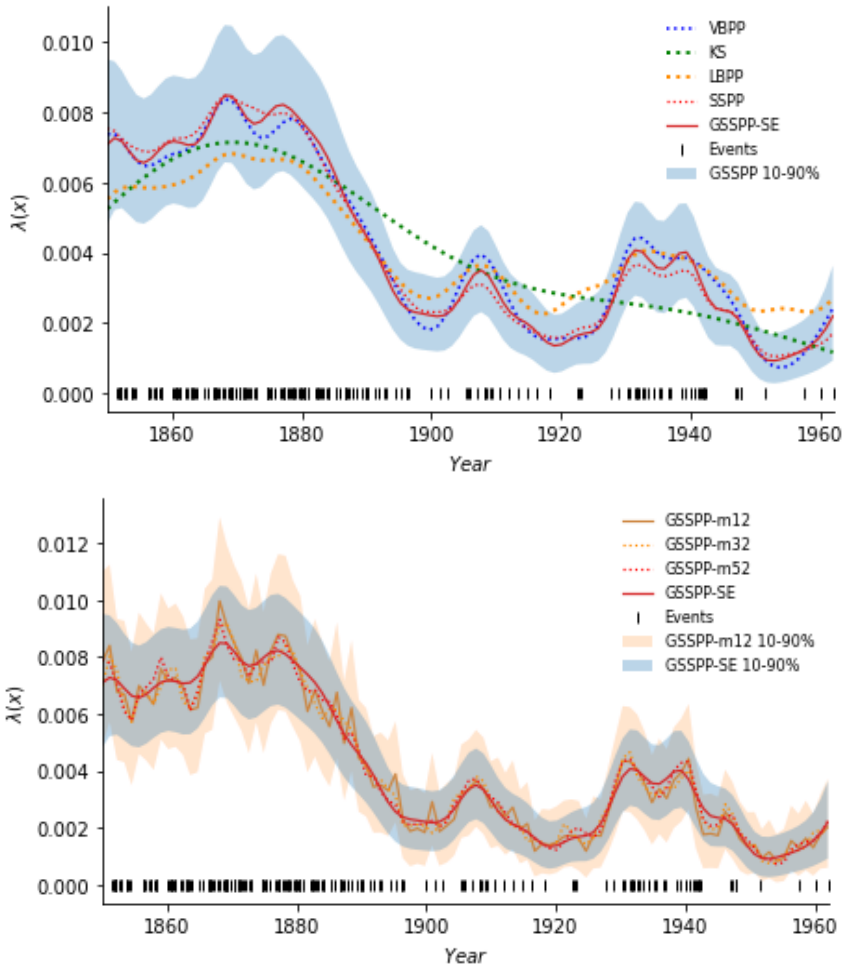
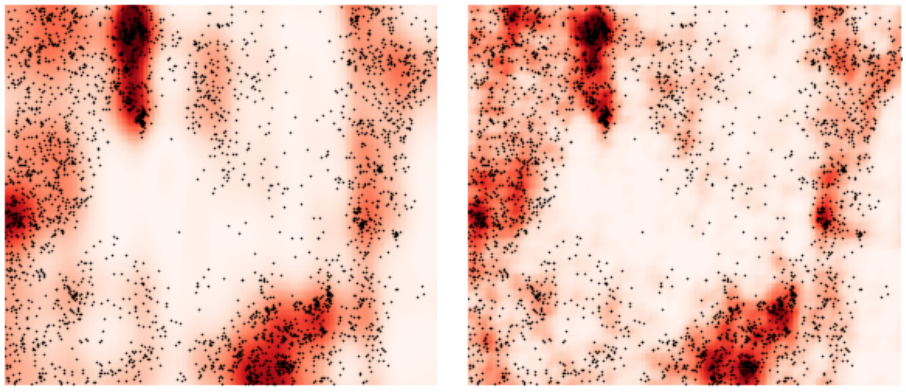


Figure 4.3: Predictive mean intensity for the coal mine accident, with highest 80% credible intervals.

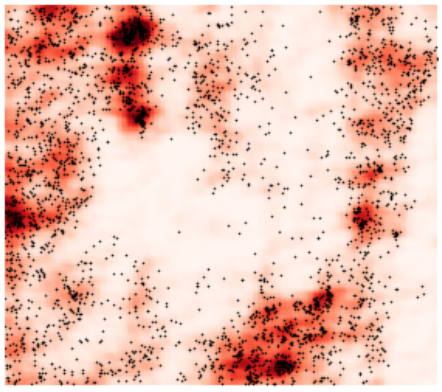
used by Møller and Waagepetersen (2007) in the context of point process modeling. The information contained in the Bei data set is a valuable resource for researchers studying tropical rainforest ecology, as it can be used to test hypotheses about the factors that influence the spatial distribution of a single species of tree in a highly diverse ecosystem

For these data, we evaluate predictive performance using 100 random partitions of the original sample into train and test subsamples of approximately equal size, with p now ranging from 15 to 150. Table 4.4 presents the results. An illustration of a single fit is provided in Figure 4.4. Additionally, Figure 4.8 shows the performances

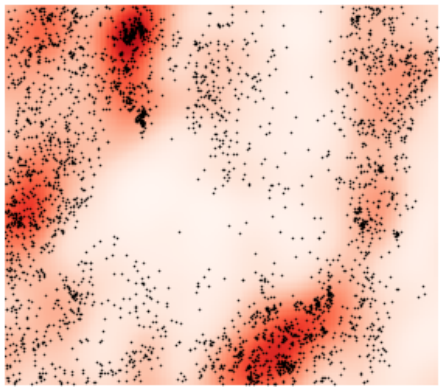


(a) GSSPP-SE

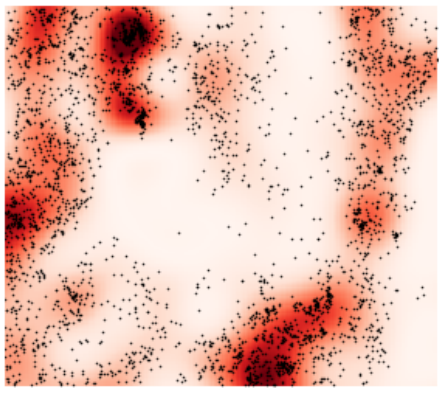
(b) GSSPP-m32



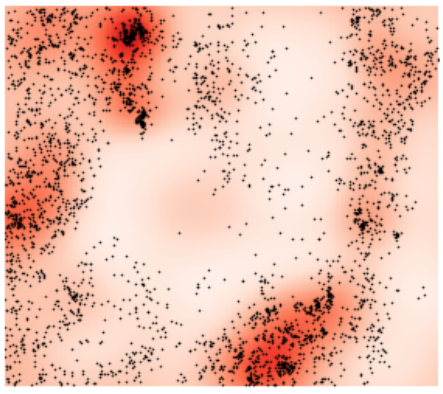
(c) GSSPP-m12



(d) SSPP



(e) VBPP



(f) LBPP

Figure 4.4: Heat map of the predictive mean intensity with $p = 150$ for the *Bei* data set. The black dots are the 3605 input data points.

Table 4.4: Results on Bei data experiment with standard errors in brackets.

	Beil data (2D)	
	$\mathcal{L}_{\text{test}}$	time(s)
GSSPP-SE	763.49 (±3.81)	20.48
GSSPP-m12	760.82 (±4.31)	20.31
GSSPP-m32	**764.73** (±2.60)	20.85
GSSPP-m52	763.82 (±1.00)	20.55
SSPP	751.50 (±2.55)	17.54
KS	735.78 (±1.49)	4.22
VBPP	757.95 (±3.14)	28.41
LBPP	711.72 (±1.35)	1.35

per number of spectral points or inducing points.

Porto Taxi data set The *Porto Taxi* data set (Moreira-Matias et al., 2013) is the third data set that we analyze in this study. It consists of 1.7×10^6 taxi journey trajectories that took place in Porto, Portugal, between 2013 and 2014, including pickup and drop-off locations and timestamps. Although Porto has two taxi companies, each operating a fleet, the authors of the data set used information only from the largest company, which had 441 registered taxis equipped with mobile data terminals.

Porto is a medium-sized urban area with a population of 1.3 million, and according to a recent aerial survey of the road traffic of the city (Moreira-Matias et al., 2013; Ferreira et al., 2009), the number of vacant taxis is greater than the passenger demand, resulting in intense competition between the two companies and drivers. The regulations in place prohibit drivers from randomly searching for passengers and require them to select one of the 63 available taxi stands in the city and wait for the next service immediately after dropping off their last passenger. A map of the stand spatial distribution is presented in the Figure 4.5. The data set provides valu-

Figure 4.5: Taxi Stand spatial distribution over the city of Porto, Portugal. Graph presented is obtained from Moreira-Matias et al. (2013, section 4).

able insights into the urban mobility patterns of the city. Researchers have used this data set to explore various aspects of urban mobility. For example, Aglietti et al. (2019) and John and Hensman (2018) leveraged the data set to test their Cox process methodology.

We consider the pick-up locations as observations of a point process. As in Aglietti et al. (2019), we restrict the analysis to 7000 events selected with (latitude, longitude) pairs bounded by $(41.147, -8.58)$ and $(41.18, -8.65)$. We select 1400 events at random as training set and use the rest as testing set. We set p ranging from 15 to 200 Table 4.5 presents the results. Figure 4.7 illustrates a single fit to the full data set for four models. Figure 4.8 shows the performances per number of spectral points or inducing points.

For each of the real-data applications, GSSPP performs best. The average fitting time using GSSPP is comparable to that of VBPP. Figure 4.3 shows the effect of different choice of GSSPP kernel function g. When choosing g to be a Matérn kernel, with parameter ν, spectral points $\mathbf{z}$ are drawn from a Student-t distribution with 2ν degree of freedom as discussed in Section C.2 of Appendix C. Compared to GSSPP-SE, these models tend to produce feature mappings with larger coefficients

(a) (b)

Figure 4.6: Sample of trajectories for the *Taxi* data, with 20,000 random pick-up events selected. The light blue dots indicate the input points, while the red rectangle represents the window with (latitude, longitude) pairs bounded by $(41.147, -8.58)$ and $(41.18, -8.65)$ considered in our experiments.

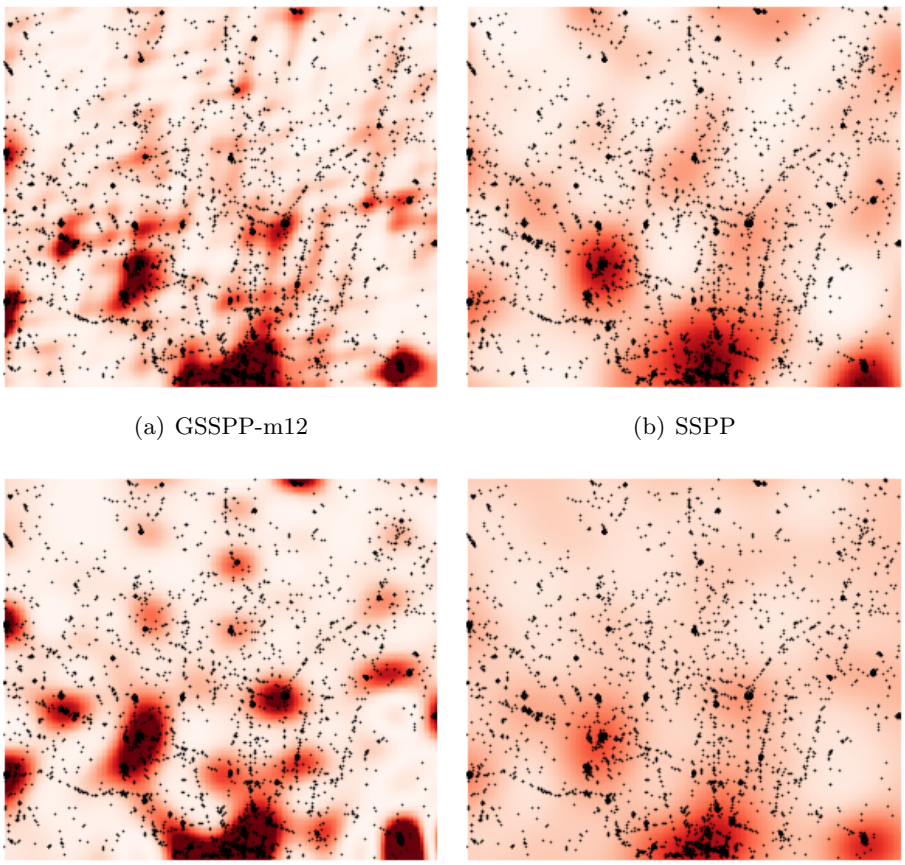
Table 4.5: Results on Taxi data experiment with standard errors in brackets.

	Taxi data (2D)	
	$\mathcal{L}_{\text{test}}$	time(s)
GSSPP-SE	278.54 (±1.64)	23.91
GSSPP-m12	**283.23** (±1.11)	23.86
GSSPP-m32	280.48 (±0.72)	19.60
GSSPP-m52	281.18 (±1.43)	23.94
SSPP	268.32 (±0.65)	12.13
KS	262.13 (±0.26)	2.55
VBPP	281.02 (±0.63)	20.45
LBPP	254.45 (±0.17)	1.04

for trigonometric components; hence the resulting intensity appears less smooth.

4.7 Conclusion

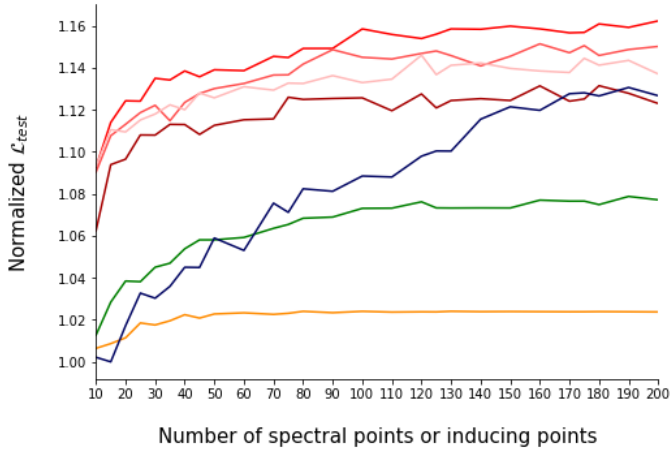
We introduce a novel Bayesian framework to infer the intensity function of a permanental process. Our approach uses a Laplace-based inference exploiting *generalized*



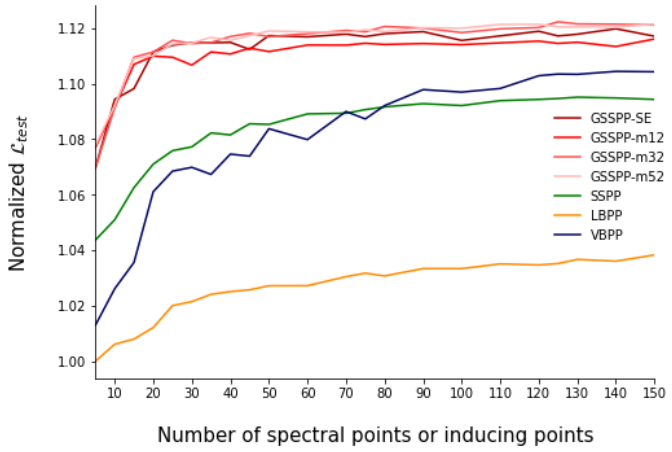
(a) GSSPP-m12 (b) SSPP

(c) VBPP (d) LBPP

Figure 4.7: Heat map of the predictive mean intensity for the *Taxi* data set scaled to a unit square. The black dots are the input data points.



(a) Taxi data



(b) Bei data

Figure 4.8: Average normalized test expected log-likelihood ($\mathcal{L}_{\text{test}}$) of the different methods on 2D real data, as a function of the number of spectral points or inducing points.

kernels and random Fourier features (RFFs). The approach requires no discretization of the domain, allows kernel designs, and provides better predictive accuracy than the alternative Laplace-based approach of Walder and Bishop (2017). The performance of our scheme also compares favorably with other standard methods on both real temporal and large spatial data sets.

Part II

Chapter 5

Background review : Gaussian Process time series models

Temporal dynamics are ubiquitous in numerous domain, encompassing fields such as biology, engineering, econometrics, and the social sciences. These dynamics reflect the ever-evolving and changing nature of systems over time. Time series data serves as a crucial instrument to understand these temporal phenomena. Formally, time series data are defined as a mapping time ($t \in \mathcal{T}$) to measurements ($x_t \in \mathbb{R}$), either in discrete ($\mathcal{T} = \mathbb{Z}$) or continuous ($\mathcal{T} = \mathbb{R}$) time, depending on sampling frequency and precision. The primary objective of time series analysis is to uncover the underlying relationships among data points and to facilitate predictive modeling of future observations.

Conventional time series modeling relies on probabilistic models marked by stringent assumptions. When dealing with dynamically dependent data, practitioners commonly turn to linear models due to their computational simplicity and approximation capabilities. The autoregressive (AR) and state space models (SSM) are notable examples. AR models project future values as linear combinations of past values, possibly with external factors. They extend into autoregressive moving average (ARMA) models when a moving average component (MA) is incorporated, eventually converging to AR models under specific conditions (notably with a sufficiently large lag parameter) (Brockwell and Davis, 2016). In contrast, SSM treats

time series as a latent process evolving over time under underlying state dynamics and measurement equations. Past values inform latent state estimates, facilitating predictions of future values.

However, modeling dynamically dependent data presents unique challenges that simplistic probabilistic models someties fail to address. A significant challenge is the presence of nonlinearity, as real-world phenomena frequently exhibit intricate nonlinear dependencies, adding complexity to the modeling process. To tackle this challenge, researchers have developed parametric methods tailored to nonlinear dynamics. These methods encompass the use of basis function expansion, nonlinear extensions of AR models (e.g., Threshold Autoregressive Models (Tong and Lim, 1980) or Markov Switching Models (Hamilton, 1989)), and extensions of state space models (e.g., Extended Kalman filter (Athans et al., 1968; Julier et al., 1995)).

A more recent approach involves extending state space models and AR models to GPs. GPs offer enhanced flexibility by eliminating the need for explicit parameterization. They adeptly capture nonlinear paterns within time serie data while providing analytical predictive densities under reasonable assumptions. This presents a contrast to many nonlinear Bayesian approaches for predictive density estimation, which often rely on computationally demanding techniques such as Monte Carlo sampling, local expansions, or variational methods.

Another significant challenge in working with time series data is non-stationarity, which involves abrupt changes in generative parameters. The failure to detect these specific *change points*, where the underlying distribution undergoes shifts, can profoundly disrupt the predictive performance of stationary parametric models. Common models used in Change point detection (CPD) include Poisson processes with varying intensity (Ritov et al., 2002), Gaussian models with changing variance (Johnson et al., 2003), and Markov models with time-varying transition matrices (Braun and Müller, 1998). These methods have demonstrated their value across various domains, including finance Chib (1998); Koop and Potter (2004); Kummerfeld and Danks (2013), quality control Aroian and Levene (1950), climate modelling Manogaran and Lopez (2018), cybersecurity Polunchenko et al. (2012), genetics Caron et al. (2012a) and

speech recognition Panda and Nayak (2016).

Bayesian Online Change Point Detection (BOCPD) methods stands out due to their fully Bayesian methodology for change point analysis, which bypasses the need for approximate inference. It was independently introduced by Adams and MacKay (2007) and Fearnhead and Liu (2007). Traditionally, Bayesian approaches to Change Point Detection (CPD) primarily relied on retrospective inference techniques, often resorting to segmentation methods to obtain samples from the posterior distribution, which represents the likelihood of change point locations Barry and Hartigan (1993); Green (1995); Xuan and Murphy (2007). In contrast, BOCPD offers real-time or online inference by continuously generating a predictive distribution for the next data point, considering the already observed information.

BOCPD has naturally been extended to incorporate GP-based underlying time series models, as proposed by Saatçi et al. (2010), addressing the non-stationarity issue while allowing for flexible and nonlinear dependencies. In subsequent sections of this document, we introduce an enhanced version of Saatçi et al. (2010)'s approach, aimed at improving both robustness and computational efficiency while enhancing prediction quality. This chapter provides a comprehensive review of prior research, setting the stage for our unique contribution.

In Section 5.1, we thoroughly delve into time series modeling with GPs, beginning with an introduction to two primary approaches for GP-based time series modeling: the GP time series model (GPTS) and the Gaussian Process Autoregressive (GPAR) models. Our exploration will provide a comprehensive analysis of their distinctive features. Moving on to Section 5.2, we shift our focus to the core principles of BOCPD. This section offers an understanding of BOCPD and further examines its GP-based extensions, as originally proposed by Saatçi et al. (2010). These will serve as essential building blocks and benchmarks for assessing our contribution.

5.1 GP time series

GP-based time series models leverage the inherent flexibility of GPs to adeptly capture intricate temporal dependencies within data, while concurrently facilitating the application of Bayesian principles to the domain of time series modeling. This section serves to provide a formal exposition and analysis of several GP-based time series models, which will play an integral role in subsequent sections and the forthcoming chapter.

Firstly, in Section 5.1.1, we introduce the GPTS model (Roberts et al., 2013). The GPTS model considers the temporal index as an input variable and possesses the noteworthy attribute of generalizing numerous classical linear models, encompassing MA, AR, ARMA, and Kalman filter classes. Our discussion will delve into the intricacies of this generalization. Subsequently, in Section 5.1.2, we present the GPAR model (Quiñonero-Candela et al., 2003), an adaptation of the autoregressive model within the Gaussian Process framework. Lastly, in Section , we offer a concise presentation of GP-based state space models. Although slightly beyond the primary scope of our work, we include it for completeness.

5.1.1 GPTS

The Gaussian Process Time Series (GPTS) model, initially introduced by Roberts et al. (2013), treats the time index $t \in [1, T]$ as an input variable. Given a sequence of observations $\mathbf{x}_{1:T} = \{x_i\}_{i=1}^{T}$, the GPTS model takes the following form:

$$x_t = f(t) + \varepsilon_t. \tag{5.1}$$

Here, f is a Gaussian process with a mean of zero and covariance function k, i.e., $f \sim \mathcal{GP}(0, k)$, and ε_t is Gaussian noise with zero mean and standard deviation σ_n, i.e., $\varepsilon_t \sim \mathcal{N}(\varepsilon_t|0, \sigma_n)$. An illustration of the model is provided in Figure 5.1.

Numerous approaches have been proposed in the literature for designing the GPTS model. One approach involves treating the problem as a regression task and using

a combination of Gaussian and periodic kernels to account for time-dependent decay and periodicity in the signal, respectively (Roberts et al., 2013). By combining these kernels, the GPTS model effectively captures underlying patterns and dynamics within time series data.

Notably, GPTS exhibits distinctive attributes. First, as a GP, GPTS inherently facilitates principled Bayesian inference with tractable posterior and predictive distributions. Second, the GPTS prior can be seen as a generalization of other classical linear approaches in time series modeling, such as AR models, MA models, and the Kalman filter for state-space modeling. Many of these linear models can be viewed as instances of GPTS with specific covariance functions (Murray-Smith and Girard, 2001; Turner, 2011) [1]. Finally, it is crucial to recognize that GPTS is fundamentally a linear model. Consequently, akin to other linear models, GPTS faces a significant limitation — it cannot effectively capture nonlinear dynamics.

We now briefly discuss the correspondence between GPTS and linear models, following the content and notation introduced by Turner (2011, Chapter 3). While the intricacies of these linear models are beyond the scope of this thesis, readers interested in further understanding can refer to time series modeling textbooks, such as Box et al. (1994); Brockwell and Davis (1991).

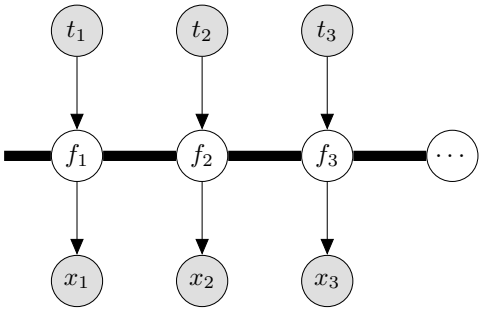
MA processes The MA model is a popular approach for modeling univariate time series in time series analysis (Brockwell and Davis, 1991, Chapter 3). This model posits that the output variable exhibits cross-correlalation with a distinct random variable, which is not identical to itself. In essence, the MA model can be regarded as a finite impulse response filter applied to white noise. The MA(∞) process is defined as follows:

$$x_t = \mu + \sigma_\varepsilon \varepsilon_t + \sum_{i=1}^{t-1} m_i \varepsilon_{t-i} \tag{5.2}$$

where μ is a constant drift, $\boldsymbol{\varepsilon}_{1:t} = [\varepsilon_1, \cdots, \varepsilon_t]$ are i.i.d normally distributed white noise error terms with mean 0 and variance 1, and $\boldsymbol{m}_{1:t-1} = [m_1, \cdots, m_{t-1}]$ denotes

[1] Note that this does not directly translate to the posterior, as the GPTS posterior function can theoretically converge to non-random functions, rendering points independent, unlike in MA or AR models.

Figure 5.1: Graphical model for linear-Gaussian process time series model (GPTS). The arrows directions suggest directions of influence. Grey circles represent the observed variables. The bold horizontal bar represents a set of fully connected nodes. Note that an observation x_i is conditionally independent of all other nodes given the corresponding latent variable, f_i.



the model's parameters. We can represent the MA process more concisely as:

$$x_t = \mu + \mathbf{M}(t, 1:t) \cdot \boldsymbol{\varepsilon}_{1:t} \tag{5.3}$$

where $\mathbf{M}$ is a parameters matrix. Its t-th row, denoted as $\mathbf{M}(t, 1:t)$, equals the vector $\left[(\mathbf{E}_{t-1} \cdot \boldsymbol{m}_{1:t-1})^\top, \quad \sigma_\varepsilon\right]$, with $\mathbf{E}_{t-1}$ being the $(t-1) \times (t-1)$ *exchange matrix* with i, j elements

$$\mathbf{E}_{i,j} = \begin{cases} 1, & i + j = t \\ 0, & i + j \neq t. \end{cases} \tag{5.4}$$

In a typical MA(q) model, where the lag size is set to the last q noise terms, we set $\mathbf{M}(t, t-q:t)$ to $\left[(\mathbf{E}_q \boldsymbol{m}_{1:q})^\top, \quad \sigma_\varepsilon\right]$ for all $t > q$.

Any MA process can be represented as a GPTS model with a suitable choice of the covariance function. Similar to the 'weight' view of GPs introduced in Section 2.4.3, for a finite set of input vectors $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, a GP $f \sim \mathcal{GP}(0, k)$ can be represented using Cholesky factorization:

$$f(\mathbf{X}) = \text{chol}(\mathbf{K}_{n,n})\mathbf{w}, \quad \text{with} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}_n) \tag{5.5}$$

where $K_{n,n}$ is the $n \times n$ Gram matrix with entries $k(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j = 1, \ldots, n$.

In the context of GPTS, where the observations correspond to the input time t, this leads to the following representation:

$$x_t = \mathbf{L}(t, 1:t)\boldsymbol{\varepsilon}_{1:t}, \quad \text{with} \quad \boldsymbol{\varepsilon}_{1:t} \sim \mathcal{N}(\boldsymbol{\varepsilon}_{1:t}|\mathbf{0}, \mathbf{I}_t) \tag{5.6}$$

where $\mathbf{L} := \text{chol}(\mathbf{K}_{1:t,1:t})$ is the $t \times t$ lower triangular matrix obtained from the Cholesky factorization of the matrix $\mathbf{K}_{1:t,1:t}$ corresponding to the first t input vectors.

To simplify the model without loss of generality, we can assume $\mu = 0$. Therefore, by equating Equation (5.3) and Equation (5.6), we can transform a GPTS into an MA model and vice versa, yielding $\mathbf{L} = \mathbf{M} \implies \mathbf{K}_{1:t,1:t} = \mathbf{M}\mathbf{M}^\top$.

AR processes Autoregressive (AR) models are widely employed in fields such as econometrics, signal processing, and statistics for describing time-varying processes. The fundamental concept behind AR models is that the current value of a time series is determined by its past values and a stochastic term. Mathematically, an AR(∞) process can be represented as follows:

$$x_t = \mu + \sigma_\varepsilon \varepsilon_t + \sum_{i=1}^{t-1} a_i x_{t-i} \tag{5.7}$$

where ε_t is a normally distributed white noise error term, and $\boldsymbol{a}_{1:t-1} = [a_1, \cdots, a_{t-1}]$ are the parameters of the model. Alternatively, an AR model can be expressed in matrix form as::

$$p(x_t|\mathbf{x}_{1:t-1}) = \mathcal{N}(x_t|\mathbf{A}(t, t-1)\mathbf{x}_{1:t-1}, \mathbf{A}(t,t)^2) \tag{5.8}$$

where $\mathbf{A}$ is a matrix of model parameters such that its t-th row, denoted as $\mathbf{A}(t, 1:t)$, is equal to the vector $[(\mathbf{E}_{t-1}\boldsymbol{a}_{1:t-1})^\top, \quad \sigma_\varepsilon]$. In a standard AR($p$) model, the p most recent values of the time series are used to predict the next value. To represent this, we set $\mathbf{A}(t, t-p:t)$ to be equal to $\left[(\mathbf{E}p \cdot \boldsymbol{a}_{1:p})^\top, \quad \sigma_\varepsilon\right]$ for $t > p$.

We can convert a GPTS model into an AR model, starting from the Cholesky

representation of the GP:

$$\begin{aligned} x_t &= \mathbf{L}(t, 1:t)\boldsymbol{\varepsilon}_{1:t} \\ &= \mathbf{A}(t, t-1)\mathbf{x}_{1:t-1} + \mathbf{A}(t, t)^2\varepsilon_t \\ &= \mathbf{A}(t, t-1)\mathbf{L}(1:t-1, 1:t-1)\boldsymbol{\varepsilon}_{1:t-1} + \mathbf{A}(t, t)^2\varepsilon_t. \end{aligned}$$

This implies a recursive relation:

$$\mathbf{L}(t, 1:t) = \left[\mathbf{A}(t, t-1)\mathbf{L}(1:t-1, 1:t-1), \quad \mathbf{A}(t, t)\right]. \tag{5.9}$$

Finally, we can obtain the equivalent covariance matrix from the Cholesky factorization, i.e., $\mathbf{K}_{1:t,1:t} = \mathbf{L}\mathbf{L}^\top$.

ARMA processes A standard ARMA(p, q) combines an AR component and an MA component. The parameters p and q specify the number of past values and past errors used in the model, respectively. Mathematically, the ARMA(p,q) model is represented as:

$$x_t = \mu + \sigma_\varepsilon\varepsilon_t + \sum_{i=1}^{p} a_i x_{t-i} + \sum_{i=1}^{q} m_i\varepsilon_{t-i}. \tag{5.10}$$

This model can also be expressed in matrix form as:

$$x_t = \mu + \sigma_\varepsilon\varepsilon_t + \mathbf{M}(t, 1:t) \cdot \boldsymbol{\varepsilon}_{1:t} + \mathbf{A}(t, 1:t-1) \cdot \mathbf{x}_{1:t-1}. \tag{5.11}$$

To compute the equivalent covariance matrix, we can use the following recursive equation:

$$\begin{aligned} x_t &= \mathbf{L}(t, 1:t)\boldsymbol{\varepsilon}_{1:t} \\ &= \mathbf{M}(t, 1:t)\boldsymbol{\varepsilon}_{1:t} + \mathbf{A}(t, 1:t-1)\mathbf{y}_{1:t-1} \\ &= \left(\mathbf{M}(t, 1:t-1) + \mathbf{A}(t, t-1)\mathbf{L}(1:t-1, 1:t-1)\right)\boldsymbol{\varepsilon}_{1:t-1} + \mathbf{M}(t, t)\varepsilon_t \end{aligned} \tag{5.12}$$

This allows us to compute $\mathbf{L}$ recursively using the following equations:

$$\mathbf{L}(t, 1:t-1) = \mathbf{M}(t, 1:t-1) + \mathbf{A}(t, t-1)\mathbf{L}(1:t-1, 1:t-1)$$
$$\mathbf{L}(t, t) = \mathbf{M}(t, t)$$

Finally, the equivalent covariance matrix is obtained as before: $\mathbf{K}_{1:t,1:t} = \mathbf{L}\mathbf{L}^\top$.

Kalman filter In their work, Hartikainen and Sarkka (2010) demonstrated the reformulation of GPTS as Kalman filtering and smoothing applied to linear state space models. We provides an overview of their approach, which treats the GPTS model as an estimation of the state of a multi-dimensional continuous-time Gauss-Markov process.

We start by considering an m-th order scalar stochastic differential equation (SDE) given by:

$$\frac{\partial^m f(t)}{t^m} + a_{m-1}\frac{\partial^{m-1} f(t)}{t^{m-1}} + \cdots + a_1\frac{\partial f(t)}{t} + a_0 f(t) = w_t \tag{5.13}$$

Here, $a_0, \cdots, a_{m-1}$ are known constants, and w_t denotes a white noise process with a spectral density equal to the constant q. This equation can be expressed as a first-order Markov process by rewriting it as:

$$\frac{\partial \mathbf{u}_t}{\partial t} = \mathbf{F}\ \mathbf{u}_t + \mathbf{G}w_t \tag{5.14}$$

where $\mathbf{u}_t$ is a vector defined as $\mathbf{u}_t = [f(t), \frac{\partial f(t)}{\partial t}, \cdots, \frac{\partial^{m-1} f(t)}{\partial t^m}]^\top$, and the matrices $\mathbf{F}$ and $\mathbf{G}$ are given by

$$\mathbf{F} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & \dots & -a_{m-2} & -a_{m-1} \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

It's worth noting that in Equation (5.14), a white noise representation is used for SDEs. However, to conform to the standard notation for SDEs, dW_t should be used instead of w_t, where W_t represents a Brownian motion.

To obtain $f(t)$ from $\mathbf{u}_t$, we define the vector $\mathbf{H} = [1, 0, \cdots, 0]^T \in \mathbb{R}^{m+1}$, which picks out the first component of $\mathbf{u}_t$, i.e., $f(t) = \mathbf{H}^\top \mathbf{u}_t$. Substituting this expression into Equation (5.14) and applying the Fourier transform to both sides, we obtain an expression for the spectral density $S(\omega)$ of $f(t)$ given by:

$$S(\omega) = \mathbf{H}^\top (\mathbf{F} + i\omega\mathbf{I})^{-1} \mathbf{G}\, q\, \mathbf{G}^\top \left[(\mathbf{F} - i\omega\mathbf{I})^{-1}\right]^\top \mathbf{H}. \tag{5.15}$$

We assume that covariance function of $f(t)$ is stationary covariance function $k(\tau) = k(t - s) = k(\tau)$ with $\tau = t - s$. In the stationary state, i.e. after the process has run for an infinite amount of time, the covariance function is given by the inverse Fourier transform of $S(\omega)$, which yields

$$k(\tau) = \mathbf{H}^\top \mathbf{\Sigma}_\infty \exp(\mathbf{F}|\tau|)\mathbf{H}. \tag{5.16}$$

Here $\mathbf{\Sigma}_\infty$ is the stationary covariance of $\mathbf{u}_t$, which satisfies the matrix Riccati equation:

$$\frac{\partial \mathbf{\Sigma}_\infty}{\partial t} = \mathbf{F}\mathbf{\Sigma}_\infty + \mathbf{\Sigma}_\infty \mathbf{F}^\top + \mathbf{G}\, q\, \mathbf{G}^\top. \tag{5.17}$$

To represent a given GPTS process $f(t)$ with a stationary covariance function $k(\tau)$ as a Markov process similar to Equation (5.14), we need to find a consistent transition matrix $\mathbf{F}$ and scalar q that satisfy Equation (5.15) for the spectral density of f. The authors show that this is possible when the spectral density of $k(\tau)$ can be expressed as a rational function of ω^2 in the form:

$$S(\omega) \propto (\text{polynomial in } \omega^2)^{-1}. \tag{5.18}$$

In this case, we can write $S(\omega) = H(i\omega)\, q\, H(-i\omega)$, where $H(\cdot)$ is a transfer function and q is a scalar, similar to Equation (5.15) (with $H(i\omega) = \mathbf{H}^\top (\mathbf{F} + i\omega\mathbf{I})^{-1}\mathbf{G}$). By applying the Fourier transform, we obtain the time-domain equation:

$$\frac{\partial^m f(t)}{t^m} + h_{m-1}\frac{\partial^m f(t)}{t^m} + \cdots + h_1 \frac{\partial f(t)}{t} + h_0 f(t) = w_t \tag{5.19}$$

where $h_0, \cdots, h_{m-1}$ are the coefficients of polynomial in the denominator of $H(i\omega)$ and w_t is a white noise with spectral density equal to q. Thus, we can construct a consistent transition matrix $\mathbf{F}$ by setting its coefficients to $h_0, \cdots, h_{m-1}$.

The authors provide an explicit reformulation for the Matérn class of kernel, whose covariance function is of the form given in Equation (5.18). For the Squared Exponential kernel, however, the spectral density cannot be expressed as a rational function. Instead, the authors use a spectral Taylor approximation to obtain an equivalent finite-dimensional Markov process.

We can convert the continuous-time model described in Equation (5.14) into a discrete-time state-space model with the following process equation:

$$\mathbf{u}_t = \mathbf{A}_{t-1}\mathbf{u}_{t-1} + \mathbf{q}_{t-1}, \quad \mathbf{q}_{t-1} \sim \mathcal{N}(\mathbf{u}_{t-1}|\mathbf{0}, \mathbf{Q}_{t-1}) \tag{5.20}$$

where $\mathbf{A}_{t-1} = \exp(\mathbf{F}\Delta t)$ and $\mathbf{Q}_{t-1}$ is given by the integral expression

$$\mathbf{Q}_{t-1} = \int_0^{\Delta t-\tau} \exp(\mathbf{F}\Delta t)\mathbf{G}\ q\ \mathbf{G}^\top \exp(\mathbf{F}\delta t - \tau)d\tau.$$

with Δt denoting the time difference between t and $(t-1)$. The measurement equation is given by

$$\mathbf{x}_t = \mathbf{H}^\top \mathbf{u}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\varepsilon_t|0, \sigma_n). \tag{5.21}$$

We have demonstrated that GPTS can be represented as linear state space models. Remarkably, the converse is also true: any linear state space model can be expressed as a GPTS. To illustrate this, let's consider a general linear state space model with a representation akin to Equation (5.20) in discrete time or (5.14) in continuous time. In this model, we have transition matrices $\mathbf{F}$ and $\mathbf{A}$, as well as transition vectors $\mathbf{G}$ and $\mathbf{H}$. The latent process $\mathbf{u}_t$ is modeled as an Ornstein-Uhlenbeck (OU) process. Consequently, we can derive its stationary covariance $\mathbf{\Sigma}_\infty(s,t)$, which provides information about the relationship between entries of $\mathbf{u}_t$ and $\mathbf{u}_s$. Furthermore, we can use this covariance structure to express the stationary covariance for the Gaussian measurements $\mathbf{x}_t$. Intriguingly, this equivalence is akin to having a GPTS prior, complete

with a consistent kernel $k(s,t)$.

5.1.2 GPAR

In addition to GPTS, another powerful approach to time series modeling using GPs is the Autoregressive Gaussian Process (GPAR) (Quiñonero-Candela et al., 2003; Turner, 2011). A GPAR model of order p incorporates the preceding p values $\mathbf{x}_{1:t-1}$ as inputs at time t, leading to the following equation:

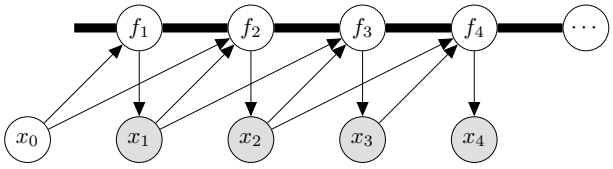
$$x_t = f(\mathbf{x}_{t-p:t-1}) + \varepsilon_t \tag{5.22}$$

Here, as in GPTS, the function f follows a GP with a zero mean and covariance function k, denoted as $f \sim \mathcal{GP}(0,k)$. Additionally, ε_t represents Gaussian noise with a mean of zero and standard deviation σ_n, signifying $\varepsilon_t \sim \mathcal{N}(\varepsilon_t|0,\sigma_n)$. This GPAR formulation can be seen as a specific instance of GP-based Nonlinear AutoRegressive models with eXogenous inputs (GP-NARX), where exogenous inputs are incorporated into the model (Kocijan et al., 2005; Ažman and Kocijan, 2011; Worden and Green, 2014; Worden et al., 2018).

An illustration of this model is presented in Figure 5.2. The model depicted in the graph follows a second-order autoregressive structure, meaning each latent variable f_i depends on the previous two observed variables x_{i-1} and x_{i-2}, as well as the previous latent functions. An observation x_i is conditionally independent of all other nodes given the corresponding latent variable f_i.

The GPAR methodology is often regarded as more versatile than GPTS. While GPTS can be viewed as a linear AR time series model, GPAR is able to capture nonlinear relationships, making it better suited for modeling complex dynamics. However, it's important to acknowledge that GPAR comes with a higher computational cost compared to GPTS. Furthermore, GPTS retains some elegant properties that are not preserved when using GPAR. Specifically, GPAR assumes directionality in the dynamics, which may not be suitable for reversible physical systems. Moreover, GPTS can handle continuous time, whereas GPAR is limited to cases where observations

Figure 5.2: Graphical model for second order auto-regressive Gaussian process time serie model (GPAR). The arrows directions suggest directions of influence. Grey circles represent the observed variables. The bold horizontal bar represents a set of fully connected nodes. Note that an observation x_i is conditionally independent of all other nodes given the corresponding latent variable, f_i.



are uniformly sampled.

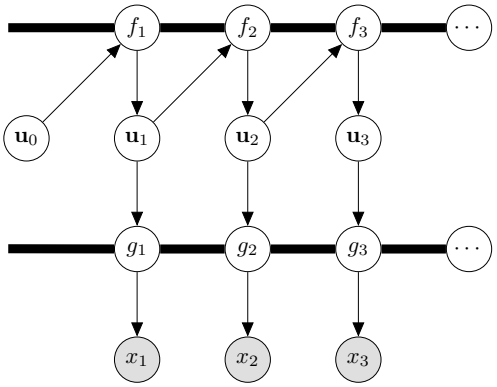
Numerous inference methods have been proposed in the literature for GPAR. Gregorcic and Lightbody (2002) and Kocijan et al. (2005) introduced inference techniques that maximize the marginal likelihood $p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(x_t|\mathbf{x}_{1:t-1})$. Other studies, such as those by Quiñonero-Candela et al. (2003) and Girard et al. (2002), have explored GP-NARX models with uncertain inputs for multi-step ahead forecasting, enabling the propagation of predictive uncertainty over several time steps. More recently, the GP-NARX approach for multi-step ahead prediction has been combined with different sparse GP approximations Gutjahr et al. (2012) or variational GP techniques to overcome computational challenges associated with large datasets.

5.1.3 GP-SSM

State space models (SSMs) have become a fundamental tool for analyzing time series data. These models propose that the observed time series results from an underlying process involving hidden or latent states that evolve over time in accordance with specific dynamics and measurement equations. The foundational linear SSMs with Gaussian noise were initially introduced by Kalman (1960). Subsequently, researchers have explored various extensions to accommodate non-linear and non-Gaussian scenarios.

One notable extension involves the integration of Gaussian processes to introduce non-linear functions into SSMs, giving rise to what are known as Gaussian Process

Figure 5.3: Graphical model for Gaussian process state-space model (GP-SSM). The arrows directions suggest directions of influence. Grey circles represent the observed variables. The bold horizontal bar represents a set of fully connected nodes. Note that an observation x_i is conditionally independent of all other nodes given the corresponding latent variable, g_i.



State Space Models (GP-SSMs) (Wang et al., 2005). GP-SSMs are characterized by system dynamics expressed as follows:

$$\mathbf{u}_t = f(\mathbf{u}_{t-1}) + \boldsymbol{\varepsilon}_f$$

$$x_t = g(\mathbf{u}_t) + \boldsymbol{\varepsilon}_g.$$

In this formulation, $\mathbf{u}_{1:T}$ represents a latent state, and a Gaussian process prior is imposed on the function f. The noise terms $\boldsymbol{\varepsilon}_f$ and $\boldsymbol{\varepsilon}_g$ are assumed to follow Gaussian distributions. For a visual representation of the model, please refer to Figure 5.3.

Various methodologies have been proposed for performing inference and learning in GP-SSMs, addressing different aspects of state estimation and system identification. For real-time state estimation, Ko and Fox (2009) and Deisenroth et al. (2009) introduced filtering methods, which aim to deduce the current system state based on available measurements. Techniques for estimating past states, known as *smoothing*, have been explored by Deisenroth and Mohamed (2012) and Deisenroth et al. (2012). They employ both deterministic and stochastic approximations to incorporate both current and future observations for estimating past states.

Learning the dynamics of GP-SSMs can be approached in different ways. Turner et al. (2010) and Dempster et al. (1977) have employed MLE with Expectation-Maximization (EM) approximations to identify system parameters. A more comprehensive Bayesian approach is presented by Frigola et al. (2013), which employs particle sampling (PMC). A hybrid approach that combines learning and inference has been proposed by Frigola et al. (2014b). This method utilizes variational GP approximation and particle sampling, striking a balance between flexibility and computational feasibility.

Most of these methods assume that the measurement function g is known and parametric. An innovative approach introduced by Eleftheriadis et al. (2017) allows for joint learning of both latent transitions and measurement functions using GP priors through variational inference. This approach enhances modeling capabilities, particularly in scenarios where the measurement function is not well-defined or predetermined.

In this section, we discussed various GP-based time series models, which, unfortunately, do not accommodate non-stationarity. In the next section, we introduce the BOCPD methodology, designed explicitly to address non-stationarity. What sets BOCPD apart is its model-agnostic nature, allowing us to seamlessly integrate it into our GP-based time series models. In the following chapter, we'll explore how this integration enhances our ability to analyze both stationary and non-stationary time series data, offering a comprehensive and flexible approach.

5.2 Bayesian change point detection

The conventional approach to Bayesian inference for Change Point Detection (CPD), especially when dealing with an unknown number of change points, typically involves offline segmentation. In this context, the posterior distribution of change point locations is estimated using various methods, such as Markov Chain Monte Carlo (MCMC) (Chib, 1998), reversible jump MCMC (Green, 1995; Punskaya et al., 2002), or direct simulation employing exact techniques for posterior means computa-

tion (Barry and Hartigan, 1993; Liu and Lawrence, 1999; Fearnhead, 2005, 2006b).

In contrast, BOCPD, introduced by Adams and MacKay (2007) and Fearnhead and Liu (2007), offers a unique capacity for online inference. In simpler terms, BOCPD excels at detecting changes in data acquired incrementally over time. It does so by constructing a predictive distribution for the next data point based on the observed data up to the present moment. BOCPD has received considerable recent interest, , with research efforts focusing on performance enhancement Saatçi et al. (2010), model selection Knoblauch and Damoulas (2018), hyperparameter learning Turner et al. (2009); Wilson et al. (2010); Caron et al. (2012a) and change point prediction Agudelo-España et al. (2019).

In this section, we provide a concise overview of the BOCPD algorithm. We start by introducing the original iteration of BOCPD, as conceived by Adams and MacKay (2007), which assumes that observations are independent and identically distributed within each data segment. Notably, the algorithm has undergone significant evolution in the literature, expanding to incorporate more advanced *underlying predictive model* (UPM) that serves as base generative models.

In particular, we delve deeper into the various extensions of BOCPD introduced by Saatçi et al. (2010), which integrate nonparametric Gaussian Process-based UPMs to specifically capture temporal structures within data segments. Throughout this section, we provide in-depth insights into the distinct components and procedural steps of the algorithm, with the aim of enhancing readers' understanding.

5.2.1 BOCPD algorithm

In BOCPD, we assume a sequence of observations $\mathbf{x}_{1:T} = \{x_i\}_{i=1}^{T}$ that can be partitioned into sub-groups separated by possible change points.

Run length To quantify the time elapsed since the last change point, BOCPD employs the concept of *run length*. Denoted as $r_t \in \mathbb{N}$ at time t, it represents the

length of the current run at time t^+ and follows this rule:

$$r_t = \begin{cases} 0, & \text{if changepoint occurs at time } t^+ \\ r_{t-1} + 1, & \text{otherwise.} \end{cases} \tag{5.23}$$

The run length increases by one or resets to zero when a changepoint occurs. If $r_t = n > 0$, then x_{t+1} will be the $(n+1)$-th instance of an existing run $(x_{t-r_t+1}, \cdots, x_t)$. On the other hand, if $r_t = 0$, it means that x_{t+1} is the first instance of a new run starting at t^+. We assume that $r_0 = 0$.

The transition probabilities are modeled using a conditional changepoint prior $p(r_t|r_{t-1})$. This assumes that r_t is independent of everything given r_{t-1} and follows this rule:

$$p(r_t|r_{t-1}) = \begin{cases} H(r_{t-1}) & \text{if } r_t = 0 \\ 1 - H(r_{t-1}) & \text{if } r_t = r_{t-1} + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.24}$$

Here, the *hazard function* $H(t)$ is calculated as:

$$H(\tau) = \frac{P_{\text{change}}(\tau)}{\sum_{t=\tau}^{\infty} P_{\text{change}}(t)}$$

where P_{change} represents the probability distribution over the interval between changepoints. A simple case arises when $P_{\text{change}}(\cdot)$ is a discrete exponential geometric distribution with a scale parameter of $1/h$, resulting in $H(\tau) = h$.

Predictive distribution Our primary concern is to recursively estimate the predictive distribution of the current run length $r_t \in [1, t]$, for each time step t within the range $[1, T]$. To calculate this predictive distribution for $\mathbf{x}_{1:T}$, BOCPD incorporates run length variables as follows:

$$\begin{aligned} p(x_t|\mathbf{x}_{1:t-1}) &= \sum_{r_{t-1}} p(x_t|\mathbf{x}_{t-1}, r_{t-1}) p(r_{t-1}|\mathbf{x}_{1:t-1}) \\ &:= \sum_{r_{t-1}} p(x_t|\mathbf{x}_{t-1}^{(r)}) p(r_{t-1}|\mathbf{x}_{1:t-1}). \end{aligned} \tag{5.25}$$

Here, $\mathbf{x}_{t-1}^{(r)}$ refers to the last r_{t-1} observations before x_t. An *underlying predictive model* (UPM) is defined to evaluate the posterior predictive distribution of the next datum given the possible previous run length, i.e., $p(x_t|\mathbf{x}_{t-1}^{(r)})$ for all $r \in [0, t-1]$ and $t \in [1, T]$. The UPM serves as a base model with parameters that change for each run length.

The procedure for inferring the run length involves recursively computing the joint probability of the current run length r_t and the observed sequence up to time t, denoted as $\mathbf{x}_{1:t}$. Given the hazard function and the UPM, this is achieved as follows:

$$\begin{aligned} p(r_t, \mathbf{x}_{1:t}) &= \sum_{r_{t-1}} p(x_t, r_t|r_{t-1}, \mathbf{x}_{1:t-1}) p(r_{t-1}, \mathbf{x}_{1:t-1}) \\ &= \sum_{r_{t-1}} p(r_t|r_{t-1}) p(x_t|r_{t-1}, \mathbf{x}_{1:t-1}) p(r_{t-1}, \mathbf{x}_{1:t-1}) \\ &= \sum_{r_{t-1}} \underbrace{p(r_t|r_{t-1})}_{\text{Hazard}} \underbrace{p(x_t|\mathbf{x}_{t-1}^{(r)})}_{\text{UPM}} \underbrace{p(r_{t-1}, \mathbf{x}_{1:t-1})}_{\text{Message}}. \end{aligned} \tag{5.26}$$

The normalizing constant of $p(r_t|\mathbf{x}_{1:t})$ is obtained by summing up all its evaluation instances since r_t is a discrete random variable, i.e.

$$p(r_t|\mathbf{x}_{1:t}) = \frac{p(r_t, \mathbf{x}_{1:t})}{\sum_{r_t} p(r_t, \mathbf{x}_{1:t})} \tag{5.27}$$

The predictive distribution of $\mathbf{x}_{1:T}$ is derived by recursively computing the probabilities $p(r_t|\mathbf{x}_{1:t})$ for all feasible values of r_t within the interval $[0, t]$, as defined in Equation (5.27). These computed probabilities are subsequently employed in the predictive distribution calculation outlined in Equation (5.25). The process is visually illustrated in Figure 5.4.

Alternative formulation Some variations of the original BOCPD algorithm in the literature employ a message passing recursion on the conditional distribution $p(r_t|\mathbf{x}_{1:t-1})$ directly, resulting in the following formulation:

$$p(r_t|\mathbf{x}_{1:t}) \propto \sum_{r_{t-1}} \underbrace{p(r_t|r_{t-1})}_{\text{Hazard}} \underbrace{p(x_t|\mathbf{x}_{t-1}^{(r)})}_{\text{UPM}} \underbrace{p(r_{t-1}|\mathbf{x}_{1:t-1})}_{\text{Message}}. \tag{5.28}$$

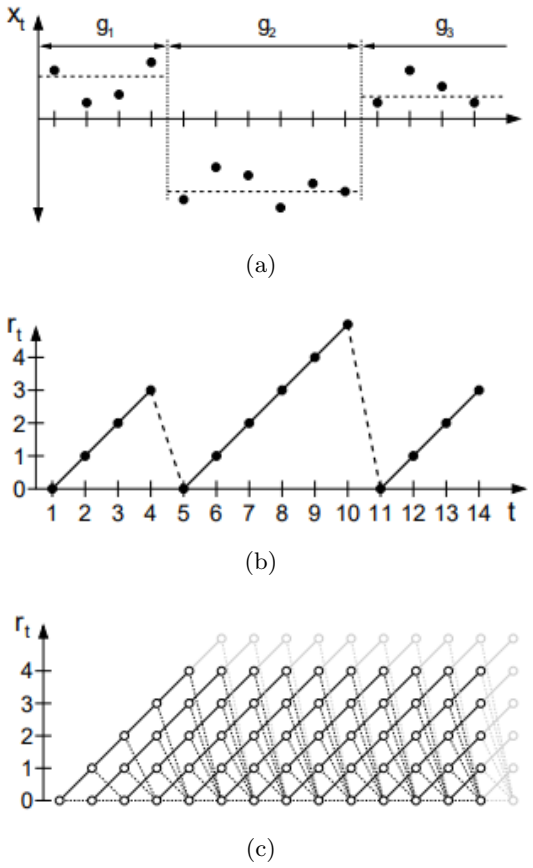


Figure 5.4: BOCPD model description in terms of run lenghts. Figure (a) shows hypothetical univariate data divided by change points into three segments of lenghts $g_1 = 4, g_2 = 6$. Figure (b) shows r_t as a function of time. Figure (c) shows the trellis on which the message passing algorithm lives. Solid lines indicates that probability mass is being passed "upwards". Dotted lines indicate the possibility that the current run is truncated and the run lenght drops to zero. The Figures are taken from Adams and MacKay (2007).

Interestingly, the normalizing factor is equal to $p(\mathbf{x}_{1:t-1})$, which can be determined by summing all evaluation instances over r_t.This provides an immediate evaluation of the marginal likelihood up to time step $t-1$, which proves beneficial for the estimation of model hyperparameters, as presented in Section 5.2.3. Moreover, the conditional distributions delineated in Equation (5.28) can be directly leveraged for computing the predictive distribution outlined in Equation (5.25).

5.2.2 Time independent UPM

In the original BOCPD algorithm (Adams and MacKay, 2007), an assumption is made that the observations within each run length are independent and identically distributed (i.i.d), following an exponential family distribution. This assumption, particularly in the context of Gaussian i.i.d. observations, is referred to as TIM-UPM, with TIM standing for time independent model.

Normal-Inverse-Gamma conjugate model For a time series $\mathbf{x}_{1:T} = \{x_i\}_{1:T}$, the prior distribution is specified as follows:

$$x_t \sim \mathcal{N}(x_t|\mu, \tau^{-1}) \tag{5.29}$$

where

$$\mu \sim \mathcal{N}(\mu|\mu_0, (\tau\kappa_0)^{-1}) \quad \text{and} \quad \tau \sim \text{Gamma}(\alpha_0, \beta_0) \tag{5.30}$$

Here, $\eta_0 := (\alpha_0, \beta_0, \kappa_0, \mu_0)$ represents hyperparameters, and $\{x_i\}_{1:t}$ are assumed independent given μ and τ. This joint prior is referred to as the Normal-Inverse-Gamma (NIG) distribution $\mathcal{NIG}(\mu, \tau|\eta_0)$.

The NIG distribution possesses the conjugacy property, allowing us to compute a conjugate posterior over parameters at time t given a series of observations within a run length r_{t-1} as:

$$p(\mu, \tau|\mathbf{x}_{t-1}^{(r)}, \eta_0) \sim \mathcal{NIG}(\eta_t^{(r)}) \tag{5.31}$$

where parameters $\eta_t^{(r)} = (\kappa_t^{(r)}, \mu_t^{(r)}, \alpha_t^{(r)}, \beta_t^{(r)})$ can be expressed recursively as follows:

$$\mu_t^{(r)} = \frac{\kappa_0\mu_0 + \sum \mathbf{x}_{t-1}^{(r)}}{\kappa_0 + r_{t-1}} \tag{5.32}$$

$$\kappa_t^{(r)} = \kappa_0 + r_{t-1} \tag{5.33}$$

$$\alpha_t^{(r)} = \alpha_0 + r_{t-1}/2 \tag{5.34}$$

$$\beta_t^{(r)} = \beta_0 + \frac{1}{2}\sum {\mathbf{x}_{t-1}^{(r)}}^2 - \frac{1}{2r_{t-1}}\left(\sum \mathbf{x}_{t-1}^{(r)}\right)^2 + \frac{\kappa_0 r_{t-1}(\sum \mathbf{x}_{t-1}^{(r)} - \mu_0)^2}{2(\kappa_0 + r_{t-1})} \tag{5.35}$$

Algorithm 5 Original BOCPD with TIM-UPM

for $t = 1$ **to** T **do**
  Evaluate predictive probabilities:
  $\pi_t^{(r)} = p(x_t|\eta_t^{(r)})$
  Update growth probabilities:
  $p(r_t = r_{t-1} + 1, \mathbf{x}_{1:t}) = p(r_{t-1}, \mathbf{x}_{t-1})\pi_t^{(r)}(1 - H(r_{t-1}))$
  Update change point probabilities:
  $p(r_t = 0, \mathbf{x}_{1:t}) = \sum_{r_{t-1}} p(r_{t-1}, \mathbf{x}_{1:t-1})H(r_{t-1})$
  Perform prediction:
  update the sample moments and parameters $\eta_{t+1}^{(r)}$ given x_t
  $p(\mathbf{x}_{1:t}) = \sum_{r_t} p(r_t, \mathbf{x}_{1:t})$.
  $p(r_t|\mathbf{x}_{1:t-1}) = p(r_t, \mathbf{x}_{1:t})/p(\mathbf{x}_{1:t})$.
  $p(x_{t+1}|\mathbf{x}_{1:t-1}) = \sum_{r_t} p(x_{t+1}|\eta_{t+1}^{(r)})p(r_t|\mathbf{x}_{1:t})$
end for

These updates yield sufficient statistics $\sum {\mathbf{x}_{t-1}^{(r)}}^2$ and $\sum \mathbf{x}_{t-1}^{(r)}$), which can be stored and updated in a streaming manner across time and run lengths.

TIM UPM Integrating over the parameters μ and τ leads to a UPM given by:

$$\begin{aligned} p(x_t|\mathbf{x}_{t-1}^{(r)}) &:= p(x_t|\eta_t^{(r)}) \\ &\sim St_{2\alpha}\left(\mu_t, \frac{\beta_t}{\alpha_t}\frac{\kappa_t + 1}{\kappa_t}\right) \end{aligned} \tag{5.36}$$

where $St_{2\alpha}$ denotes a Student-t distribution with 2α degrees of freedom. Note that in our notation, $\eta_t^{(r)}$ represents the set of parameters associated with the predictive for x_t and depends on r_{t-1}.

Applying this predictive posterior over each run length results in Algorithms 5 and 6. Algorithm 5 corresponds to a recursion applied to Equation (5.26), as seen in the original BOCPD (Adams and MacKay, 2007). Conversely, Algorithm 6 is a vectorized version of BOCPD where iteration is applied to Equation (5.28).

5.2.3 GP-based UPM

As discussed, the standard BOCPD algorithm relies on the assumption that data within each segment follows an i.i.d. distribution. However, real-world datasets often violate this assumption, as they exhibit temporal smoothness within each regime, in-

Algorithm 6 Vectorized BOCPD Run length estimation with TIM-UPM

Initialize the recursion
$\Xi_0 \leftarrow 1$
for $t = 1$ **to** T **do**
  $\boldsymbol{\pi}_t \leftarrow \texttt{vec}(\{p(x_t|\eta_t^{(r)})\}_{r_t})$
  $\mathbf{h} \leftarrow H(1:t)$
  Update growth probabilities:
  $\Xi_t[2:t+1] \leftarrow \Xi_{t-1} \odot \boldsymbol{\pi}_t \odot (1-\mathbf{h})$
  Update change point probabilities:
  $\Xi_t[1] \leftarrow \sum \Xi_{t-1} \odot \boldsymbol{\pi}_t \odot \mathbf{h}$
  Perform prediction:
  $p(x_t|\mathbf{x}_{1:t-1}) = \sum \Xi_t$
  $\Xi_t \leftarrow$ normalized Ξ_t
  $p(r_t|\mathbf{x}_{1:t-1}) = \Xi_t$
  update the sample moments and parameters $\eta_{t+1}^{(r)}$ given x_t
end for
Compute the evidence
$p(\mathbf{x}_{1:T}) = \sum \Xi_T$

troducing dependencies between data points. This violation of the i.i.d. assumption can lead to inaccurate change point detections. Additionally, the standard BOCPD algorithm treats hyperparameters as fixed and known, but the algorithm's performance is highly sensitive to their settings.

To address these limitations, several extensions to the BOCPD algorithm have emerged, leveraging more adaptable UPMs capable of capturing temporal dependencies within each regime. For instance, Knoblauch and Damoulas (2018) introduced a BOCPD extension that employs a linear parametric Bayesian Vector Autoregressive (BVAR) model to describe the processes between changepoints.

Notably, Saatci et al. (2010) introduced a nonparametric UPM based on GPs. As discussed in Section 5.1, these GP-based UPM excels in modeling intricate dynamics and temporal smoothness within each segment of a time series, all without relying on explicit parametric distributions. These extensions naturally build upon the GPTS, previously described as a generalized linear model, and GPAR for capturing nonlinear dynamics. These UPMs inherently inherit the various properties of GPTS and GPAR discussed in previous sections (Section 5.1.1 and Section 5.1.2), respectively. We now proceed to delve into these properties in more detail:

GPTS-UPM Given a time series $\mathbf{x}_t$ and a prediction target time t, the GPTS-UPM provides a predictive distribution in the form of:

$$p(x_t|\mathbf{x}_{t-1}^{(r)}) \sim \mathcal{N}(x_t|m_{t,r}, v_{t,r}) \tag{5.37}$$

where

$$\begin{aligned} m_{t,r} &= \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1} \mathbf{x}_{t,r} \\ v_{t,r} &= k(t,t) - \mathbf{k}_*^\top \left(\mathbf{K} + +\sigma_n^2\mathbf{I}\right)^{-1} \mathbf{k}_*. \end{aligned} \tag{5.38}$$

In the equations above, $\mathbf{x}_{t,r} := \mathbf{x}_{t-r+1:t-1}$, $\mathbf{K}$ is the matrix with i,j entries $k(t-r+i, t-r+j)$ for $i,j = 1,\cdots,r-1$ and $\mathbf{k}_*$ is an $(r-1)$-dimensional vector with the ith entry being $k(t, t-r+i)$ for $i = 1,\cdots,r-1$.

GPAR-UPM For a GPAR-UPM of order p, where at time t, the past p values $\mathbf{x}_{t-p:t-1}$ are considered as input, and x_t as the observation, the model yields an auto-regressive predictive distribution:

$$p(x_t|\mathbf{x}_{t-1}^{(r)}) \sim \mathcal{N}(x_t|m_{t,r}, v_{t,r}) \tag{5.39}$$

where

$$\begin{aligned} m_{t,r} &= \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1} \mathbf{x}_{t,r} \\ v_{t,r} &= k(x_t,x_t) - \mathbf{k}_*^\top \left(\mathbf{K} + +\sigma_n^2\mathbf{I}\right)^{-1} \mathbf{k}_*. \end{aligned} \tag{5.40}$$

In these equations, $\mathbf{x}_{t,r} = \mathbf{x}_{t-r+1:t-1}$, $\mathbf{K}_{i,j} = k(\mathbf{x}_{i-p+1:i}, \mathbf{x}_{j-p+1:j})$ for $i,j = t-r,\cdots,t-2$ and $\mathbf{k}_*$ is an $(r-1)$-dimensional vector with the ith entry being $k(\mathbf{x}_{t-p:t-1}, \mathbf{x}_{i-p+1:i})$ for $i = t-r,\cdots,t-2$.

Edges effects In the case of AR models, edge effects emerge when $t \leq p$. To address this, we can establish artificial initial conditions by setting $\mathbf{x}t = 0$. Similarly, in the context of GPAR-UPMs, when calculating the predictive distribution in Equation (5.39), we nullify the subvector $\mathbf{x}_{t-r-p+1:t-r}$.

Efficient implementation Efficient implementation is paramount to the practicality of GP-based BOCPD. A straightforward GP inference approach, which retains all run lengths, can lead to a daunting computational complexity of $\mathcal{O}(T^5)$ if GP predictions are recalculated at each time step. Fortunately, several implementation techniques have been proposed in the literature, including those by Saatçi et al. (2010), to mitigate this computational burden.

When dealing with uniformly sampled data in GPTS-UPMs, the covariance matrices exhibit a Toeplitz structure. Saatçi et al. (2010) astutely leveraged this property to efficiently solve predictions using recursive relationships similar to the Yule-Walker equations found in linear AR models (Golub and Van Loan, 1996). This approach significantly reduces the computational complexity to $\mathcal{O}(T)$ per step and $\mathcal{O}(T^3)$ in total.

In the case of GPAR-UPMs, Saatçi et al. (2010) introduced rank one updates to the precision matrix of the covariance matrix, allowing for *vertical updates* at every time step and *horizontal updates* across run lengths. This optimization significantly lowers the complexity to $\mathcal{O}(T^4)$.

Another valuable modification, introduced in the original BOCPD paper (Adams and MacKay, 2007), and also employed by Saatçi et al. (2010); Knoblauch and Damoulas (2018), involves the practice of *pruning* some run length probability estimates. This technique focuses on removing estimates with a total mass below a specific threshold or considering only the $R_{\max}$ most probable values. When combined with the previously mentioned methods, this results in a running complexity of $O(R_{\max}^2 T^2)$ for GPAR-UPMs and $O(R_{\max}^2 T)$ for GPTS models.

Learning hyperparameters Hyperparameters play a pivotal role in determining the performance of GP-based BOCPD. Their determination is a challenge in itself, particularly in BOCPD, where a single run can be computationally intensive. These hyperparameters are represented as $\Theta = (\theta, \sigma_n, H(\cdot))$, with θ denoting the GP kernel hyperparameters.

(Saatçi et al., 2010, 2015) proposed to learn the hyperparameters of BOCPD by maximizing the log marginal likelihood $\log p(\mathbf{x}_{1:T'}|\Theta)$ using the derivatives of one-step-ahead predictive likelihoods over a test set of data on a test set $[0, T']$, where $T' < T$. To achieve this, Saatçi et al. (2010) introduced the relationship

$$\log p(\mathbf{x}_{1:T'}|\Theta) = -\sum_{i=1}^{T'} \log p(x_t|\mathbf{x}_{1:t-1}, \Theta), \tag{5.41}$$

They then proposed a recursive algorithm, described in (Saatçi et al., 2015), for evaluating the gradients $\nabla_\Theta \log p(x_t|\mathbf{x}_{1:t-1}, \Theta)$ for $t \in [0, T']$. However, it's important to note that this approach can be computationally demanding as it requires multiple BOCPD runs to obtain optimal hyperparameters. A similar approach has been employed in Knoblauch and Damoulas (2018). We will provide a more detailed description of a modification to the BOCPD algorithm that addresses the computation of these marginal derivatives in the next chapter.

Alternatively, Caron et al. (2012b) suggest an online method for hyperparameter learning. This approach involves the iterative use of gradient descent, expressed as

$$\Theta_{t+1} = \Theta_t + \alpha_t \nabla \Theta_t \log p(x_{t+1}|\mathbf{x}_{1:t}) \quad \text{for } t \in [0, T]. \tag{5.42}$$

This method allows for incremental learning of hyperparameters as new data points become available.

5.3 Summary

This chapter provided an insightful exploration of time series data modeling with GPs, with an emphasis on their significance. In contrast to conventional methods for time series analysis, GPs derive a notable advantage from their non-parametric nature, offering remarkable flexibility without the need for explicit probabilistic assumptions.

We introduced a range of GP-based models, including GPTS and GPAR, as powerful tools for addressing the challenges posed by time series data. GPTS was presented as a general linear model, while GPAR emerged as a versatile nonlinear counterpart.

This distinction is crucial, as it enables us to effectively tackle the issue of nonlinearity in time series data, a challenge that traditional models often struggle with.

Additionally, we introduced BOCPD as a valuable approach for identifying changes in statistical characteristics within time series data, especially when dealing with non-stationary data. Our emphasis on the work of Saatçi et al. (2010) in extending BOCPD through the incorporation of GPTS and GPAR models serves as the backdrop for our work presented in in Chapter 6. These foundational concepts equip us with the necessary knowledge to comprehend and appreciate our contribution discussed in the following chapter.

Chapter 6

Bayesian online change point detection with Hilbert space approximate Student-t process

In this chapter, we introduce a variant of Bayesian online change point detection (BOCPD) with a reduced-rank Student-t process (TP) and dependent Student-t noise, as a nonparametric time series model. Our method builds and improves upon the state-of-the-art Gaussian process (GP) change point model benchmark of Saatçi et al. (2010). The Student-t process generalizes the concept of a GP and hence yields a more flexible alternative. Additionally, unlike a GP, the predictive variance explicitly depends on the training observations, while the use of an entangled Student-t noise model preserves analytical tractability. Our approach also uses a *Hilbert space* reduced-rank representation of the TP kernel, derived from an eigenfunction expansion of the Laplace operator Solin and Särkkä (2020), to alleviate its computational complexity. Improvements in prediction and training time are demonstrated with real-world data sets.

6.1 Introduction

The original BOCPD algorithm makes the assumptions that observations are i.i.d. within each run length. In Chapter 5, we introduced an extended BOCPD approach that incorporates a more flexible nonparametric UPM based on GPs, as proposed by Saatçi et al. (2010). This extension includes a non-linear auto-regressive GP-based model (GPAR) and a time-deterministic GP model (GPTS) with change points, both of which significantly enhance predictive performance.

In this chapter, we build on the GP-based approach of Saatçi et al. (2010), introducing an alternative UPM based on a Student-t process (TP) with Student-t noise Shah et al. (2014) and Hilbert space reduced-rank kernel proposed by Solin and Särkkä (2020). Benefiting from its fatter tails, a TP offers inherent robustness against outliers, surpassing GPs in this regard. Specifically, in the context of BOCPD, TPs exhibit a lower propensity for generating false alarms when detecting change points caused by outliers. Additionally, TPs offer more adaptive predictive variance in comparison to GPs, adjusting more effectively to the variance of past observations. We will explore this aspect further in Section 6.3.1. Lastly, a TP introduces greater flexibility compared to a GP, as it represents the most general elliptical process with a tractable density Shah et al. (2014)

The first mention of a TP can be found in Rasmussen and Williams (2005) and early applications in Archambeau and Bach (2011) and Yu et al. (2007). However, Rasmussen and Williams (2005) concluded that a TP is not practicable, due to the intractability of the posterior when adding noise (since the Student-t distribution is not closed under addition). TPs have received greater recent attention since Shah et al. (2014) proposed a derivation from a Wishart prior, and introduced a dependent Student-t noise preserving tractability. The benefit of a TP compared to a GP has since been demonstrated for regression Shah et al. (2014); Tang et al. (2016, 2017); Li and Ma (2021), state-space models Solin and Särkkä (2015) and Bayesian optimization Tracey and Wolpert (2018).

To overcome the GP computational complexity, several schemes have been proposed in the literature. Reduced-rank approximation methods which approximate the kernel Gram matrix with another matrix of smaller rank have been popular (see Chapter 8 Rasmussen and Williams, 2005). Common examples include the *Nyström* method (see Williams and Seeger, 2001b) and *Random Fourier Features* Rahimi and Recht (2007). Solin and Särkkä (2020) introduced an *Hilbert space* method for reduced-rank which approximates the eigendecomposition of stationary kernels in terms of an eigenfunction expansion of the Laplace operator. In their original paper, Solin and Särkkä (2020) adapt the method for a GP approximation referred as HSGPs, that has been used in the context of GP regression Solin and Särkkä (2020); Riutort-Mayol et al. (2022) and GP-based state-space models Svensson et al. (2016); Svensson and Schön (2017). The choice of an *Hilbert space* approach is particularly convenient in a BOCPD context. In the approximation, features vectors are independent of the covariance function, yielding computational advantages detailed later.

Combining a Student-t process predictor model and Hilbert space reduced-rank kernels, our Hilbert space TP-UPM (HSSPAR) shows systematic improvement in predictive performance and hyperparameter learning time for real-world data sets presented in Section 6.6.

6.2 Preliminaries

In this section, we begin by offering a concise overview of the foundational concepts essential to understanding our research contribution. In Section 6.2.1, we breifly revisit the basic component of BOCPD discussed in more details in Section 5.2. Additionally, Section 6.2.2 offers a comprehensive introduction to the Student-t process, along with an analysis of its properties when combined with Student-t noise.

6.2.1 BOCPD algorithm

Given an *hazard function* and an UPM, the inference about the run is done recursively at every time step. This inference process is described by the following expressions,

based on Equations (**??**) and (5.25):

$$p(r_t|\mathbf{x}_{1:t}) \propto \sum_{r_{t-1}} \underbrace{p(r_t|r_{t-1})}_{\text{Hazard}} \underbrace{p(x_t|\mathbf{x}_{t-1}^{(r)})}_{UPM} p(r_{t-1}|\mathbf{x}_{1:t-1}) \tag{6.1}$$

where $\mathbf{x}_{t-1}^{(r)}$ indicates the last r_{t-1} observations prior to x_t. The conditional prior, expressed as $p(r_t|r_{t-1})$, is defined by Equation (5.24). The normalizing constant of $p(r_t|\mathbf{x}_{1:t})$ in Equation (6.1) is obtained by summing up all its evaluation instances, since r_t is a discrete random variable. The marginal predictive distribution $p(x_t|\mathbf{x}_{1:t-1})$, is subsequently determined by Equation 5.25.

As outlined in Section 5.2, Adams and MacKay (2007) adopt Gaussian i.i.d. assumptions with a Normal-Inverse-Gamma prior on parameters. On the other hand, Saatçi et al. (2010) introduced two GP UPM variants. GPTS effectively utilizes time as an index, allowing it to adapt to irregular time intervals. Moreover, as demonstrated in Section 5.1.1, GPTS has been shown to possess an equivalent linear AR representation. On the other hand, GPAR, an auto-regressive UPM of order p, takes $\mathbf{x}_{t-p:t-1}$ as input at time t. Although it imposes a uniform time step constraint, GPAR generalizes GPTS, making it highly suitable for handling non-linearity and modeling more complex data dynamics.

6.2.2 Student-t Processes (TP)

We review the properties of the Student-t distribution and process, which serves in later sections as our predictive model.

Definition 6.2.1. *An n-dimensional vector $\mathbf{y}$ is multivariate Student-t-distributed with ν degrees of freedom, mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, if its joint probability density is given by*

$$St(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K}, \nu) = \frac{\Gamma((\nu+n)/2)}{\Gamma(\nu/2)((\nu-2)\pi)^{n/2}|\mathbf{K}|^{1/2}} \times \left(1 + \frac{1}{\nu-2}(\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}}. \tag{6.2}$$

As for the Gaussian distribution, the conditional distribution for a multivariate Student-t has an analytical form. The following result can be found in Kotz and Nadarajah (2004) and Shah et al. (2014).

Lemma 6.2.2. *Let* $\mathbf{y} \sim St(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K}, \nu)$ *and partition* $\mathbf{y}$ *into two sub-vectors* $\mathbf{y}_1 \in \mathbb{R}^{n_1}$ *and* $\mathbf{y}_2 \in \mathbb{R}^{n_2}$ *such that* $\boldsymbol{\mu}_p = \mathbb{E}[\mathbf{y}_p]$ *and* $\mathbf{K}_{p|p} = cov[\mathbf{y}_p, \mathbf{y}_p]$ *for* $p = 1, 2$. *Then the conditional density for* $\mathbf{y}_1|\mathbf{y}_2$ *has an analytical form* $\mathbf{y}_1|\mathbf{y}_2 \sim St(\boldsymbol{\mu}_{1|2}, \mathbf{K}_{1|2}, \nu_{1|2})$ *with* $\boldsymbol{\mu}_{1|2} = \mathbf{K}_{1,2}\mathbf{K}_{1,2}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)$, *covariance* $\mathbf{K}_{1|2} = \frac{\nu-2+\beta}{\nu-2+n_2}(\mathbf{K}_{1,1} - \mathbf{K}_{1,2}\mathbf{K}_{2,2}^{-1}\mathbf{K}_{1,2})$, $\beta = (\mathbf{y}_2 - \boldsymbol{\mu}_2)^\top \mathbf{K}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)$ *and* $\nu_{1|2} = \nu + n_2$ *degrees of freedom.*

TP construction As described in Shah et al. (2014), we can construct a Student-t process by placing an inverse Wishart process prior on the kernel of a Gaussian process. The Wishart distribution is a probability distribution over $\Pi(n)$, the set of real-valued, $n \times n$, symmetric, positive definite matrices.

Definition 6.2.3. *A random matrix* $\Sigma \in \Pi(n)$ *is inverse Wishart distributed with parameters* $\nu \in \mathbb{R}^+$, $\mathbf{K} \in \Pi(n)$ *and we write* $\Sigma \sim \mathcal{IW}(\nu, \mathbf{K})$ *if its density is given by*

$$p(\Sigma|v, \mathbf{K}) \propto |\Sigma|^{-\frac{v+2n}{2}} \exp\left(-\frac{1}{2} tr(\mathbf{K}\boldsymbol{\Sigma}^{-1})\right). \tag{6.3}$$

Dawid (1981) shows that the inverse Wishart distribution is consistent under marginalization. Thus we can defined a Wishart process for some input space $\mathcal{X}$ and a positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Definition 6.2.4. *The process* σ *is a inverse Wishart process (IWP) on* $\mathcal{X}$ *with parameter* v *and kernel function* $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ *if for any finite collection* $x_1, \cdots, x_n \in \mathcal{X}$, $\sigma(x_1, \cdots, x_n) \sim \mathcal{IW}(\nu, \mathbf{K})$ *where* $\mathbf{K} \in \Pi(n)$ *is the Gram matrix with* i, j *entries* $k(x_i, x_j)$ *for* $i, j = 1, \ldots, n$. *We write* $\sigma \sim \mathcal{IWP}(\nu, k)$.

For some kernel function k parametrized by θ and a mean function $\mu : \mathcal{X} \to \mathbb{R}$, Shah et al. (2014) propose deriving the Student-t process f as a hierarchical model

such that

$$\begin{aligned}\sigma &\sim \mathcal{IWP}(\nu, k_\theta)\\ f|\sigma &\sim \mathcal{GP}(\mu, (\nu-2)\sigma).\end{aligned} \tag{6.4}$$

For any collection $\mathbf{f} = (f(x_1), \cdots, f(x_n))^\top$ with $\boldsymbol{\mu} = (\mu(x_1), \cdots, \mu(x_n))^\top$ and $\boldsymbol{\Sigma} = \sigma(x_1, \cdots, x_n)$, we see that

$$\begin{aligned}p(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}, v) &= \int p(\mathbf{f}|\boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}|v, K) d\boldsymbol{\Sigma}\\ &\propto \frac{\exp\left(-\frac{1}{2} tr\left(\boldsymbol{K} + \frac{(\mathbf{f}-\boldsymbol{\mu})(\mathbf{f}-\boldsymbol{\mu})^\top}{\nu-2}\right)\right)}{|\boldsymbol{\Sigma}|^{(\nu+2n+1)/2}}\\ &\propto \left(1 + \frac{1}{v-2}(\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)^{-\frac{v+n}{2}}\end{aligned} \tag{6.5}$$

which is a multivariate Student-t distribution $St(\boldsymbol{\mu}, \mathbf{K}, \nu)$. Since the multivariate Student-t distribution is consistent under marginalisation, Shah et al. (2014) conclude that Equation (6.5) is the finite-dimensional distribution of a well defined stochastic process f. We write $f \sim \mathcal{TP}(\nu, k)$.

Definition 6.2.5. *A random real-valued function $f : \mathcal{X} \to \mathbb{R}$ is said to follow a Student-t process $f \sim \mathcal{TP}(0, k, \nu)$, with ν degrees of freedom, mean function $\boldsymbol{\mu} \in \mathcal{X}$ and covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, if any collection of function values has a joint multivariate Student-t distribution such that*

$$(f(x_1), \cdots, f(x_n)) \sim St(\boldsymbol{\mu}, \mathbf{K}, \nu) \tag{6.6}$$

where $\mathbf{K}$ is a covariance matrix with entries $\mathbf{K}_{i,j} = k(x_i, x_j)$ for $i, j = 1, \cdots, n$.

Student-t noise model Unfortunately, with a TP, adding Student-t noise removes analytical tractability. To overcome this issue, Shah et al. (2014) and Zhang and Yeung (2010) propose to add an uncorrelated but dependent noise term, which preserves tractability.

We assume each observation in $\mathbf{y} = \{y_i\}_{i=1}^n$ is to be modelled from a latent process

$\mathbf{f} = \{f(x_i)\}_{i=1}^n$ and a noise vector $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^n$ such that

$$y_i = f_i + \varepsilon_i \qquad \text{for } i = 1, \cdots, n \tag{6.7}$$

where

$$\begin{bmatrix} \mathbf{f} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim St\left(\mathbf{0}_{2n}, \begin{pmatrix} \mathbf{K} & 0 \\ 0 & \sigma_n^2 \mathbf{I}_n \end{pmatrix}, \nu\right). \tag{6.8}$$

Utilizing the linear transformation properties of the multivariate Student-t distribution, applied to the expression:

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon} = \begin{bmatrix} \mathbf{I}_n, \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \boldsymbol{\varepsilon} \end{bmatrix},$$

we can deduce that $\mathbf{y} \sim \mathcal{TP}(\mathbf{0}_n, \mathbf{K} + {\sigma_n}^2 \mathbf{I}_n, \nu)$. Therefore, we obtain a tractable distribution for $\mathbf{y}$, simply incorporating the noise variance into the kernel. Note that $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ in Equation (6.8) are not independent since the scaling parameter ν has an effect on both $\mathbf{f}$ and $\boldsymbol{\varepsilon}$.

Tang et al. (2016) gives a probabilistic interpretation to this noise incorporation. Equation (6.8) can be shown to be equivalent to a noise model following

$$p(\boldsymbol{\varepsilon}|\mathbf{f}, \sigma_n) \sim St\left(\boldsymbol{\varepsilon}\middle|\nu + n, \mathbf{0}, \sigma_n \left(\frac{\nu}{\nu + n}\right) \cdot \left(1 + \frac{1}{\nu}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f}\right)\mathbf{I}\right). \tag{6.9}$$

Thus, the variance of the noise model adjusts to the data fit term $\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f}$ present in the noise-free model marginal log likelihood $-\log p(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}, v)$ derived from Equation (6.2.2). This means that when the noise-free model fits the data well, the added noise will have a smaller variance, and vice versa.

Relation to GPs A TP can be seen as a generalization of a GP. As the parameter ν approaches infinity, the GP converges to the TP in the following sense: If we have $f \sim \mathcal{TP}(\mu, k, \nu)$, where μ represents the mean function and k denotes the covariance function, then the distribution of f tends towards $\mathcal{GP}(\mu, k)$ as ν tends to infinity (Shah et al., 2014, Lemma 2). A TP is in fact the most general elliptical process with

Algorithm 7 BOCPD run length estimation

$(\Xi_0, \nabla_h \Xi_0, \nabla_\theta \Xi_0) \leftarrow (1, 0, 0)$ {Initialize the recursion, set hazard and UPM derivatives to 0.}
Compute the eigenfunctions evaluation $\mathbf{\Phi}$.
Define $\tilde{\Xi}_t$ as $\Xi_t[2 : t+1]$.
for $t = 1$ **to** T **do**
  $(\boldsymbol{\pi}_t, \nabla_\theta \boldsymbol{\pi}_t) \leftarrow$ `UPM_predictive`$(\mathbf{x}_t, t, \mathbf{\Phi})$
  $\mathbf{h} \leftarrow H(1 : t)$
  Update growth probabilities:
  $\tilde{\Xi}_t \leftarrow \Xi_{t-1} \odot \boldsymbol{\pi}_t \odot (1 - \mathbf{h})$
  $\nabla_\theta \tilde{\Xi}_t \leftarrow (1 - \mathbf{h}) \odot (\nabla_\theta \Xi_{t-1} \odot \boldsymbol{\pi}_t + \nabla_\theta \boldsymbol{\pi}_t \odot \Xi_{t-1})$
  $\nabla_h \tilde{\Xi}_t \leftarrow \boldsymbol{\pi}_t \odot (\nabla_h \Xi_{t-1} \odot (1 - \mathbf{h}) - \Xi_{t-1} \odot \nabla_h \mathbf{h})$
  Update change point probabilities:
  $\Xi_t[1] \leftarrow \sum \Xi_{t-1} \odot \boldsymbol{\pi}_t \odot \mathbf{h}$
  $\nabla_\theta \Xi_t[1] \leftarrow \sum \mathbf{h} \odot (\nabla_\theta \Xi_{t-1} \odot \boldsymbol{\pi}_t + \nabla_\theta \boldsymbol{\pi}_t \odot \Xi_{t-1})$
  $\nabla_h \Xi_t[1] \leftarrow \sum \boldsymbol{\pi}_t \odot (\nabla_h \Xi_{t-1} \odot \mathbf{h} + \Xi_{t-1} \odot \nabla_h \mathbf{h})$
  Perform prediction:
  $p(r_t | \mathbf{x}_{1:t-1}) \leftarrow$ normalized Ξ_t.
end for
$p(\mathbf{x}_{1:T}) = \sum \Xi_T$
Compute the evidence
$\nabla p(\mathbf{x}_{1:T}) = (\sum \nabla_h \Xi_T, \nabla_\theta \Xi_T)$
return $(p(\mathbf{x}_{1:T}), \nabla p(\mathbf{x}_{1:T}))$

an analytically-representable density Shah et al. (2014). Furthermore, Tang et al. (2016) argue that a TP with noise incorporated in the kernel as in Equations (6.8) and (6.9) tends to a GP with i.i.d Gaussian noise as $\nu \to \infty$.

6.3 Model

In this Section, we present two variations of our extensions. Section 6.3.1 introduces BOCPD with TP-based UPM, while Section 6.3.3 provides a more detailed exposition of BOCPD with Hilbert space approximate TP UPM (RRSPAR-CP).

6.3.1 BOCPD with TP-based UPM

We propose a BOCPD extension where the UPM is based on a TP process with Student-t noise. We first introduce a TP auto-regressive model of order p where at

time t, the past p values $\mathbf{x}_{t-p:t-1}$ are taken as input and x_t as the observation, i.e.

$$x_t = f(\mathbf{x}_{t-p:t-1}) + \varepsilon_t \tag{6.10}$$

where $f \sim \mathcal{TP}(0, k, \nu)$ and $\boldsymbol{\varepsilon}$ is a dependent Student-t noise with scale parameter σ_n described in Equation (6.8).

Interestingly, by Lemma 6.2.2, we can marginalize out the latent f, to yield an marginal predictive distribution. This yields an auto-regressive TP-based UPM of the form

$$p(x_t|\mathbf{x}_{t-r:t-1}) \sim St(x_t|m_{t,r}, v_{t,r}, \nu + r - 1) \tag{6.11}$$

where

$$\begin{aligned} m_{t,r} &= \mathbf{k}_*^\top \tilde{\mathbf{K}}^{-1} \mathbf{x}_{t,r} \\ v_{t,r} &= \alpha_{t,r} \left(k(x_t, x_t) - \mathbf{k}_*^\top \tilde{\mathbf{K}}^{-1} \mathbf{k}_* \right) \\ \alpha_{t,r} &= \frac{v - 2 + \beta_{t,r}}{v - 3 + r} \\ \beta_{t,r} &= \mathbf{x}_{t,r}^\top \tilde{\mathbf{K}}^{-1} \mathbf{x}_{t,r}. \end{aligned} \tag{6.12}$$

Here $\mathbf{x}_{t,r} = \mathbf{x}_{t-r+1:t-1}$, $\tilde{\mathbf{K}}_{i,j} = k(\mathbf{x}_{i-p+1:i}, \mathbf{x}_{j-p+1:j}) + \sigma_n^2 \delta_{i,j}$ for $i, j = t - r, \cdots, t - 2$, $\delta_{i,j}$ denotes the Kronecker delta and $\mathbf{k}_*$ is an $(r-1)$-dimensional vector with the ith entry being $k(\mathbf{x}_{t-p:t-1}, \mathbf{x}_{i-p+1:i})$ for $i = t - r, \cdots, t - 2$.

The predictive mean $m_{t,r}$ has the same form as for a GP (assuming the same kernel and hyperparameters). However, due to the differing marginal likelihood between TP and GP, the predictive mean differs after learning the hyperparameters. Unlike a GP, the TP model exhibits more adaptive predictive volatility based on the training observations. The parameter $\beta_{t,r}$ explicitly depends on $\mathbf{x}_{t,r}$. When $\beta_{t,r}$ exceeds $(r-1)$, TP's predictive variance surpasses that of a GP, and vice versa. In fact, assuming $\mathbf{x}_{t,r}$ is drawn from a GP prior $\mathcal{N}(0, \tilde{\mathbf{K}})$, $\beta_{t,r}$ follows a Chi-squared distribution with mean $(r - 1)$. Consequently, if observations have similar variance as expected under a GP prior, TP's covariance is comparable to that of a GP. However, significantly larger or smaller variability in the observations leads to higher or lower posterior uncertainty

in TP, respectively.

Note that this TP-UPM inherits the uniform time step constraint from its GPAR counterpart. Consequently, we are limited to uniformly sampling discrete observations. At present, to our knowledge, there is no immediate solution available for handling missing data.

6.3.2 Hilbert space approximate Student-t processes

The TP UPM in Equation (6.11) inherits the same cubic computational cost of GPs, which is prohibitive for most applications. We propose a reduced-rank implementation of the Student-t Process UPM based on the novel *Hilbert space* method for reduced-rank kernel approximation Solin and Särkkä (2020). Solin and Särkkä (2020) obtain approximate eigendecompositions of stationary covariance functions in terms of an eigenfunction expansion of the Laplace operator in a compact subset of $\mathbb{R}^d$.

The *Hilbert space reduced-rank* method provides a different advantage in our case compared to other reduced-rank approximations :

(i) The Laplace-based feature vectors are independent of the particular choice of kernel, including the kernel hyperparameters. Gradient computation is thus facilitated, which in turn speeds up the learning phase. We refer to Section 6.5 for more details.

(ii) The decay of the expansion coordinates is fast. Hence, a good approximation can be obtained with relatively few basis points. As an example, Solin and Särkkä (2020) obtains a good approximation to univariate RBF kernels with only 12 eigenfunctions. They argue that adding more eigenfunctions has negligible effect on the approximation accuracy.

Hilbert space reduced-rank kernel

In this section, we present a summary of the mathematical details of the Hilbert space based reduced-rank kernels introduced by Solin and Särkkä (2020).

Bochner representation Hilbert space methods for reduced-rank kernels are constructed via the Bochner's theorem (Bochner, 1932; Rudin, 2017), which was introduced in the previous sections (Theorem 4.3.1). The theorem states that any bounded, continuous, and shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') := k(\boldsymbol{\tau})$ with $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, is the inverse Fourier transform of a bounded positive measure. More precisely, the kernel k can be represented as follows:

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^d} \exp(iw^\top \boldsymbol{\tau}) d\mu(w) \tag{6.13}$$

where $\mu(w)$ is a positive definite measure.

If the measure $\mu(w)$ admits a *spectral density* $S(w)$, we can furthermore express the following Fourier identities

$$\begin{aligned} k(\boldsymbol{\tau}) &= \frac{1}{(2\pi)^d} \int \exp(iw^\top \boldsymbol{\tau}) S(w) dw \\ S(w) &= \int k(\boldsymbol{\tau}) \exp(-iw^\top \boldsymbol{\tau}) d\boldsymbol{\tau}. \end{aligned} \tag{6.14}$$

In the isotropic case where the covariance function only depends on the Euclidian norm $||\boldsymbol{\tau}||$ such that $k(\boldsymbol{\tau}) = k(||\boldsymbol{\tau}||)$, the spectral density is also only dependent on the norm of w i.e. $S(w) = S(||w||)$.

Covariance operator as a pseudo-differential operator We can define a covariance operator $\mathcal{K}$ associated with each covariance function k as

$$\mathcal{K}f = \int k(\cdot, \mathbf{x}') f(\mathbf{x}') d\mathbf{x}' \tag{6.15}$$

for any regular functions f. When k is stationary then $\mathcal{K}$ is translation invariant. Thus, we can express the Fourier transform of $\mathcal{K}$ as a transfer function, which is the spectral density $S(\cdot)$ itself. Indeed, one can verify that $\mathcal{F}[\mathcal{K}f](w) = S(w)\mathcal{F}[f](w)$ where $\mathcal{F}[\cdot]$ denotes the Fourier transform of its argument.

We consider the isotropic case $S(w) = S(||w||)$. We further assume that $S(\cdot)$ is

regular enough to be represented as a polynomial expansion i.e.

$$S(||w||) = a_0 + a_1||w||^2 + a_2(||w||^2)^2 + a_3(||w||^2)^3 + \cdots \tag{6.16}$$

Recall that the transfer function of the Laplace operator ∇^2 is $-||w||^2$ i.e. $\mathcal{F}[\nabla^2 f](w) = -||w||^2\mathcal{F}[f](w)$. Thus from Equation (6.16), we have

$$\begin{aligned}
\mathcal{F}[\mathcal{K}f](w) &= S(||w||)\mathcal{F}[f](w) \\
&= \left[a_0 + a_1||w||^2 + a_2(||w||^2)^2 + a_3(||w||^2)^3 + \cdots\right]\mathcal{F}[f](w) \\
&= a_0\mathcal{F}[f](w) - a_1\mathcal{F}[\nabla^2 f](w) - a_2\mathcal{F}[(\nabla^2)^2 f](w) - a_3\mathcal{F}[(\nabla^2)^3 f](w) + \cdots \\
&= \mathcal{F}[a_0 - a_1\nabla^2 f - a_2(\nabla^2)^2 f - a_3(\nabla^2)^3 f + \cdots](w).
\end{aligned} \tag{6.17}$$

From the equality (6.17), we get the following representation of $\mathcal{K}$, which defines a pseudo-differential operator as a series of Laplace operator

$$\mathcal{K} = a_0 - a_1\nabla^2 - a_2(\nabla^2)^2 - a_3(\nabla^2)^3 + \cdots . \tag{6.18}$$

Hilbert space approximation of $\mathcal{K}$ We now form a Hilbert space approximation for the pseudo-differential operator defined in Equation (6.18). Consider the eigenvalue problem for the Laplace operator ∇^2 in the compact subset $\Omega \subset \mathbb{R}^d$ and with Dirichlet boundary conditions

$$\begin{aligned}
-\nabla^2\phi_j(\mathbf{x}) &= \lambda_j\phi_j(\mathbf{x}) \quad \text{if} \quad \mathbf{x} \in \Omega, \\
\phi_j(\mathbf{x}) &= 0 \qquad\qquad \text{if} \quad \mathbf{x} \in \partial\Omega
\end{aligned} \tag{6.19}$$

where $\{\phi_j\}_{j=1}^{\infty}$ and $\{\lambda_j\}_{j=1}^{\infty}$ are the set of eigenvalues and eigenfunctions of the Laplacian operator. Because $-\nabla^2$ is a positive definite Hermitian operator, the set of eigenfunction $\{\phi_j\}_{j=1}^{\infty}$ is orthonormal with respect to the inner product

$$\langle f, g\rangle = \int_{\Omega} f(\mathbf{x})g(\mathbf{x})d\mathbf{x} \tag{6.20}$$

that is

$$\int_{\Omega} \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x} = \delta_{i,j} \tag{6.21}$$

and all eigenvalues $\{\lambda_j\}$ are real and positive.

The Laplace operator can be assigned a formal kernel

$$l(\mathbf{x}, \mathbf{x}') = \sum_j \lambda_j \phi_j(\mathbf{x})\phi_j(\mathbf{x}') \tag{6.22}$$

in a sense that

$$\begin{aligned} -\nabla^2 f(\mathbf{x}) &= \sum_j \lambda_j \langle f, \phi_j \rangle \phi_j(\mathbf{x}) \qquad \textit{(spectral decomposition)} \\ &= \int_{\Omega} l(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mathbf{x}'. \end{aligned}$$

Similarly, we can define the kernel of the power representation the Laplace operator as

$$l^s(\mathbf{x}, \mathbf{x}') = \sum_j \lambda_j^s \phi_j(\mathbf{x})\phi_j(\mathbf{x}') \tag{6.23}$$

for $s = 1, 2, \cdots$, in a sense that due the orthonormality of the basis

$$-(\nabla^2)^s f(\mathbf{x}) = \int_{\Omega} l^s(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mathbf{x}'.$$

This implies that we also have

$$\begin{aligned} &\left[a_0 + a_1(-\nabla^2) + a_2(-\nabla^2)^2 + \cdots\right] f(\mathbf{x}) \\ &= \int_{\Omega} \left[a_0 + a_1 l^1(\mathbf{x}, \mathbf{x}') + a_2 l^2(\mathbf{x}, \mathbf{x}') + \cdots\right] f(\mathbf{x}') d\mathbf{x}'. \end{aligned} \tag{6.24}$$

The left hand side is $\mathcal{K}f$ as defined in Equation (6.18). Thus from Equation (6.15), we conclude that

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &\approx a_0 + a_1 l^1(\mathbf{x}, \mathbf{x}') + a_2 l^2(\mathbf{x}, \mathbf{x}') + \cdots \\ &= \sum_j \left[a_0 + a_1\lambda_j + a_2\lambda_j^2 + \cdots\right] \phi_j(\mathbf{x})\phi_j(\mathbf{x}'). \end{aligned} \tag{6.25}$$

By letting $||w||^2 = \lambda_j$ the spectral density in Equation (6.16) becomes

$$S(||w||) = a_0 + a_1\lambda_j + a_2\lambda_j^2 + a_3\lambda_j^3 + \cdots .$$

and substituting in Equation (6.25) leads to the final approximation

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} S(\sqrt{\lambda_j})\phi_j(\mathbf{x})\phi_j(\mathbf{x}') \tag{6.26}$$

In one dimension For one dimension within a closed interval $\Omega = [-L, L] \subset \mathbb{R}$ where L is some positive real number, the solution to the Laplacian eigenvalue problem in Equation (6.19) is independent of the specific choice of covariance function and is given by

$$\lambda_j = \left(\frac{\pi j}{2L}\right)^2, \tag{6.27a}$$

$$\phi_j(x) = \frac{1}{\sqrt{L}} \sin\left(\sqrt{\lambda_j}(x + L)\right). \tag{6.27b}$$

for $j = 1, \cdots, m$ where m denotes the number of basis functions.

In d dimensions In the d-dimensional case, we consider rectangular domain $\Omega = [-L_1, L_1] \times \cdots \times [-L_d, L_d]$ with Dirichlet boundary conditions. The number of eigenfunctions and eigenvalues in the approximation is equal to the number of d-tuples, that is, possible combinations of univariate eigenfunctions over all dimensions.

Every k-th dimension has a number of univariate basis functions equal to m_k with indices ranging from $1, \cdots, m_k$. Let $\mathbf{S} \in \mathbb{N}^{m^* \times d}$ be the matrix of all these d-tuples indices with $m^* = \prod_{k=1}^{d} m_k$.

Each multivariate eigenfunction $\phi_j^* : \Omega \to \mathbb{R}$ corresponds to the product of the univariate eigenfunctions whose indices corresponds to the j-th element of the d-tuples $\mathbf{S}_{j,\cdot}$, and each multivariate eigenvalue $\boldsymbol{\lambda}_j^*$ is a d-vector with elements that are the univariate eigenvalues whose indices corresponds to the j-th elements of the d-tuples $\mathbf{S}_{j,\cdot}$. Thus for $\mathbf{x} = (x_1, \cdots, x_d) \in \Omega$ and $j = 1, \cdots, m^*$, we have

$$\boldsymbol{\lambda}_j^* = \{\lambda_{\mathbf{S}_{j,k}}\}_{k=1}^d = \left\{ \left(\frac{\pi \mathbf{S}_{j,d}}{2L} \right)^2 \right\}_{k=1}^d, \tag{6.28}$$

$$\phi_j^*(\mathbf{x}) = \prod_{k=1}^d \phi_{\mathbf{S}_{j,k}}(x_k) = \prod_{k=1}^d \frac{1}{\sqrt{L_k}} \sin\left(\sqrt{\lambda_{\mathbf{S}_{j,k}}} (x_d + L_k) \right) \tag{6.29}$$

for $j = 1, \cdots, m^*$.

The approximate covariance function is then

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^m S\left(\sqrt{\boldsymbol{\lambda}_j^*} \right) \phi_j^*(\mathbf{x}) \phi_j^*(\mathbf{x}') \tag{6.30}$$

where S denotes the d-dimensional spectral density of the covariance functions with argument $\sqrt{\boldsymbol{\lambda}_j^*}$ that denotes the element-wise square root of the vector $\boldsymbol{\lambda}_j^*$.

Comparison to RFF This *Hilbert space reduced-rank* method exhibits similarities with the RFF detailed in Section 4.3.1, as they both originate from Bochner's theorem and thus possess spectral characteristics. However, they differ fundamentally in their construction. While RFF is inherently random, being generated by direct sampling from the spectral density of the kernel, the method presented here is deterministic and relies on an approximation of the spectral density. Specifically, this is accomplished by conducting a Taylor expansion of the spectral density as outlined in Equation (6.16), thereby establishing a parallel with the spectral decomposition of the Laplace operator.

Student-t processes with Hilbert space kernel

We now assume a Student-t process as defined in definition 6.2.5 with a stationary kernel with hyperparameters θ. For the univariate case with observations within a closed interval $\Omega = [-L, L] \subset \mathbb{R}$, where L is some positive real number, we can approximate the stationary kernel k with hyperparameters θ with a kernel representation given by Equation 6.26 where $S_\theta(\cdot)$ is the spectral density of k.

In particular, for a Gaussian kernel $k(x - x') = \sigma^2 \exp\left(-(x - x')^2 / 2\ \ell\right)$ with

scaling parameter σ and length-scale parameter ℓ, the corresponding spectral density is defined as $S_\theta(w) = \sigma\sqrt{2\pi\ell^2}\exp(-\frac{\pi^2\ell^2 w^3}{2})$, where $\{\phi_j\}_{j=1}^{\infty}$ and $\{\lambda_j\}_{j=1}^{\infty}$ are the sets of eigenfunctions and eigenvalues of the Laplace operator ∇^2 in Ω as described in Equations (6.27a) and (6.27b).

As discussed, the eigenvalues λ_j are monotonically increasing with j and for bounded kernel the spectral density goes to zero with higher frequencies. Thus, a good approximation is obtained by truncating the expansion in Equation (6.26) to the first m terms. This results in an approximate covariance described by:

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^{m} S_\theta(\sqrt{\lambda_j})\phi_j(\mathbf{x})\phi_j(\mathbf{x}'). \tag{6.31}$$

Similarily, we can form an approximate eigendecomposition of the covariance matrix

$$\mathbf{K} \approx \mathbf{\Phi}^\top \mathbf{\Lambda} \mathbf{\Phi} \tag{6.32}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries $\{S_\theta(\sqrt{\lambda_j})\}_{j=1}^{m}$ and $\mathbf{\Phi}$ is a matrix of eigenfunction evaluations such that $\mathbf{\Phi}_{i,j} = \phi_i(x_j)$. The quality on the approximation also relies on the choice of closed interval $[-L, L]$. Solin and Särkkä (2020) simply normalized the data and adjust L accordingly.

6.3.3 BOCPD with Hilbert space approximate TP UPM

Using the reduced-rank Hilbert method, the predictive distribution of Equation (6.11) becomes $p(x_t|\mathbf{x}_{t-r:t-1}) \sim St(x_t|m_{t,r}, v_{t,r}, \nu + r - 1)$. Further, using the Woodbury matrix inversion formula

$$\begin{aligned}
m_{t,r} &= \phi(x_{t-1})^\top \mathbf{Q}_{t,r} \mathbf{\Phi}_{t,r} \mathbf{x}_{t,r} \\
v_{t,r} &= \alpha_{t,r} \left(\phi(x_{t-1})^\top \mathbf{Q}_{t,r} \phi(x_{t-1}) \right) \\
\alpha_{t,r} &= \frac{(\nu - 2)\sigma_n^2 + \beta_{t,r}}{\nu - 1} \\
\beta_{t,r} &= ||\mathbf{x}_{t,r}||_2^2 - (\mathbf{x}_{t,r}\mathbf{\Phi}_{t,r})^\top \mathbf{Q}_{t,r} (\mathbf{\Phi}_{t,r}\mathbf{x}_{t,r})
\end{aligned} \tag{6.33}$$

Algorithm 8 RRSPAR-CP UPM implementation

Function `UPM_predictive`$(\mathbf{x}_t, t, \mathbf{\Phi})$:
Inputs: $r_{\max}, \mathbf{Q}_{t-1,r_{\max}}$, $\mathbf{u}_{t-1} := \mathbf{\Phi}_{t-1,r_{\max}}\mathbf{x}_{t-1,r_{\max}}$ {Inputs from previous iterations}
$\mathbf{Q} \leftarrow$ `h_update`$(\mathbf{Q}_{t-1,r_{\max}}, \phi(x_{t-2}))$ {Eq.(6.37)}
$\mathbf{u} \leftarrow \mathbf{u}_{t-1} + x_{t-1}\phi(x_{t-2})$
if $r_{\max} + 1 > R_{\max}$ **then**
  $\mathbf{Q} \leftarrow$ `h_downdate`$(\mathbf{Q}, \phi(x_{t-r_{\max}}))$
  $\mathbf{u} \leftarrow \mathbf{u} - x_{t-r_{\max}}\phi(x_{t-r_{\max}-1})$
end if
$r_{\max} \leftarrow \min(R_{\max}, r_{\max} + 1)$
$\mathbf{Q}_{t,r_{\max}} \leftarrow \mathbf{Q}$
$\mathbf{u}_t \leftarrow \mathbf{u}$
for $r = r_{\max}$ **to** $t = 1$ **do**
  $m_{t,r} \leftarrow \phi(x_{t-1})^\top \mathbf{Q}\mathbf{u}$
  $\beta_{t,r} \leftarrow ||\mathbf{x}_{t,r}||_2^2 - \mathbf{u}^\top \mathbf{Q}\mathbf{u}$
  Compute $\alpha_{t,r}$ and $v_{t,r}$ {Eq (6.33)}
  $\pi_{t,r} \leftarrow p(x_t|\mathbf{x}_{t-r:t-1})$ {Eq.(6.11)}
  Compute $\nabla m_{t,r}, \nabla \beta_{t,r}, \nabla \alpha_{t,r}$ and $\nabla v_{t,r}$
  Compute $\nabla \pi_{t,r}$ with chain rule
  $\mathbf{Q} \leftarrow$ `v_downdate`$(\mathbf{Q}, \phi(x_{t-r}))$ {Eq.(6.36) }
  $\mathbf{u} \leftarrow \mathbf{u} - x_{t-r}\phi(x_{t-r-1})$
end for
return $(\boldsymbol{\pi}_t, \nabla \boldsymbol{\pi}_t)$.

with $\mathbf{x}_{t,r} = \mathbf{x}_{t-r+1:t-1}$. Here, $\mathbf{\Phi}_{t,r}$ is a $m \times (r-1)$ matrix of eigenfunctions with i, j entry $\phi_i(\mathbf{x}_j)$ for $i = 1, \cdots, m$ and $j = t-r, \cdots, t-2$, and $\mathbf{Q}_{t,r}$ is the $m \times m$ precision matrix such that

$$\mathbf{Q}_{t,r} = \left(\mathbf{\Phi}_{t,r}\mathbf{\Phi}_{t,r}^\top + \sigma_n^2 \mathbf{\Lambda}^{-1}\right)^{-1}. \tag{6.34}$$

6.4 Implementation details

As noted in Equation (6.27), the basis function in the reduced-rank approximation does not depends on covariance function hyperparameters. Thus the eigenfunctions can be evaluated once and stored in a cached $T \times m$ matrix $\mathbf{\Phi}$ through the learning process, with $O(Tm)$ space complexity. If the number of observations T is so large that storing is not feasible, evaluation can be carried out in blocks or only when necessary.

Pruning the run length distribution In a naive implementation, all the run lengths are retained and the posterior $p(r_t|\mathbf{x}_{1:t})$ for $r_t = \{1, \cdots, t\}$ forms a vector of size t at every update step. In practice the run length distribution is highly peaked. A modification of the algorithm is to *prune* out the run length probability estimates with a total mass below a certain threshold, i.e $\leq 1/R_{\max}$; or to only consider the $R_{\max}$ most probable values, i.e $|r_t| \leq R_{\max}$ Adams and MacKay (2007). This yields a running complexity of $O(TR_{\max}^2 m^2)$ for the reduced-rank TP-based UPM.

Vertical Rank-One update We can improve the implementation further, by performing a rank-one update of the precision matrix $\mathbf{Q}_{t,r}$ across run lengths. Indeed, at time t, the product $\mathbf{\Phi}_{t,r}\mathbf{\Phi}_{t,r}^\top$ in Equation (6.34) can be updated across run lengths as

$$\mathbf{\Phi}_{t,r}\mathbf{\Phi}_{t,r}^\top = \mathbf{\Phi}_{t,r-1}\mathbf{\Phi}_{t,r-1}^\top + \phi(x_{t-r})\phi(x_{t-r})^\top. \tag{6.35}$$

Thus, knowing $\mathbf{Q}_{t,r_{\max}}$, where $r_{\max}$ stands for the maximum run length size at time t, we can use the Sherman-Morrison inversion formula to obtain the following recursion

$$\mathbf{Q}_{t,r-1} = \left(\mathbf{I}_m + \frac{\mathbf{Q}_{t,r}\phi(x_{t-r})\phi(x_{t-r})^\top}{1 - \phi(x_{t-r})^\top \mathbf{Q}_{t,r}\phi(x_{t-r})}\right)\mathbf{Q}_{t,r}. \tag{6.36}$$

Equation (6.36) can be implemented as an outer product of two matrix-vector products. To make the evaluation fast, we used the specialized BLAS routines for rank-one update (i.e. the scipy method `linalg.blas.dger` for Python). While reports have mentioned potential numerical instability with rank-one updates to the precision matrix (Schölkopf and Smola, 2018, Chapter 10), our implementation has not exhibited any such issues.

This proposed rank-one update of $\mathbf{Q}_{t,r}$, together with an efficient update of the product $\mathbf{\Phi}_{t,r}\mathbf{x}_{t,r}$ in Equation (6.33) yields a running complexity of $O(TR_{\max}m^2)$.

Horizontal Rank-One Update We can also perform a *horizontal* update of the precision matrices across time t. Let $\mathbf{Q}_{t,r_{\max}}$ denote the precision matrix associated with the largest run length at time t (i.e. for $r_t = |r_t|$). Using Equation (6.35), we

obtain

$$\mathbf{Q}_{t,r_{\max}} = \left(\mathbf{I}_m - \frac{\mathbf{Q}_{t-1,r_{\max}}\phi(x_{t-2})\phi(x_{t-2})^\top}{1+\phi(x_{t-2})^\top\mathbf{Q}_{t-1,r_{\max}}\phi(x_{t-2})}\right)\mathbf{Q}_{t-1,r_{\max}}. \tag{6.37}$$

To maintain consistency with *pruning*, an additional rank-one downdate is necessary when $|r_{t-1}|+1 > R_{\max}$, to remove the information carried by $\phi(x_{t-|r_{t-1}|})$, as given below

$$\mathbf{Q}_{t,r_{\max}} = \left(\mathbf{I}_m + \frac{\mathbf{Q}_{t,r_{\max}}\phi(x_{t-|r_{t-1}|})\phi(x_{t-|r_{t-1}|})^\top}{1-\phi(x_{t-|r_{t-1}|})^\top\mathbf{Q}_{t,r_{\max}}\phi(x_{t-|r_{t-1}|})}\right)\mathbf{Q}_{t,r_{\max}}. \tag{6.38}$$

Maximum A-Posteriori (MAP) segmentation For the identification of change points, we used a variation of the *MAP segmentation* algorithm proposed by Knoblauch and Damoulas (2018). We compute MAP_t, an estimator of the density of the run length MAP estimate before t with the recursion

$$\mathrm{MAP}_t = \max_r\{p(r_t = r|\mathbf{x}_{1:t})\mathrm{MAP}_{t-r-1}\}. \tag{6.39}$$

For r_t^*, the maximizer of Equation (6.39) at time t, the MAP segmentation is $S_t = S_{t-r_t^*-1} \cup \{(t-r_t^*)\}$, $S_0 = \emptyset$, where $t' \in S_t$ means a CP occurs at $t' \leq t$.

6.5 Hyperparameter learning

Following Saatçi et al. (2010), the hyperparameters $\Theta := (\theta, \nu, \sigma_n)$ where θ refers to the kernel hyperparameters, are learned by minimizing the marginal negative log likelihood

$$\log p(\mathbf{x}_{1:T}|\Theta) = -\sum_{i=1}^{T}\log p(x_t|\mathbf{x}_{1:t-1},\Theta). \tag{6.40}$$

Saatçi et al. (2010) optimize the hyperparameters on a test subset $\{\mathbf{x}_{1:T'}\}$ by running the BOCPD multiple times to find $\tilde{\Theta} = \arg\min_\Theta\{\log p(\mathbf{x}_{1:T'}|\Theta)\}$. The gradient of the log likelihood is obtained from the gradient of the one-step-ahead predictor gradi-

ents $\nabla p(x_t|\mathbf{x}_{1r:t-1})$. The terms $\nabla p(x_t|\mathbf{x}_{1:t-1})$ are themselves computed by iteratively calculating the gradient of the UPM, $\nabla p(x_t|\mathbf{x}_{t-r:t-1})$, the gradient of the hazard rate $\nabla p(r_t|r_{t-1})$ and then propagating forward using the chain rule Saatçi et al. (2010). These computations are consistent with hyperparameter learning in other on-line GP methods Ranganathan et al. (2011).

For GP-based UPM, the computation and forward propagation of the gradient is particularly expensive and accounts for most of the training time. In our case, computation of the UPM gradient is easier since the feature vectors $\mathbf{\Phi}_{t,r}$ are independent of the hyperparameters Θ. We are left from Equation (6.33) with

$$\begin{aligned}
\nabla_\Theta m_{t,r} &= \phi(x_{t-1})^\top \nabla_\Theta \mathbf{Q}_{t,r} \mathbf{\Phi}_{t,r} \mathbf{x}_{t,r} \\
\nabla_\Theta v_{t,r} &= \frac{v_{t,r}}{\alpha_{t,r}} \nabla \alpha_{t,r} + \alpha_{t,r} \left(\phi(x_{t-1})^\top \nabla_\Theta \mathbf{Q}_{t,r} \phi(x_{t-1}) \right) \\
\nabla_\Theta \beta_{t,r} &= -(\mathbf{x}_{t,r} \mathbf{\Phi}_{t,r})^\top \nabla_\Theta \mathbf{Q}_{t,r} (\mathbf{\Phi}_{t,r} \mathbf{x}_{t,r})
\end{aligned} \tag{6.41}$$

where

$$\begin{aligned}
\nabla_\theta \mathbf{Q}_{t,r} &= \sigma_n^2 \mathbf{Q}_{t,r} \left(\mathbf{\Lambda}^{-2} \nabla_\theta \mathbf{\Lambda} \right) \mathbf{Q}_{t,r} \\
\nabla_\nu \mathbf{Q}_{t,r} &= 0 \\
\nabla_{\sigma_n} \mathbf{Q}_{t,r} &= \frac{1}{2\sigma_n} \mathbf{Q}_{t,r} \mathbf{\Lambda}^{-1} \mathbf{Q}_{t,r}.
\end{aligned} \tag{6.42}$$

The term $\mathbf{\Lambda}^{-2} \nabla_\theta \mathbf{\Lambda}$ in Equation (6.42) is independent of the observations and thus can be computed once at the beginning of each optimizing step and reused throughout the BOCPD iterations. Equation (6.41) and (6.42) provide a simple computational routine for the gradient, once the precision matrix $\mathbf{Q}_{t,r}$ update is obtained. The gradient of the UPM, $\nabla p(x_t|\mathbf{x}_{t-r:t-1})$ is then derived from the gradient of UPM parameters.

6.6 Experiments

We compare our scheme to the two GP-based UPM variants introduced in Saatçi et al. (2010), namely ARGP and GPTS. We also include as baseline the normal i.i.d UMP (TIM) of Adams and MacKay (2007). We use the acronyms HSSPAR-CP to refer to our reduced-rank Student-t process-based UPM as in Equation (6.33), and HSGPAR-CP for an equivalent reduced-rank GP-based UPM. We test the algorithms on four real data sets (3 in 1D and 1 in 3D). The average one-step-ahead negative log likelihood (NLL) and the mean squared error relative to the predictive mean (MSE) are used as evaluation metrics. Results are presented in Table 6.4.

6.6.1 Settings

We use a hazard function with a trainable constant hazard rate h initialized at 100, which yields a conditional prior $p(r_t|r_{t-1})$ with probability of a change point equal to 0.01. Following Saatçi et al. (2010), for GPTS we used a rational quadratic kernel and a Gaussian kernel for the auto-regressive variants. The GPTS execution time is improved by assuming uniform discrete observation time and exploiting the *Toeplitz* structure of the covariance function (Saatçi et al., 2010). The ARGP implementation includes horizontal and vertical rank-one Cholesky updates.

For HSSPAR, the trainable hyperparameters consist of the UPM parameters $\Theta := (\theta, \nu, \sigma_n)$ where θ refers to the kernel hyperparameters. Our implementations of HSSPAR and HSGPAR use the Hilbert space reduced-rank kernel derived from Gaussian kernels with the number of basis functions m ranging from 5 to 15. For auto-regressive UPM (GPAR and HSSPAR variants), we use lag parameter $p = 1, 2, 3$. We observed that for larger p, the computational advantage of HSSPAR reduces, since as discussed earlier, the number of multivariate basis functions increases exponentially with dimension. Other authors make similar observations for multivariate HSGP regression (Riutort-Mayol et al., 2022).

Table 6.1: Results of predictive performance on Nile data. The results are provided with 95% error bars and the p-value testing the null hypothesis that methods are equivalent to the best performing method, according to NLL, using one sided t-test. (·)-CP refers to the BOCPD variant of the respective method.

Method	Negative Log Likelihood	p-value	MSE	p-value	time(s)
Nile Data (200 training points, 463 Test points)					
RRGPAR-CP	1.1480 (± 0.0564)	0.072	0.5756 (±0.0977)	0.480	43.18
RRSPAR-CP	**1.0984** (± 0.0653)	N/A	0.5783 (±0.0995)	N/A	44.52
GPTS	1.2313 (±0.0449)	< 0.001	0.6050 (±0.0942)	0.301	2.74
GPTS-CP	1.1468 (±0.0533)	0.067	**0.5381** (±0.0890)	0.208	5.20
GPAR	1.1729 (±0.0527)	0.020	0.5587 (±0.0978)	0.355	142.66
GPAR-CP	1.1481 (±0.0587)	0.079	0.5792 (±0.0964)	0.493	267.17
TIM	1.1769 (±0.0852)	0.065	0.6644 (±0.1029)	0.081	N/A

6.6.2 Nile data

The *Nile* data set records the lowest annual water levels of the Nile river during the period 622-1284. The data has been used for change point detection in Garnett et al. (2009) and Saatçi et al. (2010). Following Saatçi et al. (2010), we learn the hyper-parameters on the first 200 entries and evaluate the performance on the remaining period 822-1284. A structural change in the data is known to occur in year 715 due to an upgrade in ancient sensor technology to the nilometer. Results are given in Table 6.4. The run length posterior for HSSPAR is displayed in Figure 6.1. We can see by comparing HSGPAR-CP to GPAR-CP that the reduced-rank approximation does not alter the performance significantly. HSSPAR-CP outperforms both GPTS-CP and GSPAR-CP in terms of NLL. The error bars tend to be larger than desired, but this is something that was also observed in Saatçi et al. (2010), and attributable to the small test size (463 points). In Figure 6.1 we can also see that HSSPAR correctly captures the known change point at the year 715. While Saatçi et al. (2010) identified 18 CPs, our algorithm is more robust in that it only detects 9 CPs.

6.6.3 Well Log data

The *Well Log* data set contains 4050 measurements of radioactivity taken during the drilling of a well. These data have been studied in the context of change point de-

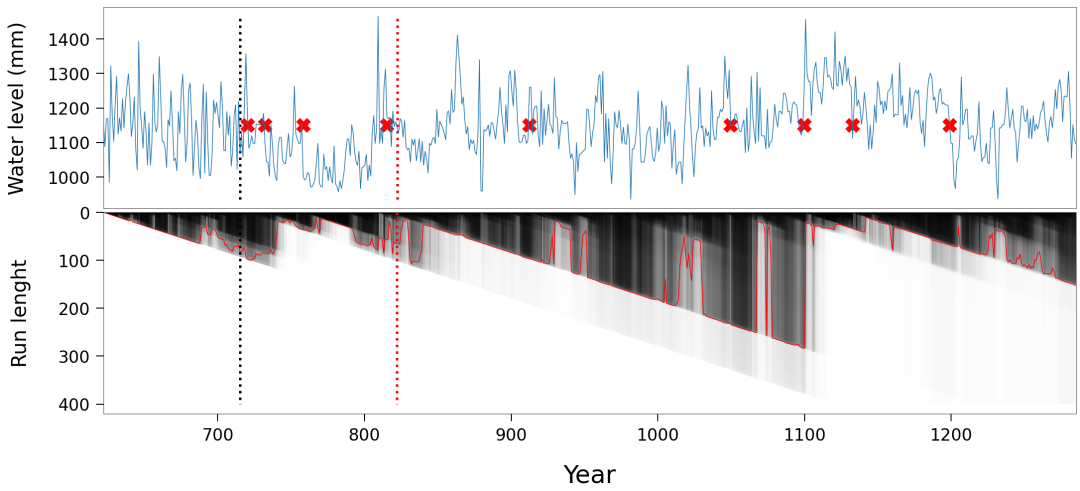


Figure 6.1: Results for the Nile Record data with RRSPAR-CP. **Top:** The vertical dashed red line represents the boundary between train set and test sets. The vertical dashed black line marks the installation of the nilometer in 715. The small red crosses represents alert locations obtained from MAP segmentation. **Bottom:** The run length CDF and its median in solid red.

tection by Ruanaidh and Fitzgerald (2012) and by Fearnhead and Clifford (2003). The data set contains many outliers. Some authors, e.g. Adams and MacKay (2007); Levy-leduc and Harchaoui (2007) remove these before running the change point algorithms; however, outliers are retained by other authors, e.g. Fearnhead and Rigaill (2019) and Knoblauch et al. (2018). In our case, we use the data unfiltered. Results in Table 6.4 show slightly better performance for HSSPAR-CP compared to HSGPAR-CP and GPAR-CP even though this advantage might lack statistical significance. For this data set, we see the effectiveness of the reduced-rank formulation when the training set becomes relatively large (≥ 1000). The fitting of HSSPAR-CP and HSGPAR-CP is $> 20\times$ faster than that of GPAR-CP in our experiment. In terms of alerted change points, on the unfiltered data, HSSPAR-CP identifies 25 CPs compared to 44 for GPAR-CP in Saatçi et al. (2010). Notably, when the data is filtered, the number of CPs reduces to 22 for HSSPAR-CP, indicating that applying filtering results in only 3 additional CPs.

Table 6.2: Results of predictive performance on Well-Log. The results are provided with 95% error bars and the p-value testing the null hypothesis that methods are equivalent to the best performing method, according to NLL, using one sided t-test. (·)-CP refers to the BOCPD variant of the respective method.

Method	Negative Log Likelihood	p-value	MSE	p-value	time(s)
Well-Log Data (1000 training points, 3047 Test points)					
RRGPAR-CP	0.1927 (±0.0343)	0.390	**0.1165** (±0.0109)	0.312	528.75
RRSPAR-CP	**0.1875** (±0.0321)	N/A	0.1194 (±0.0123)	N/A	659.20
GPTS	0.5557 (±0.0480)	< 0.001	0.1575 (±0.0199)	0.007	17.88
GPTS-CP	0.2489 (±0.0446)	< 0.001	0.1201 (± 0.0115)	0.460	78.24
GPAR	0.3001 (±0.0383)	< 0.001	0.1704 (±0.0380)	0.023	11,596.64
GPAR-CP	0.1926 (±0.0342)	0.392	0.1166 (±0.0110)	0.316	13,610.75
TIM	0.2562 (±0.0287)	0.003	0.1921 (±0.0275)	0.002	N/A

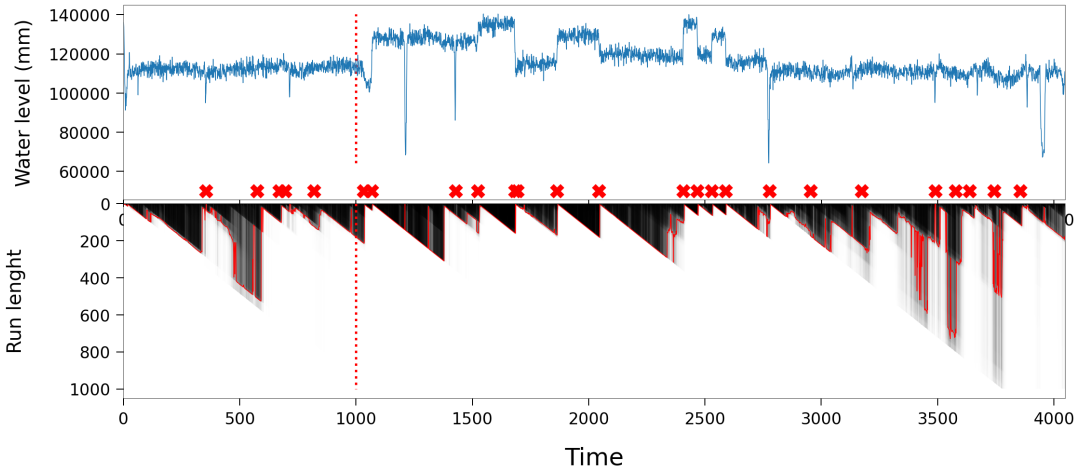


Figure 6.2: Results for the unfiltered Well Log data with HSSPAR-C. **Top:** The vertical dashed red line represents the boundary between train and test sets. The small red crosses represents alert locations obtained from MAP segmentation. **Bottom:** The run length CDF (black) and its median (red).

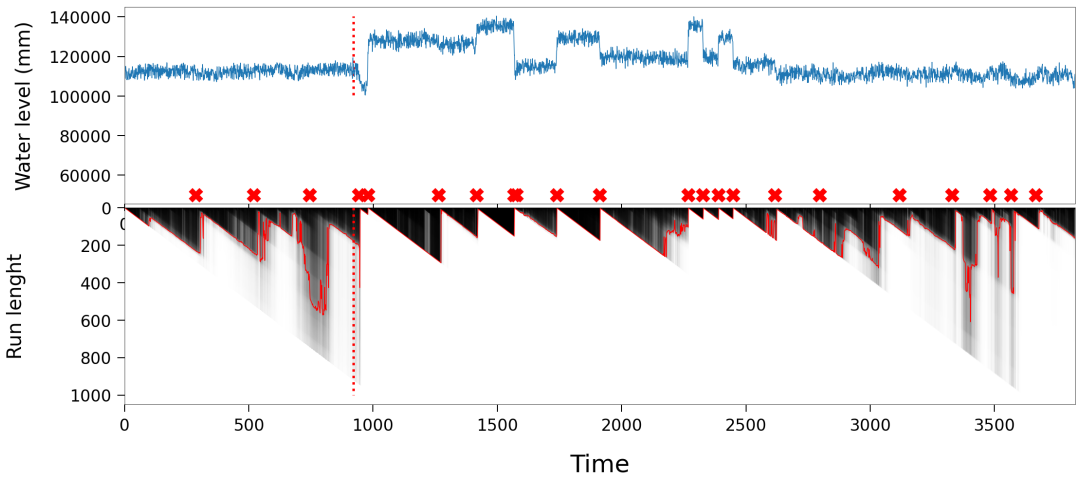


Figure 6.3: Results for the filtered Well Log data with HSSPAR-CP. **Top:** The vertical dashed red line represents the boundary between train and test sets. The small red crosses represents alert locations obtained from MAP segmentation. **Bottom:** The run length CDF (black) and its median (red).

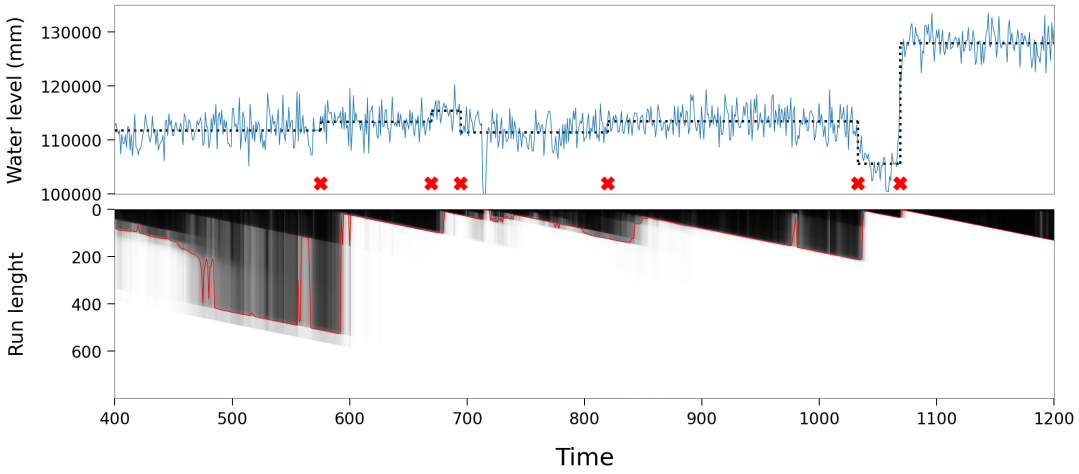


Figure 6.4: Results for the unfiltered Well Log data with HSSPAR-CP, considering measurements ranging from 400 to 1200 (in time units). **Top:** Alert locations obtained from MAP segmentation are represented by small red crosses. The horizontal dashed black line indicates the mean of observations between change points. **Bottom:** The run length CDF (black) and its median (red).

We provide visualizations of the run length posterior for HSSPAR-CP on both

[0]Here error bars are ±1.96× standard error.

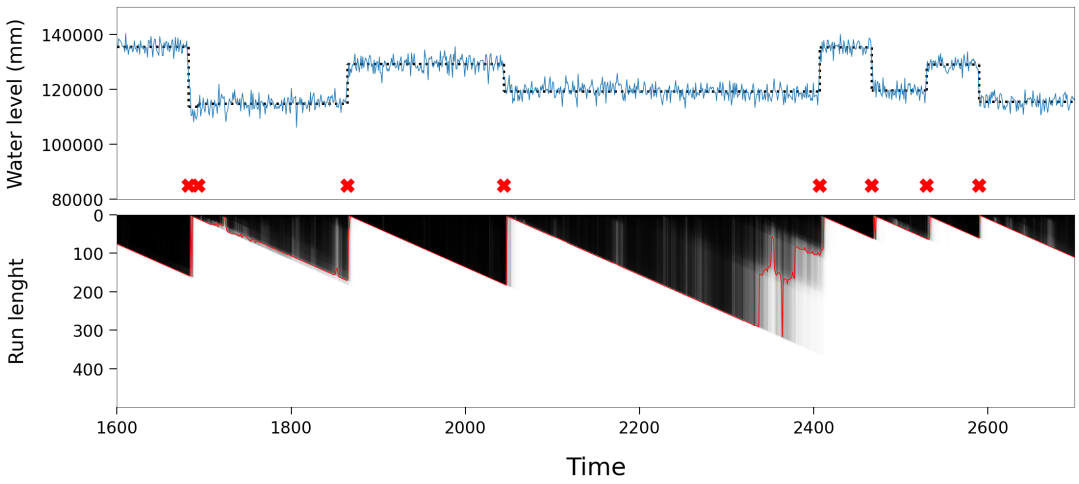


Figure 6.5: Results for the unfiltered Well Log data with HSSPAR-CP, considering measurements ranging from 1600 to 2700 (in time units). **Top:** Alert locations obtained from MAP segmentation are represented by small red crosses. The horizontal dashed black line indicates the mean of observations between change points. **Bottom:** The run length CDF (black) and its median (red).

filtered and unfiltered data, as depicted in Figures 6.2 and 6.3. Specifically, Figure 6.2 presents results obtained using unfiltered data, similar to the experiment described in Table 6.4. To explore the effects of outlier processing, Figure 6.3 showcases results obtained from a filtered version of the Well Log data set. Furthermore, Figure 6.4 and Figure 6.5 display similar outcomes as Figure 6.2, but with a reduced measurement range of 400 to 1200 and 1600 to 2700 (time units), respectively. Consequently, Figure 6.2 allows for a direct comparison to Figure 2 in Adams and MacKay (2007).

6.6.4 Bee Waggle Dance data

The waggle dance is bees' method of communicating the location of forage (direction, distance and profitability of food source) to each other. Entomologists have been interested in identifying change points in different stages in the bee dance. The *Bee Waggle Dance* data set contains the bee's x-coordinate position, y-coordinate position and head angle at each frame of 6 video sequences of bee waggle dances. Following Saatçi et al. (2010), we examine the first video sequence only, and consider angle differences for the angle sequence. HSSPAR-CP outperforms in terms of NLL and

Table 6.3: Results of predictive performance on Bee Waggle data. The results are provided with 95% error bars and the p-value testing the null hypothesis that methods are equivalent to the best performing method, according to NLL, using one sided t-test. (·)-CP refers to the BOCPD variant of the respective method.

Method	Negative Log Likelihood	p-value	MSE	p-value	time(s)
Bee Waggle Data (250 training points, 806 Test points)					
RRGPAR-CP	-0.9249 (±0.1574)	0.006	0.8623 (±0.1670)	0.034	225.63
RRSPAR-CP	**-1.2291** (±0.1099)	N/A	**0.6646** (±0.1071)	N/A	315.41
GPTS	1.2786 (±0.2440)	< 0.001	1.6688 (±0.2321)	< 0.001	13.58
GPTS-CP	0.0766 (±0.1737)	< 0.001	1.1911 (±0.1856)	< 0.001	20.91
GPAR	-0.4948 (±0.2976)	< 0.001	0.7757 (±0.1115)	0.054	412.66
GPAR-CP	-1.0430 (±0.1175)	0.013	0.7238 (±0.1275)	0.202	485.46
TIM	1.3853 (±0.1106)	< 0.001	1.3670 (±0.1943)	< 0.001	N/A

MSE. Figure 6.6 shows the run length posterior and change point alerts for HSSPAR-CP. The HSSPAR-CP model correctly identifies 16 of the 19 known CPs.

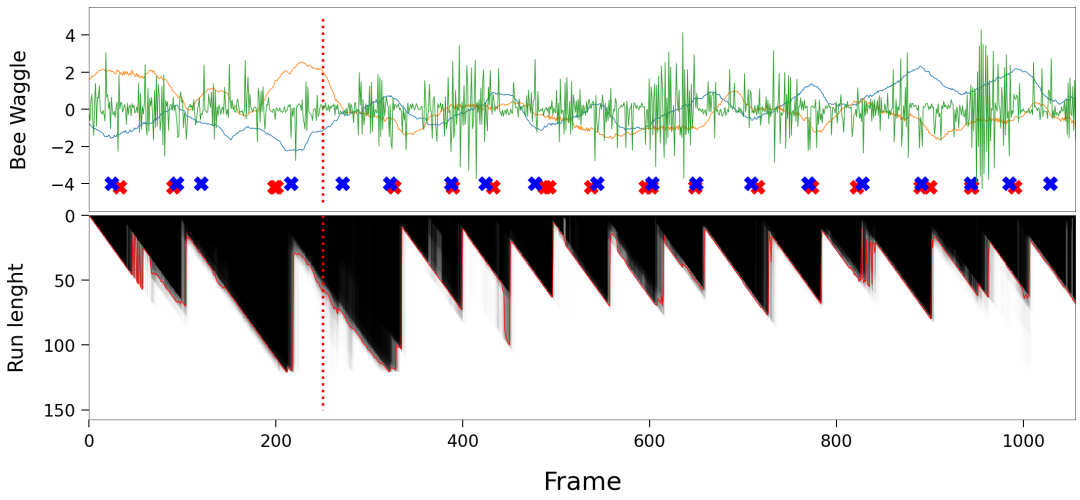


Figure 6.6: Results for the Bee Waggle Dance data with HSSPAR-CP. **Top:** The time-series are the bee's x-location (blue), y-location (orange) and angular difference (green). The vertical dashed red line represents the boundary between train and test sets. The small red crosses represents alert locations obtained from MAP segmentation. The small blue crosses represents the known true change point. **Bottom:** The run length CDF (black) and its median (red).

6.6.5 Snowfall data

The *Snowfall* data report the historical daily snowfall level in Whistler BC (Canada) from 1972 to 2008. We train the model on the first 1000 entries of the data (corresponding to approximately three years) and test on the 12,880 remaining points. The HSSPAR-CP model performs significantly better in terms of both NLL and MSE compared to its competitors. Fitting of HSSPAR-CP is also $> 20\times$ faster than that of GPAR-CP.

The Student-t UPM outperforms other GP-based CP algorithms in terms of NLL in all experiments. We attribute this performance to the generalization property (compared to a GP) and to the fatter predictive distribution of a TP. The reduced-rank approximation yields significantly faster training while maintaining good performance for applications with larger training sets, i.e. *Well Log* and *Snowfall*.

Table 6.4: Results of predictive performance on Whithler Snowfall data. The results are provided with 95% error bars and the p-value testing the null hypothesis that methods are equivalent to the best performing method, according to NLL, using one sided t-test. $(\cdot)$-CP refers to the BOCPD variant of the respective method.

Method	Negative Log Likelihood	p-value	MSE	p-value	time(s)
Whistler Snowfall Data (1000 training points, 13380 Test points)					
RRGPAR-CP	-0.0278 ($\pm$0.0531)	< 0.001	1.3040 ($\pm$0.0962)	< 0.001	605.64
RRSPAR-CP	**-0.52425** ($\pm$0.0393)	N/A	**0.9785** ($\pm$0.0900)	N/A	591.06
GPTS	1.2965 ($\pm$0.0495)	< 0.001	1.1828 ($\pm$0.0774)	0.002	18.15
GPTS-CP	0.6143 ($\pm$0.0693)	< 0.001	1.1701 ($\pm$0.0807)	0.003	59.10
GPAR	1.1708 ($\pm$0.1453)	< 0.001	1.1195 ($\pm$0.1013)	0.021	12,150.95
GPAR-CP	-0.1890 ($\pm$0.0433)	< 0.001	1.1959 ($\pm$0.0994)	0.004	14,493.47
TIM	0.3374 ($\pm$0.0264)	< 0.001	0.9912 ($\pm$0.0769)	0.381	N/A

6.7 Conclusion

We introduce a Bayesian online change point detection framework that combines a Student-t process with dependent Student-t noise as a time-series model, and *Hilbert space* reduced-rank kernel approximation for mitigating computation complexity. We illustrate the use of our scheme on a diverse set of real world examples. Our method compares favorably to other GP-based alternatives in terms of both prediction and hyperparameter learning time.

Chapter 7

Conclusion

7.1 Contributions

In this thesis, we explored advanced methods that leverage the spectral sparse representation of GPs for analyzing time series and spatial data. Our contributions are twofold: Firstly, we were particularly interested in developing fast and flexible inference schemes for Gaussian Cox processes. Secondly, our goal has been to develop a fast and robust alternative to existing GP-based Bayesian change point detection methods (BOCPD), specifically designed to address non-stationary patterns in time series data.

In Chapter 4, we introduced our new approach for Bayesian inference in the context of permanental processes. Permanental processes belong to a special class of Gaussian Cox processes, where the Poisson intensity is modeled as the square of a Gaussian process. This unique characteristic allows for the analytical computation of intensity integrals, particularly when employing a Gaussian kernel. Our methodology combines a random Fourier representation of the Gaussian process kernel with a fast Laplace approximation to the intractable posterior distribution. Notably, this approach extends to *generalized kernels*. These kernels are known to be dense within the family of stationary kernel, , which implies that they can approximate any stationary kernel with arbitrary precision given an adequate number of spectral components. Furthermore, the use of generalized kernels simplifies the process of learning

the underlying latent function's differentiability. In essence, we offered an efficient inference technique that eliminates the need for numerical integration, accommodates customized kernel design, exhibits linear scalability concerning the number of events, and, importantly, outperforms the alternative Laplace-based approach proposed by Walder and Bishop (2017). We demonstrated the superior performance of our approach on various datasets, including synthetic data, real-world temporal data, and extensive spatial datasets.

In Chapter 6, we proposed an approach that improve the existing BOCPD with GPs Saatçi et al. (2010). Our proposed method incorporates an Auto-Regressive Student-t Process underlying predictive model (UPM), which includes dependent Student-t noise, resulting in a solution that is more flexible but also more robust. The intractability inherent to this problem is resolved by introducing dependent noise, inspired by the work of Shah et al. (2014). Student-t processes are renowned for their ability to generalize the concept of GPs, providing us with greater flexibility. In our specific context, TP UPMs also prove to be more robust to data noise, reducing the likelihood of producing false change point alarms. However, BOCPD with GPs is recognized for its computational complexity, typically demanding $O(T^5)$ operations for a straightforward implementation. To address this challenge, we integrate a *Hilbert space* reduced-rank approximation, developped by Solin and Särkkä (2020). This method approximates the spectral decomposition of stationary kernels through an eigenfunction expansion of the Laplace operator. Combining this technique with others results in a reduced computational complexity of $O(TR_{\max}m^2)$, where $R_{\max}$ represents the pruning threshold, and m signifies the number of basis functions. In practical experiments, our method outperforms other GP-based alternatives in terms of both prediction accuracy and hyperparameter learning efficiency.

7.2 Future work

A promising avenue for future research involves extending our existing spatial approach for permanental processes into a spatio-temporal context, where data is collected across both space and time. These processes find applications in diverse do-

mains, ranging from the study of disease occurrence and its temporal propagation (Balderama et al., 2012; Dong et al., 2023) to urban mobility patterns (Du et al., 2016) and criminal activities (Flaxman et al., 2019; Rosser and Cheng, 2019), among others. The recent surge in data availability has invigorated interest and research in this domain. For instance, several cities, including Chicago, Seattle, Detroit, and Baltimore, have made their spatio-temoral data accessible for research purposes in the study of criminal activity.

A common objective in these areas is the development of models that can effectively capture both spatial and temporal dependencies and their effects on propagation. However, introducing a temporal dimension introduces computational complexities, resulting in the prevailing but limited strategy of using separate models for event time and space. Various methods have been employed, with kernel density estimation (Lee and Mitchell, 2014) being the most prevalent approach, alongside the use of log-Gaussian Cox Processes (Diggle et al., 2005, 2013), and self-exciting point process models (Mei and Eisner, 2017; Rosser and Cheng, 2019). Notably, Flaxman et al. (2019) introduced a methodology in a spatio-temporal setting that bears some similarity to our approach, incorporating a Random Feature Fourier (RFF) representation of log-Gaussian Cox Processes.

In our context, our aim is to expand a permanental spatial model into a temporal spatial framework, with a dual focus on tractability properties and the potential for enhanced generalization, leveraging the use of generalized kernels.

Appendices

Appendix A

Matrix Algebra

Throughout the thesis, we have utilized several Matrix algebra identities, which we include here for reference. For a more detailed discussion, we recommend consulting Golub and Van Loan (1996) or Harville (2008).

A.1 Woodbury identity

In linear algebra, the *Woodbury matrix identity*, says that the inverse of a rank-m pertubation of some $n \times n$ matrix can be computed by doing a rank-m correction to the inverse of the original matrix.

The identity is given by the following equation:

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}. \tag{A.1}$$

Here, $\mathbf{A}$ and $\mathbf{C}$ are matrices of sizes $n \times n$ and $m \times m$, respectively, and $\mathbf{U}$ and $\mathbf{V}$ are arbitrary matrices of sizes $n \times m$ and $m \times n$, respectively. The left-hand side of the equation can be computed directly, but requires a complexity of $\mathcal{O}(n^3)$. However, the right-hand side can be computed more efficiently in $\mathcal{O}(m^2 n)$, which is particularly advantageous when $m \ll n$.

A.2 Cholesky factorization

The Cholesky decomposition of a real positive-definite matrix $\mathbf{A}$, is a decomposition of the form:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\top} \tag{A.2}$$

where $\mathbf{L}$ is a lower triangular matrix with real and positive diagonal entries, and $\mathbf{L}^{\top}$ denotes the conjugate transpose of $\mathbf{L}$. Every real-valued symmetric positive-definite matrix has a unique Cholesky decomposition (Golub and Van Loan, 1996, p.143).

The Cholesky factorization method computes the Cholesky factor $\mathbf{L}$ of an $n \times n$ matrix in $\mathcal{O}(n^3/6)$ time and is a highly stable operation from a numerical standpoint. This factorization is particularly useful for solving linear systems of the form $\mathbf{Ax} = \mathbf{b}$, since

$$\mathbf{Ax} = \mathbf{b} \iff \mathbf{x} = \mathbf{L}^{\top} \backslash (\mathbf{L} \backslash \mathbf{b}) \tag{A.3}$$

where $\mathbf{L}\backslash\mathbf{b}$ denotes the solution to the linear system $\mathbf{Lx} = \mathbf{b}$. The two linear systems can be solved using forward and backward substitution in $\mathcal{O}(n^2/2)$ time each, which is faster and more accurate than directly solving $\mathbf{Ax} = \mathbf{b}$.

A.3 Matrix Product Trace Invariance

The trace of a matrix product is invariant under cyclic permutations, meaning that:

$$\mathrm{tr}(\mathbf{ABCD}) = \mathrm{tr}(\mathbf{BCDA}) = \mathrm{tr}(\mathbf{CDAB}) = \mathrm{tr}(\mathbf{DABC}). \tag{A.4}$$

This relationship holds true regardless of the dimensions of the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$.

A.4 Exchange Matrix

The exchange matrix $\mathbf{E}$ are special cases of permutation matrices, with ones on the anti-diagonal and zeros on all other elements.

Definition A.4.1. *If* $\mathbf{E}_n$ *is the* $n \times n$ *exchange matrix with* i, j *elements*

$$\mathbf{E}_{i,j} = \begin{cases} 1, & i+j = n+1 \\ 0, & i+j \neq n+1. \end{cases} \tag{A.5}$$

In particular,

$$\mathbf{E}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{E}_3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{A.6}$$

Thus, the exchange matrix $\mathbf{E}_n$ is a permutation matrix of order n that interchanges rows and columns of the identity matrix $\mathbf{I}_n$ along the main diagonal. Algebraically, $\mathbf{E}_n$ has the property that for any vector $\mathbf{x} = [x_1, \cdots, x_n]^\top \in \mathbb{R}^n$, the product $\mathbf{E}_n\mathbf{x}$ results in a vector with the same entries as $\mathbf{x}$ but in reverse order ,

$$\mathbf{E}_n\mathbf{x} = \mathbf{x}(n : -1 : 1) = [x_n, \cdots, x_1]. \tag{A.7}$$

This can be interpreted as flipping the vector around the middle entry (if n is odd) or between the two middle entries (if n is even).

A.5 Matrix derivatives

The elements of the inverse matrix $(\mathbf{A})^{-1}$ have a derivative with respect to the parameters θ given by:

$$\frac{\partial}{\partial\theta}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\left(\frac{\partial\mathbf{A}}{\partial\theta}\right)\mathbf{A}^{-1}. \tag{A.8}$$

If $\mathbf{A}$ is a positive definite symmetric matrix, the derivative of the log determinant with respect to the parameters θ is given by:

$$\frac{\partial}{\partial\theta}\log|\mathbf{A}| = -\mathrm{tr}\left(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\theta}\right). \tag{A.9}$$

Appendix B

Gaussian Identities

In this section, we review a few identities regarding multivariate Gaussian distributions that are relevant to the thesis. For more comprehensive information, we refer to the work of Mardia et al. (1979).

Multivariate Gaussian distribution The multivariate normal distribution is an extension of the one-dimensional (univariate) normal distribution to higher dimensions. The probability density function (PDF) of a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by the following equation:

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &= (2\pi)^{-1/n}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \end{aligned} \tag{B.1}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of the covariance matrix.

B.1 Conditional rule of the Gaussian distribution

Suppose that we partition $\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ into two random vectors i.e $\mathbf{x} = \left(\begin{smallmatrix}\mathbf{x}_1 \\ \mathbf{x}_2\end{smallmatrix}\right)$. Further, suppose that we partition the mean vector and covariance matrix in a corresponding manner. That is, $\boldsymbol{\mu} = \left(\begin{smallmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{smallmatrix}\right)$. and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{smallmatrix}\right)$.

Marginalisation Any distribution for a subset of variables from a multivariate normal, is a multivariate normal distribution i.e.

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Conditioning rule Any distribution for a subset of variables from a multivariate normal, conditional on known values for another subset of variables, is a multivariate normal distribution i.e.

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \tag{B.2}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &:= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_1),\\ \boldsymbol{\Sigma}_{1|2} &:= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}^{-1})\boldsymbol{\Sigma}_{21}\end{aligned}$$

B.2 Integral of the product of two Gaussians

Consider two n-dimensional Gaussian distributions $\mathcal{N}(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\Sigma}_{11})$ and $\mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are n-dimensional vectors, $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are $n \times n$ covariance matrices, and $\boldsymbol{\mu}_2$ is an n-dimensional mean vector.

The integral of the product of these two Gaussian verifies

$$\int \mathcal{N}(\mathbf{x}_1|\mathbf{x_2}, \boldsymbol{\Sigma}_{11})\mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})d\mathbf{x}_1 = \mathcal{N}(\mathbf{x}_1|\mathbf{x}_2, \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22}) \tag{B.3}$$

where the integral is taken over the entire n-dimensional space.

This integral is used, for example, in Bayesian inference to compute the posterior distribution of a parameter $\mathbf{x}_2$ from observations $\mathbf{x}_1$, Gaussian likelihood $p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\mathbf{x_2}, \boldsymbol{\Sigma}_{11})$ and Gaussian prior $p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. In this context, the resulting posterior distribution is also a Gaussian distribution.

B.3 Linear transformations

If $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the linear transformation $\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x} + \boldsymbol{\alpha}$ is also Gaussian, with mean and covariance given by:

$$\mathbb{E}[\tilde{\mathbf{x}}] = \mathbf{A}\boldsymbol{\mu} + \boldsymbol{\alpha}, \tag{B.4}$$

$$\mathrm{Cov}(\tilde{\mathbf{x}}) = \mathbb{E}[(\tilde{\mathbf{x}} - \mathbb{E}[\tilde{\mathbf{x}}])(\tilde{\mathbf{x}} - \mathbb{E}[\tilde{\mathbf{x}}])^\top] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top. \tag{B.5}$$

Appendix C

Proofs

C.1 Proof of proposition (4.2.1) : Integral Expression for LBPP with Nyström

In Proposition 4.2.1 we express the integral term $\int_B f(\mathbf{x})^2 d\mathbf{x}$ under the Nyström approximation both because it is not available in Walder and Bishop (2017) and to demonstrate the similarities with the corresponding derivation of our proposed method in Proposition 4.3.3.

C.1.1 Integral Calculation

Let f be approximated by the Nyström-based approach defined in Equation (4.4) i.e. $f(\mathbf{x}) \approx {\mathbf{w}^{(m)}}^\top \boldsymbol{\varphi}^{(m)}(\mathbf{x})$. The integral expression $\int_B \lambda(\mathbf{x})^2 \, d\mathbf{x}$ can be written as

$$
\begin{aligned}
\int_B f(\mathbf{x})^2 d\mathbf{x} &\approx \sum_{i=1}^{m}\sum_{j=1}^{m} w_i^{(m)} w_j^{(m)} \int_B \varphi_i^{(m)}(\mathbf{x})\varphi_j^{(m)}(\mathbf{x}) \, d\mathbf{x} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{w_i^{(m)} w_j^{(m)}}{\sqrt{\lambda_i^{(m)}\lambda_j^{(m)}}} \int_B \left(k(\mathbf{x}, \mathbf{X}_m)\mathbf{e}_i^{(m)}\right)\left(k(\mathbf{x}, \mathbf{X}_m)\mathbf{e}_j^{(m)}\right) \, d\mathbf{x} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{w_i^{(m)} w_j^{(m)}}{\sqrt{\lambda_i^{(m)}\lambda_j^{(m)}}} {\mathbf{e}_i^{(m)}}^\top \left(\int_B k(\mathbf{X}_m, \mathbf{x})k(\mathbf{x}, \mathbf{X}_m) \, d\mathbf{x}\right) \mathbf{e}_j^{(m)} \\
&= {\mathbf{w}^{(m)}}^\top \underbrace{\boldsymbol{\Lambda}^{-\frac{1}{2}} \left[{\mathbf{U}^{(m)}}^\top \Psi^{(m)} \, \mathbf{U}^{(m)}\right] \boldsymbol{\Lambda}^{-\frac{1}{2}}}_{:=\mathbf{M}^{(m)}} \mathbf{w}^{(m)}
\end{aligned}
$$

where $\Psi^{(m)} = \int_B k(\mathbf{X}_m, \mathbf{x})k(\mathbf{x}, \mathbf{X}_m)\ d\mathbf{x}$ is the integral statistic already defined in Lloyd et al. (2015) and John and Hensman (2018). In particular, for the separable Gaussian kernel defined in Equation (C.3),

$$\begin{aligned}\Psi_{i,j}^{(m)} &= \sigma^4 \int_B \prod_{k=1}^{d} \exp\left(-\frac{(x_{k,i}^{(m)} - x_{k,j}^{(m)})^2}{4{\ell_k}^2}\right) \exp\left(-\frac{(x_k - \bar{x}_{k,i,j}^{(m)})^2}{{\ell_k}^2}\right)\ d\mathbf{x} \\ &= \sigma^4 \prod_{k=1}^{2} \frac{\ell_k\sqrt{\pi}}{d} \exp\left(-\frac{(x_{k,i}^{(m)} - x_{k,j}^{(m)})^2}{4{\ell_k}^2}\right) \times \left[\operatorname{erf}\left(\frac{\bar{x}_{k,i,j}^{(m)} - B_k^{\min}}{\ell_k}\right) - \operatorname{erf}\left(\frac{\bar{x}_{k,i,j}^{(m)} - B_k^{\max}}{\ell_k}\right)\right]\end{aligned}$$

where σ and $\boldsymbol{\ell} := (\ell_1, \ell_2)$ are respectively the scaling and length-scale parameters of the covariance function, $x_{k,i}^{(m)}$ is the kth coordinate of the ith Nyström-sampled point $\mathbf{x}_i^{(n)}$ and $\bar{x}_{k,i,j}^{(m)} := (x_{k,i}^{(n)} + x_{k,j}^{(m)})/2$.

Offset Term Adding an offset term β to the intensity i.e $\lambda(\cdot) = (f(\cdot) + \beta)^2$ yields

$$\int_B (f(\mathbf{x}) + \beta)^2\, d\mathbf{x} = \int_B f(\mathbf{x})^2 d\mathbf{x} + \ 2\beta \int_B f(\mathbf{x}) d\mathbf{x} + \ \beta^2 |B|$$

with

$$\int_B f(\mathbf{x}) d\mathbf{x} \approx \left(\int_B k(\mathbf{x}, \mathbf{X}_m) d\mathbf{x}\right) \mathbf{U}^{(m)} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{w}^{(m)} = {\boldsymbol{\psi}^{(m)}}^\top \mathbf{U}^{(m)} {\boldsymbol{\Lambda}^{(m)}}^{-\frac{1}{2}} \mathbf{w}^{(m)}$$

where $\boldsymbol{\psi}^{(m)} := \int_B k(\mathbf{X}_m, \mathbf{x}) d\mathbf{x}$ is a m-vector such that in the separable Gaussian kernel case above, we have

$$\boldsymbol{\psi}_i^{(m)} = \sigma^2 \prod_{d=1}^{2} \frac{\ell_d\sqrt{\pi}}{\sqrt{2}} \times \left[\operatorname{erf}\left(\frac{\mathbf{x}_{d,i}^{(m)} - B_d^{\max}}{\ell_d\sqrt{2}}\right) - \operatorname{erf}\left(\frac{x_{d,i}^{(m)} - B_d^{\min}}{\ell_d\sqrt{2}}\right)\right].$$

C.2 Proof of proposition (4.3.3) : Integral Expression via RFF

Let f be approximated by a RFF-based approach as defined in Equation (4.13) of the main text i.e. $f^{(r)}(\mathbf{x}) \approx {\mathbf{w}^{(r)}}^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x})$ where the feature map $\boldsymbol{\varphi}^{(r)}$ follows Equation (4.8) of the main text.

C.2.1 Real Valued Feature Mapping

We first detail the derivation of the real valued Fourier features described in Equation (4.9) of the main text. The imaginary part of Equation (4.8) of the main text can be discarded as follows

$$
\begin{aligned}
k(\mathbf{x}-\mathbf{x}') &= \sigma^2 \mathbb{E}_{\mathbf{z}}\left[\exp(-i\mathbf{z}^\top(\mathbf{x}-\mathbf{x}'))\right] \\
&= \sigma^2 \mathbb{E}_{\mathbf{z}}\left[\cos(\mathbf{z}^\top(\mathbf{x}-\mathbf{x}')) + i\sin(\mathbf{z}^\top(\mathbf{x}-\mathbf{x}'))\right] \qquad \text{(C.1)} \\
&= \sigma^2 \mathbb{E}_{\mathbf{z}}\left[\cos(\mathbf{z}^\top(\mathbf{x}-\mathbf{x}'))\right] \\
&= \sigma^2 \mathbb{E}_{\mathbf{z}}\left[\cos(\mathbf{z}^\top\mathbf{x})\cos(\mathbf{z}^\top\mathbf{x}') + \sin(\mathbf{z}^\top\mathbf{x})\sin(\mathbf{z}^\top\mathbf{x}')\right] \qquad \text{(C.2)} \\
&\approx \frac{\sigma^2}{r}\sum_{i=1}^{r} \cos(\mathbf{z}_i^\top\mathbf{x})\cos(\mathbf{z}_i^\top\mathbf{x}') + \sin(\mathbf{z}_i^\top\mathbf{x})\sin(\mathbf{z}_i^\top\mathbf{x}') \\
&= \frac{\sigma^2}{r}\boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top\boldsymbol{\varphi}^{(r)}(\mathbf{x}')
\end{aligned}
$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_r$ are independent samples with density $S(\mathbf{z})$ and the explicit feature mapping $\boldsymbol{\varphi}^{(r)}(\cdot)$ is defined as

$$
\boldsymbol{\varphi}^{(r)}(\mathbf{x}) := \frac{\sigma}{\sqrt{r}}\begin{bmatrix} \cos(\mathbf{z}_1^\top\mathbf{x}) \\ \cdots \\ \cos(\mathbf{z}_r^\top\mathbf{x}) \\ \sin(\mathbf{z}_1^\top\mathbf{x}) \\ \cdots \\ \sin(\mathbf{z}_r^\top\mathbf{x}) \end{bmatrix}.
$$

Gaussian Kernel Specifically, without approximation, for a Gaussian kernel k_g with $\mathcal{X} = \mathbb{R}^d$ and where

$$k_g(\mathbf{x} - \mathbf{x}') = \sigma^2 \prod_{i=1}^{d} \exp\left(-\frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{2\,\ell_i^2}\right) \tag{C.3}$$

with scaling parameter σ and length-scale vector $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_d]^\top$, the corresponding spectral density $S(\mathbf{z})$ is a multivariate normal $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\gamma}))$ with $\boldsymbol{\gamma} := [1/\ell_1^2, \ldots, 1/\ell_d^2]^\top$.

Matérn Kernel For a Matérn class of kernel function k_m such that

$$k_m(\mathbf{x} - \mathbf{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(\mathbf{x} - \mathbf{x}')}{\ell}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}(\mathbf{x} - \mathbf{x}')}{\ell}\right) \tag{C.4}$$

where $\sigma \in \mathbb{R}^+$, $\ell \in \mathbb{R}^+$, $\nu \in \mathbb{R}^+$ and K_ν is a modified Bessel function, the corresponding spectral density $S(\mathbf{z})$ is a d dimension multivariate Student-t distribution $St(\mathbf{0}, \boldsymbol{\Sigma}, 2\nu)$ with covariance function $\boldsymbol{\Sigma} = (1/\ell)\ \mathbf{I}_d$ and degree of freedom 2ν. The spectral locations $\mathbf{Z}$ are sampled as

$$\mathbf{Z} = \sqrt{u/2\nu\ell}\ \mathbf{G} \qquad \text{where} \quad u \sim \chi^2(2\nu) \tag{C.5}$$

and $\mathbf{G}$ is a $d \times r$ matrix of of i.i.d. standard normal random variables.

C.2.2 Integral Calculation

We now detail the integral expression of proposition (4.3.3) for the real valued Fourier features in Equation (4.9) of the main text. We consider without loss of generality the spatial case where $B = [-a, a]^2$. $x_{d,i}$ refers to the dth coordinate of the ith training input $\mathbf{x}_i$ for $i = 1, \ldots, N$ and $z_{d,i}$ to the dth coordinate of the ith spectral point $\mathbf{z}_i$ for $i = 1, \ldots, r$. The integral of f over B becomes

$$\int_{[-a,a]^2} f(\mathbf{x})^2 d\mathbf{x} = \sum_{i,j} w_i^{(r)} w_j^{(r)} \int_{[-a,a]^2} \boldsymbol{\varphi}_i^{(r)}(\mathbf{x}) \boldsymbol{\varphi}_j^{(r)}(\mathbf{x})\, d\mathbf{x} \tag{C.6}$$

where

$$\boldsymbol{\varphi}_i^{(r)}(\mathbf{x})\boldsymbol{\varphi}_j^{(r)}(\mathbf{x}) = \frac{\sigma^2}{r}\begin{cases}\cos(\mathbf{z}_i^\top\mathbf{x})\cos(\mathbf{z}_j^\top\mathbf{x}) & \text{if } (i,j)\in[1,r]^2\\ \sin(\mathbf{z}_i^\top\mathbf{x})\sin(\mathbf{z}_j^\top\mathbf{x}) & \text{if } (i,j)\in[r+2,2r]^2\\ \cos(\mathbf{z}_i^\top\mathbf{x})\sin(\mathbf{z}_j^\top\mathbf{x}) & \text{if } (i,j)\in[1,r]\times[r+2,2r]\\ \sin(\mathbf{z}_i^\top\mathbf{x})\cos(\mathbf{z}_j^\top\mathbf{x}) & \text{if } (i,j)\in[r+2,2r]\times[r,r].\end{cases} \tag{C.7}$$

Thus, $\int_B f(\mathbf{x})^2 d\mathbf{x} = \mathbf{w}^\top\mathbf{M}^{(r)}\mathbf{w}$, where $\mathbf{M}^{(r)}$ is the matrix with i,j entry obtained by integrating Equation (*C*.7). The *'cos'*, *'sin'* and *'cos-sin'* expressions can be written as

$$\int_{[-a,a]^2}\cos(\mathbf{z}_i^\top\mathbf{x})\cos(\mathbf{z}_j^\top\mathbf{x})\,d\mathbf{x} = \frac{1}{2}\int_{[-a,a]^2}\left[\cos((\mathbf{z}_i-\mathbf{z}_j)^\top\mathbf{x}) + \cos((\mathbf{z}_i+\mathbf{z}_j)^\top\mathbf{x})\right]\,d\mathbf{x}$$

$$\int_{[-a,a]^2}\sin(\mathbf{z}_i^\top\mathbf{x})\sin(\mathbf{z}_j^\top\mathbf{x})\,d\mathbf{x} = \frac{1}{2}\int_{[-a,a]^2}\left[\cos((\mathbf{z}_i-\mathbf{z}_j)^\top\mathbf{x}) - \cos((\mathbf{z}_i+\mathbf{z}_j)^\top\mathbf{x})\right]\,d\mathbf{x}$$

and

$$\int_{[-a,a]^2}\cos(\mathbf{z}_i^\top\mathbf{x})\sin(\mathbf{z}_j^\top\mathbf{x})\,d\mathbf{x} = \frac{1}{2}\int_{[-a,a]^2}\left[\sin((\mathbf{z}_i-\mathbf{z}_j)^\top\mathbf{x}) + \sin((\mathbf{z}_i+\mathbf{z}_j)^\top\mathbf{x})\right]\,d\mathbf{x} = 0.$$

Thus, since the off-diagonal blocks of $\mathbf{M}^{(r)}$ are null, we can rewrite Equation (C.6) as

$$\begin{aligned}\int_B f(\mathbf{x})^2 d\mathbf{x} &= \frac{\sigma^2}{r}\left[{\mathbf{w}_{:r}^{(r)}}^\top(\mathbf{A}+\mathbf{B})\mathbf{w}_{:r}^{(r)} + {\mathbf{w}_{r:}^{(r)}}^\top(\mathbf{A}-\mathbf{B})\mathbf{w}_{r:}^{(r)}\right]\\ &= \frac{\sigma^2}{r}\,{\mathbf{w}^{(r)}}^\top\left[\mathbf{D}_l^\top(\mathbf{A}+\mathbf{B})\mathbf{D}_l + \mathbf{D}_r^\top(\mathbf{A}-\mathbf{B})\mathbf{D}_r\right]\mathbf{w}^{(r)}\end{aligned} \tag{C.8}$$

where $\mathbf{w}_{:r}^{(r)} := [w_1^{(r)},\dots,w_r^{(r)}]^\top$, $\mathbf{w}_{r:}^{(r)} := [w_{r+1}^{(r)},\dots,w_{2r}^{(r)}]^\top$, $\mathbf{D}_l := \begin{bmatrix}\mathbf{I}_r & 0\end{bmatrix}$, $\mathbf{D}_r := \begin{bmatrix}0 & \mathbf{I}_r\end{bmatrix}$ and $\mathbf{A}$ and $\mathbf{B}$ are two $r\times r$ matrices defined as

$$\mathbf{A}_{i,j} = \frac{1}{2}\int_{[-a,a]^2}\cos((\mathbf{z}_i-\mathbf{z}_j)^\top\mathbf{x})\,dx \quad \text{and} \quad \mathbf{B}_{i,j} = \frac{1}{2}\int_{[-a,a]^2}\cos((\mathbf{z}_i+\mathbf{z}_j)^\top\mathbf{x})\,d\mathbf{x}.$$

$\mathbf{A}$ and $\mathbf{B}$ can be evaluated as follows

Case 1: $\mathbf{z}_i \neq \mathbf{z}_j$ We define again $\bar{z}_{d,i,j} := z_{d,i} + z_{d,j}$. Then,

$$\mathbf{A}_{i,j} = \frac{\cos[a(\tilde{z}_{1,i,j} - \tilde{z}_{2,i,j})] - \cos[a(\tilde{z}_{1,i,j} + \tilde{z}_{2,i,j})]}{\tilde{z}_{1,i,j}\bar{z}_{2,i,j}} = \frac{2}{\tilde{z}_{1,i,j}\bar{z}_{2,i,j}} \left(\sin[a\tilde{z}_{1,i,j}] \sin[a\tilde{z}_{2,i,j}]\right).$$
$$\mathbf{B}_{i,j} = \frac{\cos[a(\bar{z}_{1,i,j} - \bar{z}_{2,i,j})] - \cos[a(\bar{z}_{1,i,j} + \bar{z}_{2,i,j})]}{\bar{z}_{1,i,j}\bar{z}_{2,i,j}} = \frac{2}{\bar{z}_{1,i,j}\bar{z}_{2,i,j}} \left(\sin[a\bar{z}_{1,i,j}] \sin[a\bar{z}_{2,i,j}]\right).$$

Case 2: $\mathbf{z}_i = \mathbf{z}_j$

$$\mathbf{A}_{i,i} = 2a^2.$$
$$\mathbf{B}_{i,i} = \frac{\cos[2a(z_{1,i} - z_{2,i})] - \cos[2a(z_{1,i} + z_{2,i})]}{4z_{1,i}z_{2,i}} = \frac{1}{2z_{1,i}z_{2,i}} \left(\sin[2az_{1,i}] \sin[2az_{2,i}]\right).$$

Offset Term For the offset term β, we need to compute the integral of f, that is obtained from

$$\int_{[-a,a]^2} f(\mathbf{x})d\mathbf{x} = \frac{\sigma}{\sqrt{r}} \sum_{i=1}^{r} w_i^{(r)} \int_{[-a,a]^2} \cos(\mathbf{z}_i^\top \mathbf{x})\, d\mathbf{x} + \frac{\sigma}{\sqrt{r}} \sum_{i=r+1}^{2r} w_i^{(r)} \underbrace{\int_{[-a,a]^2} \sin(\mathbf{z}_i^\top \mathbf{x})\, d\mathbf{x}}_{=0}$$
$$= \frac{\sigma}{\sqrt{r}} \mathbf{w}^{(r)\top} \mathbf{m}^{(r)}$$

where $\mathbf{m}$ is a r-vector such that

$$\mathbf{m}_i^{(r)} = \frac{2\cos[a(z_{1,i} - z_{2,i})] - 2\cos[a(z_{1,i} + z_{2,i})]}{z_{1,i}z_{2,i}} = \frac{4}{z_{1,i}z_{2,i}} \left(\sin[az_{1,i}] \sin[az_{2,i}]\right).$$

C.3 Proof of proposition (4.3.3) : Integral Expression for GK

We assume k_{GS} to be a *Generalized kernel* given in Equation (4.10) in the main text, with a kernel g that admits a consistent RFF representation such that $g(\mathbf{x} - \mathbf{x}') \approx \boldsymbol{\varphi}_g^{(r)}(\mathbf{x})^\top \boldsymbol{\varphi}_g^{(r)}(\mathbf{x}')$ where $\boldsymbol{\varphi}_g^{(r)}$ is an explicit feature mapping $\boldsymbol{\varphi}_g : \mathcal{X} \to \mathbb{R}^r$.

C.3.1 Real Valued Feature Mapping

The *Generalized kernel* k_{GS} becomes

$$k_{GS}(\boldsymbol{\tau}) \approx \sum_{k=1}^{K} \sigma_k^2 \; \boldsymbol{\varphi}_g^{(r)}(\mathbf{x} \odot \boldsymbol{\gamma}_k)^\top \boldsymbol{\varphi}_g^{(r)}(\mathbf{x}' \odot \boldsymbol{\gamma}_k) \Psi_k(\mathbf{x}^\top \boldsymbol{\omega}_k)^\top \Psi_k(\mathbf{x}'^\top \boldsymbol{\omega}_k)$$

where $\Psi_k(\mathbf{x})$ is a map $\Psi_k : \mathcal{X} \to \mathbb{R}^2$ such that $\Psi_k(\mathbf{x}) = \begin{bmatrix} \cos(\mathbf{x}^\top \boldsymbol{\omega}_k) \\ \sin(\mathbf{x}^\top \boldsymbol{\omega}_k) \end{bmatrix}$ for $k = 1, \ldots, K$ and $\forall \mathbf{x} \in \mathcal{X}$ so that $\Psi_k(\mathbf{x}^\top \boldsymbol{\omega}_k)\Psi_k(\mathbf{x}'^\top \boldsymbol{\omega}_k) = \cos((\mathbf{x} - \mathbf{x}')^\top \boldsymbol{\omega}_k)$ for all $\boldsymbol{\omega}_k \in \mathbb{R}^d$ and $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Thus,

$$k_{GS}(\mathbf{x}, \mathbf{x}') \approx \sum_{k=1}^{K} h_k(\mathbf{x})^\top h_k(\mathbf{x}')$$

with

$$h_k(\mathbf{x}) = \sigma_k \; \boldsymbol{\varphi}_g^{(r)}(\mathbf{x} \odot \boldsymbol{\gamma}_k) \otimes \begin{bmatrix} \cos(\boldsymbol{\omega}_k^\top \mathbf{x}) \\ \sin(\boldsymbol{\omega}_k^\top \mathbf{x}) \end{bmatrix}$$

for $k = 1, \ldots, K$, where $\otimes$ denotes the Kronecker product.

In particular, when $\boldsymbol{\varphi}_g^{(r)}$ follows Equation (4.9) in the main text,

$$h_k(\mathbf{x}) = \frac{2\sigma_k}{\sqrt{r}} \begin{pmatrix} \cos(\mathbf{z}_1^\top (\mathbf{x} \odot \boldsymbol{\gamma}_k)) \begin{bmatrix} \cos(\boldsymbol{\omega}_k^\top \mathbf{x}) \\ \sin(\boldsymbol{\omega}_k^\top \mathbf{x}) \end{bmatrix} \\ \ldots \\ \sin(\mathbf{z}_r^\top (\mathbf{x} \odot \boldsymbol{\gamma}_k)) \begin{bmatrix} \cos(\boldsymbol{\omega}_k^\top \mathbf{x}) \\ \sin(\boldsymbol{\omega}_k^\top \mathbf{x}) \end{bmatrix} \end{pmatrix} \tag{C.9}$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_r$ are independent samples from $S_g(\mathbf{z})$ the spectral density of g.

The resulting approximate Gaussian process with generalized kernel can be written, in terms of a new $4rK$-size latent vector as follow

$$f(\mathbf{x}) \approx {\mathbf{w}^{(r)}}^\top \boldsymbol{\varphi}^{(r)}(\mathbf{x}) \qquad \text{with } \mathbf{w}^{(r)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{4Kr}).$$

where

$$\boldsymbol{\varphi}^{(r)}(\mathbf{x}) = [h_1(\mathbf{x})^\top, \cdots, h_K(\mathbf{x})^\top]^\top$$

with $h_k(\mathbf{x})$ defined as in Equation (C.9) for $k = 1, \cdots, K$.

C.3.2 Integral Calculation

The integral of f over B becomes

$$\int_{[-a,a]^2} f(\mathbf{x})^2 d\mathbf{x} = \sum_{i,j} w_i^{(r)} w_j^{(r)} \int_{[-a,a]^2} \boldsymbol{\varphi}_i^{(r)}(\mathbf{x}) \boldsymbol{\varphi}_j^{(r)}(\mathbf{x})\, d\mathbf{x} \tag{C.10}$$

Thus, $\int_{[-a,a]} f(\mathbf{x})^2 d\mathbf{x} = \mathbf{w}^{(r)\top} \mathbf{M}^{(r)} \mathbf{w}^{(r)}$, where $\mathbf{M}^{(r)}$ is the matrix with i, j entry obtained by integrating Equation (C.10). The computation of $M^{(r)}$ can be split into different cases expressed below as *'cos'*, *'sin'* and *'cos-sin'* terms.

Cos Terms The *'cos'* i, j terms can be written as

$$\begin{aligned}
&\int_{[-a,a]^2} \cos(\mathbf{z}_i^\top (\mathbf{x} \odot \boldsymbol{\gamma}_i)) \cos(\mathbf{x}^\top \boldsymbol{\omega}_i)) \cos(\mathbf{z}_j^\top (\mathbf{x} \odot \boldsymbol{\gamma}_j)) \cos(\mathbf{x}^\top \boldsymbol{\omega}_j))\, d\mathbf{x} \\
&\qquad = \frac{1}{8} \sum_{k=1}^{8} \int_{[-a,a]^2} \cos(\mathbf{x}^\top \boldsymbol{\eta}_{i,j}^{(k)})\, d\mathbf{x}
\end{aligned} \tag{C.11}$$

where

$$\begin{aligned}
\boldsymbol{\eta}_{i,j}^{(1)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) + \boldsymbol{\omega}_i + (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) + \boldsymbol{\omega}_j, & \boldsymbol{\eta}_{i,j}^{(5)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) - \boldsymbol{\omega}_i + (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) - \boldsymbol{\omega}_j, \\
\boldsymbol{\eta}_{i,j}^{(2)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) + \boldsymbol{\omega}_i + (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) - \boldsymbol{\omega}_j, & \boldsymbol{\eta}_{i,j}^{(6)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) - \boldsymbol{\omega}_i + (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) + \boldsymbol{\omega}_j, \\
\boldsymbol{\eta}_{i,j}^{(3)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) + \boldsymbol{\omega}_i - (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) - \boldsymbol{\omega}_j, & \boldsymbol{\eta}_{i,j}^{(7)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) - \boldsymbol{\omega}_i - (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) + \boldsymbol{\omega}_j, \\
\boldsymbol{\eta}_{i,j}^{(4)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) + \boldsymbol{\omega}_i - (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) + \boldsymbol{\omega}_j, & \boldsymbol{\eta}_{i,j}^{(8)} &= (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) - \boldsymbol{\omega}_i - (\mathbf{z}_j \odot \boldsymbol{\gamma}_j) - \boldsymbol{\omega}_j.
\end{aligned}$$

Integrating the left hand integrants in Equation (C.11) yields

$$\int_{[-a,a]^2} \cos(\mathbf{x}^\top \boldsymbol{\eta}_{i,j}^{(k)})\, d\mathbf{x} = \begin{cases} 2a^2, & \text{if } i = j \text{ and } k \in \{3,7\} \\ \frac{1}{2\eta_{1,i,j}^{(k)}\eta_{2,i,j}^{(k)}} \left(\sin[a\, \eta_{1,i,j}^{(k)}] \sin[a\, \eta_{2,i,j}^{(k)}] \right), & \text{otherwise} \end{cases}. \tag{C.12}$$

Sin Terms The *'sin' i, j* terms are

$$\begin{aligned} &\int_{[-a,a]^2} \sin(\mathbf{x}^\top(\mathbf{z}_i \odot \boldsymbol{\gamma}_i)) \sin(\mathbf{x}^\top \boldsymbol{\omega}_i)) \sin(\mathbf{x}^\top(\mathbf{z}_j \odot \boldsymbol{\gamma}_j)) \sin(\mathbf{x}^\top \boldsymbol{\omega}_j))\, d\mathbf{x} \\ &\qquad = \frac{1}{8} \sum_{k=1}^{8} (-1)^k \int_{[-a,a]^2} \cos(\mathbf{x}^\top \boldsymbol{\eta}_{i,j}^{(k)})\, d\mathbf{x} \end{aligned} \tag{C.13}$$

The left hand integrants in Equation (C.13) integrate alike to Equation (C.12) up to a $(-1)^k$ factor.

Cos-sin Terms The *'cos-sin' i, j* terms can be evaluated as follows

$$\begin{aligned} &\int_{[-a,a]^2} \sin(\mathbf{x}^\top(\mathbf{z}_i \odot \boldsymbol{\gamma}_i)) \sin(\mathbf{x}^\top \boldsymbol{\omega}_i)) \cos(\mathbf{x}^\top(\mathbf{z}_j \odot \boldsymbol{\gamma}_j)) \cos(\mathbf{x}^\top \boldsymbol{\omega}_j))\, d\mathbf{x} \\ &\qquad = \frac{1}{8} \sum_{k=1}^{8} (-1)^{m(k)} \int_{[-a,a]^2} \cos(\mathbf{x}^\top \boldsymbol{\eta}_{i,j}^{(k)})\, d\mathbf{x} \end{aligned} \tag{C.14}$$

where $m(k) = 1$ if $k = 1, \cdots, 4$ and 0 else. The left hand integrants in Equation (C.14) integrate alike to Equation (C.12) up to a $m(k)$ factor.

The remaining terms, yields sums of integrals of the type $\int_{[-a,a]} \sin(\mathbf{x}^\top \boldsymbol{\eta}) d\mathbf{x}$ with $\boldsymbol{\eta} \in \mathbb{R}^d$, that equal zero.

Offset Term For the offset term β, we need to compute the integral of f, that is $\int f(\mathbf{x}) d\mathbf{x} = \mathbf{w}^{(r)\top} \mathbf{m}^{(r)}$ where $\mathbf{m}^{(r)}$ is a $4Kr$-vector such that

$$\mathbf{m}_i^{(r)} = \int_{[-a,a]^2} \boldsymbol{\varphi}_i^{(r)}(\mathbf{x})\, d\mathbf{x}$$

The computation of $m^{(r)}$ can be split into two cases : the *'cos'* terms

$$\int_{[-a,a]^2} \cos(\mathbf{z}_i^\top (\mathbf{x} \odot \boldsymbol{\gamma}_i)) \cos(\mathbf{x}^\top \boldsymbol{\omega}_i))\, d\mathbf{x} =$$
$$\frac{1}{\eta_{1,i}^{(1)} \eta_{2,i}^{(1)}} \Big(\sin[a\ (\eta_{1,i}^{(1)})] \sin[a\ (\eta_{1,i}^{(1)})] \Big) + \frac{1}{\eta_{1,i}^{(2)} \eta_{2,i}^{(2)}} \Big(\sin[a\ (\eta_{1,i}^{(2)})] \sin[a\ (\eta_{1,i}^{(2)})] \Big)$$

and the *'sin'* terms

$$\int_{[-a,a]^2} \sin(\mathbf{z}_i^\top (\mathbf{x} \odot \boldsymbol{\gamma}_i)) \sin(\mathbf{x}^\top \boldsymbol{\omega}_i))\, d\mathbf{x} =$$
$$\frac{1}{\eta_{1,i}^{(1)} \eta_{2,i}^{(1)}} \Big(\sin[a\ (\eta_{1,i}^{(1)})] \sin[a\ (\eta_{1,i}^{(1)})] \Big) - \frac{1}{\eta_{1,i}^{(2)} \eta_{2,i}^{(2)}} \Big(\sin[a\ (\eta_{1,i}^{(2)})] \sin[a\ (\eta_{1,i}^{(2)})] \Big)$$

where

$$\boldsymbol{\eta}_i^{(1)} = (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) + \boldsymbol{\omega}_i, \qquad \boldsymbol{\eta}_{i,j}^{(2)} = (\mathbf{z}_i \odot \boldsymbol{\gamma}_i) - \boldsymbol{\omega}_i.$$

The remaining *'cos-sin'* terms equal zero.

C.4 Predictive Expected Log-likelihood

For a training set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and an held-out test set $\mathbf{X}^* = \{\mathbf{x}_i^*\}_{i=1}^{N^*}$, we can derive an approximation for the expected predictive log-likelihood

$$\mathbb{E}\left[\log p(\mathbf{X}^*|\mathbf{X})\right] \approx -\mathbb{E}_{\mathbf{w}^{(r)}}\left[\int_B (\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}) + \beta)^2 d\mathbf{x}\right] + \sum_{i=1}^{N^*} \mathbb{E}_{\mathbf{w}^{(r)}}\left[\log(\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i^*) + \beta)^2)\right]$$

where $\mathbf{w}^{(r)} \sim q(\mathbf{w}^{(r)}|\mathbf{X}, \Theta)$.

The integral term can be solved as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}^{(r)}}\left[\int_B (\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}) + \beta)^2 d\mathbf{x}\right] \\
&\quad= \int_B \mathbb{E}_{\mathbf{w}^{(r)}}[\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}) + \beta]^2 d\mathbf{x} + \int_B \mathrm{Var}[\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x})] d\mathbf{x} \\
&\quad= \int_B \left(\boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top \hat{\mathbf{w}}^{(r)}\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x})\right) d\mathbf{x} + 2\beta \int_B \left(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x})\right) d\mathbf{x} + \beta^2|B| \\
&\qquad\quad + \int_B \left(\boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top \mathbf{Q}\boldsymbol{\varphi}^{(r)}(\mathbf{x})\right) d\mathbf{x} \\
&\quad= \mathrm{tr}\Big((\hat{\mathbf{w}}^{(r)}\hat{\mathbf{w}}^{(r)\top} + \mathbf{Q}) \underbrace{\int_B \boldsymbol{\varphi}^{(r)}(\mathbf{x})\boldsymbol{\varphi}^{(r)}(\mathbf{x})^\top d\mathbf{x}}_{:=\mathbf{M}^{(r)}}\Big) + 2\beta\hat{\mathbf{w}}^{(r)\top} \underbrace{\left(\int_B \boldsymbol{\varphi}^{(r)}(\mathbf{x}) d\mathbf{x}\right)}_{:=\mathbf{m}^{(r)}} + \beta^2|B| \\
&\quad= \hat{\mathbf{w}}^{(r)\top}\mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)} + \mathrm{tr}\left(\mathbf{Q}\mathbf{M}^{(r)}\right) + 2\beta\hat{\mathbf{w}}^{(r)\top}\mathbf{m}^{(r)} + \beta^2|B|
\end{aligned}$$

where $\mathbf{M}^{(r)}$ and $\mathbf{m}^{(r)}$ are defined in Proposition 4.3.3. Note that we used the cyclical property of the trace in the last two lines. We also used the Tonelli's theorem in the first line to reverse the ordering of the integration over the positive integrand $(\mathbf{w}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}) + \beta)^2 q(\mathbf{w}^{(r)})$.

The sum-of-expectations can also be expressed analytically. It takes of form

$$\sum_i \mathbb{E}[\log z_i^2] \qquad \text{where} \qquad z_i \sim \mathcal{N}(\mu_i, \sigma_i) \tag{C.15}$$

with

$$\mu_i := \hat{\mathbf{w}}^{(r)\top}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i^*) + \beta \qquad \text{and} \qquad \sigma_i := \boldsymbol{\varphi}^{(r)}(\mathbf{x}_i^*)^\top \mathbf{Q}\boldsymbol{\varphi}^{(r)}(\mathbf{x}_i^*). \tag{C.16}$$

Following Lloyd et al. (2015, section 4.3), each summand can be expressed as

$$\mathbb{E}[\log z_i^2] = -G(-\frac{\mu_i}{2\sigma_i^2}) + \log\left(\frac{\sigma_i^2}{2}\right) - C$$

where $G(\cdot)$ is defined as

$$G(z) = 2z \sum_{j=0}^{\infty} \frac{j! z^j}{(2)_j (1/2)_j} \tag{C.17}$$

with $(\cdot)_j$ being the rising Pochhammer series. The constant $C \approx 0.57721566$ is the Euler Mascheroni constant. $G(\cdot)$ can in practice be evaluated using a large multi-resolution look-up table of pre-computed values. Accurate evaluation can be obtained by linear interpolation of the values from the table.

Bibliography

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Lecture Notes in Statistics. Dover, New Yory.

Adam, V., Eleftheriadis, S., Artemev, A., Durrande, N., and Hensman, J. (2020). Doubly sparse variational gaussian processes. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2874–2884. PMLR.

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv prepprint arXiv:0710.3742*.

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16.

Aglietti, V., Bonilla, E. V., Damoulas, T., and Cripps, S. (2019). Structured variational inference in continuous Cox process models. In *Advances in Neural Information Processing Systems*, volume 32.

Agudelo-España, D., Gómez-González, S., Bauer, S., Schölkopf, B., and Peters, J. (2019). Bayesian online detection and prediction of change points. *CoRR*, abs/1902.04524.

Aizerman, M. A., Braverman, E. A., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. In *Automation*

and Remote Control,, number 25 in Automation and Remote Control,, pages 821–837.

Archambeau, C. and Bach, F. (2011). Multiple Gaussian process models. *arXiv prepprint arxiv:1110.5238*.

Aroian, L. A. and Levene, H. (1950). The effectiveness of quality control charts. *Journal of the American Statistical Association*, 45(252):520–529.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Athans, M., Wishner, R., and Bertolini, A. (1968). Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements. *IEEE Transactions on Automatic Control*, 13(5):504–514.

Ažman, K. and Kocijan, J. (2011). Dynamical systems identification using gaussian process models with incorporated local models. *Engineering Applications of Artificial Intelligence*, 24(2):398–408.

Baddeley, A. J., Chang, Y.-M., Song, Y., and Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, 5:221–236.

Balderama, E., Schoenberg, F. P., Murray, E., and Rundel, P. W. (2012). Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107(498):467–476.

Banerjee, S., Carlin, B. P., Gelfand, A. E., and Banerjee, S. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC. 1 st edition.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US.

Billingsley, P. (1995a). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley.

Billingsley, P. (1995b). *Probability and measure (3th ed.)*. Wiley.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017a). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017b). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bochner, S. (1932). Vorlesungen über fouriersche integrale. *Akademische Verlagsgesellschaf*.

Box, G., Jenkins, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. Forecasting and Control Series. Prentice Hall.

Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, 13(2):142 – 162.

Brémaud, P. (2020). *Probability Theory and Stochastic Processes*. Universitext. Springer International Publishing.

Brix, A. and Diggle, P. (2001). Spatio-temporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society B*, 63:823–841.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Buhmann, M. D. (2001). A new class of radial basis functions with compact support. *Math. Comput.*, 70:307–318.

Campbell, N. (1909). The study of discontinuous phenomena. *Proc. Cambridge Philos. Soc*, 15:117–136.

Capinski, M. and Kopp, P. (2013). *Measure, Integral and Probability*. Springer Undergraduate Mathematics Series. Springer London.

Caponnetto, A.and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical bayesian analysis of changepoint problems. *Applied statistics*, 41:389–405.

Caron, F., Doucet, A., and Gottardo, R. (2012a). On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2):579–595.

Caron, F., Doucet, A., and Gottardo, R. (2012b). On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2):579–595.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Cox, D. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society*, Series B. 17:129–164.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Csató, L. and Opper, M. (2002). Sparse on-line gaussian processes. *Neural computation*, 14:641–68.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008a). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 192–199. Association for Computing Machinery.

Cunningham, J. P., Yu, B. M., Shenoy, K. V., and Sahani, M. (2008b). Inferring neural firing rates from spike trains using gaussian processes. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20.

Cygu, S., Dushoff, J., and Bolker, B. M. (2021). pcoxtime: Penalized cox proportional hazard model for time-dependent covariates.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I. Probability and its Applications*. Springer, New York.

Dao, T., Sa, C. D., and Ré, C. (2017). Gaussian quadrature for kernel features. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6109–6119, Red Hook, NY, USA. Curran Associates Inc.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). On nearest-neighbor gaussian process models for massive spatial data. *WIREs Computational Statistics*, 8(5):162–171.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, 68:265–274.

Deisenroth, M. and Mohamed, S. (2012). Expectation propagation in gaussian process dynamical systems. In *Advances in Neural Information Processing Systems 26 (NIPS/NeurIPS), Cambridge, MA: MIT Press.* the mit press.

Deisenroth, M. P., Huber, M. F., and Hanebeck, U. D. (2009). Analytic moment-based gaussian process filtering. In *International Conference on Machine Learning*, ICML '09, page 225–232.

Deisenroth, M. P., Turner, R. D., Huber, M. F., Hanebeck, U. D., and Rasmussen, C. E. (2012). Robust filtering and smoothing with gaussian processes. *IEEE Transactions on Automatic Control*, 57(7):1865–1871.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diggle, P. (2003a). *Statistical analysis of spatial point patterns*. Edward Arnold. 2nd edition.

Diggle, P., Rowlingson, B., and TL, S. (2005). Point process methodology for on-line spatio temporal disease surveillance. *Environmetrics*, 16(5):423–431.

Diggle, P. J. (1985). A Kernel method for smoothing point process data. *Journal of The Royal Statistical Society Series C-applied Statistics*, 34:138–147.

Diggle, P. J. (2003b). *Statistical analysis of spatial point patterns*. Hodder Education. 2nd edition.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542 – 563.

Dong, Z., Zhu, S., Xie, Y., Mateu, J., and Rodríguez-Cortés, F. J. (2023). Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2):368–386.

Donner, C. and Opper, M. (2018a). Efficient Bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19(67):1–34.

Donner, C. and Opper, M. (2018b). Efficient Bayesian inference of Sigmoidal Gaussian Cox processes. *J. Mach. Learn. Res.*, 19(1):2710–2743.

Doob, J. L. (1991). *Stochastic processes*. John Wiley & Sons, New York.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. KDD '16, page 1555–1564, New York, NY, USA. Association for Computing Machinery.

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1166–III–1174. JMLR.org.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. (2011). Additive Gaussian processes. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Dvořák, J., Møller, J., Mrkvička, T., and Soubeyrand, S. (2019). Quick inference for log gaussian cox processes with non-stationary underlying random fields. *Spatial Statistics*, 33:100388.

Eleftheriadis, S., Nicholson, T., Deisenroth, M., and Hensman, J. (2017). Identification of gaussian process state space models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fearnhead, P. (2005). Exact bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166.

Fearnhead, P. (2006a). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213.

Fearnhead, P. (2006b). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models

via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B*, 69:589–605.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Ferreira, M., Conceição, H., Fernandes, R., and Tonguz, O. K. (2009). Stereoscopic aerial photography: An alternative to model-based urban mobility approaches. In *Proceedings of the Sixth ACM International Workshop on VehiculAr InterNETworking*, VANET '09, page 53–62, New York, NY, USA. Association for Computing Machinery.

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Language*, 93:93–114.

Flaxman, S., Chirico, M., Pereira, P., and Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the NIJ "real-time crime forecasting challenge". *Annals of Applied Statistics*.

Flaxman, S., Teh, Y. W., and Sejdinovic, D. (2017). Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081 – 5104.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015a). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 607–616. PMLR.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015b). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine*

Learning, volume 37 of *Proceedings of Machine Learning Research*, pages 607–616, Lille, France. PMLR.

Frigola, R., Chen, Y., and Rasmussen, C. E. (2014a). Variational gaussian process state-space models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3680–3688, Cambridge, MA, USA. MIT Press.

Frigola, R., Chen, Y., and Rasmussen, C. E. (2014b). Variational gaussian process state-space models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, Cambridge, MA, USA. MIT Press.

Frigola, R., Lindsten, F., Schön, T. B., and Rasmussen, C. E. (2013). Bayesian inference and learning in gaussian process state-space models with particle mcmc. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Gamerman, D. and Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gardner, J. K. and Knopoff, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America*, 64(5):1363–1367.

Garnett, R., Osborne, M. A., and Roberts, S. J. (2009). Sequential bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, page 345–352.

Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Genton, M. G. (2001). *Journal of Machine Learning Research*, 2:299–312.

Ghahramani, Z. (2012). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society of London A: Mathematical,Physical and Engineering Sciences*, 371(1984).

Gilboa, E., Saatçi, Y., and Cunningham, J. P. (2013). Scaling multidimensional inference for structured gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):424–436.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Girard, A., Edward, R. C., Quiñonero-Candela, J., and Roderick, M.-S. (2002). Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 545–552, Cambridge, MA, USA. MIT Press.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123 – 214.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, third edition.

Gramacy, R. B. (2016). laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R. *Journal of Statistical Software*, 72(i01).

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Gregorcic, G. and Lightbody, G. (2002). Gaussian processes for modelling of dynamic non-linear systems. In *Proceedings of the Irish Signals and Systems Conference*, page 141–147.

Grubesic, T. H. and Mack, E. A. (2008). Spatio-temporal interaction of urban crime. *Journal of Quantitative Criminology*, 24:285–306.

Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, 103(483):1238–1247.

Gunter, T., Lloyd, C., Osborne, M., and Roberts, S. (2014). Efficient Bayesian nonparametric modelling of structured point processes. *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*.

Gutjahr, T., Ulmer, H., and Ament, C. (2012). Sparse gaussian processes with uncertain inputs for multi-step ahead prediction. *IFAC Proceedings Volumes*, 45(16):107–112. 16th IFAC Symposium on System Identification.

H., R. and L., H. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384.

Hartikainen, J. and Sarkka, S. (2010). Kalman filtering and smoothing solutions to temporal gaussian process regression models. *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 379–384.

Harville, D. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer New York.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7517–7527, Red Hook, NY, USA.

Hensman, J., Durrande, N., and Solin, A. (2017). Variational fourier features for gaussian processes. *J. Mach. Learn. Res.*, 18(1):5537–5588.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA. AUAI Press.

Hensman, J., Matthews, A. G., Filippone, M., and Ghahramani, Z. (2015). Mcmc for variationally sparse gaussian processes. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Hernandez-Lobato, D. and Hernandez-Lobato, J. M. (2016). Scalable gaussian process classification via expectation propagation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, Cadiz, Spain. PMLR.

Hjort, N., Holmes, C., Mueller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347.

Hu, S. and Papageorgiou, N. (2013). *Handbook of Multivalued Analysis: Volume I: Theory*. Mathematics and Its Applications. Springer US.

Hubbell, S. and Foster, R. (1983). Diversity of canopy trees in a neotropical forest and implications for conservation. *Tropical Rain Forest: Ecology and Management*, 11:25–41.

Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, 6(4):1499 – 1530.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193.

John, S. and Hensman, J. (2018). Large-scale Cox process inference using variational Fourier features. In *International Conference on Machine Learning*, pages 2362–2370. PMLR.

Johnson, T. D., Elashoff, R. M., and Harkema, S. J. (2003). A bayesian change-point analysis of electromyographic data: detecting muscle activation patterns and associated applications. *Biostatistics*, 4 1:143–64.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2).

Julier, S., Uhlmann, J., and Durrant-Whyte, H. (1995). A new approach for filtering nonlinear systems. In *Proceedings of 1995 American Control Conference - ACC'95*, volume 3, pages 1628–1632 vol.3.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Keerthi, S. and Chu, W. (2005). A matching pursuit approach to sparse gaussian process regression. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Knoblauch, J. and Damoulas, T. (2018). Spatio-temporal Bayesian on-line change-point detection with model selection. In *Proceedings of the 35th International Conference on Machine Learning*, ICML'18, pages 2718–2727.

Knoblauch, J., Jewson, J. E., and Damoulas, T. (2018). Doubly robust Bayesian inference for non-stationary streaming data with beta-divergences. In *Advances in Neural Information Processing Systems*, volume 31.

Ko, J. and Fox, D. (2009). GP-BayesFilters: Bayesian filtering using gaussian process prediction and observation models. *Autonomous Robots*, 27(1):75–90.

Ko, Y.-J. and Seeger, M. W. (2016). Expectation propagation for rectified linear poisson regression. In Holmes, G. and Liu, T.-Y., editors, *Asian Conference on*

Machine Learning, volume 45 of *Proceedings of Machine Learning Research*, pages 253–268, Hong Kong. PMLR.

Kocijan, J., Girard, A., Banko, B., and Murray-Smith, R. (2005). Dynamic systems identification with gaussian processes. *Mathematical and Computer Modelling of Dynamical Systems*, 11(4):411–424.

Kolmogorov, A. (1941). Interpolation und extrapolation von stationären zufälligen folgen. *USSR, Ser. Math: Bull. Acad. Sci.*, 5.

Koop, G. and Potter, S. (2004). Forecasting and estimating multiple change-point models with an unknown number of change points. *Rev. Econ. Stud.*, 74.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kummerfeld, E. and Danks, D. (2013). Tracking time-varying graphical structure. In *Advances in Neural Information Processing Systems*, volume 26.

Kuss, M. and Rasmussen, C. (2006). Assessing approximations for gaussian process classification. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Lawrence, N., Seeger, M., and Herbrich, R. (2002). Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Lawrence, N. D. (2003). Gaussian process latent variable models for visualisation of high dimensional data. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 329–336, Cambridge, MA, USA. MIT Press.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. (2009). Inter-domain gaussian processes for sparse inference using inducing features. In Bengio, Y., Schuurmans, D., Laf-

ferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.

Le, Q., Sarlos, T., and Smola, A. (2013). Fastfood - approximating kernel expansions in loglinear time. In *30th International Conference on Machine Learning*.

Lee, D. and Mitchell, R. (2014). Controlling for localised spatio-temporal autocorrelation in long-term air pollution and health studies. *Statistical methods in medical research*, 23.

Leininger, T. J. and Gelfand, A. E. (2017). Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Analysis*, 12:1–30.

Levy-leduc, C. and Harchaoui, Z. (2007). Catching change-points with Lasso. In *Advances in Neural Information Processing Systems*, volume 20.

Li, X. and Ma, J. (2021). Non-central student-t mixture of Student-t processes for robust regression and prediction. In *Intelligent Computing Theories and Application*, pages 499–511.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2021). Towards a unified analysis of random fourier features. *J. Mach. Learn. Res.*, 22(1).

Lian, W., Henao, R., Rao, V., Lucas, J. p., and Carin, L. (2015). A multitask point process predictive model. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2030–2038, Lille, France. PMLR.

Liu, H., Cai, J., Wang, Y., and Ong, Y. (2018). Generalized robust bayesian committee machine for large-scale gaussian process regression. In *In International Conference on Machine Learning*, pages 3131–3140.

Liu, H., Ong, Y., Shen, X., and Cai, J. (2020). When gaussian process meets big data: A review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*, 31:4405–4423.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics*, 15 1:38–52.

Liu, W., Principe, J. C., and Haykin, S. (2010). *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley Publishing, 1st edition.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822. PMLR.

Loeve, M. (1978). *Probability Theory II*. Springer.

Lopez-lopera, A. F., John, S., and Durrande, N. (2019). Gaussian process modulated cox processes under linear inequality constraints. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1997–2006. PMLR.

Lyu, Y. (2017). Spherical structured feature maps for kernel approximation. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2256–2264. PMLR.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA.

MaguireE, B. A., S., P. E., and Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, 39(1-2):168–180.

Manogaran, G. and Lopez, D. (2018). Spatial cumulative sum algorithm with big data analytics for climate change detection. *Computers & Electrical Engineering*, 65:207–221.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Acad. Press.

Matthews, A. G. D. G. (2016). *Scalable Gaussian process inference using variational methods*. PhD thesis, University of Cambridge.

Matérn, B. (1960). *Spatial Variation*. Lecture Notes in Statistics. pringer-Verlag, Berlin.

McCullagh, P. and Møller, J. (2006). The permanental process. *Advances in Applied Probability*, 38(4):873–888.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Melkumyan, A. and Ramos, F. (2009). A sparse covariance function for exact gaussian process inference in large datasets. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 1936–1942, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, 209:415–446.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2.

Moller, J. and Waagepetersen, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Moran, P. A. (1968). *An Introduction to Probability Theory*. Clarendon Press, Oxford.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., and Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9:541–548.

Murray-Smith, R. and Girard, A. (2001). Gaussian process priors with arma noise models. *Irish Signals and Systems Conference*, pages 147–152.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Process*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Møller, J. and Waagepetersen, R. (2016). Some recent developments in statistics for spatial point patterns. *Annual Review of Statistics and Its Application*, 4(1):317–342.

Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.

Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv: Data Analysis, Statistics and Probability*.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(67):2035–2078.

Opper, M. and Archambeau, C. (2009). The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792.

Orbanz, P. and Teh, Y. W. (2011). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer.

Panda, S. and Nayak, A. (2016). Automatic speech segmentation in syllable centric speech recognition system. *Int J Speech Technol 19*, 9(18).

Park, C. and Huang, J. Z. (2016). Efficient computation of gaussian process regression for large spatial data sets by patching local gaussian processes. *Journal of Machine Learning Research*, 17(174):1–29.

Park, M., Weller, J., Horwitz, G., and Pillow, J. (2014). Bayesian active learning of neural firing rate maps with transformed gaussian process priors. *Neural computation*, 26:1–23.

Pennington, J., Yu, F. X. X., and Kumar, S. (2015). Spherical random features for polynomial kernels. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Plataniotis, K. N. and Hatzinakos, D. (2001). *Gaussian Mixtures and their Applications to Signal Processing*, volume Advanced Signal Processing Handbook of *CRC Press*. Springer-Verlag.

Polunchenko, A. S., Tartakovsky, A. G., and Mukhopadhyay, N. (2012). Nearly optimal change-point detection with an application to cybersecurity. *Sequential Analysis*, 31:409 – 435.

Punskaya, E., Andrieu, C., Doucet, A., and Fitzgerald, W. (2002). Bayesian curve fitting using mcmc with applications to signal segmentation. *IEEE Transactions on Signal Processing*, 50(3):747–758.

Quiñonero-Candela, J., Girard, A., Larsen, J., and Rasmussen., C. E. (2003). Propagation of uncertainty in bayesian kernel models - application to multiple-step ahead forecasting. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 2, pages II–701.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, page 1177–1184.

Ranganathan, A., Yang, M.-H., and Ho, J. (2011). Online sparse Gaussian process regression and its applications. *IEEE Transactions on Image Processing*, 20(2):391–404.

Rasmussen, C. and Ghahramani, Z. (2001). Infinite mixtures of gaussian process experts. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian processes for machine learning*. MIT Press.

Riesz, F. and Nagy, B. (1990). *Functional Analysis*. Dover Books on Mathematics. Dover Publications.

Ritov, Y., Raz, A., and Bergman, H. (2002). Detection of onset of neuronal activity by allowing for heterogeneity in the change points. *Journal of Neuroscience Methods*, 122(1):25–42.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. (2022). Practical hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550.

Rosser, G. and Cheng, T. (2019). Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Applied Spatial Analysis and Policy*, 12.

Ruanaidh, J. and Fitzgerald, W. (2012). *Numerical Bayesian Methods Applied to Signal Processing*. Statistics and Computing. Springer.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill Science/Engineering/Math.

Rudin, W. (2017). *Fourier Analysis on Groups*. Courier Dover Publications.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Saatçi, Y., Turner, R., and Rasmussen, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 927–934.

Saatçi, Y., Turner, R., and Rasmussen, C. E. (2015). Adaptive sequential bayesian change point detection. *Temporal Segmentation Workshop at NIPS 2009*.

Samo, Y.-L. K. (2017). *Advances in kernel methods : towards general-purpose and scalable models*. PhD thesis, University of Oxford.

Samo, Y.-L. K. and Roberts, S. (2015a). Generalized spectral kernels. *arxiv:1506.02236*.

Samo, Y.-L. K. and Roberts, S. (2015b). Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2227–2236. PMLR.

Samo, Y.-L. K. and Roberts, S. J. (2015c). String and membrane gaussian processes. *J. Mach. Learn. Res.*, 17:131:1–131:87.

Schervish, M. (1996). *Theory of Statistics*. Springer Series in Statistics. Springer New York.

Schölkopf, B. and Smola, A. J. (2002a). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press.

Schölkopf, B. and Smola, A. J. (2002b). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Schölkopf, B. and Smola, A. J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Seeger, M. (2003). *PAC-Bayesian generalisation error bounds and sparse approximations*. PhD thesis, University of Edinburgh.

Seeger, M. and Bouchard, G. (2012). Fast variational bayesian inference for non-conjugate matrix factorization models. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1012–1018, La Palma, Canary Islands. PMLR.

Seeger, M. W., Williams, C. K. I., and Lawrence, N. D. (2003). Fast forward selection to speed up sparse gaussian process regression. In Bishop, C. M. and Frey, B. J., editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 254–261.

Shah, A., Wilson, A., and Ghahramani, Z. (2014). Student-t processes as alternatives to Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, ICML'14, pages 877–885.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Sheth, R., Wang, B., and Khardon, R. (2015). Sparse variational inference for generalized gaussian process models. In *International Conference on Machine Learning*.

Shirai, T. and Takahashi, Y. (2003). Random point fields associated with certain fredholm determinants i: fermion, poisson and boson point processes. *Journal of Functional Analysis*, 205:414–463.

Smola, A. and Bartlett, P. (2001). Sparse greedy gaussian process regression. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Solin, A. and Särkkä, S. (2020). Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30(2):419–446.

Solin, A. and Särkkä, S. (2015). State space methods for efficient inference in Student-t process regression. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, ICML'15, pages 885–893.

Sriperumbudur, B. K. and Szabó, Z. (2015). Optimal rates for random fourier features. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1144–1152.

Stein, M. L. (1999). *Interpolation of spatial data*. Springer Series in Statistics. Springer-Verlag, New York.

Steinwart, I. (2017). Convergence types and rates in generic karhunen-loève expansions with applications to sample path properties.

Steinwart, I. and Christmann, A. (2007). *Support vector machines*. Springer.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417.

Stuart, A. M. (2010). Inverse problems : A bayesian perspective. *Constructive Approximation*, 19:451–559.

Sutherland, D. and Schneider, J. (2015). On the error of random fourier features. In *Proceedings of 31st Conference on Uncertainty in Artificial Intelligence (UAI '15)*, pages 862 – 871.

Svensson, A. and Schön, T. B. (2017). A flexible state–space model for learning nonlinear dynamical systems. *Automatica*, 80:189–199.

Svensson, A., Solin, A., Särkkä, S., and Schön, T. B. (2016). Computationally efficient Bayesian learning of Gaussian process state space models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, AISTATS, page 213–221.

Tang, Q., Niu, L., Wang, Y., Dai, T., An, W., Cai, J., and Xia, S.-T. (2017). Student-t process regression with Student-t likelihood. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2822–2828.

Tang, Q., Wang, Y., and Xia, S.-T. (2016). Student-t process regression with dependent Student-t noise. In *Proceedings of the 22th European Conference on Artificial Intelligence*, ECAI'16, page 82–89.

Taylor, B. M. and Diggle, P. J. (2014). Inla or mcmc? a tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284.

Titsias, M. (2009a). Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR.

Titsias, M. (2009b). Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR.

Ton, J.-F., Flaxman, S., Sejdinovic, D., and Bhatt, S. (2018). Spatial mapping with gaussian processes and nonstationary fourier features. *Spatial Statistics*, 28:59–78.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):245–268.

Tracey, B. D. and Wolpert, D. (2018). Upgrading from gaussian processes to student-t processes. *AIAA Non-Deterministic Approaches Conference*.

Tran, D., Ranganath, R., and Blei, D. (2016). The variational gaussian process. In *The variational Gaussian process*.

Turner, R., Deisenroth, M., and Rasmussen, C. (2010). State-space inference and learning with gaussian processes. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 868–875, Chia Laguna Resort, Sardinia, Italy. PMLR.

Turner, R., Saatçi, Y., and Rasmussen, C. E. (2009). Adaptive sequential Bayesian change point detection. In *Advances in Neural Information Processing Systems*, Temporal segmentation Workshop on the Meaning.

Turner, R. D. (2011). *Gaussian Processes for State Space Models and Change Point Detection*. PhD thesis, University of Cambridge.

Valiant, L. (2006). The complexity of computing the permanent. *Theoret, Comput, Sci*, 8:189–201.

van der Wilk, M., Rasmussen, C. E., and Hensman, J. (2017). Convolutional gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vish-

wanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vedaldi, A. and Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492.

Vempati, S., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2010). Generalized RBF feature maps for efficient detection. In *British Machine Vision Conference*.

W., L. P. A. and Shedler (1979). Simulation of a nonhomogeneous Poisson process by thinning.

Waageptersen, R. (2004). Convergence of posteriors for discretized log gaussian cox processes. *Scand. J. Statist.*, 66:229–235.

Walder, C. J. and Bishop, A. N. (2017). Fast Bayesian intensity estimation for the permanental process. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3579–3588. PMLR.

Wang, J., Hertzmann, A., and Fleet, D. J. (2005). Gaussian process dynamical models. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Warren, H., Oliveira, R., and Ramos, F. (2022). Generalized bayesian quadrature with spectral kernels. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2085–2095. PMLR.

Wendland, H. (2004). *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Technology Press books in science and engineering. Technology Press of the Massachusetts Institute of Technology.

Williams, C. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342 – 1351.

Williams, C. and Seeger, M. (2001a). Using the nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Williams, C. and Seeger, M. (2001b). Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688.

Williams, C. K. I. and Rasmussen, C. E. (1995). Gaussian processes for regression. In *NIPS*.

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1067–1075.

Wilson, R., Nassar, M., and Gold, J. (2010). Bayesian online learning of the hazard rate in change points problems. *Neural computation*, 22:2452–76.

Worden, K., Becker, W., Rogers, T., and Cross, E. (2018). On the confidence bounds of gaussian process narx models and their higher-order frequency response functions. *Mechanical Systems and Signal Processing*, 104:188–223.

Worden, K. and Green, P. L. (2014). A machine learning approach to nonlinear modal analysis. In Catbas, F. N., editor, *Dynamics of Civil Structures, Volume 4*, pages 521–528, Cham. Springer International Publishing.

Xuan, X. and Murphy, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning*, ICML'07, page 1055–1062.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions*, volume 1. Springer.

Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. (2014). Quasi-monte carlo feature maps for shift-invariant kernels. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 485–493, Bejing, China. PMLR.

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., and Kumar, S. (2016). Orthogonal random features. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Yu, S., Tresp, V., and Yu, K. (2007). Robust multi-task learning with t-processes. In *Proceedings of the 24th International Conference on Machine Learning*, ICML'07, page 1103–1110.

Yue, Y. R. and Loh, J. M. (2010). Bayesian Semiparametric Intensity Estimation for Inhomogeneous Spatial Point Processes. *Biometrics*, 67(3):937–946.

Zhang, Y. and Yeung, D. (2010). Multi-task learning using generalized t process. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, AISTATS, pages 964–971.